Where the Wild Things Are: Computer Vision for Global-Scale Biodiversity Monitoring

Thesis by
Sara Meghan Beery

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
Defended May 20, 2022

ACKNOWLEDGEMENTS

First, thank you to Pietro Perona for giving me a chance, for always saying yes when I had a crazy idea, for many passionate debates, and for supporting me every step of the way. Your guidance and your friendship have been invaluable.

Next, to my thesis committee–Yisong Yue, Katie Bouman, and Serge Belongie–thank you for opening my mind to an academic career. You saw what I couldn't–that I had what it took to be a professor and that it would be the right career path for my big dreams.

Thank you to all of my other mentors: you have each changed the course of my life for the better in such significant ways. To Agnieszka Miguel for inviting me to work with her on my first ever research project at Seattle University. I would never have considered a PhD without your support and enthusiasm. To Grant van Horn, for taking me under your wing when I arrived at Caltech, bright eyed and bushy tailed and clueless. I learned so much from you while we worked together on my first paper in grad school. To Oisin Mac Aodha, for being my wise sage on the mountain and for always asking the best questions. To Dan Morris, for sharing my vision for impactful, technology-driven solutions for biodiversity and for your invaluable help alongside Siyu Yang making the MegaDetector so accessible and useful for camera trappers worldwide. To Tanya Berger-Wolf for being my champion and showing me what a successful interdisciplinary career could look like in academia, and for helping me set up my first experience with fieldwork in Kenya. To Devis Tuia and Bistra Dilkina for supporting me and helping to lead our young but growing interdisciplinary field. To Adam Wierman for being a constant champion for DEI in Caltech CMS. To Jonathan Huang, who I think of as a second academic advisor, for your guidance and friendship as we dove into interesting research problems together at Google over the past three years.

Thank you to all of my collaborators and idea-bouncer-offers in academia, industry, and everything in between. At Microsoft - Neel Joshi, Siyu Yang, Arash Norouzzadeh, Ashish Kapoor, Jim Piavis, Nebojsa Jojic. At Google - Vivek Rathod, Guanhang Wu, Vighnesh Birodkar, David Ross, Tanya Birch, Tomer Gadot, Stefan Schneider, the Visual Dynamics team, the Auto Arborist team, Google AI for Nature and Society, and the Wildlife Insights team. To the AI for Conservation community, and particularly Elizabeth Bondi, Ben Weinstein, Beni Kellenberger, Kevin Winner, Blair R. Costelloe, Alexander and Mackenzie Mathis, Silvia Zuffi, Holger Klinck, the WILDLABS team. To the Climate Change AI and CompSust communities, and particularly Priya Donti and David Rolnick. To the Visipedians that I haven't mentioned already - Christine Kaeser-Chen, Jessie Barry, Cullen Hanks, Kimberley Wilber. To my collaborators on our conservation technology deployments in Kenya - Jake Wall and the whole team at MEP, Dan Rubenstein, Rosemary Warungu and the whole team at Mpala, Chuck Stewart and the whole team at RPI, Arjun Subramonian, Shane Lubold, and Jason Parham and the team at Wild Me. To Shiori Sagawa, Pang Wei Koh, and the entire WILDS team.

Thank you to all of the amazing students I mentored throughout my PhD–Peter Kulits, Justin Kay, Tzarina Shippee, Angelina Pan, Edoardo Lanzini, Avirath Sundaresan, Enoch Luk, Arushi Agarwal, Roanne Coutzer, Tuhin Das. Working with you has been so rewarding.

To the Caltech Vision Lab++–Eli Cole, Neehar Kondapaneni, Tony Zhang, Serim Ryou, Matteo Ruggero Ronchi, Joe Marino, Ron Appel, Mason McGill, David Hall, Alvita Tran, Eyrun Eyjolfsdottir, Krzysztof Chałupka, Cristina Segalin, Jennifer Sun, Daniel Laumer, Laure Delisle, Suzanne Stathatos, Kevin Mei, Rogerio Guimaraes, plus of course Yang Liu, Angela Gao, Berthy Feng, Aviad Lewis, He Sun, and Ray Wu. You are all a big, beautiful family and I hope my future lab can be half as fun. And of course to Caroline, for making it all possible.

To my CMS cohort–Natalie Bernat, Eric Zhan, Riley Murray, Jeremy Bernstein, Nikola Kovachki, Zachary Lee, Fengyu Zhou, and Chen Liang — I never would have gotten through first year without you, and it's wonderful to see how far we all have come.

Thank you to my friends spread across the globe. To the Wet Bandits, for making grad school fun. To our friends in Santa Barbara, and particularly to Kat and Joe for being our pandemic roomies. To Rachel Boyadjis, my little monster, for knowing me so well. To Kristine Triandafilou, my oldest friend, for always leading the way.

Thank you to the Hilt Estate Winemaking team, for taking me in and letting me use the winery wifi during the pandemic. Without you guys, this thesis would be a lot shorter.

To Mama and Papa Beery, you have never stopped cheering for me. From the first day I stepped on stage as a baby ballerina to the day I stepped up to the podium to defend my PhD, you have been my constant rock, my dependable, loving support network. To my little brothers, it has been so fun to watch us all become scientists. Maybe Dad was right and we can blame it on the radio towers we grew up next to. To the Wiltons, for welcoming me to your family with open arms.

Finally, thank you Cary, for loving me and supporting me throughout all the ups and downs of life in research. This thesis is for you. I love you.

ABSTRACT

We require a real-time, modular earth observation system that unites efforts across research groups in order to provide the necessary information necessary for global-scale impact in sustainability and conservation in the face of climate change. The development of such systems requires collaborative, interdisciplinary approaches that translate diverse sources of raw information into accessible scientific insight. For example, we need to monitor species in real time and in greater detail to quickly understand which conservation efforts are most effective and take corrective action. Current ecological monitoring systems generate data far faster than researchers can analyze it, making scaling up impossible without automated data processing. However, ecological data collected in the field presents a number of challenges that current methods, like deep learning, are not designed to tackle. These include strong spatiotemporal correlations, imperfect data quality, fine-grained categories, and long-tailed distributions. Our work seeks to overcome these challenges, and this thesis includes methods which can learn from imperfect data, systematic frameworks and benchmarks for measuring and overcoming performance drops due to domain shift, and the development and deployment of efficient human-AI systems that have real-world conservation impact.

PUBLISHED CONTENT AND CONTRIBUTIONS

[1] Sara Beery. Scaling Biodiversity Monitoring for the Data Age. *ACM XRDS: Crossroads*, 27(4):14–18, 2021. doi: 10.1145/3466857. S.B. wrote the manuscript.

[2] Sara Beery* and Elizabeth Bondi*. Can Poachers Find Animals from Public Camera Trap Images? *The CV4Animals Workshop at CVPR*, 2021. doi: 10.48550/arXiv.2106.11236. S.B. participated in designing and funding the project, collecting the data, developing the method, running the experiments, and writing the manuscript. * denotes co-first authorship.

[3] Sara Beery* and Ellie Warren*. The Promise and Pitfalls of Machine Learning for Conservation. *WILDLABS Series on Technical Difficulties*, 2021. S.B. participated in the writing of the article.

[4] Sara Beery, Grant Van Horn, Oisin MacAodha, and Pietro Perona. The iWildCam 2018 Challenge Dataset. *The Fifth Fine-Grained Visual Categorization Workshop at CVPR*, 2018. doi: 10.48550/arXiv.1904.05986. S.B. participated in designing the project, curating the dataset, hosting the competition, and writing the manuscript.

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in Terra Incognita. *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. doi: 10.1007/978-3-030-01270-0_28.
S.B. participated in designing the project, curating the dataset, developing the method, running the experiments and writing the manuscript.

[6] Sara Beery, Dan Morris, and Pietro Perona. The iWildCam 2019 Challenge Dataset. *The Sixth Fine-Grained Visual Categorization Workshop at CVPR*, 2019. doi: 10.48550/arXiv.1907.07617. S.B. participated in designing the project, curating the dataset, hosting the competition, and writing the manuscript.

[7] Sara Beery, Elijah Cole, and Arvi Gjoka. The iWildCam 2020 Competition Dataset. *The Seventh Fine-Grained Visual Categorization Workshop at CVPR*, 2020. doi: 10.48550/arXiv.2004.10340. S.B. participated in designing the project, curating the dataset, hosting the competition, and writing the manuscript.

[8] Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic Examples Improve Generalization for Rare Classes. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 863–873, 2020. doi: 10.1109/WACV45572.2020.9093570. S.B. participated in designing the

project, curating the dataset, developing the method, running the experiments and writing the manuscript.

[9] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context R-CNN: Long Term Temporal Context for Per-Camera Object Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2020. doi: 10.1109/cvpr42600.2020.01309. S.B. participated in designing the project, developing the method, running the experiments and writing the manuscript.

[10] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iWildCam 2021 Competition Dataset. *The Eighth Fine-Grained Visual Categorization Workshop at CVPR*, 2021. doi: 10.48550/arXiv.2105.03494. S.B. participated in designing the project, curating the dataset, hosting the competition, and writing the manuscript.

[11] Sara Beery*, Elijah Cole*, Joseph Parker, Pietro Perona, and Kevin Winner. Species Distribution Modeling for Machine Learning Practitioners: A Review. *ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 329–348, 2021. doi: 10.1007/978-3-030-01270-0_28. S.B. participated in designing the project, conducting informational interviews, and writing the manuscript. * denotes co-first authorship.

[12] Sara Beery, Guanhang Wu, Trevor Edwards, Filip Pavetic, Bo Majewski, Shreyasee Mukherjee, Stanley Chan, John Morgan, Vivek Rathod, and Jonathan Huang. The Auto Arborist Dataset: A Large-Scale Benchmark for Multiview Urban Forest Monitoring Under Domain Shift. 2022. S.B. participated in designing the project, curating the dataset, developing the method, running the experiments and writing the manuscript.

[13] Peter Kulits, Jake Wall, Anka Bedetti, Michelle Henley, and Sara Beery. ElephantBook: A Semi-Automated Human-in-the-Loop System for Elephant Re-Identification. pages 88–98, 2021. doi: 10.1145/3460112.3471947. S.B. participated in the conception and funding of the project, led communication with ecological partners, managed the project team and led weekly meetings, and participated in the analysis of the data and writing of the manuscript.

[14] Devis Tuia*, Benjamin Kellenberger*, Sara Beery*, Blair R Costelloe*, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, Roland Kays, Holger Klinck, Martin Wikelski, Iain D. Couzin, Grant Van Horn, Margaret C. Crofoot, Charles V. Stewart, and Tanya Berger-Wolf. Perspectives in Machine Learning for Wildlife Conservation. *Nature Communications*, 13(1):1–15, 2022. doi: 10.1038/s41467-022-27980-y. S.B. participated in organizing the writing team, structured and organizing the manuscript, and writing the manuscript. * denotes co-first authorship.

TABLE OF CONTENTS

NOMENCLATURE

(visual) Descriptor. Higher-level statistics extracted from data that are supposed to summarize, or pronounce, more abstract differences within the data point to facilitate the task of the subsequent machine learning model, also called a *feature*. For example, a common descriptor used in traditional vegetation mapping on remote sensing imagery is the Normalized Difference Vegetation Index (NDVI), whose values are highly contrastive between vegetated and non-vegetated areas than bare pixel values alone. Traditional machine learning algorithms require manual definition and calculation of such features, whereas deep learning methods learn them automatically in the training process.

Artificial Intelligence (AI). The concept of a machine being able to perform higher-level, semantic reasoning.

Big data. Many definitions exist, but we cast *big data* as *information content for analyses whose volumes are too large to handle for users with conventional hardware*. Many sensors addressed produce *big data*, in particular remote sensing, social media and camera trap networks. Analysis of such volumes of data quickly becomes intractable for conventional machine learning methods, in particular if the study area of interest exceeds regional ecosystems.

Classification. Assigning an entire image or video to a single category.

Computer Vision. A field of research that seeks to enable computers to derive information from images, videos and other structured inputs, often involving methods and techniques from machine learning..

Convolutional Neural Network (CNN). Deep learning models that contain at least one convolution layer. In such layers, neurons are organized into banks of filters that are convolved with the inputs (*i.e.*, the same filter weights are applied across multiple locations in the image). This allows reducing the number of required neurons while also providing a limited amount of translation invariance.

Data science. Like *big data*, *data science* is a less-well-defined term, denoted here as an inter- or multidisciplinary research field on automated information extraction from observations or other content sources.

Deep learning. Family of prediction models that consist of neurons, grouped into three or more sequential layers, where each neuron receives the output from one (or more) previous neurons and itself predicts an output, consisting of weighted combinations of its inputs.

Detection. Localizing the area within an image that corresponds to a category of interest, usually represented by a rectangular *bounding box*–the tightest box that could be drawn around that object while still containing all of its pixels.

Detection rate. See *recall*.

Domain Adaptation. Methods to describe, evaluate, and/or tackle the challenge of out-of-domain data.

False positive. Incorrect prediction of a data point, object, or background area (*e.g.*, in an image) as a certain class.

Feature. See *(visual) Descriptor*.

Fine-grained classification. Label classes are denoted as *fine-grained* if they belong to a common supercategory (*e.g.*, "American Robin" and "Guineafowl" both belong to the supercategory "bird"). Fine-grained classification can be challenging if categories exhibit similar visual properties..

Individual identification. Recognizing unique instances of an object in an image or video (frame). Individual identification is usually performed through recognizing of unique visual cues that serve as *fingerprints* for an individual, such as the striping pattern of zebra or dot pattern on the back of whale shark individuals.

Inference. The act of performing prediction with a (trained) machine learning model.

Instance Segmentation. Grouping every pixel in an image with the other pixels corresponding to that same *instance* or object. If the image contained seven lions, each lion would be categorized with a different pixel label, even if the lions' pixel masks touch each other.

Localization. Identifying the position of an object within an image or video (frame). Unlike Detection, localization may not always include estimation of the full extents of an object, *e.g.*, through a bounding box, but might be limited to spatial coordinates of the object's center.

Loss function. Numerical criterion that measures the disagreement between an machine learning model prediction and the Ground Truth labels. For example, the *cross-entropy loss function* returns the negative log likelihood between a predicted model probability and the label class.

Machine Learning. The ability of a computer to perform prediction tasks by learning from data (*i.e.*, without primarily relying on hard-coded cascades of rules).

Object detection. See *detection*.

Open-set. Scenario where a dataset may exhibit categories at test time that were unseen during machine learning model training. For example, a model for individual identification may be presented with images of an individual that got newly introduced to the area after training, and needs to be able to recognize it as a new individual accordingly.

Out-of-domain. Data that is not drawn from the identical set that an machine learning model was trained on. A good example of this would be images from a camera trap that was not seen during training.

Overfitting. Training an machine learning model to achieve (near-) perfect accuracy on the training set, but unacceptable accuracy on the validation or test set. Overfitting can occur if the model has too many free parameters or if the training set is not representative enough. See also *underfitting*.

Pose estimation. 2D: predicting the pixel location of known parts of an object, for example, localizing the nose, eyes, joints, and tail of a lion. 3D: predicting the parts location in space, or predicting the 3D rotation of an articulated animal skeleton.

Posture Estimation. See *pose estimation*.

Precision. Class-wise measure of exactness of machine learning model predictions. A precision of 1.0 means that every prediction made by a model is correct, while one approaching 0.0 means that there is a high number of wrong predictions (see *false positive*).

Recall. Class-wise measure of completeness of machine learning model predictions. A recall of 1.0 means that every data point with a given true label class has been correctly predicted as such by the model, while a recall of 0.0 means that the model has missed all data points of that class.

Semantic Segmentation. Assigning every pixel in an image to a specific class, *i.e.*, all "lion pixels" would be labeled as such, regardless of the actual individual they belong to.

Semi-supervised learning. Training an machine learning model on data for which only a small subset contains labels.

Supervised learning. Training an machine learning model on data that consists of inputs (*e.g.*, images) and labels (*e.g.*, species names, bounding boxes).

Tracking. Localizing individual objects and correctly match them between frames throughout a video or temporal sequence of images.

Training. Altering the free (learnable) parameters of an machine learning model to optimize it to the training dataset, usually performed by minimizing values of a Loss function.

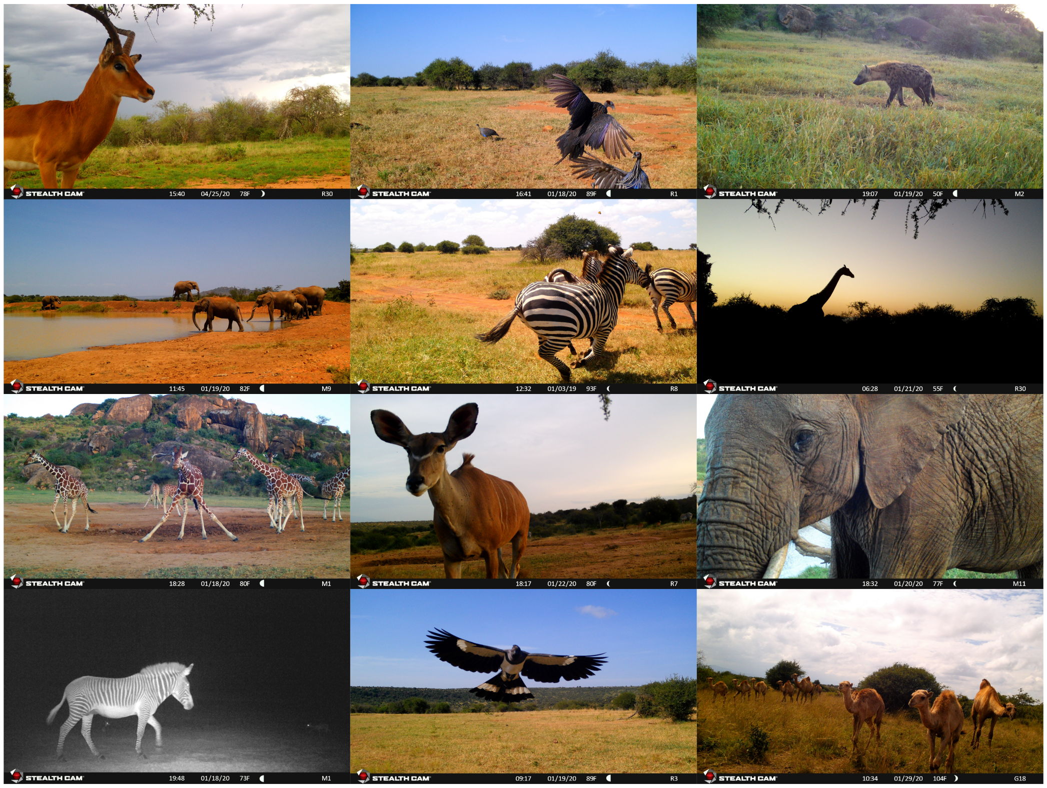
Underfitting. A machine learning model underfits the training set if it cannot appropriately capture the data distribution, resulting in unacceptable accuracy. Underfitting usually occurs if the model does not have a sufficient number of free parameters. See also *overfitting*.

Unsupervised learning. Training a machine learning model on data that only consists of inputs, but not of labels.

Chapter 1

INTRODUCTION

Biodiversity is declining at an unprecedented rate worldwide. The World Wildlife Fund 2020 Living Planet Report [6] found that between 1970 and 2016, the the number of living mammals, birds, amphibians, reptiles, and fish has decreased by an average of 68%. This figure jumps to 94% in the American tropics, some of the most highly biodiverse regions in the world. These numbers are the result of a collaborative effort from hundreds of ecological experts from around the globe based on decades of data. However, though they are the best estimates we currently have, in many cases we still do not have enough information to build an accurate understanding of the scope of our current loss and prioritize measures to counteract it. For example, 14% of the threatened species on the International Union for Conservation of Nature (IUCN) Red List are considered data deficient–they may be in even more danger than is currently known [21]. Increased biodiversity monitoring


STEALTH CAM 15:40 04/25/20 78F R30
STEALTH CAM 16:41 01/18/20 89F R1
STEALTH CAM 19:07 01/19/20 50F M2
STEALTH CAM 11:45 01/19/20 82F M9
STEALTH CAM 12:32 01/03/19 93F R8
STEALTH CAM 06:28 01/21/20 55F R30
STEALTH CAM 18:28 01/18/20 80F M1
STEALTH CAM 18:17 01/22/20 80F R7
STEALTH CAM 18:32 01/20/20 77F M11
STEALTH CAM 19:48 01/18/20 73F M1
STEALTH CAM 09:17 01/19/20 89F R3
STEALTH CAM 10:34 01/29/20 104F G18

is necessary not only to understand ecosystems and how they are changing in response to climate change and human encroachment, but also to get necessary feedback on the effectiveness of conservation actions.

Automatic and manual biodiversity monitoring is in place in many protected areas. The data collected is used to understand the effects of climate change and conservation policies on the size and diversity of species populations across the taxonomic tree. It provides feedback on land management policies and conservation interventions. This information is vital to those seeking to preserve natural resources and react quickly and appropriately to ecosystem threats. The data collected by biodiversity monitoring systems is also used in many aspects of ecological research, including predicting where a species might be found based on environmental variables (species distribution modeling), estimating the size of a population of a certain species in an area (population estimation), and understanding changes in plant and animal behavior in reaction to human encroachment and climate change. However, most data processing and analysis is done manually, which prevents these systems from scaling up spatiotemporally and taxonomically to the magnitude of data necessary to capture complex global ecosystem dynamics in near-real-time.

1.1 Biodiversity monitoring's data challenge

Biodiversity data can be time consuming and expensive to collect, as it frequently relies on humans manually collecting samples or purchasing, deploying, and maintaining networks of sensors–often cameras or microphones–in the field. In addition to the cost of data collection, experts must invest significant effort to filter, categorize, and analyze the resulting data. Even small-scale biodiversity monitoring systems can generate data far faster than researchers can analyze it. For example, it can take years for scientists to manually process and interpret a single season of data from a network of camera traps. As a concrete example, our collaborators at the Idaho Department of Fish and Game were previously five years behind in their camera trap data processing even with a team of ecologists working full-time to categorize their data. Producing real-time estimates of fish escapement upstream –necessary to maintain sustainable fisheries–requires teams of field ecologists working in shifts to watch near-continuous streams of sonar data [24]. Only a handful of monitoring points can be managed with this level of effort. The challenge is even greater for taxa that are studied by trapping and collecting, such as beetles and other insects. Entomologists can collect thousands of beetles in a few days, but it may require months or years for an expert to exhaustively identify all of the specimens

(a) **Camera traps.** Our network of 100 camera traps deployed at Mpala Research Center in Kenya collected over 10M images in 2021.

(b) **Community science.** There are over 100M observations of species that have been collected by volunteers via the iNaturalist platform [1]

(c) **Aerial surveys.** A single survey flight, such as the ones undertaken in [22], can collect up to 200TB of video.

Figure 1.1: Ecological data is collected in many different ways. Camera traps, community science projects, and aerial surveys are three types of ecological data currently collected to help monitor biodiversity. Each of these data collection types generate vast amounts of data, making it difficult to analyze the data by hand and extract insights quickly.

to the species level. This pace and scale of analysis is insufficient to keep up with impact from human activity and a rapidly changing climate. To make effective conservation decisions, policymakers need to know how different ecosystems are reacting quickly and in greater detail.

Luckily, new advances in technology, data collection, data processing, and data management are making it possible to scale and speed up biodiversity monitoring efforts worldwide. These advances provide a diverse perspective on our natural world, capturing data at different scales and across modalities.

Since the advent of global-scale earth observation missions in the 1970s, remote sensing data collected from satellites, and low-flying aircraft has been used as a proxy for direct biodiversity observations, including monitoring forest habitat intactness and estimating populations of large colonies of birds. However, the high-spatial-resolution imagery necessary for direct wildlife observation and analysis can only be collected with aircraft and more recently, drones, which are expensive, restricted in some geographies, and sometimes dangerous to operate.

Data collected from fixed networks of sensors such as camera traps, phenocams, bioacoustic sensors, and sonar provide consistent temporal sampling, allowing ecologists to monitor changes over time. Camera traps, for example, provide inexpensive high-resolution in situ imagery, even under the forest canopy, and are used by ecologists to monitor a broad set of wildlife species.

On-animal sensors, like GPS collars and radio tags, are used to track single animals and provide remarkable insight into animal behavior and movement patterns. They are also used to collect contextual environmental metadata, such as ambient temperature. Manually collected samples of insect populations, scat, or soil can be used in eDNA analysis, which is able to recognize genetic barcodes for species. This can be used to provide insight into animal behavior and species interactions, such as which species are drinking at a given watering hole. However, placing on-animal sensors is invasive, labor intensive, and sometimes very expensive.

Community scientists can collect vast sets of species observation data via images, sound recordings, or species checklists, which are community and expert curated in data repositories such as iNaturalist and eBird. These data collection methods have an impressive ability to scale, but the data can be noisy due to inexpert species identification and tends to be spatially biased towards areas of high human traffic, like cities or well-traveled nature preserves.

Recent reductions in sensor costs have allowed many of these types of data collection to scale up far beyond what was previously possible. Open-source, modular, and accessible data collection systems such as AudioMoth [20] and FieldKit [34] are helping to build strong communities of conservation technologists. These groups share resources, best-practices, and even source code. All this has led to the creation of vast collections of data which need to be stored, processed, shared, and analyzed to derive insights from the data. Each data collection method is optimal for some subset of species, areas, and monitoring needs. They are complementary–no one data collection method can capture the entire biodiversity picture. Together they may one day span the tree of life and the globe. With these already vast and growing sets of diverse data comes a significant data processing challenge that bottlenecks our ability to extract needed ecological insight from raw data streams. We need to automate ecological data processing.

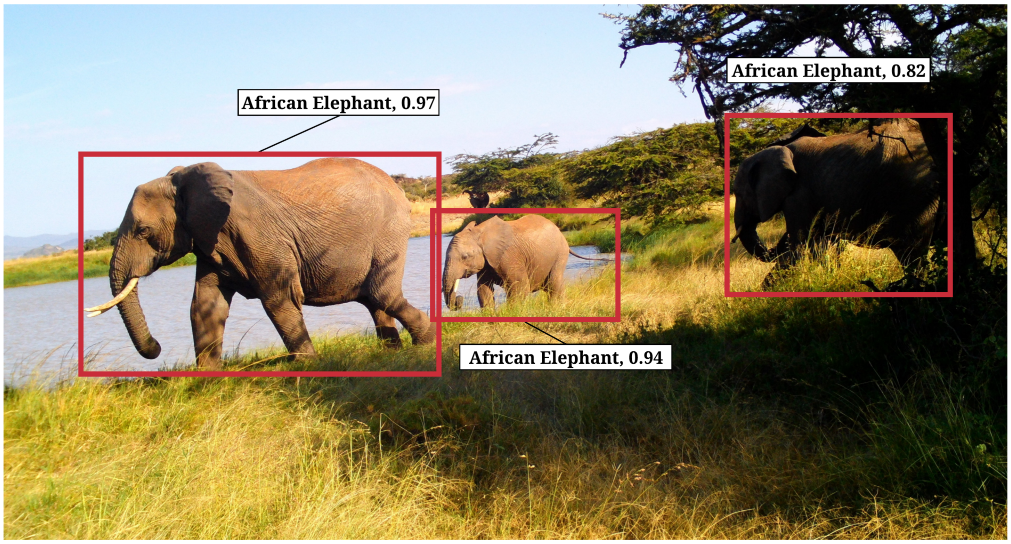


Figure 1.2: Models trained using machine learning can be used to automatically extract relevant ecological information directly from the raw data. For example, computer vision detection models can be used to find and categorize animal species in images collected by camera traps.

1.2 Biodiversity data poses new challenges for machine learning

Computer vision models trained on large repositories of data curated by teams of experts play a role in many diverse applications (for example, self-driving cars, Instagram filters, and Google image search). The success of these systems, as well as their ability to process large, complex datasets efficiently, have led ecologists and conservation technologists to explore how these methods can be used to help monitor our planet. There have been hundreds of papers in the last year alone applying machine learning and computer vision methods across the breadth of biodiversity data sources. As further incentive for ecologists to explore automated data processing methods, there is a large amount of “bycatch” hidden in the data. For example, observations of certain plant species can be extracted from imagery originally collected for wildlife monitoring. From the point of view of a wildlife researcher, the plant images are bycatch that they don’t have the capacity to label or curate from their existing datasets. Targeted and well-trained machine learning models could do so quickly and scalably, drastically increasing the accessibility of bycatch observations and the taxonomic scope of our current databases.

Machine learning is already being used in practice to process ecological data at scale. The Microsoft AI for Earth MegaDetector [10], an animal, human, and vehicle detection model for camera trap data, is used as a first data filtration step in the ecological data pipelines of over thirty organizations worldwide, including the Wildlife Conservation Society, San Diego Zoo Global, and Island Conservation. The publicly-hosted MegaDetector API is queried hundreds of thousands of times per month. The model works off-the-shelf for most camera trap data due to a combination of community building, data science, and machine learning research. The MegaDetector is trained to localize animals but not predict their species, which has been shown to be more robust to both new species and new camera deployments than species-specific models [8]. There is a significant need in the camera trapping community to filter out images containing animals, humans, or vehicles from large sets of mostly empty imagery, and this model does so efficiently and accurately.

However, biodiversity data still presents challenges that are not well addressed by existing machine learning methods:

To begin with, we must develop machine learning methods that can learn from fewer examples and handle significant bias in datasets due to data imbalance. The distribution of species worldwide is long-tailed. This means most observations are for common species, and the vast majority of species have few, if any, observations [36]. This results in highly-imbalanced datasets, with most rare species having insufficient data representation to be learned accurately by traditional machine learning frameworks. Another factor which adds to this imbalance is the use of passive ecological monitoring sensors which collect a large amount of "empty" data, that is, data without any observations of the study's target species. For example, The Snapshot Serengeti project estimates that 90% of the images from their camera traps are empty [29] (note that this data can still contain valuable information on non-target species–bycatch–that machine learning can help to extract).

Secondly, machine learning models assume that each data point is collected independently from the same underlying distribution. However, many biodiversity monitoring systems capture signals that are correlated in time and space. This correlation can result in model overfitting, particularly for static sensors or sparsely-sampled drone flights, causing poor model generalization to new deployments.

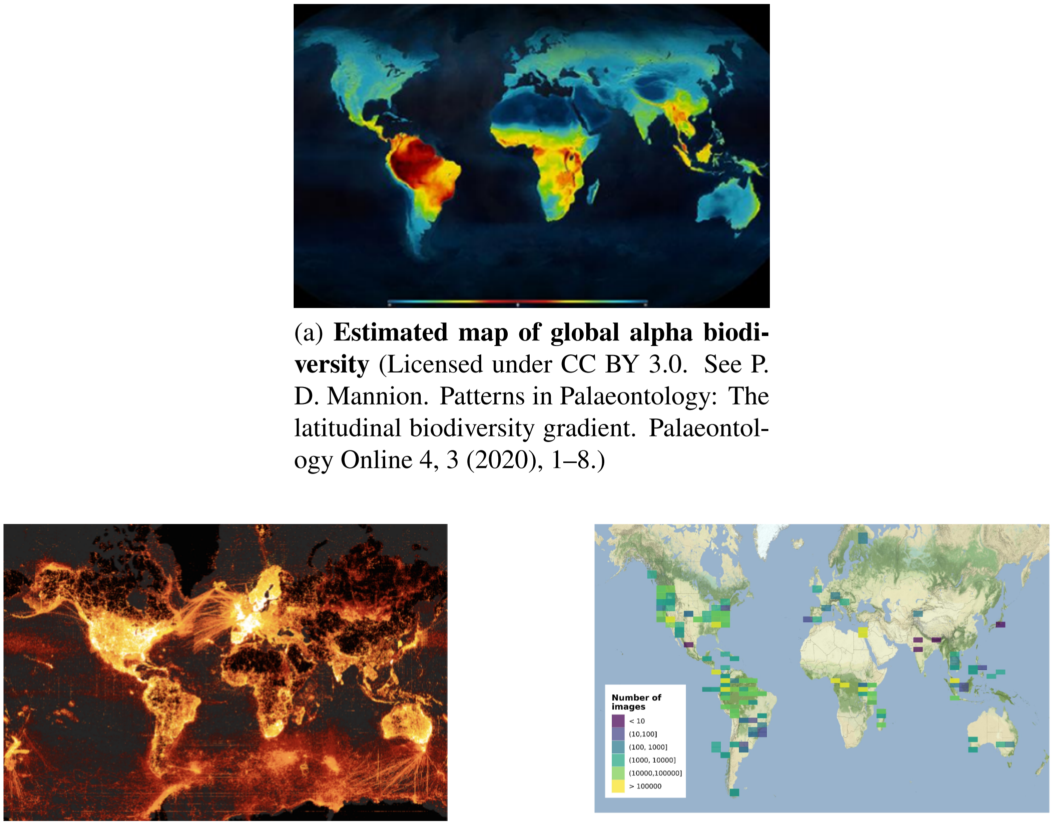
Further, the computer vision and machine learning communities usually work on high-quality datasets curated by human experts, with well-framed objects of interest and clean, accurate data labels. By contrast, biodiversity monitoring data is collected from sensors with limited intelligence, such as camera traps which collect data based on motion triggers. This leads to observations of interest that are too close or too far from the sensor, low resolution, or obscured by noise.

Finally, ecological data can also require expertise to categorize correctly. This involves challenging tasks like distinguishing between species of gulls or identifying species-specific behaviors. Labels curated from community scientists or non-experts frequently contain errors. Unlike relatively simple classification tasks, such as identifying stop signs on a street, building machine learning models for biodiversity monitoring requires technologists to connect task-specific expert knowledge with machines and data.

These challenges provide an exciting opportunity for the computer vision and machine learning research communities to develop creative solutions. New methodologies are being developed to tackle these challenges, often incorporating expert domain knowledge via ecologists collecting and labeling additional data, and providing guidance around the structure inherent to data collected from different sensor types and across taxa. Exciting recent work includes using synthesized data for rare species to improve rare-class performance [12], incorporating learned geospatial priors to improve species identification by letting the model know which species are most likely to be seen in a given area at a specific time [28], and building models that can share information across data collected by a given static sensor, helping the model adapt to previously unseen environments [13].

1.3 Making data accessible

This increase in the amount and variety of data being collected and processed has necessitated the creation of data standards, data management tools, data sharing repositories, data aggregation, and analysis platforms. All of these share a similar goal: to help ecologists and conservationists easily and effectively share data and insights. Large-scale data repositories such as the Global Biodiversity Information Facility (GBIF) and the Macaulay Library pull together occurrence records and media from scientific studies and other large-scale but more targeted data collection and management platforms. These latter include iNaturalist for community science species observations (currently at 64M observations worldwide) and Wildlife Insights for networks of static camera traps (currently at 12.6M global camera trap images). Analysis platforms such as Map of Life seek to aggregate and analyze data from global repositories, governmental surveys, and scientific publications in order to produce ecological insights. Map of Life finetunes IUCN global range maps for most species on earth. Aggregating data allows researchers to share the cost and scale up, in collection effort, data processing effort, and across jurisdictions. These repositories provide the framework for researchers to combine their efforts.



(a) **Estimated map of global alpha biodiversity** (Licensed under CC BY 3.0. See P. D. Mannion. Patterns in Palaeontology: The latitudinal biodiversity gradient. Palaeontology Online 4, 3 (2020), 1–8.)

(b) **Species occurrence data in GBIF** (Occurrence download. GBIF.org. April 9, 2020; https://doi.org/10.15468/dl.wyv3d4. Biodiversity heatmap image licensed under CC BY 3.0)

(c) **Camera trap records in Wildlife Insights** (Created by Fabiola Iannarilli, Yale University, on behalf of Wildlife Insights; https://www.wildlifeinsights.org/home)

Figure 1.3: Global data curation platforms such as GBIF are curating data from many different types of sensors and across ecological studies, and aggregate data from sensor-specific data management and hosting platforms such as Wildlife Insights. However, there are biases in where and when data was collected and how much data is available for a given species, and these biases are not consistent across sensor types. There are currently many more species occurrence records available in the United States and Europe, but if we look at a global heatmap of biodiversity we see that there is greater diversity in the Amazon, Subsaharan Africa, and East Asia. This means that for the areas with greatest available biodiversity we have less access to information about how to protect the species that are there, and how those species are being affected by climate change and human encroachment.

There are many challenges that biodiversity monitoring and modeling systems still face. As mentioned above, different data types have biases based on where the data is collected, or which species are likely to be seen. One of the biggest open challenges in ecological modeling is how to understand and compensate for the sampling biases of each of these types of data, while still benefiting from their complementary geospatial and taxonomic coverage in order to build an accurate, unbiased picture of our world's biodiversity. By understanding bias and associated uncertainty in ecological models, ecologists could target their data collection efforts to reduce these uncertainties and optimize their data collection practices to choose the type and placement of sensors that maximize coverage geospatially and taxonomically given the available resources.

1.4 Looking to the future

Complementary and parallel technological advances in data collection, data processing, and data management are driving the field of biodiversity monitoring forward every day, and there is a growing interdisciplinary community of researchers sharing resources, best-practices, and skills. The machine learning community has shown ever-growing interest in tackling biodiversity data challenges, with increasing numbers of biodiversity-focused workshops [2–4] and competitions [14, 17] seen each year. Looking to the future, the accessibility and standardization of ecological data and the expansion of reliable, automated ecological data processing will allow us to build systems that can efficiently answer increasingly detailed ecological questions at scale. They will go beyond merely identifying the species in a given image to answering questions about their number, their age, the behaviors they are exhibiting, and how they are interacting with the environment [37]. As we build robust data collection, processing, and management systems across different data modalities we can share context and fill in spatial, temporal, and taxonomic data gaps. We can aggregate information from remote sensing, passive and active monitoring sensors, ecological samples, and the natural history record to paint a cohesive picture of global biodiversity in order to help conservation efforts become more effective.

1.5 How my research fits into the big picture

I develop computer vision methods that enable scientific understanding of life on earth, and I pioneer and solve novel challenges for computer vision. Working jointly with stakeholders, I deploy my methods to improve sustainability and conservation worldwide.

Computer vision (CV) can play a fundamental role in sustainability and conservation. We are currently witnessing an unprecedented loss of biodiversity [6], yet biodiversity is vital to sustainable development [16], public health [31], and mitigating climate change [39]. Earth observation, the gathering of information about the biological, physical, and chemical systems of the planet, is necessary for conservation and sustainability across spatial and temporal scales–micro to macro. I posit that CV, along with machine learning (ML) and data science (DS), will prove crucial to extract scientific insight from quickly-growing repositories of natural world data.

In order to make a difference in the fields of conservation and sustainability we must shift the CV paradigm. Currently, CV research focuses on designing methods tailored to highly curated datasets coming mostly from consumer applications. These curated datasets frequently fail to capture the complexity of the real world, resulting in methods that fall short when deployed even within these consumer applications. Further, consumer application datasets do not mirror the statistics of ecological data which presents a number of obstacles that current methods struggle to master, including strong spatiotemporal correlations, imperfect data quality, fine-grained categories, and long-tailed distributions. These challenges are shared by many real-world applications, and my methodological contributions confer benefit across domains [13].

My research program aims to empower AI-assisted scientific discovery in a changing world. I collaborate with diverse stakeholders–including governmental agencies, non-governmental organizations, conservation land managers, and local communities–allowing me to identify universal challenges in conservation and sustainability. This understanding allows me to create benchmarks that matter–that truly capture the complexity of these real-world problems and enable the research community to come together and tackle them. Notable improvements on the benchmarks I develop translate to real-world gains, and I design novel methods that make progress on these challenges. I work with industry partners to build accessible and inclusive human-AI systems enabling the widespread use of my methods and models by end users from geographic locations worldwide, variable access to computational resources, with diverse sets of prior expertise.

Learning from imperfect, limited data

Data captured by human photographers, the norm in computer vision benchmarks, usually contains well-framed, in-focus objects of interest. In contrast, biodiversity monitoring data is often collected from sensors with limited intelligence, such as camera traps which collect data based on motion triggers. This leads to pictures where the animals are too close or too far from the sensor, low resolution, or obscured. Humans use temporal information, sometimes over long time horizons, to confidently ID species in challenging images. However, camera trap data is collected with a motion trigger and low frame rate, leading to failures from most video-based temporal approaches which have heavy reliance on the significant frame-to-frame visual similarity that comes with high frame rates. We proposed Context R-CNN [13], which learns to attend to relevant context from up to a month of data at a single static sensor. Our model improves upon baseline methods by up to 17% mAP, and is able to correctly label cases with occlusion, poor lighting, even in severe fog. It is applicable to any static sensor–e.g., traffic cameras or home security systems. We are implementing Context R-CNN within the global-scale Wildlife Insights platform [5], where I serve as a core member of the AI team. At the time of writing this thesis Wildlife Insights provides data management and AI-assisted categorization for >900 users from 40 countries and has ingested over 20M camera trap images globally, and its use is growing rapidly. The distribution of species worldwide is long-tailed: most observations are of common species, and the vast majority of species have few, if any, observations. This results in highly-imbalanced datasets, with insufficient data to learn rare species accurately [8]. Rare, at-risk species can be the most important to accurately categorize, but their rarity makes collecting additional training images for those species challenging–14% of the International Union for Conservation of Nature (IUCN) Red List is considered data deficient [21]. We built a system to generate synthetic examples for rare species using modern game engines and used this synthetic data to decrease rare-class error by 70% without affecting performance on common species [12, 18, 27].

Measuring domain generalization under distribution shift

Generalization to novel domains poses a fundamental challenge for computer vision. Near-perfect accuracy on benchmarks is now common [29], but these models do not work as expected when deployed. For example, models trained on a set of static cameras do not generalize to new cameras. We built a systematic framework and benchmark for analyzing generalization performance [8]. Our work promoted a paradigm shift in how CV researchers build image recognition benchmarks for real-world applications, and our evaluation protocols have become the standard across data types including passive monitoring cameras [38], aerial surveys [23], and bioacoustics [19]. Challenges stemming from domain shift are ubiquitous in real-world problems, from medical diagnosis to code autocompletion. To better understand the generality of methods built to tackle these challenges *across domains*, we published the WILDS benchmark suite [25, 33].

Evaluation protocols that match as closely as possible the intended use of a system in its target domain are crucial for understanding potential impact. Via our evaluation framework, we found class-agnostic animal detection generalizes far better to new deployments than species categorization [8]. Motivated by the community's need, we built a robust, generalizable animal detection model for camera trap data, the MegaDetector [10]. Our open-source API has processed over 100M images to date, and has been integrated into the wildlife monitoring workflows of over 50 organizations, including The Nature Conservancy and San Diego Zoo Global. Our model is used as a key component in many computer vision papers for camera trap data [12, 13, 30, 32].

Building human-AI systems for efficient, active, lifelong learning

Effective human-AI systems make human experts efficient, allowing them to extract scientific insight from large datasets with minimal manual labeling. The human in the loop provides quality control, probing model performance in new regions and correcting mislabeled rare or out-of-sample categories. My work in this space focuses in three main directions. First, we explore active learning to efficiently categorize species in new static sensor deployments [30]. Our method matches fully supervised accuracy on a 3.2M image dataset with as few as 14K manual labels, decreasing manual labelling effort on new deployments by over 99.5%. Second, when a user is provided with a CV-based automated solution, such as species identification, they immediately come back with more questions: How old is it? Is it healthy? What is it doing? We must build computer vision systems that can work with experts to efficiently answer *new questions*. Towards this goal, we compared supervised and self-supervised representations learned from 2.7M iNaturalist species observation images on 164 novel ecologically-relevant tasks [37]. Third, we built and deployed ElephantBook [26], an AI-assisted elephant re-identification system that combines robust contour matching, metric learning, and human-in-the-loop attribute labeling to enable long-term monitoring of the shifting, open-set elephant population in the Greater Mara Ecosystem.

Connecting the AI and conservation research communities

Interdisciplinary communication and understanding is vital when developing automated methods for scientific fields. As interest in AI for conservation grew, I saw a need for a space where practitioners on both sides could share opportunities and find collaborators, so I launched a community, called AI for Conservation, that has grown to include over 1000 interdisciplinary researchers and conservation technologists worldwide as of August 2022. Dan Morris, head of Microsoft AI for Earth, says *"the 'AI for Conservation' community that Sara Beery launched has become the de facto rallying point for energy in this area, and it's where we point everyone that comes to us asking how they can get involved."* To bring conservation challenges to the attention of the CV community, I designed four distinct iWildCam challenges for the Fine-Grained Visual Categorization Workshop at CVPR [7, 9, 11, 14]. Each competition defined a difficult, multimodal task and over 500 teams of CV researchers have taken part to date. Bridging the gap between two communities requires an understanding of both, and the ability to translate fundamental concepts for both audiences. In the last year I have written a review of species distribution modeling aimed at machine learning practitioners [15], and a review of CV for biodiversity monitoring aimed at ecologists [35].

1.6 Breaking down the thesis, chapter by chapter

I will briefly summarize the chapters in this thesis and the relevant contributions. This work was done in collaboration with many others across academia, industry, governmental organizations, and nonprofits, and would not have been possible without the significant efforts of these collaborators. The relevant collaborators for each chapter are listed in the co-author lists for the associated publications included in that chapter, which appear at the beginning of the chapter texts.

In Chapter 2, my collaborators and I provide a broad overview of the interdisciplinary field of machine learning for wildlife conservation. We cover newly-developed sensors for collecting wildlife data, machine learning approaches for processing that data automatically, and attention points and opportunities moving forward.

In Chapter 3, we provide a review of ecological monitoring techniques, and species distribution modeling (SDM) in particular, aimed at machine learning practitioners. We cover different methods to represent the distribution of species, historic and current, commonly-used environmental covariates, properties of and algorithms for SDMs, and discuss the challenges of evaluation and open problems. We also contribute a useful guide to which datasets and covariates are publicly accessible and ready to be used to develop novel ML-based SDM approaches.

In Chapter 4, we formalize the challenge of generalization to new sensor location and present the Caltech Camera Traps dataset along with a systematic framework for evaluating gaps in generalization performance by providing two separate validation and test sets, one from sensor locations seen during training and one from novel sensor locations. We present baseline comparisons that demonstrate a significant drop in performance on the out-of-domain test data for species categorization on full images and on object-centric crops, even when temporally aggregating data across short time scales. We additionally demonstrate that class-agnostic object detection generalizes surprisingly well, which led to the development and deployment of the MegaDetector[10].

In Chapter 5, we present the iWildCam competition dataset and its evolution over the past 5 years. iWildCam was designed to provide yearly novel challenges for the computer vision community centered in real ecological needs. From empty/not empty categorization in 2018, open-set species categorization in 2019, multi-modal generalization in 2020, to species-specific counting in 2021, the competition has been run yearly at the Fine-Grained Visual Categorization Workshop at CVPR and has engaged over 600 competition teams worldwide. In recent years, iWildCam has expanded to be included as a core challenge in the WILDS benchmark, the first to provide real-world domain shift challenges across application domains, and later in WILDS 2.0, which extends WILDS to include unsupervised data.

In Chapter 6, we address the few-shot learning challenge for rare species and discuss our development of a novel camera trap data synthesizer to produce diverse synthetic data for rare species. We demonstrate that this synthetic data can be used to significantly improve rare-species categorization accuracy, without hurting the performance of the model on common classes. In follow up work led by students I mentored and not included in this thesis [18, 27], we further increased the efficacy of the synthetic data using generative adversarial networks to mimic the low-level image statistics of the real data and adapting a method for domain adaptation to handle the single-class synthetic data.

In Chapter 7, we present a novel method for incorporating long-term temporal context for per-camera object detection inspired by how ecologists label challenging images in camera traps. Instead of looking at just one image at a time, ecologists look at all of the data for one camera trap in chronological order. As they start to learn where to look and what to look for for a specific sensor, they become much better at identifying species in images with heavy occlusion from foliage or poor weather conditions. We build an attention-based approach that enables the model to make predictions with access to up to a month of context for any one sensor, and show that our method significantly improves performance over single-frame baselines and baselines designed for video.

In Chapter 8, we present the Auto Arborist Dataset, a multiview fine-grained visual categorization dataset that contains over 2 million trees in 23 cities across the US and Canada built to foster the development of robust methods for large-scale urban forest monitoring. Our Auto Arborist dataset contains over 2.5M trees and 344 genera and is >2 orders of magnitude larger than the closest dataset in the literature. We introduce baseline results on our dataset across modalities as well as metrics for the detailed analysis of generalization with respect to geographic distribution shifts, vital for such a system to be deployed at-scale.

In Chapter 9, we present ElephantBook, a human-AI system for elephant population monitoring that combines human attribute labeling with computer vision approaches in order to robustly and accurately identify individual elephants. Our system has been deployed in the Greater Mara Ecosystem by the Mara Elephant Project since January 2021, and we can currently identify over 1000 elephants.

In Chapter 10, we discuss pitfalls and risks when using automated data processing methods such as computer vision or machine learning for ecological applications and present a case study on the risks of publishing biodiversity data that contains images of at-risk species.

Finally, in Chapter 11, I discuss potential directions for future work.

BIBLIOGRAPHY

[1] iNaturalist. https://www.inaturalist.org/.

[2] Workshop and Challenge on Computer Vision for Wildlife Conservation at ICCV. https://cvwc2019.github.io/, 2019.

[3] AI for Animal Re-Identification Workshop at WACV. https://sites.google.com/corp/view/wacv2020animalreid/, 2020.

[4] Fine Grained Visual Categorization Workshop at CVPR. http://www.fgvc.org/, 2022.

[5] Jorge A. Ahumada, Eric Fegraus, Tanya Birch, Nicole Flores, Roland Kays, Timothy G. O'Brien, Jonathan Palmer, Stephanie Schuttler, Jennifer Y. Zhao, Walter Jetz, et al. Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. *Environmental Conservation*, 47(1):1–6, 2020.

[6] Rosamund Almond, Monique Grooten, and Tom Peterson. Living Planet Report 2020-Bending the curve of biodiversity loss. *World Wildlife Fund*, 2020.

[7] Sara Beery, Grant Van Horn, Oisin MacAodha, and Pietro Perona. The iWildCam 2018 challenge dataset. *The Fifth Fine-Grained Visual Categorization Workshop at CVPR*, 2018.

[8] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[9] Sara Beery, Dan Morris, and Pietro Perona. The iWildCam 2019 challenge dataset. *The Sixth Fine-Grained Visual Categorization Workshop at CVPR*, 2019.

[10] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *In the Data Mining and Artificial Intelligence for Conservation Workshop at Knowledge Discovery in Databases (KDD)*, 2019. *selected to be featured at KDD Earth Day.

[11] Sara Beery, Elijah Cole, and Arvi Gjoka. The iWildCam 2020 competition dataset. *The Seventh Fine-Grained Visual Categorization Workshop at CVPR*, 2020.

[12] Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 863–873, 2020.

[13] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context R-CNN: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2020.

[14] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iWildCam 2021 competition dataset. *The Eighth Fine-Grained Visual Categorization Workshop at CVPR*, 2021.

[15] Sara Beery*, Elijah Cole*, Joseph Parker, Pietro Perona, and Kevin Winner. Species distribution modeling for machine learning practitioners: A review. *Proceedings of the 4th ACM SIGCAS Conf. on Computing and Sustainable Societies*, 2021.

[16] ODDS Cf. Transforming our world: The 2030 agenda for sustainable development. *United Nations*, 2015.

[17] Elijah Cole, Benjamin Deneu, Titouan Lorieul, Maximilien Servajean, Christophe Botella, Dan Morris, Nebojsa Jojic, Pierre Bonnet, and Alexis Joly. The geolifeclef 2020 dataset. *arXiv preprint arXiv:2004.04192*, 2020.

[18] Tuhin Das, Robert-Jan Bruintjes, Attila Lengyel, Jan van Gemert, and Sara Beery. Domain adaptation for rare classes augmented with synthetic samples. *arXiv preprint arXiv:2110.12216*, 2021.

[19] Félix Gontier, Vincent Lostanlen, Mathieu Lagrange, Nicolas Fortin, Catherine Lavandier, and Jean-François Petiot. Polyphonic training set synthesis improves self-supervised urban sound classification. *The Journal of the Acoustical Society of America*, 149(6):4309–4326, 2021.

[20] Andrew P. Hill, Peter Prince, Jake L. Snaddon, C. Patrick Doncaster, and Alex Rogers. Audiomoth: A low-cost acoustic device for monitoring biodiversity and the environment. *HardwareX*, 6:e00073, 2019.

[21] SSC IUCN. The IUCN red list of threatened species. *Version 3*, 2017.

[22] Benjamin Kellenberger, Diego Marcos, and Devis Tuia. Detecting mammals in uav images: Best practices to address a substantially imbalanced dataset with deep learning. *Remote Sensing of Environment*, 216:139–153, 2018.

[23] Benjamin Kellenberger, Diego Marcos, Sylvain Lobry, and Devis Tuia. Half a percent of labels is enough: Efficient animal detection in uav imagery using deep cnns and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):9524–9533, 2019.

[24] Brandon Key, James Miller, and Jiaqi Huang. Operational plan: Kenai river chinook salmon sonar assessment at river mile 13.7, 2020–2022, 2020.

[25] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[26] Peter Kulits, Jake Wall, Anka Bedetti, Michelle Henley, and Sara Beery. Elephantbook: A semi-automated human-in-the-loop system for elephant re-identification. In *ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 88–98, 2021.

[27] Edoardo Lanzini and Sara Beery. Image-to-image translation of synthetic samples for rare classes. *The Computer Vision for Animals Workshop at CVPR*, 2021.

[28] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9596–9606, 2019.

[29] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25): E5716–E5725, 2018.

[30] Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution*, 12 (1):150–161, 2021.

[31] World Health Organization et al. Connecting global priorities: Biodiversity and human health. *World Health Organization and Secretariat of the Convention on Biological Diversity*, 2015.

[32] Omiros Pantazis, Gabriel Brostow, Kate Jones, and Oisin Mac Aodha. Focus on the positives: Self-supervised learning for biodiversity monitoring. *arXiv preprint arXiv:2108.06435*, 2021.

[33] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the WILDS benchmark for unsupervised adaptation. *International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2112.05090.

[34] Shah Selbe. Fieldkit: Open environmental sensing for everyone. https://www.fieldkit.org/, 2022.

[35] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature Communications*, 13(1):1–15, 2022.

[36] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

[37] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. *arXiv preprint arXiv:2103.16483*, 2021.

[38] Robin C Whytock, Jędrzej Świeżewski, Joeri A Zwerts, Tadeusz Bara-Słupski, Aurélie Flore Koumba Pambo, Marek Rogala, Laila Bahaa-el din, Kelly Boekee, Stephanie Brittain, Anabelle W Cardoso, et al. Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution*, 12(6):1080–1092, 2021.

[39] Kathy J Willis and Shonil A Bhagwat. Biodiversity and climate change. *Science*, 326(5954):806–807, 2009.

Chapter 2

OVERVIEW AND LITERATURE REVIEW OF MACHINE LEARNING FOR WILDLIFE CONSERVATION

Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature Communications*, 13(1):1–15, 2022.

2.1 Abstract

Inexpensive and widely-available sensors are accelerating data acquisition in animal ecology. These technologies hold great potential for large-scale ecological understanding, but are limited by current processing approaches–such as manual expert data categorization–which inefficiently convert raw data into relevant and usable information. We argue that animal ecologists can capitalize on large datasets generated by modern sensors by combining machine learning approaches with domain knowledge. Incorporating machine learning into ecological workflows could improve inputs for ecological models and lead to integrated hybrid modeling tools. This approach will require close interdisciplinary collaboration to ensure the quality of novel approaches and train a new generation of data scientists in ecology and conservation. This chapter provides a broad overview and literature review of the field as it currently stands, features several existing successful projects and emphasizes what made them successful, and highlights future directions across disciplines that have potential to make significant impact.

2.2 Technology to accelerate ecology and conservation research

Animal diversity is declining at an unprecedented rate [33]. This loss comprises not only genetic, but also ecological and behavioral diversity, and is currently not well understood: out of more than 120,000 species prioritized for monitoring by the IUCN Red List of Threatened Species, up to 17,000 have a ‘Data deficient’ status [36]. We urgently need tools for rapid assessment of wildlife diversity and population dynamics at large scale and high spatiotemporal resolution, from individual animals to global densities. In this chapter we aim to build bridges across ecology and machine learning to highlight how relevant advances in technology can be leveraged to rise to this urgent challenge in animal conservation. We also point out major challenges and opportunities for future technological development in this space.

How are animals currently monitored? Conventionally, management and conservation of animal species are based on data collection carried out by human field workers who count animals, observe their behavior, and/or patrol natural reserves. Such efforts are time-consuming, labor-intensive and expensive [176]. They can also result in biased datasets due to challenges in controlling for observer subjectivity and assuring high inter-observer consistency, and often unavoidable responses of animals to observer presence [30, 113]. Human presence in the field also poses risks to wildlife [58, 95], their habitats [171], and humans themselves: as an example, many wildlife and conservation operations are performed from aircraft, and plane crashes are the primary cause of mortality for wildlife biologists [138]. Finally, the physical and cognitive limitations of humans unavoidably constrain the number of individual animals that can be observed simultaneously, the temporal resolution and complexity of data that can be collected, and the extent of physical area that can be effectively monitored [5, 82].

These limitations considerably hamper our understanding of geographic ranges, population densities and community diversity globally, as well as our ability to assess the consequences of their decline. For example, humans conducting counts of seabird colonies [73] and bats emerging from cave roosts [20] tend to significantly underestimate the number of individuals present. Furthermore, population estimates based on extrapolation from sparsely sampled locations have large uncertainties and can fail to capture the spatiotemporal variation in ecological relationships, resulting in erroneous predictions or extrapolations [135]. Failure to monitor animal populations impedes rapid and effective management actions [176]. For example, insufficient monitoring, due in part to the difficulty and cost of collecting the necessary data, has been identified as a major challenge in evaluating the impact of primate conservation actions [78] and can lead to the continuation of practices that are harmful to endangered species [146]. Similarly, poaching prevention requires intensive monitoring of vast protected areas, a major challenge with existing technology. Protected area managers invest heavily in illegal intrusion prevention and the detection of poachers. Despite this, rangers often arrive too late to prevent wildlife crime from occurring [119]. In short, while a rich tradition of human-based data collection provides the basis for much of our understanding of where species are found, how they live, and why they interact, modern challenges in wildlife ecology and conservation are highlighting the limitations of these methods.

Advances in sensor technologies are drastically increasing data collection capacity by reducing costs and expanding coverage relative to conventional methods (see Section 2.3), thereby opening new avenues for ecological studies at scale (Figure 2.1) [96]. Many previously inaccessible areas of conservation interest can now be studied through the use of high-resolution remote sensing [59], and large amounts of data are being collected non-invasively by digital devices such as camera traps [149], consumer cameras [70] and acoustic sensors [152]. New on-animal bio-loggers, including miniaturized tracking tags [15, 173] and sensor arrays featuring accelerometers, audiologgers, cameras, and other monitoring devices document the movement and behavior of animals in unprecedented detail [68], enabling researchers to track individuals across hemispheres and over their entire lifetimes at high temporal resolution and thereby revolutionizing the study of animal movement (Figure 2.1c) and migrations.

Modern ecology studies produce more data than ecologists can analyze manually. Effectively, ecology has entered the age of big data and is increasingly reliant on sensors, advanced methodologies, and computational resources [52]. Central challenges to efficient data analysis are the sheer volume of data generated by modern collection methods and the heterogeneous nature of many ecological datasets, which preclude the use of simple automated analysis techniques [52]. Crowdsourcing platforms like eMammal (emammal.si.edu), Agouti (agouti.eu) and Zooniverse (www.zooniverse.org) function as collaborative portals to collect data from different projects and provide tools to volunteers to annotate images *e.g.*, with species labels of the individuals therein. Such platforms drastically reduce the cost of data processing (*e.g.*, [97] reports a reduction of seventy thousand dollars, a 23.6% decrease in cost), but the rapid increase in the volume and velocity of data collection is making such approaches unsustainable. For example, in August 2021 the platform Agouti hosted 31 million images, of which only 1.5 million were annotated. This is because images are still annotated by hand, with a human reviewing every image, and this manual effort does not scale at the same rate as data collection. We need automated methods for cataloging, searching, and converting data into relevant information in order to increase human efficiency and keep up with incoming data. These have the potential to broaden and enhance animal ecology and wildlife conservation in scale and accuracy and pave the way forward towards new, integrated research directives.

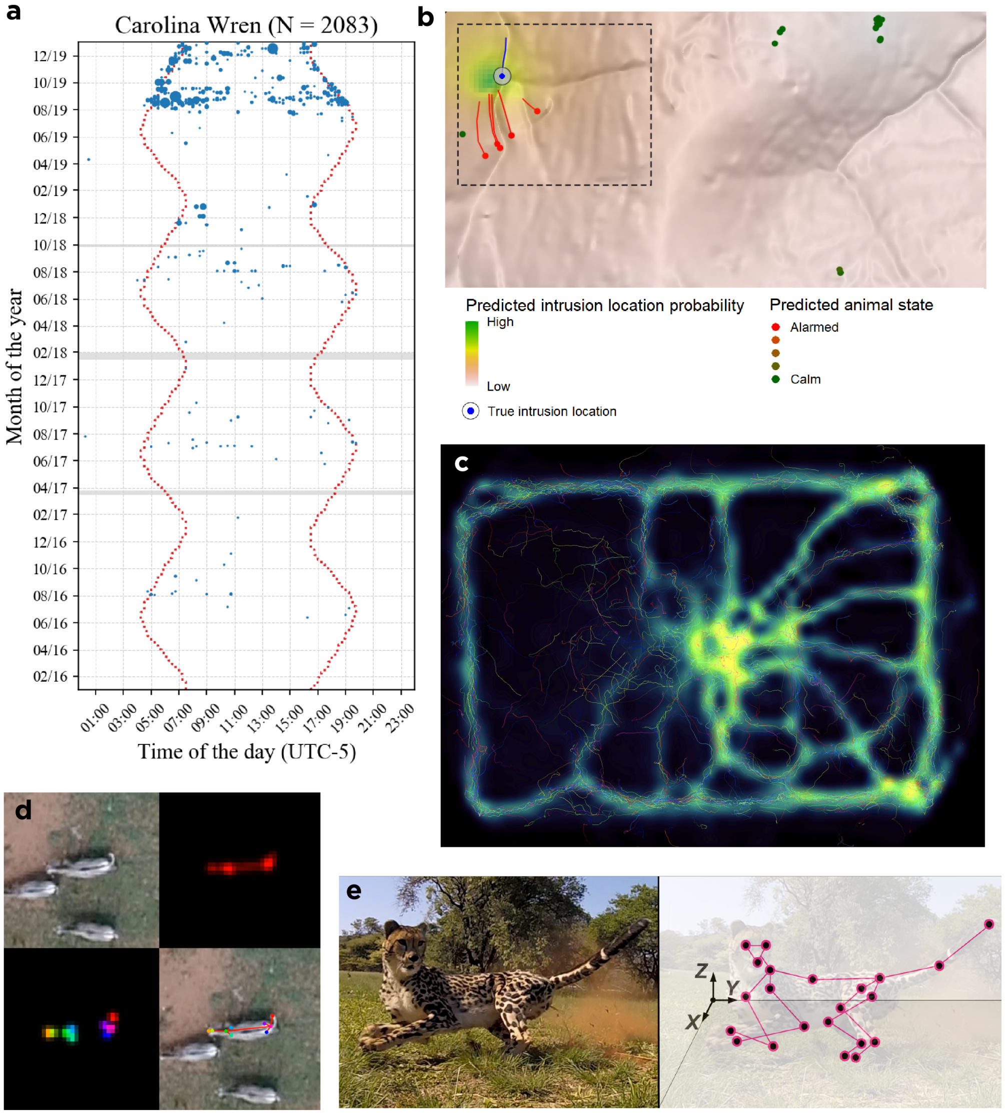


Figure 2.1: Examples of research acceleration by machine learning-based systems in animal ecology. **a:** The BirdNET algorithm [79] was used to detect Carolina wren vocalizations in more than 35,000 hours of passive acoustic monitoring data from Ithaca, New York, allowing researchers to document the gradual recovery of the population following a harsh winter season in 2015 (H. Klinck, unpublished). **b:** Machine-learning algorithms were used to analyze movement of Savannah herbivores fitted with bio-logging devices in order to identify human threats. The method can localize human intruders to within 500m, suggesting 'sentinel animals' may be a useful tool in the fight against wildlife poaching [41]. **c:** TRex, a new image-based tracking software, can track the movement and posture of hundreds of individually-recognized animals in real-time. Here the software has been used to visualize the formation of trails in a termite colony [164]. **d, e:** Pose estimation software, such as DeepPoseKit (d) [60] and DeepLabCut (e) [77, 110] allows researchers to track the body position of individual animals from video imagery, including drone footage, and estimate 3D postures in the wild. See Acknowledgements for credits and permissions.

Machine learning (ML) deals with learning patterns from data [69]. Presented with large quantities of inputs (*e.g.*, images) and corresponding expected outcomes, or labels (*e.g.*, the species depicted in each image), a supervised ML algorithm learns a mathematical function leading to the correct outcome prediction when confronted with new, unseen inputs. When the expected outcomes are absent, the (this time unsupervised) ML algorithm will use solely the inputs to extract groups of data points corresponding to typical patterns in the data. ML has emerged as a promising means of connecting the dots between big data and actionable ecological insights [35] and is an increasingly popular approach in ecology [93, 94]. A significant share of this success can be attributed to deep learning (DL [98]), a family of highly versatile ML models based on artificial neural networks that have shown superior performance across the majority of ML use cases (see Table 2.2). Significant error reduction of ML and DL with respect to traditional generalized regression models has been reported routinely for species richness and diversity estimation [87, 130]. Likewise, detection and counting pipelines moved from rough rule of thumb extrapolations from visual counts in national parks to ML-based methods with high detection rates. Initially, these methods proposed many false positives which required further human review [133], but recent methods have been shown to maintain high detection rates with significantly fewer false positives [12]. As an example, large mammal detection in the Kuzikus reserve in 2014 was improved significantly by improving the detection methodologies, from a recall rate of 20% [133], to 80% [84] (for a common 75% precision rate). Finally, studies involving human operators demonstrated that ML enabled massive speedups in complex tasks such as individual and species recognition [45, 142] and large-scale tasks such as animal detection in drone surveys [86]. Recent advances in ML methodology could accelerate and enhance various stages of the traditional ecological research pipeline (see Figure 2.2), from targeted data acquisition to image retrieval and semi-automated population surveys. As an example, the initiative Wildlife Insights [4] is now processing millions of camera trap images automatically (17 million in August 2021), providing wildlife conservation scientists and practitioners with the data necessary to study animal abundances, diversity and behavior. Besides pure acceleration, use of ML also massively reduces analysis costs, with reduction factors estimated between 2 and 10 [49].

A growing body of literature supports the use of ML in ecology by educating domain experts about ML approaches [35, 156, 169] for analyzing big data [52, 128], and for ecological inference (*e.g.*, understanding the processes underlying ecological

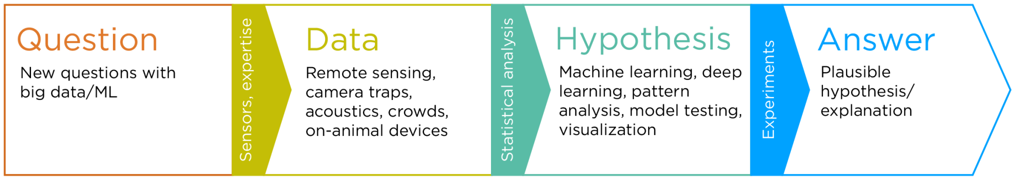

Question
New questions with big data/ML
Sensors, expertise
Data
Remote sensing, camera traps, acoustics, crowds, on-animal devices
Statistical analysis
Hypothesis
Machine learning, deep learning, pattern analysis, model testing, visualization
Experiments
Answer
Plausible hypothesis/ explanation

Figure 2.2: Traditional ecological research pipeline (colored text and boxes) and contributions of ML to the different stages discussed in this paper (black text).

patterns, rather than only predicting the patterns themselves) [107, 181]. Ecological data is challenging, it has strong spatiotemporal biases and is often sparsely sampled, the distribution of species in the data is long-tailed, the signal to noise ratios in the data can be very low, and many different modalitites of data are collected from diverse sensing platforms across spatial scales (each requiring different expertise and contextual understanding to process). These challenges are compounded by the size of the datasets generated by larger-scale sensor deployments and novel sensors–*e.g.*, increasingly small and lightweight animal tags [76], DNA sequencing on the edge via Nanopores [47]. Methods that address these challenges will require a *collaborative approach* that harnesses the expertise of both the ML and animal ecology communities. The rising interdisciplinary field of hybrid environmental algorithms (leveraging both deep learning and bio-physical models [31, 132]) and, more broadly, by theory-guided data science [80] has generated models which are less data-intensive, avoid incoherent predictions, and are generally more interpretable than purely data-driven models. The relation between ecology and ML should similarly not be unidirectional: integrating ecological domain knowledge into ML methods is essential to designing models that are accurate in the way they describe animal life. With this objective in mind, we review recent efforts at the interface of the two disciplines, present success stories of such symbiosis in animal ecology and wildlife conservation, and sketch an agenda for the future of the field.

2.3 New sensors expand available data types for animal ecology

Sensor data provide a variety of perspectives to observe wildlife, monitor populations and understand behavior. They allow larger studies in space, time, and across the taxonomic tree and, thanks to open science projects (Table 2.1), to share data across parks, geographies and the globe [122]. Sensors generate diverse data types, including imagery, soundscapes, and positional data (Figure 2.3). They can be mobile or static, and can be deployed to collect information on individuals or species

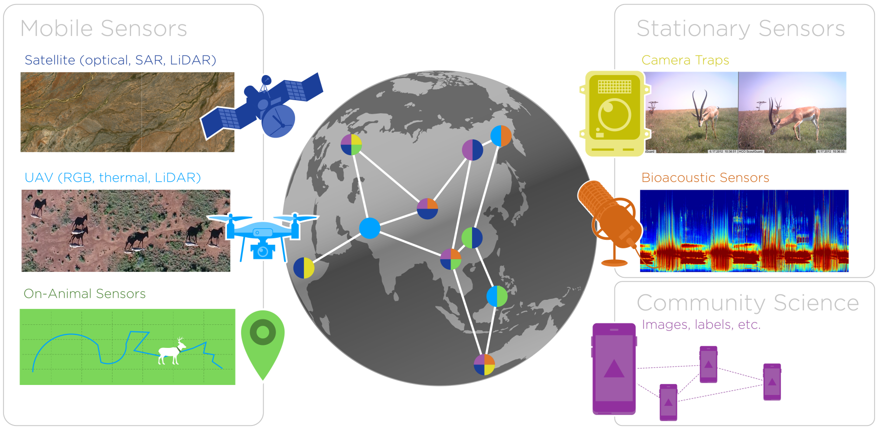


Figure 2.3: A variety of sensors used in animal ecology. Studies frequently combine data from multiple sensors at the same geographic location, or data from multiple locations to achieve deeper ecological insights.

of interest (*e.g.*, bio-loggers, drones), monitor activity in a particular location (*e.g.*, camera traps and acoustic sensors), or document changes in habitats or landscapes over time (satellites, drones). Finally, they can also be opportunistic, as in the case of community science. Below, we discuss the different categories of sensors and the opportunities they open for ML-based wildlife research.

Stationary sensors. Stationary sensors provide close-range continuous monitoring over long time scales. Common stationary sensors used in ecology include motion triggered or timelapse camera traps, passive and active bioacoustic sensors, and environmental sensors such as rain gauges and temperature sensors. These sensors collect data vith varying temporal frequency, from high frame rate video to daily captures, and are used to record species presence/absence, identify individuals, analyze behavior, and study predator-prey interaction. However, because of their stationary nature, their data is highly spatiotemporally correlated. Based on where and when in the world the sensor is placed, there is a limited number of species that can be captured. Furthermore, many animals are highly habitual and territorial, leading to very strong correlations between data taken days or even weeks apart from a single sensor [13].

- *Camera traps* are among the most used sensors in recent ML-based animal ecology papers, with more than a million cameras already used to monitor biodiversity worldwide [149]. Camera traps are inexpensive, easy to install, and provide high-resolution image sequences of the animals that trigger them, sufficient to identify the species, sex, age, health, behavior, and predator-prey interactions. Coupled with population models, camera trap data has also been used to estimate species occurrence, richness, distribution and density [149]. Software tools such as TimeLapse [62] and AIDE [85] (see Table 2.2) have been developed to help ecologists more quickly annotate their camera trap datasets, which are often large (for example, the SnapShot Serengeti network collects 1M images/year [2]). Many ecologists have already incorporated open source ML approaches for filtering out blank images (such as the Microsoft AI4Earth MegaDetector [12], see Table 2.2 and Box 2.4) into their camera trap workflows [13, 61, 118, 141]. However, automated species identification systems struggle to generalize to new sensor deployment locations and new sensor types, and require large numbers of labeled examples for every species [11]. Organizations like Wildlife Insights (www.wildlifeinsights.org) and LILA.science (www.lila.science) are making it easier for researchers to share their data, which is making it easier to curate diverse, large-scale ML training datasets across regions and taxa.

- *Bioacoustic sensors* are an alternative to camera traps, using microphones and hydrophones to study vocal animals and their habitats [152]. Networks of static bioacoustic sensors, used for passive acoustic monitoring (PAM), are increasingly applied to monitor wildlife in terrestrial [178], aquatic [44], and marine [40] ecosystems. Compared to camera traps, PAM is mostly unaffected by light and weather conditions (some factors like wind still play a role), senses the environment omnidirectionally, and is cost-effective when data needs to be collected at large spatial and temporal scales with high resolution [177]. While ML has been extensively applied to camera trap images, its application to long-term PAM datasets is still in its infancy and the first deep-learning-based studies are only starting to appear (see Fig 2.1a, [79]). Significant challenges remain when utilizing PAM. First and foremost among these challenges is the size of data acquired. Given continuous and high-frequency acquisition rates, datasets often exceed the terabyte scale–the National Centers for Environmental Information (NCEI) are a recently-established U.S. national archive for passive acoustic data which already contain over 100TB of data [163]. Handling and analyzing these datasets efficiently requires access to advanced computing infrastructure and solutions. Second, the inherent complexity of soundscapes requires noise-robust algorithms that generalize well and can separate and identify many animal sounds of interest from confounding natural and anthropogenic signals in a wide variety of acoustic environments [150]. The third challenge is the lack of large and diverse labeled datasets. As for camera trap images, species- or region-specific characteristics (*e.g.*, regional dialects [55]) affect algorithm performance. Robust, large-scale datasets have begun to be curated for some animal groups (*e.g.*, www.macaulaylibrary.org and www.xeno-canto.org for birds), but for many animal groups as well as relevant biological and non-biological confounding signals, such data is still nonexistent.

Remote sensing. Collecting data on free ranging wildlife has been restricted traditionally by the limits of manual data collection (*e.g.*, extrapolating transect counts), but have increased greatly through the automation of remote sensing [133]. Using remote sensing, *i.e.*, sensors mounted on moving platforms such as drones, aircraft, or satellites–or attached to the animals themselves–allows us to monitor large areas and track animal movement over time.

- *On-animal sensors* are the most common remote sensing devices deployed in animal ecology [82]. They are primarily used to acquire movement trajectories (*i.e.*, GPS data) of animals, which can then be classified into activity types that relate to the behavior of individuals or social groups [75, 82]. Secondary sensors, such as microphones, video cameras, heart rate monitors and accelerometers, allow researchers to capture environmental, physiological, and behavioral data concurrently with movement data [175]. However, power supply and data storage and transmission limitations of bio-logging devices are driving efforts to optimize sampling protocols or pre-process data in order to conserve these resources and prolong the life of the devices. For example, on-board processing solutions can use data from low-cost sensors to identify behaviors of interest and engage resource-intensive sensors only when these behaviors are being performed [89]. Other on-board algorithms classify raw data into behavioral states to reduce the volume of data to be transmitted [180]. Various supervised ML methods have shown their potential in automating behavior analysis from accelerometer data [28, 106], identifying behavioral state from trajectories [165] and predicting animal movement [172].

- *Satellite data* is widely available globally, and machine learning and deep learning methods have been developed for satellite data analysis across many applications, including land cover and land use categorization, agricultural monitoring, and chlorophyl concentration estimation [31, 183]. Public programs such as Landsat and Sentinel provide free and open imagery at medium resolution (between 10 and 30 m per pixel). This resolution, though usually not sufficient for direct wildlife observations, can be useful for studying their habitats [43, 87]. Commercial very high resolution imagery (with resolutions less than one meter per pixel), is opening up opportunity for direct observation of large animals such as whales [63] or elephants [48], though there are significant challenges due to occlusion from, for example, weather or trees. When focusing on smaller-bodied species, studies often focus on the detection of proxies instead of the detection of the animal itself (*e.g.*, the detection of penguin droppings to locate colonies [56]). Image bands beyond the visible spectrum are also available from many satellite imaging systems, and are widely utilized in plant ecology [27]. Multi- and hyperspectral deep learning approaches [8] are yet to be widely used in animal ecology. They could potentially contribute to the characterization of habitats and the detection of wildlife colonies such as walrus haulouts [54]. The main bottlenecks to the widespread use of commercial very high resolution imagery is the cost of purchasing the data at large scale, though platforms such as Planet sometimes donate imagery to ecological research, and the computational and systems infrastructure required to process high resolution satellite imagery from large regions.

- *Unmanned aerial vehicles (UAVs)*, or drones, capture imagery at lower-altitudes and significantly higher resolutions (with some platforms providing sub-millimeter resolution), and have been highlighted as a promising technology for animal conservation [72, 104]. Recent studies have shown the promise of UAVs and deep learning for posture tracking [60, 110, 111], semi-automatic detection of large mammals [49, 83], birds [86] and, in low altitude flight, even identification of individuals [6]. Drones are agile platforms that can be deployed rapidly–theoretically on demand–and with limited cost, making them useful for local population monitoring. Lower altitude flights in particular can provide oblique view points that partially mitigate occlusion by vegetation. The reduced costs and operation risks of UAVs further make them an increasingly viable alternative to low-flying manned aircraft.

  Common multi-rotor UAV models are built using inexpensive hardware and consumer-level cameras, and only require a trained pilot with flight permissions to perform the survey. Fully autonomous UAV platforms are also being explored in order to remove the need for a trained pilot [6]. However, multi-rotor drone-based surveys have limited spatial footprint primarily due to battery limitations (which become even more stringent in cold climates like Antarctica). Though these are of course dependent on the UAV platform, as an example the widely commercially available DJI Mavic 3 series can fly a maximum of 19 horizontal meters per second and has a maximum flight time of 46 minutes without wind and at moderate temperatures [46]. Local legislation further restricts the possible horizontal footprint of multi-rotor drone-based surveys [121]. Using drones also has a risk of modifying the behavior of the animals. A recent study [144] showed that flying at lower altitudes (*e.g.*, lower than 150 m) can have a significant impact on group and individual behavior of mammals, although the severity of wildlife disturbance from drone deployments will depend heavily on the focal species, the equipment used, and characteristics of the drone flight (such as approach speed and altitude) [17]–this is a rapidly changing field and advances that will limit noise are likely to come. More research to quantify and qualify such impacts in different ecosystems is timely and urgent, to avoid both biased conclusions and increased levels of animal stress. Combustion-driven fixed wing UAVs flying at high altitudes and human-piloted airplane-based acquisitions reduce possible behavioral impacts and can cover much larger horizontal footprints, but capture lower-resolution imagery due to the increased distance as well as encounter weather and vegetation-based occlusion that can limit high-resolution visual measurements on animals.

Community science for crowd-sourcing data

An alternative to traditional sensor networks (static or remote) is to engage community members as wildlife data collectors and processors [24, 114, 162], which also increases public engagement in science and conservation. In this case, community participants (often volunteers) work to collect the data and/or create the labels necessary to train ML models. Models trained this way can then be used to bring image recognition tasks to larger scale and complexity, from filtering out images without animals in camera trap sequences to identifying species or even individuals. Several annotation projects based on community science have appeared recently (Table 2.1). For simple tasks like animal detection, community science effort can be open to the public, while for more complex ones such as identifying bird species with subtle appearance differences ("fine-grained classification," also see the glossary), communities of experts are needed to provide accurate labels. A particularly interesting case is Wildbook (see the text box on Page 35 and Table 2.2), which routinely screens videos from social media platforms with computer vision models to identify individuals; community members (in this case video posters) are then queried in case of missing or uncertain information. Recent research shows that ML models trained on community data can perform as well as annotators [155]. However, it is prudent to note that the viability of community science services may be limited depending on the task and that oftentimes substantial efforts are required to verify volunteer-annotated data [157]. This is due to annotator errors, including misdetected or mislabeled animals due to annotator fatigue or insufficient knowledge about the annotation task, as well as systematic errors from adversarial annotators [23, 90, 147]. Another form of community science is the usage of images acquired by volunteers: in this case, volunteers replace camera traps or UAVs and provide the raw data used to train the ML model. Although this approach sacrifices control over image acquisitions and is likewise prone to inducing significant noise to datasets, for example through low-quality imagery, it provides a substantial increase in the number of images and the chances of photographing species or single individuals in different regions, poses and viewing angles. The Great Grevy's Rally, a community science-based wildlife census effort occurring every two years in Kenya [124], is a successful demonstration of the power of community science-based wildlife monitoring via volunteer-acquired images.

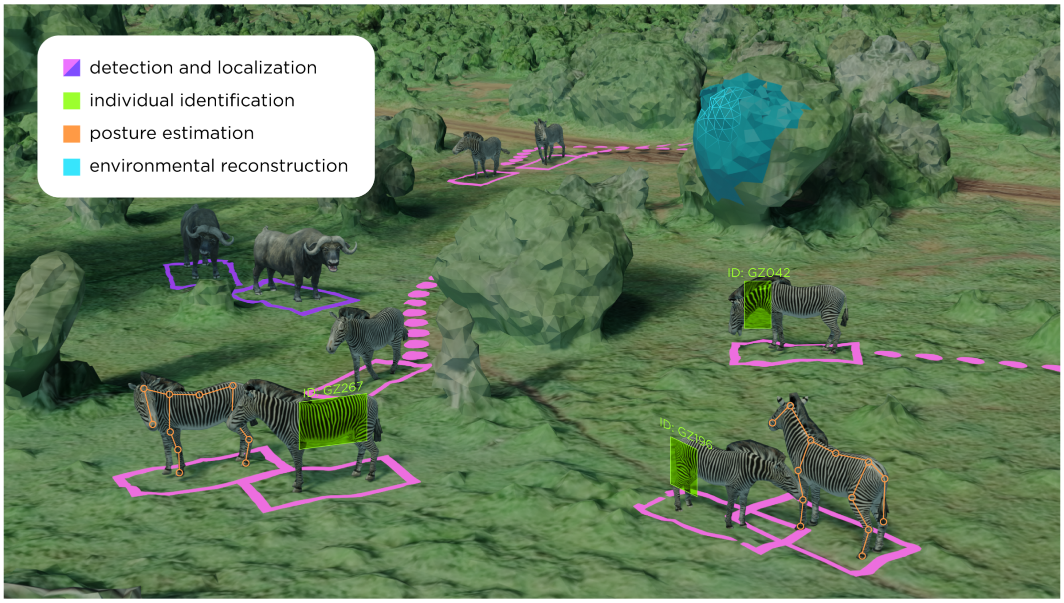

detection and localization
individual identification
posture estimation
environmental reconstruction
ID: GZ042
ID: GZ267
ID: GZ196

Figure 2.4: **Setting a common vocabulary:** ecology tasks vs corresponding ones in computer vision. Imagery can be used to capture a range of behavioral and ecological data, which can be processed into usable information with ML tools. Aerial imagery (from drones, or satellites for large species) can be used to localize animals and track their movements over time (pink and purple), and model the 3D structure of landscapes using photogrammetry (blue). Posture estimation tools allow researchers to estimate animal postures (orange), which can then be used to infer behaviors using clustering algorithms. Finally, computer vision techniques allow for the identification and re-identification of known individuals across encounters (green).

Table 2.1: Examples of community science projects in digital wildlife conservation

Name	Spatial coverage	Sensor	Task	Impact
iNaturalist[158]	Global	Human photographers	Classification, Detection	>100M observations collected by volunteers, near-global coverage, primary data contributor to GBIF [1]
SAVMAP[120]	Kuzikus Reserve, Namibia	UAV images	Detection	Near real-time ultrahigh-resolution drone imaging and data labeling to monitor biodiversity and land use at a large nature reserve.
Zooniverse[147]	Global	Images, Text, Video	Classification, Detection	Enables >2.5M volunteers to contribute >700M data labels and counting for hundreds of different projects [167].
iRecord[131]	United Kingdom	Photographic records	Classification	Collects images of species from volunteers across the UK, has >2M observations of >15K species.
Great Grevy's Rally[124]	Northern Kenya	Safari pictures	Classification, Detection, Identification	One of the only complete population censuses of a species, via the combination of CV and targeted volunteer data collection. Used by the Kenyan government to set conservation policy and allocate resources [18].

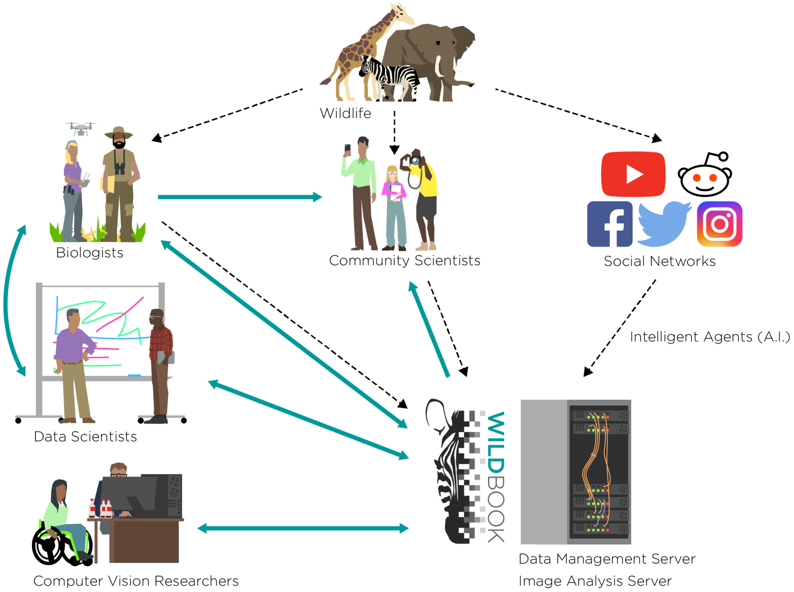


Figure 2.5: The Wildbook Ecosystem. Wildbook allows scientists and wildlife managers to leverage the power of communities and machine learning to monitor wildlife populations. Images of target species are collected via research projects, community science events (*e.g.*, the Great Grevy's Rally; see text), or by scraping social media platforms using Wildbook AI tools. Wildbook software uses computer vision technology to process the images, yielding species and individual identities for the photographed animals. This information is stored in databases on Wildbook data management servers. The data and biological insights generated by Wildbook facilitates exchange of expertise between biologists, data scientists, and stakeholder communities around the world.

Box 1: Wildbook: successes at the interface between community science and deep learning

Wildbook, a project of the non-profit Wild Me, is an open source software platform that blends structured wildlife research with artificial intelligence, community science, and computer vision to speed population analysis and develop new insights to help conservation (Figure 2.5). Wildbook supports collaborative mark-recapture, molecular ecology, and social ecology studies, especially where community science and artificial intelligence can help scale up projects. The image analysis of Wildbook can start with images from any source — scientists, camera traps, drones, community scientists, or social media — and use ML and computer vision to detect multiple animals in the images [125] to not only classify their species, but identify individual animals applying a suite of different algorithms [7, 168]. Wildbook provides a technical solution for wildlife research and management projects for non-invasive individual animal tracking, population censusing, behavioral and social population studies, community engagement in science, and building a collaborative research network for global species. There are currently Wildbooks for over 50 species, from sea dragons to zebras, spanning the entire planet. More than 80 scientific publications have been enabled by Wildbook. Wildbook data has become the basis for the IUCN Red List global population numbers for several species, and supported the change in conservation status for whale sharks from "vulnerable" to "endangered." Wildbook's technology also enabled the Great Grevy's Rally, the first ever full species census for the endangered Grevy's zebra in Kenya, using photographs captured by the public. Hosted for the first time in January 2016, it has become a regular event, held every other year. Hundreds of people, from school children and park rangers, to Nairobi families and international tourists, embark on a mission to photograph Grevy's zebras across its range in Kenya, capturing approximately 50,000 images over the two-day event. With the ability to identify individual animals in those images, Wildbook can enable an accurate population census and track population trends over time. The Great Grevy's Rally has become the foundation of the Kenya Wildlife Service's Grevy's zebra endangered species management policy and generates the official IUCN Red List population numbers for the species. Wildbook's AI enables science, conservation and global public engagement by bringing communities together and working in partnership to provide solutions that people trust.

2.4 Machine learning to scale-up and automate animal ecology and conservation research

The sensor data described in the previous section and collected at increasingly large spatial and temporal scales has the potential to unlock ecological understanding on a much larger scale, moving from local studies of a single species in a single protected area at a single point in time to (potentially one day) global-scale studies across hundreds of thousands of species in near-real time. To move towards this goal, we need systems that automatically interpret data and convert it to usable information for ecological research. For example, such conversion can take the form of abundance mapping, individual animal re-identification, herd tracking, or digital reconstruction (three dimensional, phenotypical) of the environment the animals live in. The measures yielded by this conversion, reviewed in this section, are also sometimes referred to as animal biometrics [91]. Interestingly, the tasks involved in the different approaches show similarities with traditional tasks in ML and computer vision (*e.g.*, detection, localization, identification, pose estimation), for which we provide a matching example in animal ecology in Figure 2.4.

Wildlife detection and species-level classification

Conservation efforts of endangered species require knowledge on how many individuals of the species in question are present in a study area. Such estimations are conventionally realized with statistical occurrence models that are informed by sample-based species observations often collected via imaging sensors (camera traps, UAVs, *etc.*), and converting the raw sensor data into species observations is a significant bottleneck. Early works attempted to automate this process with classical supervised ML algorithms such as support vector machines [71] (see Supplementary Table 2): these algorithms were used to make the connection between a set of characteristics of interest extracted from the image (visual descriptors, *e.g.*, color histograms, spectral indices, *etc.*, also see the nomenclature) and human-labeled annotations (presence of an animal, species, *etc.*) [133, 182]. Particularly in camera trap imagery, motion-based foreground (animal) segmentation was occasionally performed as a pre-processing step to discard image parts that were potentially confusing for a classifier [115]. These classical approaches suffered from two major limitations: first, the visual descriptors often needed to be hand-crafted for the problem and dataset at hand, and frequently could not be used effectively outside of that specific dataset without significant manual parameter tuning as they were specific to the associated environmental conditions (*e.g.*, camera type, background foliage amount and movement type), a challenge known in ML as "domain adaptation" or "generalization." Secondly, these methods were often computationally expensive, which limited the amount of data that could be used to train the models at the expense of limiting variations in data (temporal, seasonal, *etc.*), thus further reducing the generalization capabilities to new sensor deployments or regions.

Modern computer vision approaches such as deep convolutional neural networks (CNNs [123], which often outperform classical machine learning approaches by a significant margin and remove the need for handcrafted feature engineering) are widely used for species detection and classification in images [51, 118, 140, 154, 161, 174], acoustic spectrograms [79, 108], and videos [74, 139]. Models that have been shown to perform accurately and robustly across projects and do not need to be retrained for every new user–such as the MegaDetector, a model which detects animals in camera trap data; see Table 2.2 and the text box on page 39–are used widely and integrated within open-source systems helping ecologists to efficiently process and label their data. Modern computer vision models still struggle to generalize outside of their “domain" (the locations, distributions, and types of data that they were trained on) [11, 88], and much of the current research in automated wildlife detection and species-level classification focuses on building benchmark datasets which enable the study of how and why these models fail to generalize and working to build methods and models that work reliably out-of-domain [12, 13, 81, 170].

Table 2.2: Resources for machine and deep learning-based wildlife conservation

Name	Description	URL
AIDE [85]	**Tasks: Annotation; detection; classification; segmentation** Free, open source, web-based, collaborative labeling platform specifically designed for large-scale ecological image analyses. Users can concurrently annotate up to billions of images with labels, points, bounding boxes, or pixel-wise segmentation masks. AIDE tightly integrates ML models through Active Learning [145], where annotators are asked to provide inputs where the model is the least confident. AIDE further offers functionality to share and exchange trained ML models with other users of the system for collaborative annotation efforts in image campaigns across the globe.	GitHub
MegaDetector [12]	**Tasks: Detection** Free and open source detector based on deep learning hosted by Microsoft AI4Earth. The current model is trained with the TensorFlow Object Detection API using several hundred thousand camera trap images labeled with bounding boxes from a variety of ecosystems. The model identifies animals (not species-specific), humans, and vehicles, and is robust to novel sensor deployment locations and taxa not seen during training. Updates of the model, trained with additional data, are periodically released. Microsoft AI4Earth provides support to assist ecologists in using the model, including a public API for batch inference, and integration with commonly-used camera trap data management platforms such as TimeLapse and Camelot.	GitHub
Wildbook [19]	**Tasks: Individual Re-Identification** Wildbook blends structured wildlife research with artificial intelligence, community science, and computer vision to speed population analysis and develop new insights to help fight extinction. They host community-run individual re-identification systems and global data repositories for a broad and expanding set of species, including Grevy's Zebra, Whale sharks, Manta Rays, and many more.	URL
Wildlife Insights [4]	**Tasks: Filtering** Large-scale platform for camera trap data management with computer vision in the backend. Currently open for whitelisted users, extensible via a waitlist. Wildlife Insights filters blank images and provides species identification for images that the computer vision model scores highly, allowing expert ecologists to focus on labeling only challenging images.	URL
DeepLabCut [110]	**Tasks: Pose estimation and behavioral analysis.** Free and open source pose estimation toolbox based on deep learning. Pre-trained models (for instance for primate faces and bodies, as well as quadruped) as well as a light-weight, real-time version are available.	GitHub
DeepPoseKit [60]	**Tasks: Pose estimation and behavioral analysis.** Free and open source pose estimation toolbox based on deep learning.	GitHub

Box 2: Box 2. AI for Wildlife Conservation in Practice: the MegaDetector

One highly-successful example of open source AI for wildlife conservation is the Microsoft AI for Earth MegaDetector [12] (Figure 2.6). This generic, global-scale human, animal, and vehicle detection model works off-the-shelf for most camera trap data, and the publicly-hosted MegaDetector API has been integrated into the wildlife monitoring workflows of over 30 organizations worldwide, including the Wildlife Conservation Society, San Diego Zoo Global, and https://www.islandconservation.org/. We would like to highlight two MegaDetector use cases, via Wildlife Protection Solutions (WPS) and the Idaho Department of Fish and Game (IDFG). WPS uses the MegaDetector API in real-time to detect threats to wildlife in the form of unauthorized humans or vehicles in protected areas. WPS connect camera traps to the cloud via cellular networks, upload photos, run them through the MegaDetector via the public API, and return real-time alerts to protected area managers. They have over 400 connected cameras deployed in 18 different countries, and that number is growing rapidly. WPS used the MegaDetector to analyze over 900K images last year alone, which comes out to 2.5K images per day. They help protected areas detect and respond to threats as they occur, and detect at least one real threat per week across their camera network. Idaho is required to maintain a stable population of protected wolves. IDFG relies heavily on camera traps to estimate and monitor this wolf population, and need to process the data collected each year before the start of the next season in order to make informed policy changes or conservation interventions. They collected 11 million camera trap images from their wolf cameras last year, and with the MegaDetector integrated into their data processing and analysis pipeline, they were able to fully automate the analysis of 9.5 million of those images, using model confidence to help direct human labeling effort to images containing animals of interest. Using the Megadetector halved their labeling costs, and allowed IDFG to label all data before the start of the next monitoring season, whereas manual labeling previously resulted in a lag of approximately five years from image collection to completion of labeling. The scale and speed of analysis required in both cases would not be possible without such an AI-based solution.

Individual re-identification

Another important biometric is animal identity. The standard for identification of animal species and identity is DNA profiling [9], which can be difficult to scale to large, distributed populations [91, 141]. As an alternative to gene-based identification, manual tagging can be used to keep track of individual animals [82, 91]. Similar to counting and reconstruction (see next section), computer vision recently emerged as a powerful alternative for automatic individual identification [19, 125, 141, 159]. Identifying individuals from images is more challenging than species recognition, since the distinctive body patterns of individuals might be subtle or not be visible due to occlusion, motion blur, or overhead viewpoint in the case of aerial imagery.

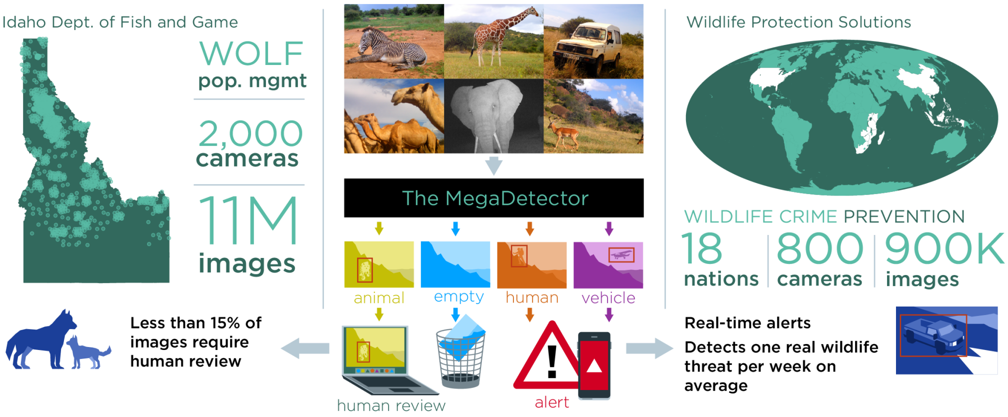


Figure 2.6: AI for Wildlife Conservation in Practice: the MegaDetector. The near-universal need of all camera trap projects to efficiently filter empty images and localize humans, animals, and vehicles in camera trap data, combined with the robustness to geographic, hardware, and species variability the MegaDetector provides due to its large, diverse training set makes it a useful, practical tool for many conservation applications out of the box. The work done by the Microsoft AI for Earth team to provide assistance running the model via hands-on engineering assistance, open source tools, and a public API have made the MegaDetector accessible to ecologists and a part of the ecological research workflow for over 60 organizations worldwide.

Conventional [168] and more recently deep-learning-based [29, 141, 142] methods have reached strong performance for some taxa, especially across small populations. Some species have individually-unique coat or skin markings that assist with re-identification: for example, tigers [102], whalesharks [7], or zebra [126]. However, effective re-identification is also possible in the absence of patterned markings: a study of a small group of twenty-three chimpanzees in Guinea applied facial recognition techniques to a 14-year video dataset comprising 20,000 tracked faces exctracted from 50 hours of video (a total of 10M face images) and achieved $> 90\%$ accuracy [142] on tracks from held-out years of data not seen during training, though it should be noted that this is the best-case scenario for evaluation: a small, closed set of individuals in a single location (*i.e.* no unknown individuals were evaluated against) with very large sets of labeled data for each individual, and multiple images over a short period of time which provide different angles and illumination on the face for each individual video track. This study compared their model to manual re-identification by humans on a subset of 100 images from their dataset: where humans achieved identification accuracy between 20% (novices) and 42% (experts), the model achieved an identification accuracy of 84% on the subset.

Animal (re)-identification in open, wild populations has several particular challenges. It is difficult to curate representative, accurately manually labeled datasets for training and evaluation, population sizes are large, animals change in appearance (*e.g.*, due to scars, growth), and there are often very few sightings per individual. Perhaps most significantly, the populations change over time due to birth, death, and immigration, therefore creating an “open-set” problem [16] wherein the model must deal with “classes” (individuals) unseen during training. The methods must have the ability to identify not only animals that have been seen just once or twice but also recognize new, previously unseen animals, as well as adjust decisions that have been made in the past, reconciling different views and biological stages of an animal.

Animal synthesis and reconstruction

3D shape recovery and pose estimation of animals can provide valuable, non-invasive insights on wild species in their natural environment. The 3D shape of an individual can be related to its health, age or reproductive status; the 3D pose of the body can provide finer information with respect to posture attributes and allows, for instance, kinematic as well as behavioral analyses. For pose estimation, marker-less methods based on DL have tremendously improved over the last years and already impacted biology [112]. Various user-friendly toolboxes are available to extract the 2D posture of animals from videos (Fig. 2.1d,e), while the user can define which body parts should be estimated (reviewed in [111]). Extracting a dense set of body surface points is also possible, as elegantly shown in [137], where the DensePose technique originally developed for humans was extended to chimpanzees. The reconstruction of the 3D shape and pose of animals from images often follows a model-based paradigm, where a 3D model of the animal is fit to visual data. Recent work defines the SMAL (Skinned Multi Animal Linear) model, a 3D articulated shape model for a set of quadruped families [184]. Biggs et al. built on this work for 3D shape and motion of dogs from video [21] and for recovery of dog shape and pose across many different breeds [22]. In [185] the SMAL model has been used in a DL approach to predict 3D shape and pose of the Grevy’s zebra from images. 3D shape models have been recently defined also for birds [166]. Image-based 3D pose and shape estimation methods provide rich information about individuals but require prior knowledge about the animal’s shape and 3D motion.

Reconstructing the environment

Wildlife behavior and conservation cannot be dissociated from the environment animals evolve and live in. Studies have shown that animal observations like trajectories highly benefit from additional cues included in the environmental context [67]. Satellite remote sensing has become an integral part to study animal habitats, biological diversity and spatio-temporal changes of abiotic conditions [129], since it allows to map quantities like land cover, soil moisture or temperature at scale. Reconstructing the 3D shape of the environment has also become central in behavior studies: for example, 3D reconstructions of kill sites for lions in South Africa revealed novel insights into the predator-prey relationships and their connection to ecosystem stability and functioning [39], while 3D spatial reconstructions shed light on the impact of forest structures on bat behavior [57]. Such spatial reconstructions of the environment can either be extracted by using dedicated sensors such as LiDAR [134] or can be reconstructed from multiple images, either by stitching the images into a unified two-dimensional panorama (*e.g.*, mosaicking [66]) or by computing the three-dimensional environment from partially overlapping images (*e.g.*, Structure from Motion [143] or Simultaneous Localization and Mapping [116]). All these approaches have strongly benefited from recent ML advancements [92], but have seldom been applied for wildlife conservation purposes, where they could greatly help when dealing with images acquired by moving or swarms of sensors [105]. However, applying these techniques to natural wildlife imagery is not trivial. For example, unconstrained continuous video recordings at potentially high frame-rates will result in large image sets which require efficient image processing [66]. Moreover, ambiguous environmental appearances and structural errors such as drift accumulate over time and therefore decrease the reconstruction quality [143]. Last but not least, a variety of inappropriate camera motions or environmental geometries can result in so-called critical configurations which cannot be resolved by the existing optimization schemes [179]. As a consequence, cues from additional external sensors are usually integrated to achieve satisfactory environmental reconstructions from video data [53].

Modeling species diversity, richness and interactions

Analyses of biodiversity, represented by such measures as species abundance and richness, are foundational to much ecological research and many conservation initiatives. Spatially explicit linear regression models have been conventionally used to predict species and community distribution based on explanatory variables such as climate and topography [64, 100]. Non-parametric ML techniques like Random Forest [26] have been successfully used to predict species richness and have significantly reduced error with respect to the traditional counterparts used in ecology, for example in the estimation of richness distributions of fishes [127, 148], spiders [32], and small mammals [10]. Tree-based techniques have also been used to predict species interactions: for example, regression trees significantly outperformed classical generalized linear models in predicting plant-pollinator interactions [130]. Tree-based methods are well-suited to these tasks because they perform explicit feature ranking (and thus feature selection) and are able to model nonlinear relationships between covariates and species distribution. More recently, graph regression techniques were deployed to reconstruct species interaction networks in a community of European birds with promising results, including better causality estimates of the relations in the graph [50].

2.5 Attention points and opportunities

Machine and deep learning are becoming necessary accelerators for wildlife research and conservation actions in natural reserves. We have discussed success stories of the application of approaches from ML into ecology and highlighted the major technical challenges ahead. In this section, we want to present a series of "attention points" that highlight new opportunities between the two disciplines.

What can we focus on now?

State-of-the-art ML models are now being applied to many tasks in animal ecology and wildlife conservation. However, while an out-of-the-box application of existing open tools is tempting, there are a number of points and potential pitfalls that must be carefully considered to ensure responsible use of these approaches.

- *Inherent model biases and generalization.* Most ecological datasets suffer from some degree of geographic bias. For example, many open imagery repositories such as Artportalen.se, Naturgucker.de and Waarneming.nl collect images from specific regions, and most contributions on iNaturalist [158] (see Table 2.1) come from the Northern hemisphere. Such biases need to be understood, acknowledged and communicated to avoid incorrect usage of methods or models that by design may only be accurate in a specific geographic region. Biases are not limited to the geographical provenance of images: the type of sensors used (RGB *vs.* infrared or thermal), the species they depict and the imbalance in the number of individuals observed per species [11, 158] must also be considered when training or using models to avoid potentially catastrophic drop-offs in accuracy, and transparency around the training data and the intended model usage is a necessity [37].

- *Curating and publishing well-annotated benchmark datasets without doing harm.* The long-term advancement of the field will ultimately require the curation of large, diverse, accurately labeled, and publicly available datasets for ecological tasks with defined evaluation metrics and maintained code repositories. However, opening up existing datasets (and especially when using private-owned images acquired by non-professionals as in [124]) is both a necessary and difficult challenge for the near future. Fostering a culture of individual and cross-institutional data sharing in ecology will allow ML approaches to improve in robustness and accuracy. Furthermore, proper credit has to be given to the data collectors, for example through appropriate data attribution and Digital Object Identifiers (DOIs) for datasets [37].

- *Understanding the ethical risks involved.* Computer scientists must also be aware of the ethical and environmental risks of publishing certain types of datasets. It is important to understand the limits of open data sharing in animal conservation in nature parks. In some cases it is imperative that the privacy of the data be preserved, for instance to avoid giving poachers access to locations of animals in near-real-time [101]. Security of rangers themselves is also at stake; for example the flight path of drones might be backtracked to reveal their location.

- *Standards of quality control are urgently needed.* Accountability for open models needs to be better understood. The estimations of models remain approximations and need to be treated as such: population counts without uncertainty estimation can lead to erroneous and potentially devastating conclusions. Increased quality control on the adequacy of a model to a new scientific question or study area is important and can be achieved by close co-operation between model developers (who have the ability to design, calibrate, and run the models at their best) and practitioners (who have the domain and local knowledge). Without such quality control measures, relying on model-based results is risky and could have difficult-to-evaluate impacts on research in animal ecology, as incorrect results hidden in a suboptimally trained model will become more and more difficult to detect. Computer scientists must be aware that errors by their models can lead to erroneous decisions on site that can be catastrophic for the population they are trying to preserve or for the populations that live at the border of human/wildlife conflicts.

- *Environmental and financial costs of machine learning.* ML is not free. Training and running models with millions of parameters on large volumes of data requires powerful, somewhat specialized hardware. Purchasing prices of such machines alone are often prohibitively high especially for budget-constrained conservation organizations; programming, running and maintenance costs further add to the bill. Although cloud computing services exist that forgo the need of hardware management, they likewise pose per-resource costs that quickly scale to several thousands of dollars per month for a single virtual machine. Besides monetary costs, ML also uses significant amounts of energy: recently, it has been estimated that training a large, state-of-the-art model for understanding natural language emits as much carbon as several cars in their entire lifetime [151]. It is of course important to put these carbon costs into perspective, and consider the societal benefit posed by a model to determine whether this cost is justified. However, as many of these large language models have been shown to carry biases that pose significant risk to minoritized groups [3, 103, 117, 160], the benefit to society of a model can be both controversial and difficult to quantify. Further, the environmental costs of AI are often disregarded or ignored, as energy consumption of large calculations is often considered an endless resource (assuming that the money to pay for it is available). The models currently used in animal ecology are far from such a carbon footprint, but as model and data size grown it is important to take the environmental costs into account–we do not want to exchange one source environmental harm (loss of biodiversity) for another (increase of emissions and energy consumption). Particular care needs to be paid to designing models that are not oversized and that can be trained efficiently. Smaller models are not only less expensive to train and use, their lighter computational costs allow them to be run on smaller devices, opening opportunities for real-time ML “on the edge”–*i.e.*, within the sensors themselves.

What’s new: vast scientific opportunities lie ahead

In the previous sections, we describe the advances in research at the interface of ML, animal ecology and wildlife conservation. The maturity of the various detection, identification and recognition tools opens a series of interesting perspectives for genuinely novel approaches that could push the boundaries towards true integration of the disciplines involved.

- *Involving domain knowledge from the start.* The ML and DL fields have focused mainly on black box models that learn correlations from data directly, and domain knowledge has been repeatedly ignored in favor of generic approaches that could fit to any kind of dataset. Such universality of ML is now strongly questioned and the inductive bias of traditional DL models is challenged by new approaches that bridge domain knowledge, fundamental laws and data science. This "hybrid models" paradigm [80, 132] is one of the most exciting avenues in modern ML and promises real collaboration between domains of application and ML, especially when coupled with algorithmic designs that allow interpretation and understanding of the visual cues that are being used [136]. This line of interdisciplinary research is small but growing, with several studies published in recent years. A representative one is Context R-CNN [13] for animal detection and species classification, which leverages the prior knowledge that backgrounds in camera trap imagery exhibit little variation over time and that camera traps acquire data with low sampling frequency and occasional dropouts. By integrating image features over long time spans (up to a month), the model is able to increase mean species identification precision in the Snapshot Serengeti dataset [153] by 17.9%. In another example [42], the hierarchical structure of taxonomies, as well as locational priors, are leveraged to constrain plant species classification from iNaturalist in Switzerland, leading to improvements of state-of-the-art models of about 5%. Similarly, [109] incorporate knowledge about the distribution of species as well as photographer biases into a DL model for species classification in images and report accuracy improvements of up to 12% in iNaturalist over a baseline without such priors. Finally, [65] used expert knowledge of park rangers to augment sparse and noisy records of poaching activity, thereby improving predictions of poaching occurrence and enabling more efficient use of limited patrol resources in a Chinese nature reserve. These approaches challenge the dogma of ML models learning exclusively from data and achieve more efficient model learning (since base knowledge is available from the start and does not have to be re-learnt) and enhanced plausibility of the solutions (because the solution space can be constrained to a range of ecologically plausible outcomes).

- *Laboratories as development spaces.* In recent years modern ML has rapidly changed laboratory-based non-invasive observation of animals [111, 112]. Neuroscience studies in particular have embraced novel tools for motion tracking, pose estimation (Figure 2.1d, e) and behavioral classification (*e.g.*, [38]). The high level of control (*e.g.*, of lighting conditions, sensor calibration, and environment) afforded by laboratory settings facilitated the rapid development of such tools, many of which are now being adopted for use in field studies of free-moving animals in complex natural environments [60, 77]. Additionally, algorithmic insights gained in the lab can be transferred back into the wild–studies on short videos or camera traps can leverage lab-generated data that is arguably less diverse, but easier to control.This opens interesting research opportunities for the adaptation of lab-generated simulation to real world conditions, similar to what has been observed in the field of image synthesis for self driving [34] and robotics [99] in the last decade. Thus, laboratories rightly serve as the ultimate development space for such in-the-wild applications.

- *Towards a new generation of biodiversity models.* Statistical models for species richness and diversity are routinely used to estimate abundances and study species co-occurrence and interactions. Recently, DL methods have also started to be employed to model species' ecological niches [25, 43], facilitated by the development of machine-learning-ready datasets such as GeoLifeCLEF. GeoLifeCLEF curated a dataset of 1.9 million iNaturalist observations from North America and France depicting over 31,000 species, together with environmental predictors (land cover, altitude, climatic data, *etc.*), and asked users to predict a ranked list of likely species per geospatial grid cell. The task is complex: only positive counts are provided, no absence data are available, and predictions are counted as correct if the ground truth species is among the 30 predicted with highest confidence. This challenging task remains an open challenge–the winners of the 2021 edition achieved only an approximate 26% top-30 accuracy.

  A recent review of species distribution modeling aimed at ML practitioners [14] provides an accessible entry point for those interested in tackling the challenges in this complex, exciting field. Open challenges include increasing the scale of joint models geospatially, temporally, and taxonomically, building methods that can leverage multiple data types despite bias from non-uniform sampling strategies, incorporating ecological knowledge such as species dispersal and community composition, and expanding methods for the evaluation of these models.

Finally, we wish to re-emphasize that the vision described here cannot be achieved without interdisciplinary thinking: for all these exciting opportunities, processing big ecological data is necessitating analytical techniques of such complexity that no single ecologist can be expected to have all the technical expertise (plus domain knowledge) required to carry out groundbreaking studies [175]. Cross-disciplinary collaborations are undeniably a critical component of ecological and conservation research in the modern era. Mutual understanding of the field-specific vocabularies, of the fields' expectations and of the implications and consequences of research ethics are within reach, but require open dialogues between communities, as well as cross-domain training of new generations.

2.6 Conclusions

Animal ecology and wildlife conservation research needs to make sense of large and ever-increasing streams of data to provide accurate estimations of populations, understand animal behavior, prevent poaching and mitigate biodiversity loss. Machine learning and deep learning are tools that can help scale local studies to a global understanding of the animal world.

In this chapter we presented a series of success stories at the interface of ML and animal ecology. We highlighted a number of performance improvements that were observed when adopting solutions based on ML and new-generation sensors. Such improvements require ever-closer cooperation between ecologists and ML specialists, since recent approaches are complex and require strict quality control and detailed design knowledge. We note the existence of useful ML-based tools for ecology stemming from corporate (*e.g.*, Wildlife Insights) and research (AIDE, MegaDetector, DeepLabCut) efforts, but that there is still much room (and need) for the development of new interdisciplinary methods, in particular hybrid models and new habitat and species distribution models at scale. Inspired by these observations, we provided our perspective on the missing links between animal ecology and ML via a series of attention points, recommendations and vision on future exciting research avenues.

We hope that this chapter can provide a jumping off point for students and researchers across both fields to understand the work in this space and to

find relevant literature for diverse applications of machine learning in wildlife ecology. We strongly incite the two communities to work hand-in-hand to find digital, scalable solutions that will elucidate the loss of biodiversity and its drivers and lead to global actions to preserve nature. Computer scientists have yet to integrate ecological knowledge such as underlying biological processes into ML models, and the lack of transparency of current DL models has so far been a major obstacle to incorporating ML into ecological research. However, an interdisciplinary community of computer scientists and ecologists is emerging, which we hope will tackle this technological and societal challenge together.

Acknowledgments

We thank Mike Costelloe for assistance with figure design and execution. SB would like to thank the Microsoft AI for Earth initiative, the Idaho Department of Fish and Game, and Wildlife Protection Solutions for insightful discussions and providing data for figures. MCC and TBW were supported by the National Science Foundation (IIS 1514174 & IOS 1250895). MCC received additional support from a Packard Foundation Fellowship (2016-65130), and the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship endowed by the Federal Ministry of Education and Research. CVS and TBW were supported by the US National Science Foundation (Awards 1453555 and 1550853). SB was supported by the National Science Foundation Grant No. 1745301 and the Caltech Resnick Sustainability Institute. IDC acknowledges support from the ONR (N00014- 19-1-2556), and IDC, BRC, MW and MCC from, the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's Excellence Strategy–EXC 2117-422037984. MWM is the Bertarelli Foundation Chair of Integrative Neuroscience.

References

[1] The Global Biodiversity Information Facility. `https://www.gbif.org/`.

[2] Snapshot Serengeti. Available online at `http://www.snapshotserengeti.org/`.

[3] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, 2021.

[4] Jorge A. Ahumada, Eric Fegraus, Tanya Birch, Nicole Flores, Roland Kays, Timothy G. O'Brien, Jonathan Palmer, Stephanie Schuttler, Jennifer Y. Zhao, Walter Jetz, et al. Wildlife insights: A platform to maximize the potential of

camera trap and other passive sensor wildlife data for the planet. *Environmental Conservation*, 47(1):1–6, 2020.

[5] Jeanne Altmann. Observational study of behavior: Sampling methods. *Behaviour*, 49(3-4):227–266, 1974.

[6] William Andrew, Colin Greatwood, and Tilo Burghardt. Aerial animal biometrics: Individual friesian cattle recovery and visual identification via an autonomous uav with onboard deep inference. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 237–243. IEEE, 2019.

[7] Zaven Arzoumanian, Jason Holmberg, and Brad Norman. An astronomical pattern-matching algorithm for computer-aided identification of whale sharks rhincodon typus. *Journal of Applied Ecology*, 42(6):999–1011, 2005.

[8] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Deep learning for classification of hyperspectral data: A comparative review. *IEEE Geoscience and Remote Sensing Magazine*, 7(2):159–173, 2019.

[9] John C. Avise. *Molecular markers, natural history and evolution*. Springer Science & Business Media, 2012.

[10] Andrew P. Baltensperger and Falk Huettmann. Predictive spatial niche and biodiversity hotspot models for small mammal communities in Alaska: Applying machine-learning to conservation planning. *Landscape Ecol*, page 17, 2015.

[11] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[12] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *In the Data Mining and Artificial Intelligence for Conservation Workshop at Knowledge Discovery in Databases (KDD)*, 2019. *selected to be featured at KDD Earth Day.

[13] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context R-CNN: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2020.

[14] Sara Beery*, Elijah Cole*, Joseph Parker, Pietro Perona, and Kevin Winner. Species distribution modeling for machine learning practitioners: A review. *Proceedings of the 4th ACM SIGCAS Conf. on Computing and Sustainable Societies*, 2021.

[15] Mikhail Y. Belyaev, Oleg N. Volkov, Olga N. Solomina, Johannes Weppler, Uschi Müller, Grigori M. Tertitski, Martin Wikelski, and Wolfgang Pitz. Development of technology for monitoring animal migration on earth using scientific equipment on the iss rs. In *2020 27th Saint Petersburg International Conference on Integrated Navigation Systems (ICINS)*, pages 1–7. IEEE, 2020.

[16] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1563–1572, 2016.

[17] Emily Bennitt, Hattie L.A. Bartlam-Brooks, Tatjana Y. Hubel, and Alan M. Wilson. Terrestrial mammalian wildlife responses to unmanned aerial systems approaches. *Scientific Reports*, 9(1):1–10, 2019.

[18] T. Y. Berger-Wolf, J. Crall, J. Holmberg, J. Parham, C. V. Stewart, B. Low Mackey, P. Kahumbu, and D. I. Rubenstein. The great grevy's rally: The need, methods, findings, implications and next steps. Report to the Kenya Wildlife Service, September 2016.

[19] Tanya Y. Berger-Wolf, Daniel I. Rubenstein, Charles V. Stewart, Jason A. Holmberg, Jason Parham, Sreejith Menon, Jonathan P. Crall, Jon Van Oast, Emre Kiciman, and Lucas Joppa. Wildbook: Crowdsourcing, computer vision, and data science for conservation. *CoRR*, abs/1710.08880, 2017. URL `http://arxiv.org/abs/1710.08880`.

[20] Margrit Betke, Diane E. Hirsh, Nicholas C. Makris, Gary F. McCracken, Marianne Procopio, Nickolay I. Hristov, Shuang Tang, Angshuman Bagchi, Jonathan D. Reichard, Jason W. Horn, et al. Thermal imaging reveals significantly smaller brazilian free-tailed bat colonies than previously estimated. *Journal of Mammalogy*, 89(1):18–24, 2008.

[21] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and smal: Recovering the shape and motion of animals from video. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 3–19. Springer, 2019.

[22] Benjamin Biggs, Oliver Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out? 3D animal reconstruction with expectation maximization in the loop. In *European Conference on Computer Vision*, pages 195–211. Springer, 2020.

[23] Tomas J. Bird, Amanda E. Bates, Jonathan S. Lefcheck, Nicole A. Hill, Russell J. Thomson, Graham J. Edgar, Rick D. Stuart-Smith, Simon Wotherspoon, Martin Krkosek, Jemina F. Stuart-Smith, et al. Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, 173:144–154, 2014.

[24] Rick Bonney. Citizen science at the cornell lab of ornithology. *Exemplary science in informal education settings: Standards-based success stories*, pages 213–229, 2007.

[25] Christophe Botella, Alexis Joly, Pierre Bonnet, François Munoz, and Pascal Monestiez. Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. *Methods in Ecology and Evolution*, 12(5):933–945, 2021.

[26] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[27] Philip G. Brodrick, Andrew B. Davies, and Gregory P. Asner. Uncovering ecological patterns with convolutional neural networks. *Trends in Ecology & Evolution*, 34(8):734–745, 2019.

[28] Ella Browning, Mark Bolton, Ellie Owen, Akiko Shoji, Tim Guilford, and Robin Freeman. Predicting animal behaviour using deep learning: Gps data alone accurately predict diving in seabirds. *Methods in Ecology and Evolution*, 9(3):681–692, 2018.

[29] Clemens-Alexander Brust, Tilo Burghardt, Milou Groenenberg, Christoph Kading, Hjalmar S Kuhl, Marie L Manguette, and Joachim Denzler. Towards automated visual monitoring of individual gorillas in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2820–2830, 2017.

[30] Gordon M. Burghardt, Julia N. Bartmess-LeVasseur, Sheri A. Browning, Kathleen E. Morrison, Courtney L. Stec, Christopher E. Zachau, and Todd M. Freeberg. Perspectives: Minimizing observer bias in behavioral studies: A review and recommendations. *Ethology*, 118(6):511–517, 2012.

[31] Gustau Camps-Valls, M Reichstein, Z Xiaoxiang, and D Tuia. *Deep learning for the Earth Sciences: A comprehensive approach to remote sensing, climate science and geosciences*. Wiley & Sons, 2021. ISBN 978-1-119-64614-3.

[32] Klemen Čandek, Urška Pristovšek Čandek, and Matjaž Kuntner. Machine learning approaches identify male body size as the most accurate predictor of species richness. *BMC Biology*, 18(1):1–16, 2020.

[33] Gerardo Ceballos, Paul R. Ehrlich, and Peter H. Raven. Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. *Proceedings of the National Academy of Sciences*, 117(24):13596–13602, 2020.

[34] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE International Conference on Computer Vsion*, pages 1511–1520, 2017.

[35] Sylvain Christin, Éric Hervet, and Nicolas Lecomte. Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644, 2019.

[36] The IUCN Red List Committee. The IUCN Red List of Threatened Species - Strategic Plan 2017-2020. Technical report, IUCN, 2017.

[37] Kyle Copas, Tim Robertson, Serge Belongie, Christine Kaeser-Chen, Adam Hartwig, Chenyang Zhang, K. Chuan Tan, Yulong Liu, Denis Brulé, Cédric Deltheil, et al. Training machines to improve species identification using gbif-mediated datasets. In *AGU Fall Meeting Abstracts*, volume 2019, pages IN53C–0758, 2019.

[38] Sandeep Robert Datta, David J. Anderson, Kristin Branson, Pietro Perona, and Andrew Leifer. Computational neuroethology: A call to action. *Neuron*, 104(1):11–24, 2019.

[39] Andrew B. Davies, Craig J. Tambling, Graham I.H. Kerley, and Gregory P. Asner. Effects of vegetation structure on the location of lion kill sites in african thicket. *PloS One*, 11(2):e0149098, 2016.

[40] Genevieve E. Davis, Mark F. Baumgartner, Julianne M. Bonnell, Joel Bell, Catherine Berchok, Jacqueline Bort Thornton, Solange Brault, Gary Buchanan, Russell A. Charif, Danielle Cholewiak, et al. Long-term passive acoustic recordings track the changing distribution of north atlantic right whales (eubalaena glacialis) from 2004 to 2014. *Scientific Reports*, 7(1): 1–12, 2017.

[41] Henrik J. de Knegt, Jasper A.J. Eikelboom, Frank van Langevelde, W. François Spruyt, and Herbert H.T. Prins. Timely poacher detection and localization using sentinel animal movement. *Scientific Reports*, 11(1):1–11, 2021.

[42] Riccardo De Lutio, Yihang She, Stefano D'Aronco, Stefania Russo, Philipp Brun, Jan D Wegner, and Konrad Schindler. Digital taxonomist: identifying plant species in community scientists' photographs. *ISPRS Journal of Photogrammetry and Remote Sensing*, 182:112–121, 2021.

[43] Benjamin Deneu, Maximilien Servajean, Christophe Botella, and Alexis Joly. Evaluation of deep species distribution models using environment and co-occurrences. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 213–225. Springer, 2019.

[44] Camille Desjonquères, Toby Gifford, and Simon Linke. Passive acoustic monitoring as a potential tool to survey animal and ecosystem processes in freshwater environments. *Freshwater Biology*, 65(1):7–19, 2020.

[45] Ellen M. Ditria, Sebastian Lopez-Marcano, Michael Sievers, Eric L. Jinks, Christopher J. Brown, and Rod M. Connolly. Automating the Analysis of Fish Abundance Using Object Detection: Optimizing Animal Ecology With Deep Learning. *Frontiers in Marine Science*, 7:429, June 2020. ISSN 2296-7745. doi: 10.3389/fmars.2020.00429. URL `https://www.frontiersin.org/article/10.3389/fmars.2020.00429/full`.

[46] DJI. DJI Mavic 3 specs. `https://www.dji.com/mavic-3/specs`, 2022.

[47] Karlijn Doorenspleet, Lara Jansen, Saskia Oosterbroek, Oscar Bos, Pauline Kamermans, Max Janse, Erik Wurz, Albertinka Murk, and Reindert Nijland. High resolution species detection: Accurate long read edna metabarcoding of north sea fish using oxford nanopore sequencing. *bioRxiv*, 2021.

[48] Isla Duporge, Olga Isupova, Steven Reece, David W. Macdonald, and Tiejun Wang. Using very-high-resolution satellite imagery and deep learning to detect and count african elephants in heterogeneous landscapes. *Remote Sensing in Ecology and Conservation*, 7(3):369–381, 2021.

[49] Jasper A.J. Eikelboom, Johan Wind, Eline van de Ven, Lekishon M. Kenana, Bradley Schroder, Henrik J. de Knegt, Frank van Langevelde, and Herbert H.T. Prins. Improving the precision and accuracy of animal population estimates with aerial image object detection. *Methods in Ecology and Evolution*, 10(11):1875–1887, 2019.

[50] Ali Faisal, Frank Dondelinger, Dirk Husmeier, and Colin M. Beale. Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods. *Ecological Informatics*, 5(6):451–464, 2010.

[51] Greg Falzon, Christopher Lawson, Ka-Wai Cheung, Karl Vernes, Guy A Ballard, Peter JS Fleming, Alistair S Glen, Heath Milne, Atalya Mather-Zardain, and Paul D Meek. Classifyme: a field-scouting software for the identification of wildlife in camera trap images. *Animals*, 10(1):58, 2019.

[52] Scott S. Farley, Andria Dawson, Simon J. Goring, and John W. Williams. Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience*, 68(8):563–576, 2018.

[53] Jordi A.O.N Ferrer Plana, Armagan Elibol, Olivier Delaunoy, Nuno Gracias, and Rafael Garcia. Large-area photo-mosaics using global alignment and navigation data. In *Mts/IEEE Oceans Conference*, pages 1–9, 2007.

[54] Anthony S Fischbach and David C Douglas. Evaluation of satellite imagery for monitoring pacific walruses at a large coastal haulout. *Remote Sensing*, 13(21):4266, 2021.

[55] John K.B. Ford. Dialects. In *Encyclopedia of marine mammals*, pages 253–254. Elsevier, 2018.

[56] Peter T. Fretwell and Philip N. Trathan. Discovery of new colonies by sentinel2 reveals good and bad news for emperor penguins. *Remote Sensing in Ecology and Conservation*, 7(2):139–153, 2021.

[57] Jérémy S.P. Froidevaux, Florian Zellweger, Kurt Bollmann, Gareth Jones, and Martin K. Obrist. From field surveys to lidar: Shining a light on how bats respond to forest structure. *Remote Sensing of Environment*, 175:242–250, 2016.

[58] Melissa Giese. Effects of human activity on adelie penguin Pygoscelis adeliae breeding success. *Biological Conservation*, 75(2):157–164, 1996. ISSN 0006-3207. doi: https://doi.org/10.1016/0006-3207(95)00060-7. URL `https://www.sciencedirect.com/science/article/pii/0006320795000607`.

[59] T.K. Gottschalk, Falk Huettmann, and Manfred Ehlers. Thirty years of analysing and modelling avian habitat relationships using satellite imagery data: A review. *International Journal of Remote Sensing*, 26(12):2631–2656, 2005.

[60] Jacob M. Graving, Daniel Chae, Hemal Naik, Liang Li, Benjamin Koger, Blair R. Costelloe, and Iain D. Couzin. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife*, 8:e47994, 2019.

[61] Saul Greenberg. Automated image recognition for wildlife camera traps: Making it work for you. Technical report, 2020.

[62] Saul Greenberg, Theresa Godin, and Jesse Whittington. Design patterns for wildlife-related camera trap image analysis. *Ecology and Evolution*, 9(24): 13706–13730, 2019.

[63] Emilio Guirado, Siham Tabik, Marga L. Rivas, Domingo Alcaraz-Segura, and Francisco Herrera. Whale counting in satellite and aerial images with deep learning. *Scientific Reports*, 9(1):1–12, 2019.

[64] Antoine Guisan and Niklaus E. Zimmermann. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3):147–186, 2000.

[65] Swaminathan Gurumurthy, Lantao Yu, Chenyan Zhang, Yongchao Jin, Weiping Li, Xiaodong Zhang, and Fei Fang. Exploiting Data and Human Knowledge for Predicting Wildlife Poaching. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–8, Menlo Park and San Jose CA USA, June 2018. ACM. ISBN 978-1-4503-5816-3. doi: 10.1145/3209811.3209879. URL `https://dl.acm.org/doi/10.1145/3209811.3209879`.

[66] Lars Haalck and Benjamin Risse. Embedded dense camera trajectories in multi-video image mosaics by geodesic interpolation-based reintegration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1849–1858, 2021.

[67] Lars Haalck, Michael Mangan, Barbara Webb, and Benjamin Risse. Towards image-based animal tracking in natural environments using a freely moving camera. *Journal of neuroscience methods*, 330:108455, 2020.

[68] Roi Harel, J. Carter Loftus, and Margaret C. Crofoot. Locomotor compromises maintain group cohesion in baboon troops on the move. *Proceedings of the Royal Society B*, 288(1955):20210839, 2021.

[69] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[70] Anna Hausmann, Tuuli Toivonen, Rob Slotow, Henrikki Tenkanen, Atte Moilanen, Vuokko Heikinheimo, and Enrico Di Minin. Social media data can be used to understand tourists' preferences for nature-based experiences in protected areas. *Conservation Letters*, 11(1):e12343, 2018.

[71] Marti A. Hearst, Susan T. Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[72] Jarrod C. Hodgson, Shane M Baylis, Rowan Mott, Ashley Herrod, and Rohan H. Clarke. Precision wildlife monitoring using unmanned aerial vehicles. *Scientific Reports*, 6(1):1–7, 2016.

[73] Jarrod C. Hodgson, Rowan Mott, Shane M. Baylis, Trung T. Pham, Simon Wotherspoon, Adam D. Kilpatrick, Ramesh Raja Segaran, Ian Reid, Aleks Terauds, and Lian Pin Koh. Drones count wildlife more accurately and precisely than humans. *Methods in Ecology and Evolution*, 9(5):1160–1167, 2018.

[74] Ekram Hossain, SM Shaiful Alam, Amin Ahsan Ali, and M Ashraful Amin. Fish activity tracking and species identification in underwater video. In *2016 5th International Conference on Informatics, electronics and vision (ICIEV)*, pages 62–66. IEEE, 2016.

[75] Lacey F. Hughey, Andrew M. Hein, Ariana Strandburg-Peshkin, and Frants H. Jensen. Challenges and solutions for studying collective animal behaviour in the wild. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1746):20170005, 2018.

[76] Vikram Iyer, Rajalakshmi Nandakumar, Anran Wang, Sawyer B Fuller, and Shyamnath Gollakota. Living iot: A flying wireless platform on live insects.

In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–15, 2019.

[77] Daniel Joska, Liam Clark, Naoya Muramatsu, Ricardo Jericevich, Fred Nicolls, Alexander Mathis, Mackenzie W. Mathis, and Amir Patel. AcinoSet: A 3D pose estimation dataset and baseline models for Cheetahs in the wild, 2021.

[78] Jessica Junker, Silviu O. Petrovan, Victor Arroyo-RodrÍguez, Ramesh Boonratana, Dirck Byler, Colin A. Chapman, Dilip Chetry, Susan M. Cheyne, Fanny M. Cornejo, Liliana CortÉs-Ortiz, et al. A severe lack of evidence limits effective conservation of the world's primates. *BioScience*, 70(9): 794–803, 2020.

[79] Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236, 2021.

[80] Anuj Karpatne, Gowtham Atluri, James H Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.*, 29(10):2318–2331, 2017.

[81] Justin Kay, Peter Kulits, Suzanne Stathatos, Siqi Deng, Erik Young, Sara Beery, Grant Van Horn, and Pietro Perona. The caltech fish counting dataset: A benchmark for multiple-object tracking and counting. *arXiv preprint arXiv:2207.09295*, 2022.

[82] Roland Kays, Margaret C. Crofoot, Walter Jetz, and Martin Wikelski. Terrestrial animal tracking as an eye on life and planet. *Science*, 348(6240): aaa2478, 2015.

[83] Benjamin Kellenberger, Diego Marcos, and Devis Tuia. Detecting mammals in uav images: Best practices to address a substantially imbalanced dataset with deep learning. *Remote Sensing of Environment*, 216:139–153, 2018.

[84] Benjamin Kellenberger, Diego Marcos, and Devis Tuia. When a few clicks make all the difference: improving weakly-supervised wildlife detection in uav images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[85] Benjamin Kellenberger, Devis Tuia, and Dan Morris. Aide: Accelerating image-based ecological surveys with interactive machine learning. *Methods in Ecology and Evolution*, 11(12):1716–1727, 2020.

[86] Benjamin Kellenberger, Thor Veen, Eelke Folmer, and Devis Tuia. 21 000 birds in 4.5 h: efficient large-scale seabird detection with machine learning. *Remote Sensing in Ecology and Conservation*, 7(3):445–460, 2021.

[87] Anders Knudby, Ellsworth LeDrew, and Alexander Brenning. Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. *Remote Sensing of Environment*, 114(6):1230–1241, 2010.

[88] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[89] Joseph Korpela, Hirokazu Suzuki, Sakiko Matsumoto, Yuichi Mizutani, Masaki Samejima, Takuya Maekawa, Junichi Nakai, and Ken Yoda. Machine learning enables improved runtime and precision for bio-loggers on seabirds. *Communications biology*, 3(1):1–9, 2020.

[90] Margaret Kosmala, Andrea Wiggins, Alexandra Swanson, and Brooke Simmons. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, 14(10):551–560, 2016.

[91] Hjalmar S. Kühl and Tilo Burghardt. Animal biometrics: quantifying and detecting phenotypic appearance. *Trends in ecology & evolution*, 28(7): 432–441, 2013.

[92] Kavitha Kuppala, Sandhya Banda, and Thirumala Rao Barige. An overview of deep learning methods for image registration with focus on feature-based approaches. *International Journal of Image and Data Fusion*, 11(2):113–135, 2020.

[93] Roberta Kwok. Ai empowers conservation biology. *Nature*, 567(7746): 133–135, 2019.

[94] Roberta Kwok. Deep learning powers a motion-tracking revolution. *Nature*, 574(7776):137–139, 2019.

[95] Sophie Köndgen, Hjalmar Kühl, Paul K. N'Goran, Peter D. Walsh, Svenja Schenk, Nancy Ernst, Roman Biek, Pierre Formenty, Kerstin Mätz-Rensing, Brunhilde Schweiger, Sandra Junglen, Heinz Ellerbrok, Andreas Nitsche, Thomas Briese, W. Ian Lipkin, Georg Pauli, Christophe Boesch, and Fabian H. Leendertz. Pandemic Human Viruses Cause Decline of Endangered Great Apes. *Current Biology*, 18(4):260–264, 2008. ISSN 0960-9822. doi: https://doi.org/10.1016/j.cub.2008.01.012. URL `https://www.sciencedirect.com/science/article/pii/S0960982208000171`.

[96] José J. Lahoz-Monfort and Michael J.L. Magrath. A Comprehensive Overview of Technologies for Species and Habitat Monitoring and Conservation. *BioScience*, page biab073, July 2021. ISSN 0006-3568, 1525-3244. doi: 10.1093/biosci/

biab073. URL https://academic.oup.com/bioscience/advance-article/doi/10.1093/biosci/biab073/6322306.

[97] Monica Lasky, Arielle Parsons, Stephanie Schuttler, Alexandra Mash, Lincoln Larson, Ben Norton, Brent Pease, Hailey Boone, Lisa Gatens, and Roland Kays. Candid critters: Challenges and solutions in a large-scale citizen science camera trap project. *Citizen Science: Theory and Practice*, 6(1), 2021.

[98] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[99] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47), 2020.

[100] A Lehmann, J McC Overton, and M P Austin. Regression models for spatial prediction: their role for biodiversity and conservation. page 8.

[101] Robert J. Lennox, Robert Harcourt, Joseph R. Bennett, Alasdair Davies, Adam T. Ford, Remo M. Frey, Matt W. Hayward, Nigel E. Hussey, Sara J. Iverson, Roland Kays, et al. A novel framework to protect animal data in a world of ecosurveillance. *BioScience*, 70(6):468–476, 2020.

[102] Shuyuan Li, Jianguo Li, Weiyao Lin, and Hanlin Tang. Amur tiger re-identification in the wild. *arXiv e-prints*, pages arXiv–1906, 2019.

[103] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

[104] Julie Linchant, Jonathan Lisein, Jean Semeki, Philippe Lejeune, and Cédric Vermeulen. Are unmanned aircraft systems (uas s) the future of wildlife monitoring? A review of accomplishments and challenges. *Mammal Review*, 45(4):239–252, 2015.

[105] Jonathan Lisein, Julie Linchant, Philippe Lejeune, Philippe Bouché, and Cédric Vermeulen. Aerial surveys using an unmanned aerial system (UAS): comparison of different methods for estimating the surface area of sampling strips. *Tropical Conservation Science*, 6(4):506–520, 2013.

[106] Zac Yung-Chun Liu, Jerry H Moxley, Paul Kanive, Adrian C. Gleiss, Thom Maughan, Larry Bird, Oliver J.D. Jewell, Taylor K. Chapple, Tyler Gagne, Connor F. White, et al. Deep learning accurately predicts white shark locomotor activity from depth data. *Animal Biotelemetry*, 7(1):1–13, 2019.

[107] Tim C. D. Lucas. A translucent box: interpretable machine learning in ecology. *Ecological Monographs*, 90(4), November 2020. ISSN 0012-9615, 1557-7015. doi: 10.1002/ecm.1422. URL `https://onlinelibrary.wiley.com/doi/10.1002/ecm.1422`.

[108] Oisin Mac Aodha, Rory Gibb, Kate E. Barlow, Ella Browning, Michael Firman, Robin Freeman, Briana Harder, Libby Kinsey, Gary R. Mead, Stuart E. Newson, et al. Bat detective—deep learning tools for bat acoustic signal detection. *PLoS Computational Biology*, 14(3):e1005995, 2018.

[109] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9596–9606, 2019.

[110] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018.

[111] Alexander Mathis, Steffen Schneider, Jessy Lauer, and Mackenzie Weygandt Mathis. A primer on motion capture with deep learning: principles, pitfalls, and perspectives. *Neuron*, 108(1):44–65, 2020.

[112] Mackenzie W. Mathis and Alexander Mathis. Deep learning tools for the measurement of animal behavior in neuroscience. *Current Opinion in Neurobiology*, 60:1–11, 2020.

[113] John F. McEvoy, Graham P. Hall, and Paul G. McDonald. Evaluation of unmanned aerial vehicle shape, flight path and camera type for waterfowl surveys: Disturbance effects and species recognition. *PeerJ*, 4:e1831, 2016.

[114] Duncan C McKinley, Abe J Miller-Rushing, Heidi L. Ballard, Rick Bonney, Hutch Brown, Susan C. Cook-Patton, Daniel M. Evans, Rebecca A. French, Julia K. Parrish, Tina B. Phillips, et al. Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological Conservation*, 208:15–28, 2017.

[115] Agnieszka Miguel, Sara Beery, Erica Flores, Loren Klemesrud, and Rana Bayrakcismith. Finding areas of motion in camera trap images. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 1334–1338. IEEE, 2016.

[116] Raul Mur-Artal and Juan D. Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[117] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[118] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25): E5716–E5725, 2018.

[119] Paul O'Donoghue and Christian Rutz. Real-time anti-poaching tags could help prevent imminent species extinctions. *The Journal of Applied Ecology*, 53(1):5, 2016.

[120] Ferda Ofli, Patrick Meier, Muhammad Imran, Carlos Castillo, Devis Tuia, Nicolas Rey, Julien Briant, Pauline Millet, Friedrich Reinhard, Matthew Parkan, et al. Combining human computing and machine learning to make sense of big (aerial) data for disaster response. *Big data*, 4(1):47–59, 2016.

[121] Semonn Oleksyn, Louise Tosetto, Vincent Raoult, Karen E Joyce, and Jane E Williamson. Going batty: The challenges and opportunities of using drones to monitor the behaviour and habitat use of rays. *Drones*, 5(1):12, 2021.

[122] Ruth Y. Oliver, Carsten Meyer, Ajay Ranipeta, Kevin Winner, and Walter Jetz. Global and national trends in documenting and monitoring species distributions. *bioRxiv*, 2020.

[123] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[124] Jason Parham, Jonathan Crall, Charles Stewart, Tanya Berger-Wolf, and Daniel I. Rubenstein. Animal population censusing at scale with citizen science and photographic identification. In *AAAI Spring Symposium-Technical Report*, 2017.

[125] Jason Parham, Charles Stewart, Jonathan Crall, Daniel Rubenstein, Jason Holmberg, and Tanya Berger-Wolf. An animal detection pipeline for identification. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1075–1083. IEEE, 2018.

[126] Jason R. Parham. Photographic censusing of zebra and giraffe in the nairobi national park. M.s. thesis, Department of Computer Science, RPI, 2015. URL `http://gradworks.umi.com/10/00/10006301.html`.

[127] Valeriano Parravicini, Michel Kulbicki, D.R. Bellwood, A.M. Friedlander, J.E. Arias-Gonzalez, Pascale Chabanet, S.R. Floeter, R. Myers, Laurent Vigliola, S. D'Agata, et al. Global patterns and predictors of tropical reef fish species richness. *Ecography*, 36(12):1254–1262, 2013.

[128] Debra P. C. Peters, Kris M. Havstad, Judy Cushing, Craig Tweedie, Olac Fuentes, and Natalia Villanueva-Rosales. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology.

Ecosphere, 5(6):art67, June 2014. ISSN 2150-8925. doi: 10.1890/ES13-00359.1. URL `http://doi.wiley.com/10.1890/ES13-00359.1`.

[129] Nathalie Pettorelli, William F. Laurance, Timothy G. O'Brien, Martin Wegmann, Harini Nagendra, and Woody Turner. Satellite remote sensing for applied ecologists: opportunities and challenges. *Journal of Applied Ecology*, 51(4):839–848, 2014.

[130] Maximilian Pichler, Virginie Boreux, Alexandra-Maria Klein, Matthias Schleuning, and Florian Hartig. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution*, 11(2):281–293, 2020.

[131] Michael J.O. Pocock, Helen E. Roy, Chris D. Preston, and David B. Roy. The Biological Records Centre: A pioneer of citizen science. *Biological Journal of the Linnean Society*, 115(3):475–493, 2015.

[132] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

[133] Nicolas Rey, Michele Volpi, Stéphane Joost, and Devis Tuia. Detecting animals in african savanna with uavs and the crowds. *Remote Sensing of Environment*, 200:341–351, 2017.

[134] Benjamin Risse, Michael Mangan, Wolfgang Stürzl, and Barbara Webb. Software to convert terrestrial LiDAR scans of natural environments into photorealistic meshes. *Environmental modelling & software*, 99:88–100, 2018.

[135] Christine R. Rollinson, Andrew O. Finley, M. Ross Alexander, Sudipto Banerjee, Kelly-Ann Dixon Hamil, Lauren E. Koenig, Dexter Henry Locke, Megan Peterson, Morgan W. Tingley, Kathryn Wheeler, et al. Working across space and time: nonstationarity in ecological research and application. *Frontiers in Ecology and the Environment*, 19(1):66–72, 2021.

[136] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

[137] Artsiom Sanakoyeu, Vasil Khalidov, Maureen S McCarthy, Andrea Vedaldi, and Natalia Neverova. Transferring dense pose to proximal animal classes. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 5233–5242, 2020.

[138] D. Blake Sasse. Job-related mortality of wildlife workers in the united states, 1937-2000. *Wildlife society bulletin*, pages 1015–1020, 2003.

[139] F. Schindler and V. Steinhage. Identification of animals and recognition of their actions in wildlife videos using deep learning techniques. *Ecological Informatics*, page 101215, 2021.

[140] Stefan Schneider, Graham W. Taylor, and Stefan Kremer. Deep learning object detection methods for ecological camera trap data. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 321–328. IEEE, 2018.

[141] Stefan Schneider, Graham W. Taylor, Stefan Linquist, and Stefan C. Kremer. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 10 (4):461–470, 2019.

[142] Daniel Schofield, Arsha Nagrani, Andrew Zisserman, Misato Hayashi, Tetsuro Matsuzawa, Dora Biro, and Susana Carvalho. Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances*, 5(9): eaaw0736, 2019.

[143] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[144] Natalia M. Schroeder, Antonella Panebianco, Romina Gonzalez Musso, and Pablo Carmanchahi. An experimental approach to evaluate the potential of drones in terrestrial mammal research: A gregarious ungulate as a study model. *Royal Society open science*, 7(1):191482, 2020.

[145] Burr Settles. Active learning. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):1–114, 2012.

[146] Julie Sherman, Marc Ancrenaz, and Erik Meijaard. Shifting apes: Conservation and welfare outcomes of bornean orangutan rescue and release in kalimantan, indonesia. *Journal for Nature Conservation*, 55:125807, 2020.

[147] Robert Simpson, Kevin R Page, and David De Roure. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd international conference on world wide web*, pages 1049–1054, 2014.

[148] Szymon Smoliński and Krzysztof Radtke. Spatial prediction of demersal fish diversity in the baltic sea: comparison of machine learning and regression-based techniques. *ICES Journal of Marine Science*, 74(1):102–111, 2017.

[149] Robin Steenweg, Mark Hebblewhite, Roland Kays, Jorge Ahumada, Jason T. Fisher, Cole Burton, Susan E. Townsend, Chris Carbone, J. Marcus Rowcliffe, Jesse Whittington, et al. Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. *Frontiers in Ecology and the Environment*, 15(1):26–34, 2017.

[150] Dan Stowell, Michael D. Wood, Hanna Pamuła, Yannis Stylianou, and Hervé Glotin. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution*, 10(3): 368–380, 2019.

[151] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, 2019.

[152] Larissa Sayuri Moreira Sugai, Thiago Sanna Freire Silva, Jr. Ribeiro, José Wagner, and Diego Llusia. Terrestrial Passive Acoustic Monitoring: Review and Perspectives. *BioScience*, 69(1):15–25, 2018.

[153] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*, 2:150026, 2015.

[154] Michael A. Tabak, Mohammad S. Norouzzadeh, David W. Wolfson, Steven J. Sweeney, Kurt C. VerCauteren, Nathan P. Snow, Joseph M. Halseth, Paul A. Di Salvo, Jesse S. Lewis, Michael D. White, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590, 2019.

[155] Colin J. Torney, David J. Lloyd-Jones, Mark Chevallier, David C. Moyer, Honori T. Maliti, Machoke Mwita, Edward M. Kohi, and Grant C. Hopcraft. A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods in Ecology and Evolution*, 10(6): 779–787, 2019.

[156] John Joseph Valletta, Colin Torney, Michael Kings, Alex Thornton, and Joah Madden. Applications of machine learning in animal behaviour studies. *Animal Behaviour*, 124:203–220, February 2017. ISSN 00033472. doi: 10.1016/j.anbehav.2016.12.005. URL `https://linkinghub.elsevier.com/retrieve/pii/S0003347216303360`.

[157] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.

[158] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[159] Maxime Vidal, Nathan Wolf, Beth Rosenberg, Bradley P. Harris, and Alexander Mathis. Perspectives on individual animal identification from biology and computer vision. *arXiv preprint arXiv:2103.00560*, 2021.

[160] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

[161] Alexander Gomez Villa, Augusto Salazar, and Francisco Vargas. Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecological Informatics*, 41:24–32, 2017. doi: 10.1016/j.ecoinf.2017.07.004.

[162] Jana Wäldchen and Patrick Mäder. Machine learning for image based species identification. *Methods in Ecology and Evolution*, 9(11):2216–2225, 2018.

[163] Carrie C. Wall, Samara M. Haver, Leila T. Hatch, Jennifer Miksis-Olds, Rob Bochenek, Robert P. Dziak, and Jason Gedamke. The next wave of passive acoustic data management: How centralized access can enhance science. *Frontiers in Marine Science*, page 873, 2021.

[164] Tristan Walter and Iain D. Couzin. Trex, a fast multi-animal tracking system with markerless identification, and 2d estimation of posture and visual fields. *Elife*, 10:e64000, 2021.

[165] Guiming Wang. Machine learning for inferring animal behavior from location and movement data. *Ecological Informatics*, 49:69–76, January 2019. ISSN 15749541.

[166] Yufu Wang, Nikos Kolotouros, Kostas Daniilidis, and Marc Badger. Birds of a feather: capturing avian shape models from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14739–14749, 2021.

[167] David Watson and Luciano Floridi. Crowdsourced science: sociotechnical epistemology in the e-research paradigm. *Synthese*, 195(2):741–764, 2018.

[168] Hendrik J. Weideman, Charles V. Stewart, Jason R. Parham, Jason Holmberg, Kiirsten Flynn, John Calambokidis, D. Barry Paul, Anka Bedetti, Michelle Henley, Jerenimo Lepirei, and Frank G. Pope. Extracting identifying contours for African elephants and humpback whales using a learned appearance model. *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1276–1285, 2020.

[169] Ben G. Weinstein. A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3):533–545, 2018. ISSN 1365-2656. doi: 10.1111/1365-2656.12780. URL `https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2656.12780`.

[170] Ben G. Weinstein, Sarah J. Graves, Sergio Marconi, Aditya Singh, Alina Zare, Dylan Stewart, Stephanie A. Bohlman, and Ethan P. White. A benchmark dataset for canopy crown detection and delineation in co-registered airborne rgb, lidar and hyperspectral imagery from the national ecological observation network. *PLoS Computational Biology*, 17(7):e1009180, 2021.

[171] Matthias H. Weissensteiner, Jelmer W. Poelstra, and Jochen B.W. Wolf. Low-budget ready-to-fly unmanned aerial vehicles: An effective tool for evaluating the nesting status of canopy-breeding bird species. *Journal of Avian Biology*, 46(4):425–430, 2015.

[172] Dhanushi A Wijeyakulasuriya, Elizabeth W. Eisenhauer, Benjamin A. Shaby, and Ephraim M. Hanks. Machine learning for modeling animal movement. *Plos One*, page 30, 2020.

[173] Martin Wikelski, Roland W Kays, N Jeremy Kasdin, Kasper Thorup, James A Smith, and George W Swenson Jr. Going wild: what a global small-animal tracking system could do for experimental biologists. *Journal of Experimental Biology*, 210(2):181–186, 2007.

[174] Marco Willi, Ross T. Pitman, Anabelle W. Cardoso, Christina Locke, Alexandra Swanson, Amy Boyer, Marten Veldthuis, and Lucy Fortson. Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, 10(1):80–91, 2019.

[175] Hannah J. Williams, Lucy A. Taylor, Simon Benhamou, Allert I. Bijleveld, Thomas A. Clay, Sophie de Grissac, Urška Demšar, Holly M. English, Novella Franconi, Agustina Gómez-Laich, et al. Optimizing the use of biologgers for movement ecology research. *Journal of Animal Ecology*, 89(1):186–206, 2020.

[176] Gary W. Witmer. Wildlife population monitoring: some practical considerations. *Wildlife Research*, 32(3):259–263, 2005.

[177] Connor M. Wood, Viorel D. Popescu, Holger Klinck, John J. Keane, R.J. Gutiérrez, Sarah C. Sawyer, and M. Zachariah Peery. Detecting small changes in populations at landscape scales: A bioacoustic site-occupancy framework. *Ecological Indicators*, 98:492–507, 2019.

[178] Peter H. Wrege, Elizabeth D. Rowland, Sara Keen, and Yu Shiu. Acoustic monitoring for conservation in tropical forests: examples from forest elephants. *Methods in Ecology and Evolution*, 8(10):1292–1301, 2017.

[179] Changchang Wu. Critical configurations for radial distortion self-calibration. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25–32, 2014.

[180] Hui Yu. An evaluation of machine learning classifiers for next-generation, continuous-ethogram smart trackers. *Movement Ecology*, 9(1):14, 2021.

[181] Qiuyan Yu, Wenjie Ji, Lara Prihodko, C. Wade Ross, Julius Y. Anchang, and Niall P. Hanan. Study becomes insight: Ecological learning from machine learning. *Methods in Ecology and Evolution*, pages 2041–210X.13686, August 2021. ISSN 2041-210X, 2041-210X. doi: 10.1111/2041-210X.13686. URL `https://onlinelibrary.wiley.com/doi/10.1111/2041-210X.13686`.

[182] Xiaoyuan Yu, Jiangping Wang, Roland Kays, Patrick A Jansen, Tianjiang Wang, and Thomas Huang. Automated identification of animal species in camera trap images. *EURASIP Journal on Image and Video Processing*, 2013 (1):52, 2013.

[183] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Towards high performance video object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7210–7218, 2018.

[184] Silvia Zuffi, Angjoo Kanazawa, David W. Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6365–6373, 2017.

[185] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael J. Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images “in the wild”. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5359–5368, 2019.

Chapter 3

OVERVIEW OF SPECIES DISTRIBUTION MODELING FOR MACHINE LEARNING PRACTITIONERS

Sara Beery*, Elijah Cole*, Joseph Parker, Pietro Perona, and Kevin Winner. Species distribution modeling for machine learning practitioners: A review. *Proceedings of the 4th ACM SIGCAS Conf. on Computing and Sustainable Societies*, 2021.

3.1 Abstract

Conservation science depends on an accurate understanding of what's happening in a given ecosystem. How many species live there? What is the makeup of the populations? How is that changing over time? Species Distribution Modeling (SDM) seeks to predict the spatial (and sometimes temporal) patterns of *species occurrence*, i.e., where a species is likely to be found. The last few years have seen a surge of interest in applying powerful machine learning tools to challenging problems in ecology [13, 14, 16]. Despite its considerable importance, SDM has received relatively little attention from the computer science community. Our goal in this work is to provide computer scientists with the necessary background to read the SDM literature and develop ecologically useful ML-based SDM algorithms. In particular, we introduce key SDM concepts and terminology, review standard models, discuss data availability, and highlight technical challenges and pitfalls.

3.2 Introduction

Ecological research helps us to understand ecosystems and how they respond to climate change, human activity, and conservation policies. Much of this work starts by deploying networks of sensors (often cameras or microphones) to monitor the organisms living in a fixed study area. Ecologists must then invest significant effort to filter, label, and analyze this data. This step is often a bottleneck for ecological research. For example, it can take years for scientists to process and interpret a single season of data from a network of camera traps. In another case, building real-time estimates of salmonid escapement requires teams of field ecologists working in shifts to watch streams of sonar data 24 hours a day. The challenge is even greater for taxa that are studied by trapping specimens, such as beetles and other insects.

Entomologists can collect thousands of beetles in a few days, but it may require months or years for a suitable expert to exhaustively identify all of the specimens to the species level.

Machine learning methods can significantly accelerate the processing and analysis of large repositories of raw data [1, 2, 4, 9, 33], which can increase the speed and geographic scope of ecological analysis. For instance, ongoing collaborations between machine learning researchers and ecologists have lead to tremendous progress in automating species identification from images in community science data [18, 187] and camera trap data [4, 30]. However, unfamiliar ecological concepts and terminology can present a barrier to entry for many computer scientists who might otherwise be interested in contributing to ecological problems. This is particularly true for more involved ecological problems which may not fit neatly into existing machine learning paradigms.

One such area is **species distribution modeling** (SDM): using species observations and environmental data to estimate the geographic range of a species.[1] This problem has received significant attention from ecologists and statisticians, and there has been increasing interest in machine learning methods due to the large amounts of available data and the highly complex relationships between species and their environments. This document is meant to serve as an easy entry point for computer scientists interested in SDM. In particular, we aim to highlight the exciting technical challenges posed by SDM while also emphasizing the needs of end-users to encourage ecologically meaningful progress. Our hope is that this document can serve as a quick resource for computer science researchers interested in getting started working on conservation and sustainability applications.

The rest of this chapter is organized as follows. In Section 3.3, we discuss different ways to represent the distribution of a species. We discuss species distribution modeling in Section 3.4, and we consider other related ecological modeling problems in Section 3.5. In Section 3.6, we point out pitfalls and challenges in SDM. Finally, we provide pointers to available data (Section 3.7) and discuss open problems (Section 3.8).

[1] We will use the term “species distribution modeling” throughout this document, though sometimes the closely related term “ecological niche modeling” would be more appropriate [142].

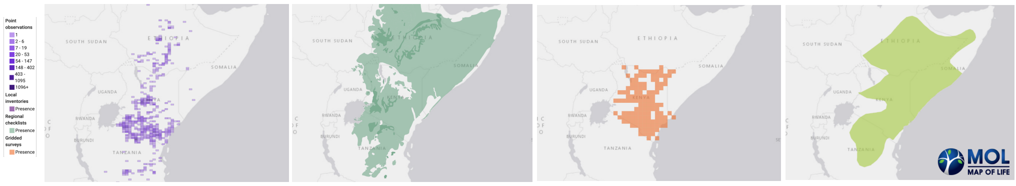


Figure 3.1: Species distribution models describe the relationship between environmental conditions and (actual or potential) species presence. However, the link between the environment and species distribution data can be complex, particularly since distributional data comes in many different forms. Above are four different sources of distribution data for the *Von Der Decken's Hornbill* [8]: (from left to right) raw point observations, regional checklists, gridded ecological surveys, and data-driven expert range maps. All images are from Map of Life [99].

Data collection method	**Example**	**Observation type**
Community science observations	iNaturalist	Presence-only
Community science checklists	eBird	Presence-absence
Static sensors	Camera traps	Presence-absence
Sample collection	Insect trapping	Presence-absence
Expert field surveys	Line transects	Presence-absence
Historic records, natural history collections	Herbarium sheets	Presence-only

Table 3.1: Sources of species observation data. Each of these examples represents a method of collecting or accessing observations of different species. One important distinction is whether the observations are *presence-only* or *presence-absence*. Presence-only data consists of locations where a species has been sighted. Presence-absence data also includes locations where a species was checked for but not observed.

3.3 Representing the distribution of species

The distribution of a species is typically represented as a *map* which indicates the spatial extent of the species. These maps can be created in a variety of ways, ranging from highly labor-intensive expert range maps to fully automatic species distribution models. We show four examples in Fig. 3.1. In this section we give a high-level overview of three important sources of maps: raw species observation data, predictions from statistical models, and expert knowledge.

Raw species observation data

Any representation of the distribution of a species begins with some sort of *species observation data*. In general, species observation data consists of records indicating whether a species is present or absent at certain locations. Species observation data can take many forms–see Table 3.1 for examples. Species observation data falls into two general categories: **presence-only** data reports known sightings, or occurrences, of a species, while **presence-absence** data also provides information on where a species did not occur. Data collection strategies define whether absence data will be available. For instance, iNaturalist collects opportunistic imagery of species from community scientists, which produces presence-only species observations. On the other hand, eBird uses species *checklists* where *all* bird species seen and/or heard within a time span at a given location are reported. Since exhaustive reporting is expected from observers, any bird species not reported is assumed to be absent. In this sense, checklists are treated as presence-absence data.

One of the simplest ways to convey the distribution of a species is to simply show all of the locations where the species is known to be present or absent on a map. However, this sort of highly simplified "species distribution" is not able to make any predictions about whether a species might be present or absent at locations which have not been sampled.

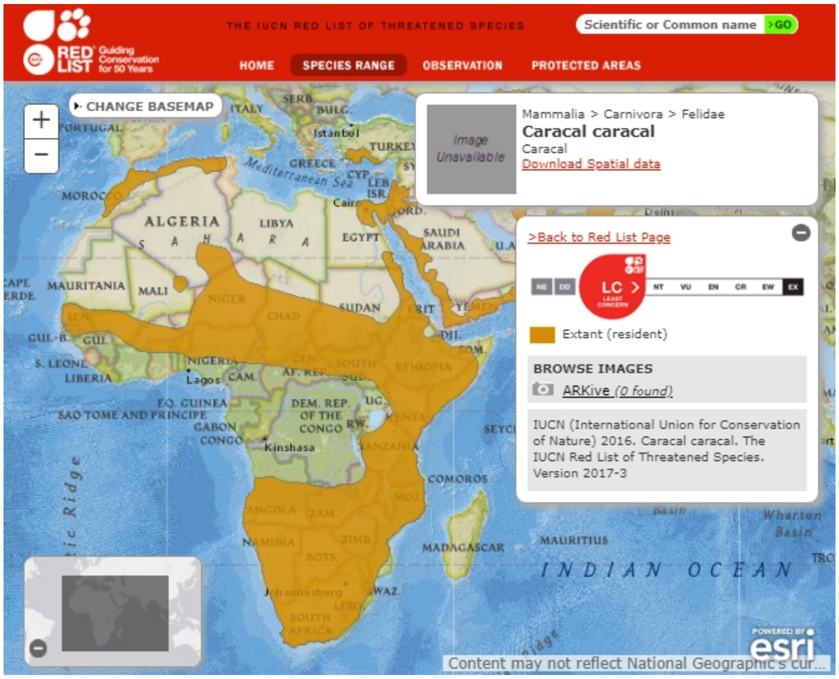
Statistical models

To create species distributions that can extrapolate beyond sampled locations, we can pair species observations with collections of environmental characteristics (altitude, land cover, humidity, temperature, etc.) and fit statistical models that use the environmental characteristics to predict species presence or absence. These models can make predictions at any place and time for which these environmental characteristics are known. Species distribution models fall into this category, and are our focus throughout this document.

Expert range maps

Species range maps have traditionally been heavily influenced by the individual scientists who study those species. These maps are often based on a complex combination of heterogeneous information sources, including personal observations, understanding of the species' habitat preferences, local knowledge/reports, etc. From our discussions with practitioners, we find that these *expert range maps* (ERMs) are often the most trusted source of distribution information. Perhaps the most widely-known expert range maps are those published by IUCN [81] as part of their *Red List* of vulnerable and endangered species. An example of the IUCN range map for the *caracal* can be seen in Fig. 3.2. Studies have shown both agreement [17] and disagreement [77, 98] between ERMs and species observation data. Expert range maps have also been found to be highly scale-dependent, tending to overestimate the occupancy area of individual species and ranges < 200km [97].

It is important to note that ERMs come in many forms, from hand-drawn maps to data-driven maps that are slightly refined by experts. In the latter case, ERMs are partially based on species observation data, so the two cannot be treated as independent sources. As we will discuss in more detail in Section 3.4, the lack of a solid "ground truth" information about the true underlying distribution of species across space and time makes it difficult to analyze the accuracy of any species distribution model, including those drawn by experts.



Mammalia > Carnivora > Felidae
Caracal caracal
Caracal
Download Spatial data

>Back to Red List Page

🟧 Extant (resident)

BROWSE IMAGES
ARKive (0 found)

IUCN (International Union for Conservation of Nature) 2016. Caracal caracal. The IUCN Red List of Threatened Species. Version 2017-3

Figure 3.2: The International Union for Conservation of Nature (IUCN) publishes expert range maps for many species, particularly those on their "Red List of Threatened Species" [193]. Here we show the IUCN Range Map for the *Caracal caracal* [22].

3.4 Species distribution models

The terminology in this area can be confusing, so we will start with a definition and a few clarifications.

Intuitive definition. A species distribution model is a function that uses the characteristics of a location to predict whether or not a species is present at that location. This can be understood as a supervised learning problem. The input is a vector of environmental characteristics for a location and the output is species presence or absence. In principle one could use almost any classification or regression technique as the basis for an SDM.

Formal definition. The key components of a simple species distribution modeling pipeline are: (1) species observation data, (2) a method for encoding locations, and (3) a function which maps location encodings to predictions. Formally, we define these components as follows:

1. A dataset of species observations. This is a collection of records indicating that a species is present or absent at given location and time. We write this as $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ where $\mathbf{x}_i \in \mathcal{X}$ is a spatiotemporal location and $y_i \in \{0, 1\}$ indicates presence (1) or absence (0). The spatiotemporal domain $\mathcal{X}$ is typically something like $\mathcal{X} = [0, 180) \times [0, 360) \times [0, 1)$ which encodes global longitude and latitude as well as the time of year.

2. A location representation $h : \mathcal{X} \to \mathcal{Z} \subset \mathbb{R}^k$. This is typically a simple "look-up" operation, where $\mathbf{x} \in \mathcal{X}$ is cross-referenced with k pre-defined geospatial data layers to produce a vector of location features $h(\mathbf{x}) \in \mathbb{R}^k$. That is, $h(\mathbf{x})$ is a representation of the location $\mathbf{x} \in \mathcal{X}$ in some environmental feature space.

3. A model $f_\theta : \mathcal{Z} \to [0, 1]$ where θ is a parameter vector. The goal is to find parameters θ of f so that $f_\theta(h(\mathbf{x})) = 1$ when the species is present and $f_\theta(h(\mathbf{x})) = 0$ otherwise. This is usually framed as a supervised learning problem on the dataset $\{(h(\mathbf{x}_i), y_i)\}_{i=1}^N$.

Note that this is a streamlined formalization meant to capture the essence of SDM. While there are many variants in practice, almost any species distribution modeling will include these core concepts.

What does an SDM actually predict? An SDM takes as input a vector of environmental features and predicts a numerical score (usually between 0 and 1) for a location. An important distinction to note regarding SDMs is *geographic space* vs. *environmental space*, elucidated in Fig. 3.3. This score is often interpreted as a prediction of habitat suitability. Typically the score *may not* be interpreted as the probability a species is present. Note that here we are only considering presence vs. absence - predicting species *abundance* is a more challenging problem, which we discuss in Section 3.5.

How is an SDM used? The most common end product is a map of the SDM predictions, which is produced by simply visualizing the SDM predictions across an area of interest. Binary predictions can be obtained by applying a threshold to the continuous predictions of the SDM.

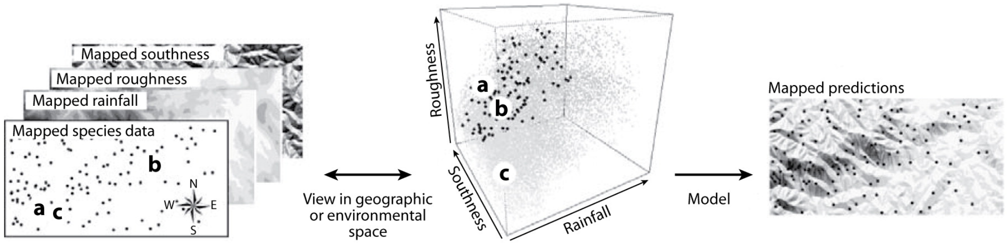
A brief history of species distribution modeling

Early predecessors for SDM include qualitative works that link patterns within taxonomic groups to environmental or geographic factors, such as Joseph Grinnel's 1904 study of the distribution of the chestnut-backed chickadee [80], among others [116, 128, 160, 196].

Modern SDMs are primarily statistical models fit to observed data. Early quantitative approaches used multiple linear regression and linear discriminant function analyses to associate species and habitat [41, 168]. The application of generalized linear models (GLMs) [20, 131] provided more flexibility by allowing non-normal error distributions, additive terms, and nonlinear relationships. The explosive proliferation of large "presence-only" datasets (see Table 3.1) in recent years has led to the development of new modeling approaches to SDMs such as the popular "Maximum Entropy Modeling" (MaxEnt) approach [144] with roots in point process modeling [152].

The first modern SDM computing package, BIOCLIM, was introduced in 1984 on the CSIRO network [35, 40]. This package took observation information, such as the species observed, location, elevation, and time, and used them to determine what environmental variables correlated with that species' occurrence. These variables were then used to map possible distributions of the species under consideration. Climate interpolation techniques developed for BIOCLIM are the basis of the existing WorldClim database [66] and are still widely used in SDMs today. Many different implementations of various SDM methods are now publicly available. We would like to highlight Wallace [106], which is a well-documented R implementation of historic and modern techniques.

As earth observation technology has improved, the scope of what is possible to include as an environmental covariate in a model has vastly increased. Improvements in weather monitoring systems gave access to high-temporal-frequency temperature, wind, and precipitation measurements. Recently, ecologists have turned to remote sensing imagery to estimate high-spatial-coverage ecological variables such as soil composition or density of sequestered carbon, as well as mapping land cover type across regions [90]. Modern SDM methods pair these covariate estimates with increasingly accurate global elevation maps, and selected high-quality but sparse in-situ measurements [111, 150].



Elith J, Leathwick JR. 2009.
Annu. Rev. Ecol. Evol. Syst. 40:677–97

Figure 3.3: **Geographic vs. environmental space.** Observation data can be associated with a geographical location, or mapped into a feature space based on environmental covariates. Most SDMs operate under the assumption that with the right set of *environmental variables* and an appropriate model, one could use environmental characteristics to map species distribution. Figure reproduced with permission, originally published in [61].

Several excellent, detailed reviews of SDMs have been published within the ecology community [61, 84, 85, 153, 163, 168]. We direct the reader to the excellent summary by Elith and Leathwick [61].

Covariates for species distribution modeling

In this section we discuss several environmental characteristics (often called *covariates*) that can be used for species distribution modeling. Here we are focused on describing the different categories of covariates–details on specific covariate datasets are available in Section 3.7. Some of the covariates we discuss are widely used in the species distribution modeling literature, while others are more recent or speculative. It is also important to keep in mind that many covariates are themselves based on sophisticated predictive models due to the cost of densely sampling any property of the earth's surface.

Climatic variables

Temperature and precipitation are critical characteristics of an ecosystem. Perhaps the most commonly used climate dataset for SDM is the WorldClim bioclimatic variables [66] dataset, which provides 19 climate-related variables averaged over the period from 1970 to 2000 at a spatial resolution of around 1km^2. We show a few examples of variables from this dataset in the top row of Fig. 3.5.

Pedologic (soil) variables

Soil characteristics are intimately related to the plant life in an area, which naturally influences the entire ecosystem. One example of a comprehensive pedologic dataset is SoilGrids250m [93], which consists of soil properties like pH, density, and organic carbon content at a 250m^2 resolution globally. We show a few examples of variables from this dataset in the bottom row of Fig. 3.5.

Vegetation indices

A *vegetation index* (VI) is a number used to measure something about the plant life in an area, and is typically computed from remote sensing data like satellite imagery. Many different VIs have been proposed. A review paper published in 1995 discussed 40 different vegetation indices that had been developed by different researchers [24]. One of the most popular examples is the *normalized difference vegetation index* (NDVI). If a remote sensing image includes the red and near-infrared (NIR) bands, then the corresponding NDVI image can be computed by applying the formula

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} \tag{3.1}$$

independently at each pixel. NDVI is meant to indicate the presence of live green plants. From a computer vision perspective, these VIs are essentially hand-designed features for remote sensing data.

Land use / land cover

The term *land cover* refers to the physical terrain at a location, while the closely related term *land use* tends to emphasize the function of a location. For instance, an area with the land cover label “dense urban” may have a land use label like “school” or “hospital.” We provide an example in Fig. 3.4, which shows RGB imagery and land cover from two different sources for the same 1km^2 area. It is not obvious what the best label set would be for species prediction, but practically speaking many of the available land use / land cover datasets are focused on relatively coarse categories related to agriculture, natural resources, or urban development. For instance, the U.S. National Land Cover Database assigns one of 20 land cover classes to every 30m^2 patch of land in the United States at a temporal resolution of 2-3 years [95]. The classes cover various general habitat types (water, snow, developed land, forests...) but are not tuned for species prediction in particular.

Measures of human influence

Humans have had a profound impact on the natural world, so it is reasonable to include measures of human influence as environmental characteristics. For instance, the Human Influence Index [159] uses eight factors (human population density, railroads, roads, navigable rivers, coastlines, nighttime lights, urban footprint, and land cover) to compute a score that is meant to quantify how much an environment has been reshaped by humans.

Remote sensing imagery

Imagery collected by satellites, planes, or drones can provide substantial information about an environment. To start with, we note that vegetation indices, land cover, land use, and many measures of human influence are all derived from some form of overhead imagery like that in Fig. 3.4. In addition, there may be more abstract patterns that can be extracted using modern computer vision techniques like convolutional neural networks. Research on the use of raw overhead imagery (instead of derived products) for SDM is in its early stages [46, 53, 175].

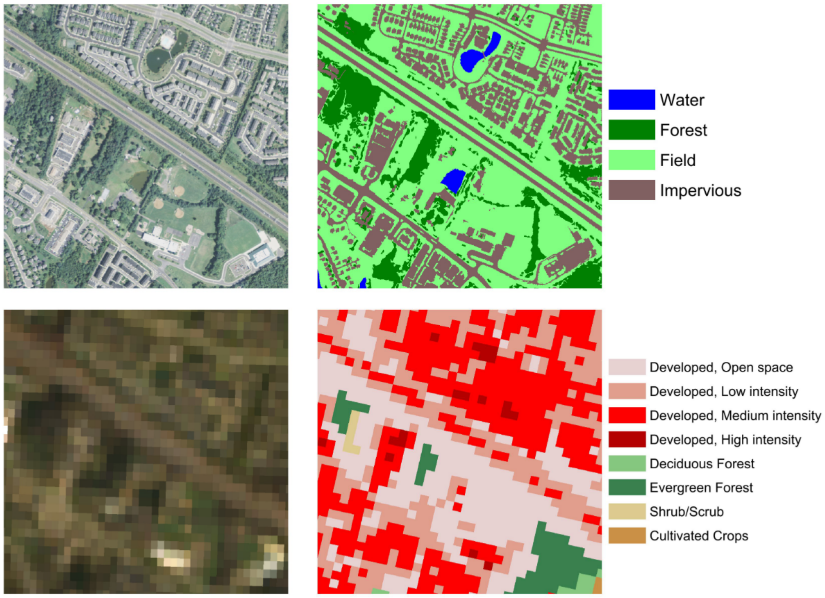


Figure 3.4: RGB imagery (left column) and land cover maps (right column) from two different remote sensing sources covering the same 1km^2 area, from [156]. RGB imagery is manually or semi-automatically annotated to produce the land cover labels. As this example demonstrates, the set of land cover labels can vary depending on the organization doing the labeling. Figure reproduced with permission, originally published in [156].

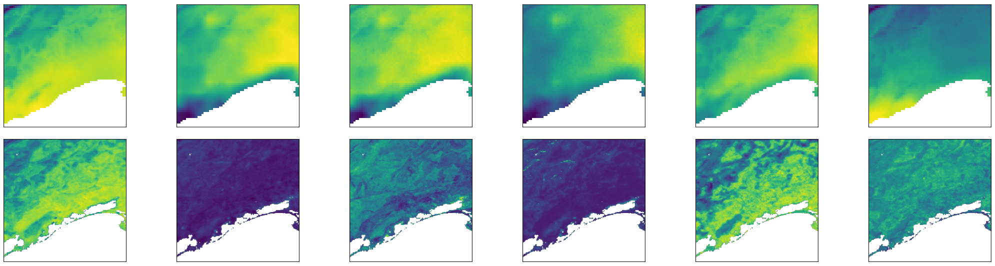


Figure 3.5: Visualizations of some of the bioclimatic variables (top row: `bio_1` - `bio_6` from left to right) and pedologic variables (bottom row: `orcdrc`, `phihox`, `cecsol`, `bdticm`, `clyppt`, `sltppt` from left to right) provided for the GeoLifeCLEF 2020 competition [46]. The area shown in each image is approximately 64 km^2 centered in Montpellier, France. While we visualize each environmental variable as a 2D raster, most species distribution modeling methods are only compatible with relatively low-dimensional vectors of environmental variables (not "stacks" of 2D patches). As is typical in a collection of covariates, we see that the pedologic variables have a different resolution than the bioclimatic variables.

Properties of species distribution models

In this section we describe important properties that can be used to categorize species distribution models. Any particular species distribution model may or may not have any of these properties. The categories we describe are in general nested or overlapping, not mutually exclusive.

Presence only vs. presence-absence models

Species observation datasets may be either presence-absence or presence-only. While presence-only data is easier to collect, the are limitations on what can be estimated from such data [89]. Typically a species distribution model is designed to handle either presence-absence or presence-only data, though there is growing interest in developing methods that can use both [70, 76, 139].

Single vs. multi-species models

Many SDMs are designed to model the distribution of a single species. This is in contrast to *multi-species* models which are meant to capture information about several species. Many of the earlier models are single-species models [61, 144], though interest in multi-species models has grown over time [88, 96, 134].

Multi-species models: stacked vs. joint

Multi-species SDMs can be classified as either *stacked* or *joint*. In a *stacked* model, a single-species SDM is fit for each species and the resulting maps are “stacked” on top of one another to provide a multi-species map. This approach is simple, but it cannot take advantage of patterns in how species co-occur. This is the motivation for *joint* SDMs, in which the estimated distribution of each species also depends on occurrence data for other species. Recent work has begun to systematically compare the results from stacked and joint species distribution models for different species and regions [92, 134, 204].

Spatially explicit models

Typically species distribution models use environmental characteristics to make predictions about the presence or absence of species. Such models represent a location in terms of these environmental features, so two different locations with the same environmental characteristics will lead to the same predictions, even though the two locations may be far apart. Models that mitigate this concern by incorporating geographical location information directly are referred to as *spatially explicit* [55] models.

Occupancy models

It is easier to confirm that a species is present than it is to confirm that a species is absent. One confident observation of a species suffices to confirm its presence at a given location. However, failing to observe a species at a location does not suffice to prove absence, since the species could have been present but not observed. *Occupancy models* are meant to account for imperfect detection by modeling the probability that a species is present but unobserved at a given location conditional on the sampling effort that has been invested [23, 117].

Understanding uncertainty and error

Species distribution models attempt to capture the behavior of a complex system from data, which is a challenging and error-prone process. [157] describes 11 sources of uncertainty and error in species distribution models, and groups them into two clusters: (i) uncertainty in the observation data itself and (ii) uncertainty due to arbitrary modeling choices. [56] studies the effect of making different reasonable modeling choices on final projections of species distribution under different future climate scenarios. Similarly, [172] considers the uncertainty introduced by the arbitrary choice of covariates while [167] analyzes the effect of uncertainty in the values of the covariates themselves. [130] focuses on the effect of uncertainty in the location of species observations. [26] reviews sources of uncertainty for different types of species distribution models, as well as best practices for minimizing uncertainty and methods for incorporating uncertainty directly into the model.

Algorithms for species distribution modeling

In this section we provide a high-level overview of the space of algorithms commonly used for species distribution modeling in the ecological community. This section draws heavily from the organization of [134], which is an excellent comparative study of different species distribution modeling techniques. We discuss several commonly used models, and note that the different methods can have very different properties, assumptions, and use cases. Unlike some classes of algorithms, different species distribution modeling methods are generally not readily interchangeable.

Presence-only methods

Perhaps the most popular approach for presence-only SDM is *MaxEnt* [144]. We follow the description given in [63]. The basic idea is to estimate the probability of observing a given species as a function of the environmental covariates. The estimate is chosen to be (i) consistent with the available species observation data and (ii) as close as possible (in KL divergence) to the marginal distribution of the covariates. Criterion (ii) is necessary because there are typically many distributions that satisfy criterion (i). Another simple approach for presence-only SDM is to introduce artificial negative observations called *pseudonegatives* or *pseudoabsences* based on some combination of domain knowledge and data. Once pseudonegatives have been generated, they are combined with the presence-only data and traditional presence/absence methods are applied.

Traditional statistical methods

Perhaps the most common methods in species distribution modeling are workhorse methods drawn from the statistics literature such as generalized linear models [71, 73, 137, 190, 194]. Important special cases include logistic regression [140] and generalized additive models [202]. Some species distribution modeling algorithms are better thought of as general frameworks whose particular realization depends on the available data sources and modeling goals. As an example, the Hierarchical

Modeling of Species Communities (HMSC) framework [137] minimally requires species occurrence data with corresponding environmental features. The species occurrences are related to environmental features by a generalized linear model. However, the framework can be extended to incorporate *e.g.*, information on species traits and evolutionary history.

Machine learning methods

The relationship between species and their environment is complex and may not satisfy traditional statistical assumptions such as linear dependence on covariates or i.i.d. sampling. For this reason, machine learning approaches have also enjoyed considerable popularity in the species distribution modeling literature. Examples include boosted regression trees [62], random forests [48], and support vector machines [58]. In addition, neural networks have been used for species distribution modeling since well before the deep learning era [37, 138, 180, 203]. Interest in joint species distribution modeling with neural networks has only grown as deep learning has come to maturity [88]. Convolutional neural networks in particular have created a new opportunity: the ability to extract features from spatial arrays of environmental features [43, 51] instead of using hand-selected environmental feature vectors.

The challenge of evaluation

How can we tell whether a species distribution model is performing well or not? The typical approach in machine learning is to use the model to make predictions on a held-out set of data and compute an appropriate performance metric by comparing the model predictions to ground-truth labels. But what is “ground truth” for a species distribution model?

Notions of Ground Truth

We describe several common approaches to the challenging problem of how to evaluate SDMs in practice. For further detail, [126] provides an excellent discussion of different metrics for evaluating SDMs and the extent to which they are ecologically meaningful.

Compare against presence-absence data. Ideally, for each location, an expert observer would determine whether each species of interest is present or absent at that location. Conducting this kind of survey for a single species in a limited area is expensive, and the survey would need to be repeated periodically to monitor change over time. These exhaustive surveys quickly become extraordinarily expensive as we expand the number of species of interest or the geographic extent of the survey. Even if the resources were available, the observations would have some degree of noise - in particular, confirming that a species is absent from an area can typically only be done up to some degree of certainty. (See the discussion of occupancy modeling in Section 3.4.) For most species and most locations on earth, this sort of ideal ground truth data is just not available. However, this kind of evaluation is possible for select species and locations at sparse time points. For instance, [64] includes presence-absence data for 226 species from 6 parts of the world collected at various time points.

Compare against presence-only data. Unfortunately, presence-absence data is often unavailable. We describe a few simple methods for comparing predictions against presence-only data along with their shortcomings.

- False negative rate: how often are locations which are known to be positive predicted to be negative? The false negative rate measures whether the model is consistent with the observed positives, but does not assess the model's behavior at other points.
- Top-k classification accuracy: how often is the observed species among the k most likely species under the model? However, there is not an obvious way to choose k. Moreover, for any fixed k it is likely that some locations will have more than k species while others will have fewer.
- Adaptive top-k classification accuracy: this is a variant of the top-k classification accuracy that assumes that the number of species is k on average, while allowing some locations to have more than k species while others may have fewer. See [46] for details. Like standard top-k classification accuracy, choosing k may be difficult.

Note that adaptive top-k and top-k are both metrics for multi-species models, while the false negative rate can be computed for single species models as well.

Compare against community science data. Community science projects like iNaturalist and eBird are generating species observation data at an extraordinary rate and frequency. iNaturalist alone generates millions of species observations per month [10]. However, the data produced by such projects can vary in terms of how easy it is to use and interpret depending on the sampling protocol [110]. For instance, iNaturalist accepts presence-only observations, which allows the user base to scale broadly but limits the utility of the data for ground truthing. iNaturalist data tells us where different species have been observed by humans, but not where those species are either absent or present without human observation. eBird uses a more rigorous sampling protocol that records both presences and absences, but their observations are limited to birds. The quality of these reports depends on the skill of the user at identifying all bird species they see or hear. Citizen science data has been found to produce results similar to those from (coarse) professional surveys under the right circumstances [94, 110, 183].

Compare against expert range maps. Another possibility is to compare the model predictions against one or more range maps that are hand-drawn by experts (see Section 3.3). However, this raises the question: how do we validate *those* range maps? A hand-drawn map may be biased by an individual's experience or by the data sources the expert prefers. It can also be hard to find a suitable expert to generate a map for a given species. Another challenging question relates to temporal progression: is each expert updating their maps according to the latest data? If so, when was that data collected? The IUCN has a published set of standards for creating species range maps [81], but not all creators of maps match these standards.

In addition, there is the methodological question of how one should evaluate a model against an expert range map, which is explored in [118]. Approaches range from very qualitative (ask an expert whether the map looks reasonable to them) to very quantitative (compute a well-defined error metric between the SDM predictions and the expert range map). Important to note here, expert range maps are most often categorical, with hard boundaries drawn representing temporal categories like "breeding", "non-breeding", "year-round", etc. On the other hand, SDM predictions are often real-valued on $[0, 1]$ over both space and time. While continuous predictions can be converted to binary maps by applying a threshold, it can be unclear how to choose this threshold if a robust validation method is not available.

Evaluation on downstream tasks. Instead of evaluating whether a species distribution model produces a faithful map of species presence, we may instead check whether it is useful for some other downstream task. For example, [18] builds a simple SDM and demonstrates that it improves accuracy on an image-based species classification task. However, it is certainly possible for an SDM to be useful whether or not it accurately reflects the true species distribution.

Evaluation pitfalls

Even when suitable ground truth data is available, there are some pitfalls that can hinder meaningful evaluation. In this section we discuss some of these pitfalls and make specific recommendations to the machine learning community for handling them.

Performance overestimation due to spatial autocorrelation. In the machine learning community it is common to sample a test set uniformly at random from the available data. However, this strategy can lead to overestimation of algorithm performance for spatial prediction tasks since it is possible to obtain high performance on a uniformly sampled test set by simple interpolation [154]. This effect is called *spatial autocorrelation*. Similar concerns are relevant for evaluating camera trap image classifiers [28]. For ecological tasks, it is important to evaluate models as they are intended to be used. In many cases, the more ecologically meaningful question is whether the model generalizes to novel locations, unseen in the training set. In these cases it is important to create a test set by holding out spatial areas. In other cases, the ecologist seeks to build a model that will perform accurately in the future at their set of monitoring sites. In these cases, instead of holding out data in space, we can split the data to hold out a test set based on time. A randomly sampled test set is not a good proxy for the use case of either scenario.

Hyperparameter selection. The performance of an algorithm typically depends on several hyperparameters. In the machine learning community these are set using cross-validation on held-out data. However, selecting and obtaining a useful validation set can be particularly challenging in SDM due to the data collection challenges described elsewhere. Recent work has also studied the sensitivity of SDMs to hyperparameters [86] and developed techniques for hyperparameter selection in the presence of spatial autocorrelation [162].

Spatial quantization. A natural first step when working with spatially distributed species observations is to define a spatial quantization scheme. By “binning” observations in this way, we can associate many species observations with a single vector of covariates. Additionally, spatially quantized data can be more natural from the perspective of many machine learning algorithms since the domain becomes discrete. However, the choice of quantization scheme (grid cell size) is difficult to motivate in a rigorous way. This is a problem because different quantization choices can result in vastly different outcomes - this is known as the *modifiable areal unit problem* [136]. It is possible to cross-validate the quantization parameters, but only in those limited cases where there is enough high-quality data for this to be a reliable procedure. Furthermore, that process may be computationally expensive.

The long tail. Many real-world datasets exhibit a *long tail*: a few classes represent a large proportion of the observations, while many classes have very few observations [28, 186]. Species observation data is no exception - for example, in the Snapshot Serengeti camera trap dataset [169] there are fewer than 10 images of gorillas out out of millions of images collected over 11 years. This presents at least two problems. The first problem is that standard training procedures will typically result in a model that perform well on the common classes and poorly on the rare classes. The second is that many evaluation metrics are averaged over all examples in the dataset, which means that the metric can be very high despite poor performance on almost all species. It is much more informative to study the performance on each class or on groups of classes (*e.g.*, common classes vs. rare classes). One common solution is to compute metrics separately for each class and then average over all classes to help avoid bias towards common classes in evaluation.

Model trust

Once a model has been built, the previously discussed challenges of model evaluation make it difficult to determine where, how much, and for how long a model is sufficiently accurate to be used. The accuracy needed may also vary by use case and subject species. In our discussions with ecologists, we find that this leads to a lack of trust in SDMs. What verification and quality control is needed to ensure a model is still valid over time? This is an open question, and an important one to answer if our models are to be used in the real world.

3.5 Other types of ecological models

Species distribution modeling is only one of many ways that ecologists seek to describe and understand the natural world. To give readers a sense of how SDM fits into the broader scope of ecological modeling, we provide a high-level overview of other common modeling tasks.

Mechanistic models

Mechanistic models make assumptions about how species depend on the environment or on other species. One example is to use an understanding of a plant's biology to predict the viable temperature range where the plant can grow [170]. Such models are useful but difficult to scale, as they require species-specific expert knowledge. Our focus in this work is on *correlative* species distribution models, which do not require mechanistic knowledge.

Abundance modeling

Abundance modeling goes beyond species presence or absence, aiming to characterize the absolute or relative number of individuals at a given location. We define abundance and related concepts in Section 3.5.

Population estimation

Population estimation is concerned with counting the total number of individuals of a species, typically within some defined area [161]. Population size is most frequently estimated using *capture-recapture models*, which require the ability to distinguish between individuals of the same species. Traditionally this individual re-identification was based on physical tags or collars [78], but some recent efforts have relied on the less invasive method of identifying visually distinctive features, such as stripe patterns or the contour of an ear [33].

Density estimation

Density estimation seeks to model *spatial abundance*, the abundance of a species per unit area, to understand where a species is densely versus sparsely populated [158, 191].

Data collection procedures for abundance

As mentioned above, capture-recapture requires an individual to be re-identifiable. In the absence of the ability to re-identify individuals, several other data collection procedures are used. One that is frequently used for insects and fish populations is the *harvest method*, where individuals are collected in traps which are open for a set amount of time and then counted [148, 164]. Sampling strategies for other taxa include:

- **Quadrat sampling:** A *quadrat* is a fixed-size area where species are to be sampled. Within the quadrat, the observer exhaustively determines the occurrence and relative abundance of the species of interest. Quadrat sampling is most commonly used for stationary species like plants. The observer will sample quadrats throughout the region of interest to derive sample variance and conduct further statistical analysis [87].

- **Line intercept sampling:** A *line intercept* or *line transect* is a straight line that is marked along the ground or the tree canopy, and is primarily used for stationary species [91]. The observer proceeds along the line and records all of the specimens intercepted by the line. Each transect is regarded as one sample unit, similar to a single quadrat.

- **Cue counting:** Cue counting is based on observing cues or signals that a species is nearby, such as whale or bird calls. It is used primarily for species that are underwater or similarly difficult to sight [119].

- **Distance sampling:** *Distance sampling* refers to a class of methods which estimate the density of a population using measured distances to individuals in the population [38]. Distance sampling can be added to line transects in order to incorporate specimens that are off the transect line but still visible. Appropriately calibrated camera traps can also benefit from distance sampling [158].

- **Environmental DNA (eDNA) sampling:** Samples of water or excrement collected in the field can be sequenced to provide species identifications. The ratios of environmental DNA for each species can be used to estimate abundance [115, 185].

Each of these procedures produces different types of data, and each method comes with its own innate collection biases. These biases can add to the challenge of evaluating ecological models, as discussed in Section 3.4.

Biodiversity measurement and prediction

While it is important to understand the distribution of particular species, in many cases the ultimate goal is to understand the health of an ecosystem at a higher level. *Biodiversity* is a common surrogate for ecosystem health, and there are many different ways to measure it [103, 104, 197]. In this section we define and discuss several biodiversity metrics and related concepts. Note that some sources give different definitions than those presented here, so caution is warranted.

We now define some preliminary notation. We let R denote an arbitrary spatial unit such as a country. Many biodiversity metrics are computed based on a *partition* of

R into N sub-units, which we denote by $\{R_i\}_{i=1}^N$. The choice of partition can have a significant impact on the value of some metrics, but for the purposes of this section we simply assume a partition has been provided.

Species richness. The species richness of R is the number of unique species in R, which we write as $S(R)$.

Absolute abundance. The absolute abundance of species k in R is the number of individuals in R who belong to species k. We write this as $A_k(R)$.

Relative abundance. The relative abundance of species k in R is the fraction of individuals in R who belong to species k, which is

$$p_k(R) = \frac{A_k(R)}{\sum_{j=1}^{S(R)} A_j(R)}. \tag{3.2}$$

Since $\sum_{j=1}^{S(R)} p_j(R) = 1$ and $p_j(R) \geq 0$ for all $j \in \{1, \ldots, S(R)\}$, the vector of relative abundances $\mathbf{p}(R) = (p_1(R), \ldots, p_{S(R)}(R))$ forms a discrete probability distribution. The species richness can then be alternately defined as the support of this distribution, given by

$$S(R) = |\{j \in \{1, \ldots, S(R)\} : p_j(R) > 0\}|. \tag{3.3}$$

Of course, we can replace p_j with A_j everywhere and get an identical quantity.

Shannon index. The Shannon index of R is the entropy of the probability distribution $\mathbf{p}(R)$, so

$$H(\mathbf{p}(R)) = -\sum_{j=1}^{S(R)} p_j(R) \log p_j(R). \tag{3.4}$$

The Shannon index quantifies the uncertainty involved in guessing the species of an individual chosen at random from R. Sometimes H is instead written as H', and sometimes the argument is written as R instead of $\mathbf{p}(R)$.

Simpson index. The Simpson index of R is the probability that two individuals drawn at random from the dataset (with replacement) are the same species, and is given by

$$\lambda(R) = \sum_{i=1}^{S(R)} p_i^2. \tag{3.5}$$

Alpha diversity. The alpha diversity of R is the average species richness across the sub-units $\{R_i\}_{i=1}^N$, given by

$$\alpha(R) = \frac{1}{N} \sum_{i=1}^{N} S(R_i). \tag{3.6}$$

Gamma diversity. The gamma diversity of R is defined as

$$\gamma(R, q) = \left(\sum_{j=1}^{S(R)} p_j^q \right)^{1/(1-q)} \tag{3.7}$$

where $q \in [0, 1) \cup (1, \infty)$ is a weighting parameter [103]. Note that gamma diversity is also commonly denoted by ${}^\gamma D_q(R)$. There are several interesting special cases:

- If $q = 0$ then gamma diversity reduces to species richness i.e. $\gamma(R, 0) = S(R)$.
- Gamma diversity is also related to the Shannon index, since $\lim_{q\to 1} \gamma(R, q) = \exp H(\mathbf{p}(R))$[103].
- If $q = 2$ then gamma diversity reduces to the inverse of the Simpson index i.e. $\gamma(R, 2) = 1/\lambda(R)$.

Beta diversity. The beta diversity of R is meant to measure the extent to which sub-units R_i are ecologically differentiated. This can be interpreted as a measure of the variability of biodiversity across sub-regions or habitats within a larger area. It is defined as

$$\beta(R, q) = \frac{\gamma(R, q)}{\alpha(R)} \tag{3.8}$$

where q is the same weighting parameter we say in the definition of gamma diversity [103, 182]. Beta diversity quantifies how many sub-units there would be if the total species diversity of the region γ and the mean species diversity per sub-unit α remained the same, but the sub-units had no species in common.

3.6 Common challenges and risks

Differences in tools

R is the dominant coding language in ecology and statistics, but Python is dominant in machine learning. This language barrier limits code sharing, which in turn limits algorithm sharing. It is also important to note that some machine learning models are extremely computationally demanding to train, and some ecologists may not have access to the necessary computational resources.

Differences in ideas and terminology

Differences in concepts and terminology can make it difficult for machine learning practitioners to find and read relevant work from the ecology community (and vice-versa). However, there is a growing body of interdisciplinary work which brings ecologists and computer scientists together [13–15]. It is important for computer scientists working in this area to establish ties with ecologists who can help them understand how to make ecologically meaningful progress.

Combining data sources

Species observation data is collected according to many different protocols, which means that effectively combining different data sources can be nontrivial [75, 109, 124, 139]. For instance, observations collected in a well-designed scientific survey have significantly different collection biases from observations collected via iNaturalist. Handling these biases in a robust, systematic way can be quite challenging, particularly for large collections of data encompassing thousands of different projects, each with their own sampling strategies. In many cases, understanding the protocols used for a specific data collection project within a larger repository requires one to delve into the literature for that project. However, for many projects there do not exist accessible, standardized definitions or quantitative analysis of bias.

Black boxes, uncertainty, and interpretability

Machine learning models are frequently "black boxes," meaning that it is difficult to understand how a prediction is being made. Ecologists are accustomed to models that are simpler to inspect and analyze, where they can confidently determine what factors are most important and what the effect of different factors might be. Because the results of ecological models are used to drive policy, being able to interpret how a model is making predictions and avoid inaccuracies due to overfitting is important. This is closely related to trust (or lack thereof) in model outputs and the need for uncertainty quantification, particularly in scenarios where models are being asked to generalize to new locations or forward in time.

Norms surrounding data sharing and open sourcing in ecology

Computer science has benefited from strong community norms promoting public data and open-sourced code. One consequence of this shift is that it is easy for computer scientists to take data for granted and to be frustrated when a scientist is unwilling to share their data publicly. However, it is important to remember that in some fields data can be extremely expensive to collect and curate. The cost of the hardware, travel to the study site, and the time needed to place the sensors and maintain the sensor network quickly adds up. Add to this the number of hours it takes for an expert to process and label the data so that it is ready for analysis, and it is easy to see why a researcher would want to publish several papers on their hard-won data before sharing it publicly. On the other hand, public datasets like those hosted on LILA.science [2] have clear benefits for the community such as promoting reproducible research. Properly attributing data to the researchers who collected it (*e.g.*, through the use of "DOIs for datasets" [155]) could encourage more open data sharing in ecology. Data sharing norms are changing and many researchers are now happy to share their data and are pushing for more open data practices [149, 151], but it is important to be aware of this cultural difference between computer science and other fields.

Model handoffs, deployment, and accessibility

Once a machine learning method has been rigorously evaluated and found to be helpful, it is important to ensure these techniques are accessible to those who can put them to good use. In computer science, we have a culture of "open code, open data" which means that for most papers, all of the data and code is publicly available. However, ecologists may be less familiar with machine learning packages like PyTorch and TensorFlow, and may not have access to the computational resources required to train models on their data. If a method is to have real impact for the ecology community, it is important to provide models and code in a format that is accessible to end-users and well-documented. If the model is meant to become an integral part of an ecology workflow, plans for model maintenance and upkeep should be discussed.

Sensitive species

It is common for ecologists to obfuscate geolocation information before publishing any data containing rare or protected species to avoid poaching or stress from ecotourism. However, it is unclear whether obfuscation of GPS signal is sufficient to obscure the location of a photograph. It may be that a better solution is to remove any photos containing sensitive species, or to restrict sensitive access to a list of verified members of the research community. Second, the obfuscation distance of GPS location in published datasets might have a large effect on the accuracy of an SDM or other ecological model, particularly when both the training and validation data have been obfuscated. This obfuscation will further effect the reproducibility of a study, as results with or without obfuscation might be quite different.

3.7 What data is available and accessible?

There is an increasing number of publicly available ecological datasets that can be used for model training and evaluation. In this section we provide a few useful data sources as a starting point. We make a distinction between "analysis-ready" datasets which package species observations and covariates together and other data sources which can be combined to produce analysis-ready datasets.

Traditional analysis-ready datasets for multi-species distribution modeling

- The comprehensive SDM comparison in [134] uses five presence-absence datasets covering different species and parts of the world. Each dataset has a different set of covariates (min 6, max 38) and a different set of species (min 50, max 242). The datasets are available for download on Zenodo [132].
- The recently released benchmark dataset [64] covers 226 species from 6 regions. Each region has a different set of covariates (min 11, max 13) and a different set of species (min 32, max 50).

Note that many "traditional" SDM datasets may not be large enough to train some of the more data-hungry machine learning methods.

Large-scale analysis-ready datasets for multi-species distribution modeling

- The GeoLifeCLEF datasets combine 2D patches of covariates with species observations from community science programs. The GeoLifeCLEF 2020 dataset [46] consists of 1.9M observations of 31k plant and animal species from France and the US, each of which is paired with high-resolution 2D covariates (satellite imagery, land cover, and altitude) in addition to traditional covariates. Previous editions of the GeoLifeCLEF dataset [36, 50] are also available, and are suitable for large-scale plant-focused species distribution modeling in France using traditional covariates. Note that all of the GeoLifeCLEF datasets are based on presence-only observations, so performance is typically evaluated using information retrieval metrics such as top-*k* accuracy.
- The eBird Reference Dataset (ERD) [127] is built around checklists collected by eBird community members. In particular, it is limited to checklists for

which the observer (i) asserts that they reported everything they saw and (ii) quantified their sampling effort. This allows unobserved species to be interpreted as absences if sufficient sampling effort has been expended. The resulting presence/absence data is combined with land cover and climate variables. Unfortunately, the ERD does not appear to be maintained or publicly available as of November 2020.

Sources for species observation data

- The Global Biodiversity Information Facility (GBIF) [1] aggregates and organizes species observation data from over 1700 institutions around the world. We discuss a few specific contributors below.
- iNaturalist [9] is a community science project that has produced over 70 million point observations of species across the entire taxonomic tree. The data can be noisy as it is collected and labeled by non-experts.
- eBird [5] is a community science project hosted by the Cornell Lab of Ornithology which has produced more than 77 million birding checklists. These checklists provide both presence and absence, but absences can be noisy as it is possible the birder did not observe every species that was present at a given location.
- Movebank [3] is a database of animal tracking data hosted by the Max Planck Institute of Animal Behavior. It contains GPS tracking data for individual animals, covering 900 taxa and including 2.2 billion unique location readings.

Sources for covariates

Earth observation datasets and their derived products can be freely obtained from many sources, including the NASA Open Data Portal [12], the USGS Land Processes Distributed Access Data Archive [11], ESA Earth Online [7], and Google Earth Engine [6]. Also see the detailed discussion of covariates in Section 3.4.

Sources for training species identification models

Species observation data can be produced by classifying the species found in geolocated images. Those who are interested in the species classification problem may be interested in the datasets below.

- The iNaturalist species classification datasets [187, 188] are curated species classification datasets built from research-grade observations in iNaturalist.
- LILA.science [2, 28, 135] hosts a number of biology-focused image classification datasets, including camera trap datasets covering diverse species and locations.
- The Fine-Grained Visual Categorization (FGVC) workshop [16] at CVPR hosts a number of competitions each year [16, 27, 29, 31, 32, 129, 174, 176, 187] which focus on species classification and related biodiversity tasks.

3.8 Open problems

There are many open problems in SDM that may benefit from machine learning tools. In this section we discuss a few of these problems which we find particularly interesting.

Scaling up, geospatially and taxonomically

One of the main challenges in modern SDMs is scale. This includes scaling up SDMs to efficiently handle large geographic regions [100, 107, 178], many-species communities [133, 145, 179, 199], and large volumes of training data [122, 179, 200]. One particularly interesting question is whether jointly modeling many species could lead to SDMs which are significantly better than those based on modeling species independently.

Incorporating ecological theory and expert knowledge

There is a considerably amount of domain knowledge and ecological theory which would ideally be incorporated into SDMs [85]. This might include knowledge about species dispersal [25, 52, 72, 125], spatial patterns of community composition [44, 49, 102], and constraints on species ranges (*e.g.*, cliffs, water) [47, 65, 69, 125]. Another area of significant interest is to factor in cross-species biological processes such as niche exclusion/competition [146, 198], predator/prey dynamics [57, 146, 181], phylogenetic niche evolution [42, 74, 141], or models linked across functional traits [45, 147, 192]. These types of “domain-aware” algorithms are an active research area in the machine learning community [34, 54, 82, 171].

Fusing data

A third open area of investigation centers on how to best incorporate and utilize data collected at different spatiotemporal scales or in heterogeneous formats. This includes combining presence-only, presence-absence, abundance, and individual data such as GPS telemetry data [67, 101, 139, 143]. It also includes multi-scale or cross-scale modeling [173, 184], such as microclimate niche vs. macroscale niche [112], individual niche variance vs. species level niche variance[67], and cross-scale ecological processes[83, 120]. Finally, it may also include models of temporal ecological processes, such as seasonal range shifts and migrations [166, 177].

Evaluation

How should we compare competing models and decide which models to trust? Naturally, fair head-to-head evaluation of different models will be important [19, 60, 134]. Future large-scale evaluations may require accounting for biases in species observation data [68, 114, 189, 195], especially that which comes from community science projects. It is important to keep in mind that there is no single metric which makes one SDM better than another. It may be significant to understand how a model's predictions change under novel climate scenarios [21, 39, 69, 113] or different conservation policies [59, 121, 165] or how well-calibrated the SDM predictions are [19, 79]. One promising avenue is to study models in increasingly realistic simulation environments [105, 123, 201], which allows for more comprehensive analysis. Many of these topics are directly related to active areas of machine learning research, such as domain adaptation and overcoming dataset bias and imbalance [108].

3.9 Conclusion

We have sought to introduce machine learning researchers to a challenging and important real-world problem domain. We have discussed common terminology and highlighted common pitfalls and challenges. To lower the initial overhead, we have inventoried some available datasets and common methods. We hope that this document is useful for any computer scientist interested in bringing machine learning expertise to species distribution modeling.

Acknowledgements

Our research for this paper included informational interviews with Meredith Palmer, Michael Tabak, Corrie Moreau, and Carrie Seltzer. Their insights into the unique challenges of species distribution modeling was invaluable. This work was supported in part by the Caltech Resnick Sustainability Institute and NSFGRFP Grant No. 1745301. The views expressed in this work are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] The Global Biodiversity Information Facility. https://www.gbif.org/.

[2] Lila.science. http://lila.science/. Accessed: 2019-10-22.

[3] Movebank. https://www.movebank.org/cms/movebank-main.

[4] Wildlife Insights. https://www.wildlifeinsights.org/home.

[5] eBird. https://ebird.org/home.

[6] Google Earth Engine. https://earthengine.google.com.

[7] ESA Earth Online. https://www.earth.esa.int/.

[8] Map of Life: Von Der Decken's Hornbill. https://mol.org/species/map/Tockus_deckeni.

[9] iNaturalist. https://www.inaturalist.org/.

[10] 50 million observations on iNaturalist! https://www.inaturalist.org/blog/40699-50-million-observations-on-inaturalist.

[11] USGS LPDAAC. https://lpdaac.usgs.gov.

[12] NASA Open Data Portal. https://www.data.nasa.gov/.

[13] Workshop and Challenge on Computer Vision for Wildlife Conservation at ICCV. https://cvwc2019.github.io/, 2019.

[14] AI for Animal Re-Identification Workshop at WACV. https://sites.google.com/corp/view/wacv2020animalreid/, 2020.

[15] OOS 64–Deep learning for image analysis in ecology, Session at the Ecological Society of America yearly meeting. https://eco.confex.com/eco/2020/meetingapp.cgi/Session/17295, 2020.

[16] Fine Grained Visual Categorization Workshop at CVPR. http://www.fgvc.org/, 2022.

[17] Bader H. Alhajeri and Yoan Fourcade. High correlation between species-level environmental data estimates extracted from iucn expert range maps and from gbif occurrence data. *Journal of Biogeography*, 46(7):1329–1341, 2019.

[18] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the International Conference on Computer Vision*, 2019.

[19] Miguel B. Araújo and Antoine Guisan. Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, 33(10):1677–1688, 2006. ISSN 1365-2699. doi: 10.1111/j.1365-2699.2006.01584.x.

[20] Michael Phillip Austin. Continuum concept, ordination methods, and niche theory. *Annual Review of Ecology, Evolution, and Systematics*, 16(1):39–61, 1985.

[21] Mike P. Austin and Kimberly P. Van Niel. Improving species distribution models for climate change studies: Variable selection and scale. *Journal of Biogeography*, 38(1):1–8, 2011. ISSN 1365-2699. doi: 10.1111/j.1365-2699.2010.02416.x.

[22] Laila Bahaa-El-Din, David Mills, Luke Hunter, and Philipp Henschel. Caracal aurata, 04 2015.

[23] Larissa L. Bailey, Darryl I. MacKenzie, and James D. Nichols. Advances and applications of occupancy models. *Methods in Ecology and Evolution*, 5 (12):1269–1279, 2014.

[24] Abdou Bannari, Daniel Morin, Ferdinand Bonn, and Alfredo R. Huete. A review of vegetation indices. *Remote Sensing Reviews*, 13(1-2):95–120, 1995.

[25] Narayani Barve, Vijay Barve, Alberto Jiménez-Valverde, Andrés Lira-Noriega, Sean P. Maher, A. Townsend Peterson, Jorge Soberón, and Fabricio Villalobos. The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, 222(11):1810–1819, June 2011. ISSN 0304-3800. doi: 10.1016/j.ecolmodel.2011.02.011.

[26] Colin M. Beale and Jack J. Lennon. Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1586):247–258, 2012.

[27] Sara Beery, Grant Van Horn, Oisin MacAodha, and Pietro Perona. The iWildCam 2018 challenge dataset. *The Fifth Fine-Grained Visual Categorization Workshop at CVPR*, 2018.

[28] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[29] Sara Beery, Dan Morris, and Pietro Perona. The iWildCam 2019 challenge dataset. *The Sixth Fine-Grained Visual Categorization Workshop at CVPR*, 2019.

[30] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *In the Data Mining and Artificial Intelligence for Conservation Workshop at Knowledge Discovery in Databases (KDD)*, 2019. *selected to be featured at KDD Earth Day.

[31] Sara Beery, Elijah Cole, and Arvi Gjoka. The iWildCam 2020 competition dataset. *The Seventh Fine-Grained Visual Categorization Workshop at CVPR*, 2020.

[32] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iWildCam 2021 competition dataset. *The Eighth Fine-Grained Visual Categorization Workshop at CVPR*, 2021.

[33] Tanya Y. Berger-Wolf, Daniel I. Rubenstein, Charles V. Stewart, Jason A. Holmberg, Jason Parham, Sreejith Menon, Jonathan P. Crall, Jon Van Oast, Emre Kiciman, and Lucas Joppa. Wildbook: Crowdsourcing, computer vision, and data science for conservation. *CoRR*, abs/1710.08880, 2017. URL `http://arxiv.org/abs/1710.08880`.

[34] Christopher M. Bishop. Model-based machine learning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20120222, 2013.

[35] Trevor H. Booth, Henry A. Nix, John R. Busby, and Michael F. Hutchinson. Bioclim: the first species distribution modelling package, its early applications and relevance to most current maxent studies. *Diversity and Distributions*, 20(1):1–9, 2014.

[36] Christophe Botella, Maximilien Servajean, Pierre Bonnet, and Alexis Joly. Overview of GeoLifeCLEF 2019: Plant species prediction using environment and animal occurrences. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum*, number 2380, 2019.

[37] David S. Broomhead and David Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.

[38] Stephen T. Buckland, David R. Anderson, Kenneth P. Burnham, and Jeffrey L. Laake. Distance sampling. *Encyclopedia of biostatistics*, 2, 2005.

[39] Laëtitia Buisson, Wilfried Thuiller, Nicolas Casajus, Sovan Lek, and Gaël Grenouillet. Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, 16(4):1145–1157, 2010. ISSN 1365-2486. doi: 10.1111/j.1365-2486.2009.02000.x.

[40] J.R. Busby. Bioclim-a bioclimate analysis and prediction system. *Plant Protection Quarterly*, 61:8–9, 1991.

[41] David E. Capen. *The use of multivariate statistics in studies of wildlife habitat*, volume 87. Rocky Mountain Forest and Range Experiment Station, U.S. Forest Service, 1981.

[42] Daniel S. Chapman, Romain Scalone, Edita Štefanić, and James M. Bullock. Mechanistic species distribution modeling reveals a niche shift during invasion. *Ecology*, 98(6):1671–1680, 2017. ISSN 1939-9170. doi: 10.1002/ecy.1835.

[43] Di Chen, Yexiang Xue, Shuo Chen, Daniel Fink, and Carla Gomes. Deep multi-species embedding. *arXiv preprint arXiv:1609.09353*, 2016.

[44] Di Chen, Yexiang Xue, Shuo Chen, Daniel Fink, and Carla Gomes. Deep Multi-Species Embedding. *arXiv:1609.09353 [cs, q-bio, stat]*, February 2017.

[45] James S. Clark, Diana Nemergut, Bijan Seyednasrollah, Phillip J. Turner, and Stacy Zhang. Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs*, 87 (1):34–56, 2017. ISSN 1557-7015. doi: 10.1002/ecm.1241.

[46] Elijah Cole, Benjamin Deneu, Titouan Lorieul, Maximilien Servajean, Christophe Botella, Dan Morris, Nebojsa Jojic, Pierre Bonnet, and Alexis Joly. The geolifeclef 2020 dataset. *arXiv preprint arXiv:2004.04192*, 2020.

[47] Jacob C. Cooper and Jorge Soberón. Creating individual accessible area hypotheses improves stacked species distribution model performance. *Global Ecology and Biogeography*, 27(1):156–165, 2018. ISSN 1466-8238. doi: 10.1111/geb.12678.

[48] D. Richard Cutler, Thomas C. Edwards Jr., Karen H. Beard, Adele Cutler, Kyle T. Hess, Jacob Gibson, and Joshua J. Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

[49] Manuela D'Amen, Jean-Nicolas Pradervand, and Antoine Guisan. Predicting richness and composition in mountain insect communities at high resolution: A new test of the SESAM framework. *Global Ecology and Biogeography*, 24(12):1443–1453, 2015. ISSN 1466-8238. doi: 10.1111/geb.12357.

[50] Benjamin Deneu, Maximilien Servajean, Christophe Botella, and Alexis Joly. Location-based species recommendation using co-occurrences and environment-GeoLifeCLEF 2018 challenge. In *Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum*, volume 2125, 2018.

[51] Benjamin Deneu, Maximilien Servajean, Christophe Botella, and Alexis Joly. Evaluation of deep species distribution models using environment and co-occurrences. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 213–225. Springer, 2019.

[52] Michele Di Musciano, Valter Di Cecco, Fabrizio Bartolucci, Fabio Conti, Anna Rita Frattaroli, and Luciano Di Martino. Dispersal ability of threatened species affects future distributions. *Plant Ecology*, 221(4):265–281, April 2020. ISSN 1573-5052. doi: 10.1007/s11258-020-01009-0.

[53] Solomon Z. Dobrowski, Hugh D. Safford, Yen Ben Cheng, and Susan L. Ustin. Mapping mountain vegetation using species distribution modeling, image-based texture analysis, and object-based classification. *Applied Vegetation Science*, 11(4):499–508, 2008.

[54] Bradley B. Doll, Dylan A. Simon, and Nathaniel D. Daw. The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 22 (6):1075–1081, 2012.

[55] Sami Domisch, Martin Friedrichs, Thomas Hein, Florian Borgwardt, Annett Wetzig, Sonja C. Jähnig, and Simone D. Langhans. Spatially explicit species distribution models: A missed opportunity in conservation planning? *Diversity and Distributions*, 25(5):758–769, 2019.

[56] Carsten F. Dormann, Oliver Purschke, Jaime R. Garcia Marquez, Sven Lautenbach, and Boris Schroeder. Components of uncertainty in species distribution analysis: A case study of the great grey shrike. *Ecology*, 89(12): 3371–3386, 2008.

[57] Carsten F. Dormann, Maria Bobrowski, D. Matthias Dehling, David J. Harris, Florian Hartig, Heike Lischke, Marco D. Moretti, Jörn Pagel, Stefan Pinkert, Matthias Schleuning, Susanne I. Schmidt, Christine S. Sheppard, Manuel J. Steinbauer, Dirk Zeuss, and Casper Kraan. Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. *Global Ecology and Biogeography*, 27(9):1004–1016, 2018. ISSN 1466-8238. doi: 10.1111/geb.12759.

[58] John M. Drake, Christophe Randin, and Antoine Guisan. Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, 43(3): 424–432, 2006.

[59] Sally Eaton, Christopher Ellis, David Genney, Richard Thompson, Rebecca Yahr, and Daniel T. Haydon. Adding small species to the big picture: Species distribution modelling in an age of landscape scale conservation. *Biological Conservation*, 217:251–258, January 2018. ISSN 0006-3207. doi: 10.1016/j.biocon.2017.11.012.

[60] Jane Elith and Catherine H. Graham. Do They? How Do They? Why Do They Differ? On Finding Reasons for Differing Performances of Species Distribution Models. *Ecography*, 32(1):66–77, 2009. ISSN 0906-7590.

[61] Jane Elith and John R. Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual review of Ecology, Evolution, and Systematics*, 40:677–697, 2009.

[62] Jane Elith, John R. Leathwick, and Trevor Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, 2008.

[63] Jane Elith, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J. Yates. A statistical explanation of maxent for ecologists. *Diversity and distributions*, 17(1):43–57, 2011.

[64] Jane Elith, Catherine Graham, Roozbeh Valavi, Meinrad Abegg, Caroline Bruce, Andrew Ford, Antoine Guisan, Robert J. Hijmans, Falk Huettmann, Lucia Lohmann, et al. Presence-only and presence-absence data for comparing species distribution modeling methods. *Biodiversity Informatics*, 15(2): 69–80, 2020.

[65] Robert M. Ewers, Charles J. Marsh, and Oliver R. Wearn. Making statistics biologically relevant in fragmented landscapes. *Trends in Ecology & Evolution*, 25(12):699–704, December 2010. ISSN 0169-5347. doi: 10.1016/j.tree.2010.09.008.

[66] Stephen E. Fick and Robert J. Hijmans. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315, 2017.

[67] John R. Fieberg, James D. Forester, Garrett M. Street, Douglas H. Johnson, Althea A. ArchMiller, and Jason Matthiopoulos. Used-habitat calibration plots: A new procedure for validating species distribution, resource selection, and step-selection models. *Ecography*, 41(5):737–752, 2018. ISSN 1600-0587. doi: 10.1111/ecog.03123.

[68] William Fithian, Jane Elith, Trevor Hastie, and David A. Keith. Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4):424–438, 2015. ISSN 2041-210X. doi: 10.1111/2041-210X.12242.

[69] Matthew C. Fitzpatrick and William W. Hargrove. The projection of species distribution models and the problem of non-analog climate. *Biodiversity and Conservation*, 18(8):2255, April 2009. ISSN 1572-9710. doi: 10.1007/s10531-009-9584-8.

[70] Robert J. Fletcher Jr., Trevor J. Hefley, Ellen P. Robertson, Benjamin Zuckerberg, Robert A. McCleery, and Robert M. Dorazio. A practical guide for combining data to model species distributions. *Ecology*, 100(6):e02710, 2019.

[71] Scott D. Foster and Piers K. Dunstan. The analysis of biodiversity using rank abundance distributions. *Biometrics*, 66(1):186–195, 2010.

[72] Janet Franklin. Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions*, 16(3):321–330, 2010. ISSN 1472-4642. doi: 10.1111/j.1472-4642.2010.00641.x.

[73] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[74] Daniel G. Gavin, Matthew C. Fitzpatrick, Paul F. Gugger, Katy D. Heath, Francisco Rodríguez-Sánchez, Solomon Z. Dobrowski, Arndt Hampe, Feng Sheng Hu, Michael B. Ashcroft, Patrick J. Bartlein, Jessica L. Blois, Bryan C. Carstens, Edward B. Davis, Guillaume de Lafontaine, Mary E. Edwards, Matias Fernandez, Paul D. Henne, Erin M. Herring, Zachary A. Holden, Woo-seok Kong, Jianquan Liu, Donatella Magri, Nicholas J. Matzke, Matt S. McGlone, Frédérik Saltré, Alycia L. Stigall, Yi-Hsin Erica Tsai, and John W. Williams. Climate refugia: Joint inference from fossil records, species distribution models and phylogeography. *New Phytologist*, 204(1): 37–54, 2014. ISSN 1469-8137. doi: 10.1111/nph.12929.

[75] Alan E Gelfand and Shinichiro Shirota. Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs*, 89(3):e01372, 2019.

[76] Andrew M. Gormley, David M. Forsyth, Peter Griffioen, Michael Lindeman, David S.L. Ramsey, Michael P. Scroggie, and Luke Woodford. Using presence-only and presence–absence data to estimate the current and potential distributions of established invasive species. *Journal of Applied Ecology*, 48(1):25–34, 2011.

[77] Catherine H. Graham and Robert J. Hijmans. A comparison of methods for mapping species ranges and species richness. *Global Ecology and biogeography*, 15(6):578–587, 2006.

[78] Annegret Grimm, Bernd Gruber, and Klaus Henle. Reliability of different mark-recapture methods for population size estimation tested against reference population sizes constructed from field data. *PLoS One*, 9(6):e98840, 2014.

[79] Liam Grimmett, Rachel Whitsed, and Ana Horta. Presence-only species distribution models are sensitive to sample prevalence: Evaluating models using spatial prediction stability and accuracy metrics. *Ecological Modelling*, 431:109194, September 2020. ISSN 0304-3800. doi: 10.1016/j.ecolmodel.2020.109194.

[80] Joseph Grinnell. The origin and distribution of the chest-nut-backed chickadee. *The Auk*, 21(3):364–382, 1904.

[81] Red List Technical Working Group et al. Mapping standards and data quality for the IUCN red list categories and criteria, 2018.

[82] Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration. In *International Conference on Machine Learning*, pages 2829–2838. PMLR, 2016.

[83] Gurutzeta Guillera-Arroita, José J. Lahoz-Monfort, Jane Elith, Ascelin Gordon, Heini Kujala, Pia E. Lentini, Michael A. McCarthy, Reid Tingley, and

Brendan A. Wintle. Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, 24 (3):276–292, 2015. ISSN 1466-8238. doi: 10.1111/geb.12268.

[84] Antoine Guisan and Wilfried Thuiller. Predicting species distribution: offering more than simple habitat models. *Ecology letters*, 8(9):993–1009, 2005.

[85] Antoine Guisan and Niklaus E. Zimmermann. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3):147–186, 2000.

[86] W. Hallgren, F. Santana, S. Low-Choy, Y. Zhao, and B. Mackey. Species distribution models can be highly sensitive to algorithm configuration. *Ecological Modelling*, 408:108719, 2019.

[87] Thomas A. Hanley. A comparison of the line interception and quadrat estimation methods of determining shrub canopy coverage. *Rangeland Ecology & Management/Journal of Range Management Archives*, 31(1):60–62, 1978.

[88] David J. Harris. Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, 6(4):465–473, 2015.

[89] Trevor Hastie and Will Fithian. Inference from presence-only data; the ongoing controversy. *Ecography*, 36(8):864–867, 2013.

[90] Kate S He, Bethany A Bradley, Anna F Cord, Duccio Rocchini, Mao-Ning Tuanmu, Sebastian Schmidtlein, Woody Turner, Martin Wegmann, and Nathalie Pettorelli. Will remote sensing shape the next generation of species distribution models? *Remote Sensing in Ecology and Conservation*, 1(1):4–18, 2015.

[91] Harold F. Heady and R.W. Gibbens. A comparison of the charting, line intercept, and line point methods of sampling shrub types of vegetation. *Rangeland Ecology & Management/Journal of Range Management Archives*, 12(4):180–188, 1959.

[92] Emilie B Henderson, Janet L Ohmann, Matthew J Gregory, Heather M Roberts, and Harold Zald. Species distribution modelling for plant communities: stacked single species or multivariate modelling approaches? *Applied vegetation science*, 17(3):516–527, 2014.

[93] Tomislav Hengl, Jorge Mendes de Jesus, Gerard BM Heuvelink, Maria Ruiperez Gonzalez, Milan Kilibarda, Aleksandar Blagotić, Wei Shangguan, Marvin N Wright, Xiaoyuan Geng, Bernhard Bauer-Marschallinger, et al. Soilgrids250m: Global gridded soil information based on machine learning. *PLoS One*, 12(2):e0169748, 2017.

[94] Motoki Higa, Yuichi Yamaura, Itsuro Koizumi, Yuki Yabuhara, Masayuki Senzaki, and Satoru Ono. Mapping large-scale bird distributions using occupancy models and citizen science data with spatially biased sampling effort. *Diversity and Distributions*, 2014.

[95] Collin Homer, Jon Dewitz, Limin Yang, Suming Jin, Patrick Danielson, George Xian, John Coulston, Nathaniel Herold, James Wickham, and Kevin Megown. Completion of the 2011 national land cover database for the conterminous united states–representing a decade of land cover change information. *Photogrammetric Engineering & Remote Sensing*, 81(5):345–354, 2015.

[96] Francis K.C. Hui, David I. Warton, Scott D. Foster, and Piers K. Dunstan. To mix or not to mix: comparing the predictive performance of mixture models vs. separate species distribution models. *Ecology*, 94(9):1913–1919, 2013.

[97] Allen H. Hurlbert and Walter Jetz. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences*, 104(33):13384–13389, 2007.

[98] Allen H. Hurlbert and Ethan P. White. Disparity between range map-and survey-based analyses of species richness: patterns, processes and implications. *Ecology Letters*, 8(3):319–327, 2005.

[99] Walter Jetz, Jana M. McPherson, and Robert P. Guralnick. Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in ecology & evolution*, 27(3):151–159, 2012.

[100] Walter Jetz, Melodie McGeoch, Guralnick Robert, Simon Ferrier, Jan Beck, Mark Costello, Miguel Fernández, Gary Geller, Petr Keil, Cory Merow, Carsten Meyer, Frank Muller-Karger, Eugenie Regan, Dirk Schmeller, and Eren Turak. Essential biodiversity variables for mapping and monitoring species populations. *Nature Ecology & Evolution*, 3, March 2019. doi: 10.1038/s41559-019-0826-1.

[101] Chris J. Johnson and Michael P. Gillingham. Sensitivity of species-distribution models to error, bias, and model design: An application to resource selection functions for woodland caribou. *Ecological Modelling*, 213(2):143–155, May 2008. ISSN 0304-3800. doi: 10.1016/j.ecolmodel.2007.11.013.

[102] Maxwell B. Joseph. Neural hierarchical models of ecological populations. *Ecology Letters*, 23(4):734–747, 2020. ISSN 1461-0248. doi: 10.1111/ele.13462.

[103] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[104] Lou Jost. Partitioning diversity into independent alpha and beta components. *Ecology*, 88(10):2427–2439, 2007.

[105] Paulo De Marco Júnior and Caroline Corrêa Nóbrega. Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. *PLOS One*, 13(9):e0202403, September 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0202403.

[106] Jamie M. Kass, Bruno Vilela, Matthew E. Aiello-Lammens, Robert Muscarella, Cory Merow, and Robert P. Anderson. Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. *Methods in Ecology and Evolution*, 9(4):1151–1156, 2018.

[107] W. Daniel Kissling, Ramona Walls, Anne Bowser, Matthew O. Jones, Jens Kattge, Donat Agosti, Josep Amengual, Alberto Basset, Peter M. van Bodegom, Johannes H. C. Cornelissen, Ellen G. Denny, Salud Deudero, Willi Egloff, Sarah C. Elmendorf, Enrique Alonso García, Katherine D. Jones, Owen R. Jones, Sandra Lavorel, Dan Lear, Laetitia M. Navarro, Samraat Pawar, Rebecca Pirzl, Nadja Rüger, Sofia Sal, Roberto Salguero-Gómez, Dmitry Schigel, Katja-Sabine Schulz, Andrew Skidmore, and Robert P. Guralnick. Towards global data products of Essential Biodiversity Variables on species traits. *Nature Ecology & Evolution*, 2(10):1531–1540, October 2018. ISSN 2397-334X. doi: 10.1038/s41559-018-0667-3.

[108] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[109] Vira Koshkina, Yan Wang, Ascelin Gordon, Robert M. Dorazio, Matt White, and Lewi Stone. Integrated species distribution models: combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, 2017.

[110] Margaret Kosmala, Andrea Wiggins, Alexandra Swanson, and Brooke Simmons. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, 14(10):551–560, 2016.

[111] Thierry Lassueur, Stéphane Joost, and Christophe F Randin. Very high resolution digital elevation models: Do they improve models of plant species distribution? *Ecological Modelling*, 198(1-2):139–153, 2006.

[112] Jonas J. Lembrechts, Ivan Nijs, and Jonathan Lenoir. Incorporating microclimate into species distribution models. *Ecography*, 42(7):1267–1279, 2019. ISSN 1600-0587. doi: 10.1111/ecog.03947.

[113] Wanwan Liang, Monica Papeş, Liem Tran, Jerome Grant, Robert Washington-Allen, Scott Stewart, and Gregory Wiggins. The effect of pseudo-absence

selection method on transferability of species distribution models in the context of non-adaptive niche shift. *Ecological Modelling*, 388:1–9, November 2018. ISSN 0304-3800. doi: 10.1016/j.ecolmodel.2018.09.018.

[114] Canran Liu, Matt White, and Graeme Newell. Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, 40(4):778–789, 2013. ISSN 1365-2699. doi: 10.1111/jbi.12058.

[115] David M. Lodge, Cameron R. Turner, Christopher L. Jerde, Matthew A. Barnes, Lindsay Chadderton, Scott P. Egan, Jeffrey L. Feder, Andrew R. Mahon, and Michael E. Pfrender. Conservation in a cup of water: Estimating biodiversity and population abundance from environmental dna. *Molecular Ecology*, 21(11):2555–2558, 2012.

[116] Robert H. MacArthur. Population ecology of some warblers of northeastern coniferous forests. *Ecology*, 39(4):599–619, 1958.

[117] Darryl I. MacKenzie, James D. Nichols, Gideon B. Lachman, Sam Droege, J. Andrew Royle, and Catherine A. Langtimm. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255, 2002.

[118] Kumar Mainali, Trevor Hefley, Leslie Ries, and William F. Fagan. Matching expert range maps with species distribution model predictions. *Conservation Biology*, 34(5):1292–1304, 2020. doi: 10.1111/cobi.13492.

[119] Tiago A. Marques, Lisa Munger, Len Thomas, Sean Wiggins, and John A Hildebrand. Estimating north pacific right whale eubalaena japonica density using passive acoustic cue counting. *Endangered Species Research*, 13(3): 163–172, 2011.

[120] Jason Matthiopoulos, John Fieberg, and Geert Aarts. *Species-Habitat Associations: Spatial Data, Predictive Models, and Ecological Insights*. University of Minnesota Libraries Publishing, December 2020. doi: 10.24926/2020.081320.

[121] William J. McSHEA. What are the roles of species distribution models in conservation planning? *Environmental Conservation*, 41(2):93–96, June 2014. ISSN 0376-8929, 1469-4387. doi: 10.1017/S0376892913000581.

[122] Cory Merow, Matthew J. Smith, and John A. Silander. A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography*, 36(10):1058–1069, 2013. ISSN 1600-0587. doi: 10.1111/j.1600-0587.2013.07872.x.

[123] Christine N. Meynard, Boris Leroy, and David M. Kaplan. Testing methods in species distribution modelling using virtual species: What have we learnt and what are we missing? *Ecography*, 42(12):2021–2036, 2019. ISSN 1600-0587. doi: 10.1111/ecog.04385.

[124] David AW Miller, Krishna Pacifici, Jamie S Sanderlin, and Brian J Reich. The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, 10(1):22–37, 2019.

[125] Jennifer A Miller and Paul Holloway. Incorporating movement in species distribution models. *Progress in Physical Geography: Earth and Environment*, 39(6):837–849, December 2015. ISSN 0309-1333. doi: 10.1177/0309133315580890.

[126] Ans M. Mouton, Bernard De Baets, and Peter L.M. Goethals. Ecological relevance of performance criteria for species distribution models. *Ecological Modelling*, 221(16):1995–2002, 2010.

[127] M. Arthur Munson, Kevin Webb, Daniel Sheldon, Daniel Fink, Wesley M. Hochachka, Marshall Iliff, Mirek Riedewald, Daria Sorokina, Brian Sullivan, Christopher Wood, et al. The ebird reference dataset. *Cornell Lab of Ornithology and National Audubon Society, Ithaca, NY [En linea]: http://www. avianknowledge. net/content. Acceso: Julio*, 2011.

[128] Andrew Murray. *The geographical distribution of mammals*. Ripol Class (translated from Russian), 1866.

[129] Ernest Mwebaze, Timnit Gebru, Andrea Frome, Solomon Nsumba, and Jeremy Tusubira. icassava 2019 fine-grained visual categorization challenge, 2019.

[130] Babak Naimi, Nicholas A.S. Hamm, Thomas A. Groen, Andrew K. Skidmore, and Albertus G. Toxopeus. Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37(2):191–203, 2014.

[131] John Ashworth Nelder and Robert W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3): 370–384, 1972.

[132] Anna Norberg. aminorberg/sdm-comparison: Norberg et al. (2019), April 2019. URL `https://doi.org/10.5281/zenodo.2637812`.

[133] Anna Norberg, Nerea Abrego, F. Guillaume Blanchet, Frederick R. Adler, Barbara J. Anderson, Jani Anttila, Miguel B. Araújo, Tad Dallas, David Dunson, Jane Elith, Scott D. Foster, Richard Fox, Janet Franklin, William Godsoe, Antoine Guisan, Bob O'Hara, Nicole A. Hill, Robert D. Holt, Francis K. C. Hui, Magne Husby, John Atle Kålås, Aleksi Lehikoinen, Miska Luoto, Heidi K. Mod, Graeme Newell, Ian Renner, Tomas Roslin, Janne Soininen, Wilfried Thuiller, Jarno Vanhatalo, David Warton, Matt White, Niklaus E. Zimmermann, Dominique Gravel, and Otso Ovaskainen. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89(3):e01370, 2019. ISSN 1557-7015. doi: 10.1002/ecm.1370.

[134] Anna Norberg, Nerea Abrego, F Guillaume Blanchet, Frederick R Adler, Barbara J Anderson, Jani Anttila, Miguel B Araújo, Tad Dallas, David Dunson, Jane Elith, et al. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89(3):e01370, 2019.

[135] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25): E5716–E5725, 2018.

[136] Stan Openshaw. *The Modifiable Areal Unit Problem*. Norwick, 1984.

[137] Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, F. Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, and Nerea Abrego. How to make more out of community data? a conceptual framework and its implementation as models and software. *Ecology Letters*, 20(5):561–576, 2017.

[138] Stacy L. Özesmi and Uygar Özesmi. An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling*, 116(1):15–31, 1999.

[139] Krishna Pacifici, Brian J. Reich, David A.W. Miller, Beth Gardner, Glenn Stauffer, Susheela Singh, Alexa McKerrow, and Jaime A. Collazo. Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, 98(3):840–850, 2017.

[140] Jennie Pearce and Simon Ferrier. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, 128(2-3):127–147, 2000.

[141] Peter B. Pearman, Antoine Guisan, Olivier Broennimann, and Christophe F. Randin. Niche dynamics in space and time. *Trends in Ecology & Evolution*, 23(3):149–158, March 2008. ISSN 0169-5347. doi: 10.1016/j.tree.2007.11.005.

[142] A. Townsend Peterson and Jorge Soberón. Species distribution modeling and ecological niche modeling: getting the concepts right. *Natureza & Conservação*, 10(2):102–107, 2012.

[143] Steven Phillips and Jane Elith. Logistic Methods for Resource Selection Functions and Presence-Only Species Distribution Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 25(1), August 2011. ISSN 2374-3468.

[144] Steven J. Phillips, Robert P. Anderson, and Robert E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4):231–259, 2006.

[145] Maximilian Pichler and Florian Hartig. A new method for faster and more accurate inference of species associations from big community data. *arXiv:2003.05331 [q-bio, stat]*, October 2020.

[146] Giovanni Poggiato, Tamara Münkemüller, Daria Bystrova, Julyan Arbel, James S. Clark, and Wilfried Thuiller. On the Interpretations of Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, February 2021. ISSN 0169-5347. doi: 10.1016/j.tree.2021.01.002.

[147] Laura J. Pollock, William K. Morris, and Peter A. Vesk. The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*, 35(8):716–725, 2012. ISSN 1600-0587. doi: 10.1111/j.1600-0587.2011.07085.x.

[148] Kevin L. Pope, Steve E. Lochmann, and Michael K. Young. Methods for assessing fish populations. *In: Hubert, Wayne A; Quist, Michael C., eds. Inland Fisheries Management in North America, 3rd edition. Bethesda, MD: American Fisheries Society: 325-351.*, pages 325–351, 2010.

[149] Stephen M Powers and Stephanie E Hampton. Open science, reproducibility, and transparency in ecology. *Ecological Applications*, 29(1):e01822, 2019.

[150] Jean-Nicolas Pradervand, Anne Dubuis, Loïc Pellissier, Antoine Guisan, and Christophe Randin. Very high resolution environmental predictors in species distribution models: Moving beyond topography? *Progress in Physical Geography*, 38(1):79–96, 2014.

[151] O. James Reichman, Matthew B. Jones, and Mark P. Schildhauer. Challenges and opportunities of open data in ecology. *Science*, 331(6018):703–705, 2011.

[152] Ian W. Renner and David I. Warton. Equivalence of maxent and poisson point process models for species distribution modeling in ecology. *Biometrics*, 69 (1):274–281, 2013.

[153] Corinne L. Richards, Bryan C. Carstens, and L. Lacey Knowles. Distribution modelling and statistical phylogeography: An integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography*, 34(11):1833–1845, 2007.

[154] David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, Jose J. Lahoz-Monfort, Boris Schoder, Wilfried Thuiller, David I. Warton, Brandan A. Wintle, Florian Hartig, and Carsten F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, pages 913–929, 2016.

[155] Tim Robertson, Serge Belongie, Adam Hartwig, Christine Kaeser-Chen, Chenyang Zhang, Kiat Chuan Tan, Yulong Liu, Denis Brulé, Cédric Deltheil, Scott Loarie, et al. Training machines to identify species using gbif-mediated datasets. *Biodiversity Information Science and Standards*, 2019.

[156] Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. Large scale high-resolution land cover mapping with multi-resolution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12726–12735, 2019.

[157] Duccio Rocchini, Joaquín Hortal, Szabolcs Lengyel, Jorge M. Lobo, Alberto Jimenez-Valverde, Carlo Ricotta, Giovanni Bacaro, and Alessandro Chiarucci. Accounting for uncertainty when mapping species distributions: the need for maps of ignorance. *Progress in Physical Geography*, 35(2): 211–226, 2011.

[158] J. Marcus Rowcliffe, Juliet Field, Samuel T. Turvey, and Chris Carbone. Estimating animal density using camera traps without the need for individual recognition. *Journal of Applied Ecology*, pages 1228–1236, 2008.

[159] Eric W. Sanderson, Malanding Jaiteh, Marc A. Levy, Kent H. Redford, Antoinette V. Wannebo, and Gillian Woolmer. The human footprint and the last of the wild: The human footprint is a global map of human influence on the land surface, which suggests that human beings are stewards of nature, whether we like it or not. *BioScience*, 52(10):891–904, 2002.

[160] Andreas Franz Wilhelm Schimper. *Plant-geography Upon a Physiological Basis...* Clarendon Press, 1903.

[161] Zoe Emily Schnabel. The estimation of the total fish population of a lake. *The American Mathematical Monthly*, 45(6):348–352, 1938.

[162] Patrick Schratz, Jannes Muenchow, Eugenia Iturritxa, Jakob Richter, and Alexander Brenning. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, 406:109–120, 2019.

[163] Boris Schroeder. Challenges of species distribution modeling belowground. *Journal of Plant Nutrition and Soil Science*, 171(3):325–337, 2008.

[164] Sebastian Seibold, Martin M. Gossner, Nadja K. Simons, Nico Blüthgen, Jörg Müller, Didem Ambarlı, Christian Ammer, Jürgen Bauhus, Markus Fischer, Jan C. Habel, et al. Arthropod decline in grasslands and forests is associated with landscape-level drivers. *Nature*, 574(7780):671–674, 2019.

[165] Steve J. Sinclair, Matthew D. White, and Graeme R. Newell. How Useful Are Species Distribution Models for Managing Biodiversity under Future Climates? *Ecology and Society*, 15(1), 2010. ISSN 1708-3087.

[166] Andrea Soriano-Redondo, Charlotte M. Jones-Todd, Stuart Bearhop, Geoff M. Hilton, Leigh Lock, Andrew Stanbury, Stephen C. Votier, and Janine B. Illian. Understanding species distribution in dynamic populations: A new approach using spatio-temporal point process models. *Ecography*, 42 (6):1092–1102, 2019. ISSN 1600-0587. doi: 10.1111/ecog.03771.

[167] Jakub Stoklosa, Christopher Daly, Scott D. Foster, Michael B. Ashcroft, and David I. Warton. A climate of uncertainty: Accounting for error in climate variables for species distribution models. *Methods in Ecology and Evolution*, 6(4):412–423, 2015.

[168] D.F. Stuffer. Linking populations and habitats: Where have we been? where are we going? *Predicting species occurrences: Issues of accuracy and scale*, 2002.

[169] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*, 2:150026, 2015.

[170] Frederick C. Sweeney and John M. Hopkinson. Vegetative growth of nineteen tropical and sub-tropical pasture grasses and legumes in relation to temperature. *Tropical Grasslands*, 9(3):209–217, 1975.

[171] Renee Swischuk, Laura Mainini, Benjamin Peherstorfer, and Karen Willcox. Projection-based model reduction: Formulations for physics-based machine learning. *Computers & Fluids*, 179:704–717, 2019.

[172] Nicholas W. Synes and Patrick E. Osborne. Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change. *Global Ecology and Biogeography*, 20(6):904–914, 2011.

[173] Matthew V. Talluto, Isabelle Boulangeat, Aitor Ameztegui, Isabelle Aubin, Dominique Berteaux, Alyssa Butler, Frédérik Doyon, C. Ronnie Drever, Marie-Josée Fortin, Tony Franceschini, Jean Liénard, Dan McKenney, Kevin A. Solarik, Nikolay Strigul, Wilfried Thuiller, and Dominique Gravel. Cross-scale integration of knowledge for predicting species ranges: A metamodelling framework. *Global Ecology and Biogeography*, 25(2):238–249, 2016. ISSN 1466-8238. doi: 10.1111/geb.12395.

[174] Kiat Chuan Tan, Yulong Liu, Barbara Ambrose, Melissa Tulig, and Serge Belongie. The herbarium challenge 2019 dataset, 2019.

[175] Luming Tang, Yexiang Xue, Di Chen, and Carla Gomes. Multi-entity dependence learning with rich context via conditional variational auto-encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[176] Ranjita Thapa, Noah Snavely, Serge Belongie, and Awais Khan. The plant pathology 2020 challenge dataset to classify foliar disease of apples, 2020.

[177] James T. Thorson, James N. Ianelli, Elise A. Larsen, Leslie Ries, Mark D. Scheuerell, Cody Szuwalski, and Elise F. Zipkin. Joint dynamic species distribution models: A tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography*, 25(9):1144–1158, 2016. ISSN 1466-8238. doi: 10.1111/geb.12464.

[178] Wilfried Thuiller, Cécile Albert, Miguel B. Araújo, Pam M. Berry, Mar Cabeza, Antoine Guisan, Thomas Hickler, Guy F. Midgley, James Paterson, Frank M. Schurr, Martin T. Sykes, and Niklaus E. Zimmermann. Predicting global change impacts on plant species' distributions: Future challenges. *Perspectives in Plant Ecology, Evolution and Systematics*, 9(3):137–152, March 2008. ISSN 1433-8319. doi: 10.1016/j.ppees.2007.09.004.

[179] Gleb Tikhonov, Li Duan, Nerea Abrego, Graeme Newell, Matt White, David Dunson, and Otso Ovaskainen. Computationally efficient joint species distribution modeling of big spatial data. *Ecology*, 101(2):e02929, 2020. ISSN 1939-9170. doi: 10.1002/ecy.2929.

[180] Tina Tirelli and Daniela Pessani. Use of decision tree and artificial neural network approaches to model presence/absence of telestes muticellus in piedmont (north-western italy). *River Research and Applications*, 25(8): 1001–1012, 2009.

[181] Anne M. Trainor, Oswald J. Schmitz, Jacob S. Ivan, and Tanya M. Shenk. Enhancing species distribution modeling by characterizing predator–prey interactions. *Ecological Applications*, 24(1):204–216, 2014. ISSN 1939-5582. doi: 10.1890/13-0336.1.

[182] Hanna Tuomisto. A diversity of beta diversities: Straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography*, 33(1):2–22, 2010.

[183] Courtney A. Tye, Robert A. McCleery, Rober J. Fletcher Jr., Daniel U. Greene, and Ryan S. Butryn. Evaluating citizen vs. professional data for modelling distributions of a rare squirrel. *Journal of Applied Ecology*, 2016.

[184] Tomáš Václavík, John A. Kupfer, and Ross K. Meentemeyer. Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM). *Journal of Biogeography*, 39(1):42–55, 2012. ISSN 1365-2699. doi: 10.1111/j.1365-2699.2011.02589.x.

[185] Alice Valentini, Pierre Taberlet, Claude Miaud, Raphaël Civade, Jelger Herder, Philip Francis Thomsen, Eva Bellemain, Aurélien Besnard, Eric Coissac, Frédéric Boyer, et al. Next-generation monitoring of aquatic biodiversity using environmental dna metabarcoding. *Molecular Ecology*, 25(4): 929–942, 2016.

[186] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

[187] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[188] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. *arXiv preprint arXiv:2103.16483*, 2021.

[189] Jeremy VanDerWal, Luke P. Shoo, Catherine Graham, and Stephen E. Williams. Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know? *Ecological Modelling*, 220(4):589–594, February 2009. ISSN 0304-3800. doi: 10.1016/j.ecolmodel.2008.11.010.

[190] William N. Venables and Brian D. Ripley. *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013.

[191] W.C.E.P. Verberk. Explaining general patterns in species abundance and distributions. *Nature Education Knowledge*, 3(10):38, 2011.

[192] Peter A. Vesk, William K. Morris, Will C. Neal, Karel Mokany, and Laura J. Pollock. Transferability of trait-based species distribution models. *Ecography*, 44(1):134–147, 2021. ISSN 1600-0587. doi: 10.1111/ecog.05179.

[193] Jean-Christophe Vié, Craig Hilton-Taylor, Caroline Pollock, James Ragle, Jane Smart, Simon N. Stuart, and Rashila Tong. The IUCN red list: A key conservation tool. *Wildlife in a changing world–An analysis of the 2008 IUCN Red List of Threatened Species*, page 1, 2009.

[194] Y.I. Wang, Ulrike Naumann, Stephen T. Wright, and David I. Warton. mvabund–an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, 3(3):471–474, 2012.

[195] Gill Ward, Trevor Hastie, Simon Barry, Jane Elith, and John R. Leathwick. Presence-Only Data and the EM Algorithm. *Biometrics*, 65(2):554–563, 2009. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2008.01116.x.

[196] Robert H. Whittaker. Vegetation of the great smoky mountains. *Ecological Monographs*, 26(1):2–80, 1956.

[197] Robert Harding Whittaker. Vegetation of the Siskiyou mountains, Oregon and California. *Ecological Monographs*, 30(3):279–338, 1960.

[198] John J. Wiens. The niche, biogeography and species interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1576): 2336–2350, August 2011. doi: 10.1098/rstb.2011.0059.

[199] David P. Wilkinson, Nick Golding, Gurutzeta Guillera-Arroita, Reid Tingley, and Michael A. McCarthy. A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*, 10(2):198–211, February 2019. ISSN 2041-210X. doi: 10.1111/2041-210X.13106.

[200] David Peter Wilkinson. A comparison of the inferential, computational, and predictive performance of joint species distribution models. *The University of Melbourne Ph.D. Theses*, 2019.

[201] Mary S. Wisz and Antoine Guisan. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecology*, 9(1):8, April 2009. ISSN 1472-6785. doi: 10.1186/1472-6785-9-8.

[202] Simon N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1): 3–36, 2011.

[203] Peggy P.W. Yen, Falk Huettmann, and Fred Cooke. A large-scale model for the at-sea distribution and abundance of Marbled Murrelets (Brachyramphus marmoratus) during the breeding season in coastal British Columbia, Canada. *Ecological Modelling*, 171(4):395–413, 2004.

[204] Damaris Zurell, Niklaus E Zimmermann, Helge Gross, Andri Baltensweiler, Thomas Sattler, and Rafael O Wüest. Testing species assemblage predictions from stacked and joint species distribution models. *Journal of Biogeography*, 47(1):101–113, 2020.

Chapter 4

RECOGNITION IN TERRA INCOGNITA

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

4.1 Abstract

It is desirable for detection and classification algorithms to generalize to unfamiliar environments, but suitable benchmarks for quantitatively studying this phenomenon are not yet available. We present a dataset designed to measure recognition generalization to novel environments. The images in our dataset are harvested from twenty camera traps deployed to monitor animal populations. Camera traps are fixed at one location, hence the background changes little across images; capture is triggered automatically, hence there is no human bias. The challenge is learning recognition in a handful of locations, and generalizing animal detection and classification to new locations where no training data is available. In our experiments state-of-the-art algorithms show excellent performance when tested at the same location where they were trained. However, we find that generalization to new locations is poor, especially for classification systems.[1]

4.2 Introduction

Automated visual recognition algorithms have recently achieved human expert performance at visual classification tasks in field biology [29, 45, 47] and medicine [10, 33]. Thanks to the combination of deep learning [12, 23], Moore's law [36] and very large annotated datasets [9, 25] enormous progress has been made during the past 10 years. Indeed, 2017 may come to be remembered as the year when automated visual categorization surpassed human performance.

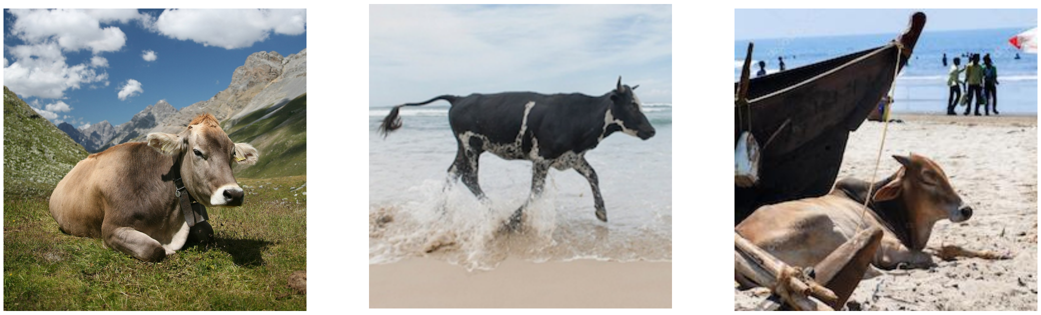
However, it is known that current learning algorithms are dramatically less data-efficient than humans [44], transfer learning is difficult [30], and, anecdotally, vision algorithms do not generalize well across datasets [43, 50] (Fig. 4.1). These observations suggest that current algorithms rely mostly on rote pattern-matching, rather than abstracting from the training set 'visual concepts' [27] that can generalize well

[1] The dataset is available at `https://beerys.github.io/CaltechCameraTraps/`

to novel situations. In order to make progress we need datasets that support a careful analysis of generalization, dissecting the challenges in detection and classification: variation in lighting, viewpoint, shape, photographer's choice and style, context/background. Here we focus on the latter: generalization to new environments, which includes background and overall lighting conditions.

Applications where the ability to generalize visual recognition to new environments is crucial include surveillance, security, environmental monitoring, assisted living, home automation, automated exploration (e.g. sending rovers to other planets). Environmental monitoring by means of camera traps is a paradigmatic application. Camera traps are heat- or motion-activated cameras placed in the wild to monitor and investigate animal populations and behavior. Camera traps have become inexpensive, hence hundreds of them are often deployed for a given study, generating a deluge of images. Automated detection and classification of animals in images is a necessity. The challenge is training animal detectors and classifiers from data coming from a few pilot locations such that these detectors and classifiers will generalize to new locations. Camera trap data is controlled for environment including lighting (the cameras are static, and lighting changes systematically according to time and weather conditions), and eliminates photographer bias (the cameras are activated automatically).

Camera traps are not new to the computer vision community
[6, 15, 24, 26, 29, 35, 40, 48, 51, 53, 54, 57, 58]. Our work is the first to identify camera traps as a unique opportunity to study generalization, and we offer the first study of generalization to new environments in this controlled setting. We make here three contributions: (a) a novel, well-annotated dataset to study visual generalization across locations, (b) a benchmark to measure algorithms' performance, and (c) baseline experiments establishing the state of the art. Our aim is to complement current datasets utilized by the vision community for detection and classification[9, 11, 21, 25] by introducing a new dataset and experimental protocol that can be used to systematically evaluate the generalization behavior of algorithms to novel environments. In this work we benchmark the current state-of-the-art detection and classification pipelines and find that there is much room for improvement.



(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Figure 4.1: **Recognition algorithms generalize poorly to new environments.** Cows in ‘common’ contexts (e.g. Alpine pastures) are detected and classified correctly (A), while cows in uncommon contexts (beach, waves and boat) are not detected (B) or classified poorly (C). Top five labels and confidence produced by ClarifAI.com shown.

4.3 Related work

Datasets

The ImageNet [9], MS-COCO [25], PascalVOC [11], and Open Images [21] datasets are commonly used for benchmarking classification and detection algorithms. Images in these datasets were collected in different locations by different people, which enables algorithms to average over photographer style and irrelevant background clutter. However, as demonstrated in Fig. 4.1, the context can be strongly biased. Human photographers are biased towards well-lit, well-focused images where the subjects are centered in the frame [32, 37]. Furthermore, the number of images per class is balanced, unlike what happens in the real world [44].

Natural world datasets such as the iNaturalist dataset [47], CUB200 [49], Oxford Flowers [28], LeafSnap [22], and NABirds700 [46] are focused on fine-grained species classification and detection. Most images in these datasets are taken by humans under relatively good lighting conditions, though iNaturalist does contain human-selected camera trap images. Many of these datasets exhibit real-world long-tailed distributions, but in all cases there is a large amount of diversity in location and perspective.

The Snapshot Serengeti dataset [40] is a large, multi-year camera trap dataset collected at 225 locations in a small region of the African savanna. It is the single largest-scale camera trap dataset ever collected, with over 3 million images. However, it is not yet suitable for controlled experiments. This dataset was collected from camera traps that fire in sequences of 3 for each motion trigger, and provides species annotation for groups of images based on a time threshold. This means that sometimes a single species annotation is provided for up to 10 frames, when in fact the animal was present in only a few of those frames (no bounding boxes are provided). Not all camera trap projects are structured in a similar way, and many cameras take shorter sequences or even single images on each trigger. In order to find a solution that works for new locations regardless of the camera trap parameters, it is important to have information about which images in the batch do or do not contain animals. In our dataset we provide annotations on a per-instance basis, with bounding boxes and associated classes for each animal in the frame.

Detection

Since camera traps are static, detecting animals in the images could be considered either a change detection or foreground detection problem. Detecting changes and/or foreground vs. background in video is a well studied problem [38], [2]. Many of these methods rely on constructing a good background model that updates regularly, and thus degrade rapidly at low frame rates. [55] and [3] consider low frame rate change detection in aerial images, but in these cases there are often very few examples per location.

Some camera traps collect a short video when triggered instead of a sequence of frames. [24, 57, 58] show foreground detection results on camera trap video. Data that comes from most camera traps take sequences of frames at each trigger at a frame rate of $\sim$ 1 frame per second. This data can be considered “video," albeit with extremely low, variable frame rate. Statistical methods for background subtraction and foreground segmentation in camera trap image sequences have been previously considered. [35] demonstrates a graph-cut method that uses background modeling and foreground object saliency to segment foreground in camera trap sequences. [26] creates background models and perform a superpixel-based comparison to determine areas of motion. [15] uses a multi-layer RPCA-based method applied to day and night sequences. [53] uses several statistical background-modeling approaches as additional signal to improve and speed up deep detection. These methods rely on a sequence of frames at each trigger to create appropriate background models, which are not always available. None of these methods demonstrate results on locations outside of their training set.

Classification

A few studies tackle classification of camera trap images. [51] showed results classifying squirrels vs. tortoises in the Mojave Desert. [54] showed classification results on data that provides image sequences of 10 frames. They do not consider the detection problem and instead manually crop the animal from the frame and balance the dataset, resulting in a total of 7,196 images across 18 species with at least 100 examples each. [6] were the first to take a deep network approach to camera trap classification, working with data from eMammal [1]. They first performed detection using the background subtraction method described in [35], then classified cropped detected regions, getting 38.31% top-1 accuracy on 20 common species. [48] show classification results on both Snapshot Serengeti and data from jungles in Panama, and saw a boost in classification performance from providing animal segmentations. [29] show 94.9% top-1 accuracy using an ensemble of models for classification on the Snapshot Serengeti dataset. None of the previous works show results on unseen test locations.

Generalization and domain adaptation

Generalizing to a new location is an instance of domain adaptation, where each location represents a domain with its own statistical properties such as types of flora and fauna, species frequency, man-made or other clutter, weather, camera type, and camera orientation. There have been many methods proposed for domain adaptation in classification [8]. [13] proposed a method for unsupervised domain adaptation by maximizing domain classification loss while minimizing loss for classifying the target classes. We generalized this method to multi-domain for our dataset, but did not see any improvement over the baseline. [14] demonstrated results of a similar method for fine-grained classification, using a multi-task setting where the adaptation was from clean web images to real-world images, and [5] investigated open-set domain adaptation.

Few methods have been proposed for domain adaptation outside of classification. [7, 19, 56] investigate methods of domain adaptation for semantic segmentation, focusing mainly on cars and pedestrians and either adapting from synthetic to real data, from urban to suburban scenes, or from PASCAL to a camera on-board a car. [17, 31, 39, 42, 52] look at methods for adapting detectors from one data source to another, such as from synthetic to real data or from images to video. Raj, et. al., [34] demonstrated a subspace-based detection method for domain adaptation from PASCAL to COCO.

4.4 The Caltech Camera Traps dataset

The Caltech Camera Traps (CCT) dataset contains 243,187 images from 140 camera locations, curated from data provided by the USGS and NPS. Our goal in this paper is to specifically target the problem of generalization in detection and classification. To this end, we have randomly selected 20 camera locations from the American Southwest to study in detail. By limiting the geographic region, the flora and fauna seen across the locations remain consistent. The current task is not to deal with entirely new regions or species, but instead to be able to recognize the same species of animals in the same region with a different camera background. In the future we plan to extend this work to recognizing the same species in new regions, and to the open-set problem of recognizing never-before-seen species. Examples of data from different locations can be seen in Fig. 4.2.

Camera traps are motion- or heat-triggered cameras that are placed in locations of interest by biologists in order to monitor and study animal populations and behavior. When a camera is triggered, a sequence of images is taken at approximately one frame per second. Our dataset contains sequences of length $1 - 5$. The cameras are prone to false triggers caused by wind or heat rising from the ground, leading to empty frames. Empty frames can also occur if an animal moves out of the field of view of the camera while the sequence is firing. Once a month, biologists return to the cameras to replace the batteries and change out the memory card. After it has been collected, experts manually sort camera trap data to categorize species and remove empty frames. The time required to sort and label images by hand severely limits data scale and research productivity. We have acquired and further curated a portion of this data to analyze generalization behaviors of state-of-the-art classifiers and detectors.

The dataset in this paper, which we call Caltech Camera Traps-20 (CCT-20), consists of $57,868$ images across 20 locations, each labeled with one of 15 classes (or marked as empty). Classes are either single species (e.g. "Coyote" or groups of species, e.g. "Bird"). See Fig. 4.4 for the distribution of classes and images across locations. We do not filter the stream of images collected by the traps, rather this is the same data that a human biologist currently sifts through. Therefore the data is unbalanced in the number of images per location, distribution of species per location, and distribution of species overall (see Fig. 4.4).

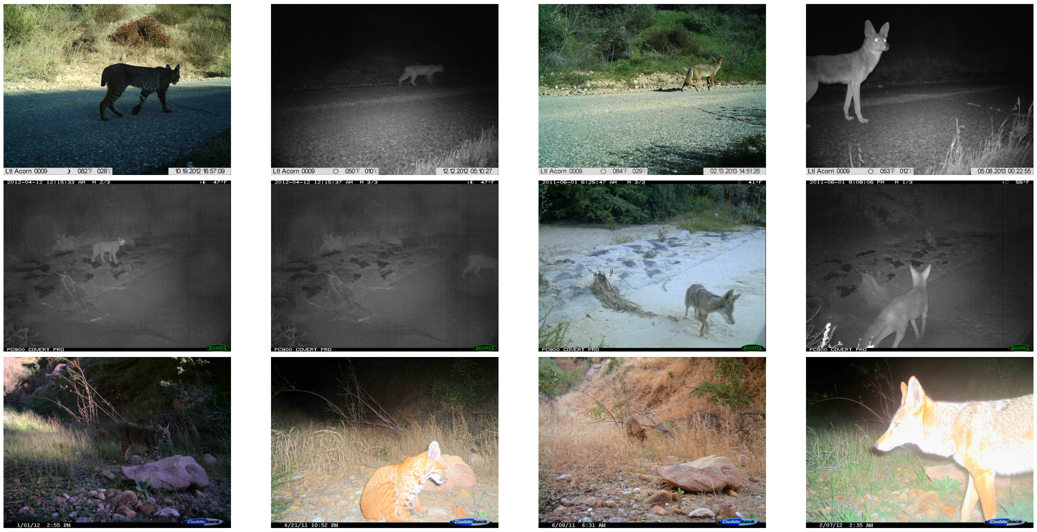
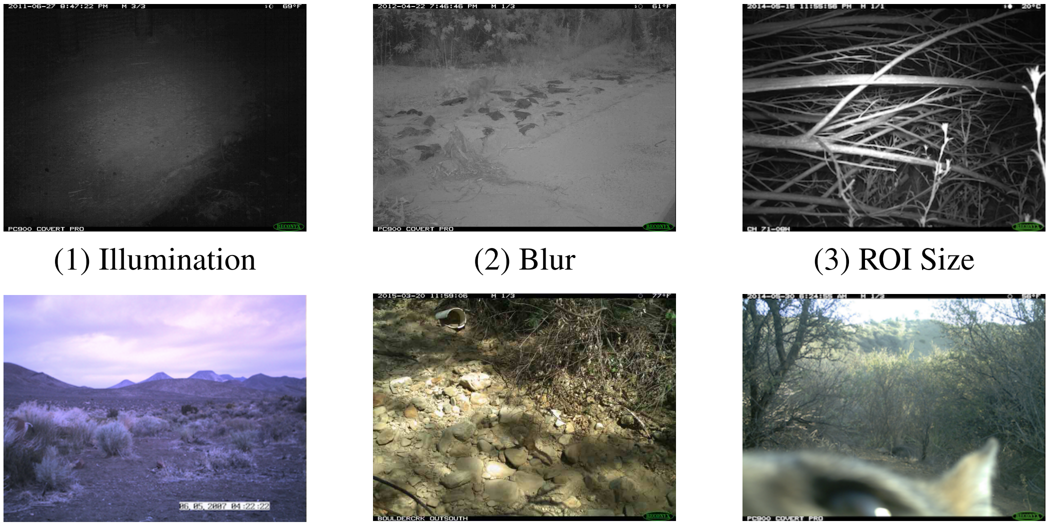


Figure 4.2: **Camera trap images from three different locations.** Each row is a different location and a different camera type. The first two cameras use IR, while the third row used white flash. The first two columns are bobcats, the next two columns are coyotes.



(1) Illumination

(2) Blur

(3) ROI Size

(4) Occlusion

(5) Camouflage

(6) Perspective

Figure 4.3: **Common data challenges**: (1) **Illumination**: Animals are not always salient. (2) **Motion blur**: common with poor illumination at night. (3) **Size of the region of interest** (ROI): Animals can be small or far from the camera. (4) **Occlusion**: e.g. by bushes or rocks. (5) **Camouflage**: decreases saliency in animals' natural habitat. (6) **Perspective**: Animals can be close to the camera, resulting in partial views of the body.

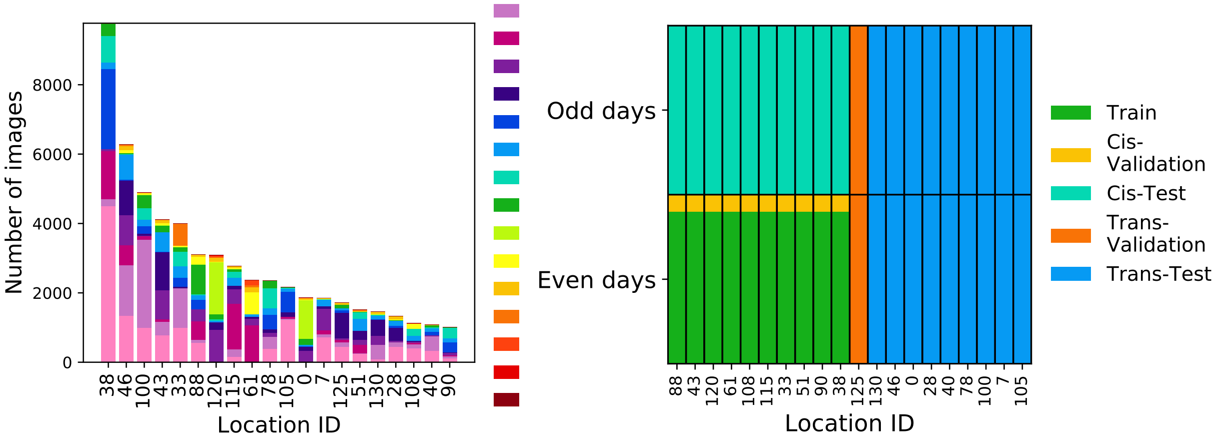


Figure 4.4: (Left) Number of annotations for each location, over 16 classes. The ordering of the classes in the legend is from most to least examples overall. The distribution of the species is long-tailed at each location, and each location has a different and peculiar distribution. (Right) Visualization of data splits. “Cis" refers to images from locations seen during training, and “trans" refers to new locations not seen during training.

Detection and labeling challenges

The animals in the images can be challenging to detect and classify, even for humans. We have determined that there are six main nuisance factors inherent to camera trap data, which can compound upon each other. Detailed analysis of these challenges can be seen in Fig. 4.3. When an image is too difficult to classify on its own, biologists will often refer to an easier image in the same sequence and then track motion by flipping between sequence frames in order to generate a label for each frame (e.g. is the animal still present or has it gone off the image plane?). We account for this in our experiments by reporting performance at the frame level and at the sequence level. Considering frame level performance allows us to investigate the limits of current models in exceptionally difficult cases.

Annotations

We collected bounding box annotations on Amazon Mechanical Turk, procuring annotations from at least three and up to ten mturkers for each image for redundancy and accuracy. Workers were asked to draw boxes around all instances of a specific type of animal for each image, determined by what label was given to the sequence by the biologists. We used the crowdsourcing method by Branson et al.[16] to determine ground truth boxes from our collective annotations, and to iteratively collect additional annotations as necessary. We found that bounding box precisions varied based on annotator, and determined that for this data the PascalVOC metric of IoU≥ 0.5 is appropriate for the detection experiments (as opposed to the COCO IoU averaging metric).

Data split: cis- and trans-

Our goal is exploring generalization to new (i.e. untrained) locations. Thus, we compare the performance of detection and classification algorithms when they are tested at the same locations where they were trained, vs new locations. For brevity, we refer to locations seen during training as ***cis-locations*** and locations not seen during training as ***trans-locations***.

From our pool of 20 locations, we selected 9 locations at random to use as trans-location test data, and a single random location to use as trans-location validation data. From the remaining 10 locations, we use images taken on odd days as cis-location test data. From within the data taken on even days, we randomly select 5% to be used as cis-location validation data. The remaining data is used for training, with the constraint that training and validation sets do not share the same image sequences. This gives us $13,553$ training images, $3,484$ validation and $15,827$ test images from cis-locations, and $1,725$ val and $23,275$ test images from trans-locations. The data split can be visualized in Fig. 4.4. We chose to interleave the cis training and test data by day because we found that using a single date to split the data results in additional generalization challenges due to changing vegetation and animal species distributions across seasons. By interleaving, we reduce noise and provide a clean experimental comparison of results on cis- and trans-locations.

4.5 Experiments

Current state-of-the-art computer vision models for classification and detection are designed to work well on test data whose distribution matches the training distribution. However, in our experiments we are explicitly evaluating the models on a different test distribution. In this situation, it is common practice to employ early stopping [4] as a means of preventing overfitting to the train distribution. Therefore, for all classification and detection experiments we monitor performance on both the cis- and trans-location validation sets. In each experiment we save two models, one that we expect has the best performance on the trans-location test set (i.e. a model that generalizes), and one that we expect has the best performance on the cis-location test set (i.e. a model that performs well on the train distribution).

Table 4.1: Classification top-1 error across experiments. Empty images are removed for these experiments.

	Cis-Locations		Trans-Locations		Error Increase	
Sequence Information	Images	Bboxes	Images	Bboxes	Images	Bboxes
None	19.06	8.14	41.04	19.56	115%	140%
Most Confident	17.7	7.06	34.53	15.77	95%	123%
Oracle	14.92	5.52	28.69	12.06	92%	118%

Classification

We explore the generalization of classifiers in 2 different settings: full images and cropped bounding boxes. For each setting we also explore the effects of using and ignoring sequence information. Sequence information is utilized in two different ways: **(1) Most Confident** we consider the sequence to be classified correctly if the most confident prediction from *all* frames grouped together is correct, or **(2) Oracle** we consider the sequence to be correctly classified if *any* frame is correctly classified. Note that (2) is a more optimistic usage of sequence information. For all classification experiments we use an Inception-v3 [41] model pretrained on ImageNet, with an initial learning rate of 0.0045, rmsprop with a momentum of 0.9, and a square input resolution of 299. We employ random cropping (containing at least 65% of the region), horizontal flipping, and color distortion as data augmentation.

Full Image.

We train a classifier on the full images, considering all 15 classes as well as empty images (16 total classes). On the cis-location test set we achieve a top-1 error of 20.83%, and a top-1 error of 41.08% on the trans-location test set with a 97% cis-to-trans increase in error. To investigate if requiring the classifier to both detect and classify animals increased overfitting on the training location backgrounds, we removed the empty images and retrained the classifiers using just the 15 animal classes. Performance stayed at nearly the same levels, with a top-1 error of 19.06% and 41.04% for cis- and trans-locations respectively. Utilizing sequence information helped reduce overall error (achieving errors of 14.92% and 28.69% on cis- and trans-locations respectively), but even in the most optimistic oracle setting, there is still a 92% increase in error between evaluating on cis- and trans-locations. See Table 4.1 for the full results.

Bounding Boxes.

We train a classifier on cropped bounding boxes, excluding all empty images (as there is no bounding box in those cases). Using no sequence information we achieve a cis-location top-1 error of 8.14% and a trans-location top-1 error of 19.56%. While the overall error has decreased compared to the image level classification, the error increase between cis- and trans-locations is still high at 140%. Sequence information further improved classification results (achieving errors of 5.52% and 12.06% on cis- and trans-locations respectively), and slightly reduced generalization error, bringing the error increase down to 118% in the most optimistic setting. See Table 4.1 for the full results. Additional experiments investigating the effect of number of images per location, number of training locations, and selection of validation location can be seen in the supplementary material.

Analysis

Fig. 4.5 provides a high level summary of our experimental findings. Namely, there is a generalization gap between cis- and trans-locations. Cropped boxes help to improve overall performance (shifting the blue lines vertically downward to the red lines), but the gap remains. In the best case scenario (red dashed lines: cropped boxes and optimistically utilizing sequences) we see a 92% increase in error between the cis- and trans-locations (with the same number of training examples), and 20x increase in training examples to have the same error rate. One might wonder whether this generalization gap is due to a large shift in class distributions between the two locations types. However, Fig. 4.7 shows that the overall distribution of classes between the locations is similar, and therefore probably does not account for the performance loss.

Detection

We use the Faster-RCNN implementation found in the Tensorflow Object Detection code base [20] as our detection model. We study performance of the Faster-RCNN model using two different backbones, ResNet-101 [18] and Inception-ResNet-v2 with atrous convolution [20]. Similar to our classification experiments we analyze the effects of using sequence information using two methods: **(1) Most Confident** we consider a sequence to be labeled correctly if the most confident detection across *all* frames has an IoU$\geq$ 0.5 with its matched ground truth box; **(2) Oracle** we consider a sequence to be labeled correctly if *any* frame's most confident detection

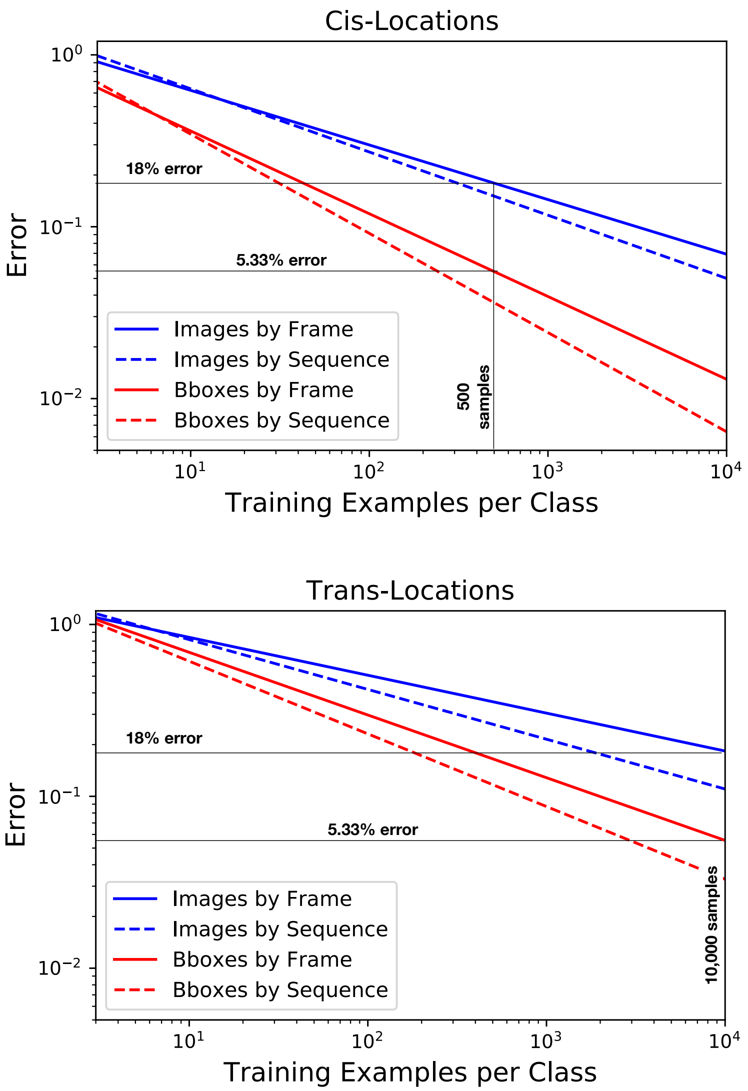


Figure 4.5: **Classification error vs. number of class-specific training examples**. Error is calculated as 1 - AUC (area under the precision-recall curve). Best-fit lines through the error-vs-n.examples points for each class in each scenario (points omitted for clarity), with average r^2 = 0.261. An example of line fit on top of data can be seen in Fig. 4.7. As expected, error decreases as a function of the number of training examples. This is true both for image classification (blue) and bounding-box classification (red) on both cis-locations and trans-locations. However, trans-locations show significantly higher error rates. To operate at an error rate of 5.33% on bounding boxes or 18% on images at the cis-locations we need 500 training examples, while we need 10,000 training examples to achieve the same error rate at the trans-locations, a 20x increase in data.

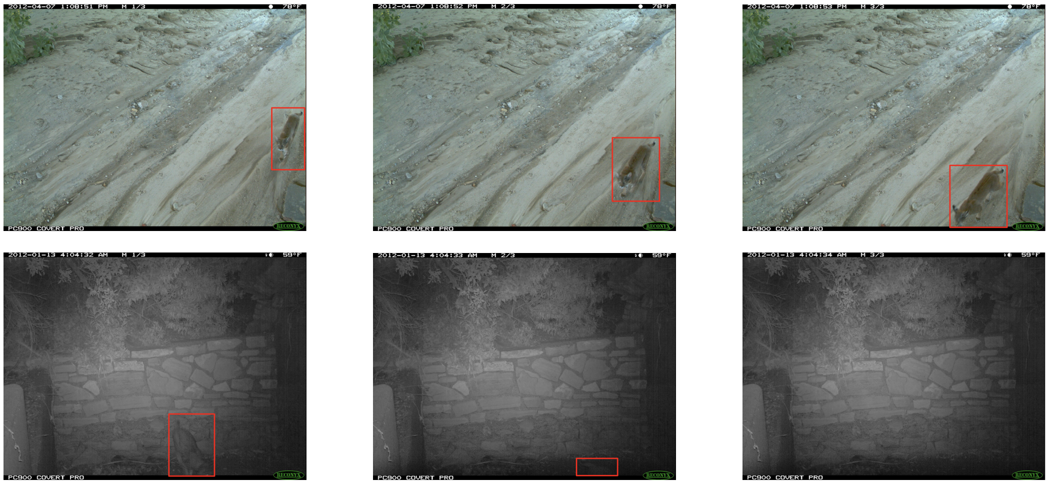


Figure 4.6: **Trans-classification failure cases at the sequence level**: (Based on classification of bounding box crops) In the first sequence, the network struggles to distinguish between ‘cat’ and ‘bobcat’, incorrectly predicting ‘cat’ in all three images with a mean confidence of 0.82. In the second sequence, the network struggles to classify a bobcat at an unfamiliar pose in the first image and instead predicts ‘raccoon’ with a confidence of 0.84. Little additional sequence information is available in this case, as the next frame contains only a blurry tail, and the last frame is empty

has IoU$\geq$ 0.5 with its matched ground truth box. Note that method (2) is more optimistic than method (1).

Our detection models are pretrained on COCO [25], images are resized to have a max dimension of 1024 and a minimum dimension of 600; each experiment uses SGD with a momentum of 0.9 and a fixed learning rate schedule. Starting at 0.0003 we decay the learning rate by a factor of 10 at 90k steps and 120k steps. We use a batch size of 1, and employ horizontal flipping for data augmentation. For evaluation, we consider a detected box to be correct if its IoU$\geq$ 0.5 with a ground truth box.

Results from our experiments can be seen in Table 4.2 and Fig 4.9. We find that both backbone architectures perform similarly. Without taking sequence information into account, the models achieve $\sim$ 77% mAP on cis-locations and $\sim$ 71% mAP on trans-locations. Adding sequence information using the most confident metric improves results, bringing performance on cis- and trans-locations to similar values at $\sim$ 85%. Finally, using the oracle metric brings mAP into the 90s for both locations. Precision-recall curves at the frame and sequence levels for both detectors can be seen in Fig. 4.9.

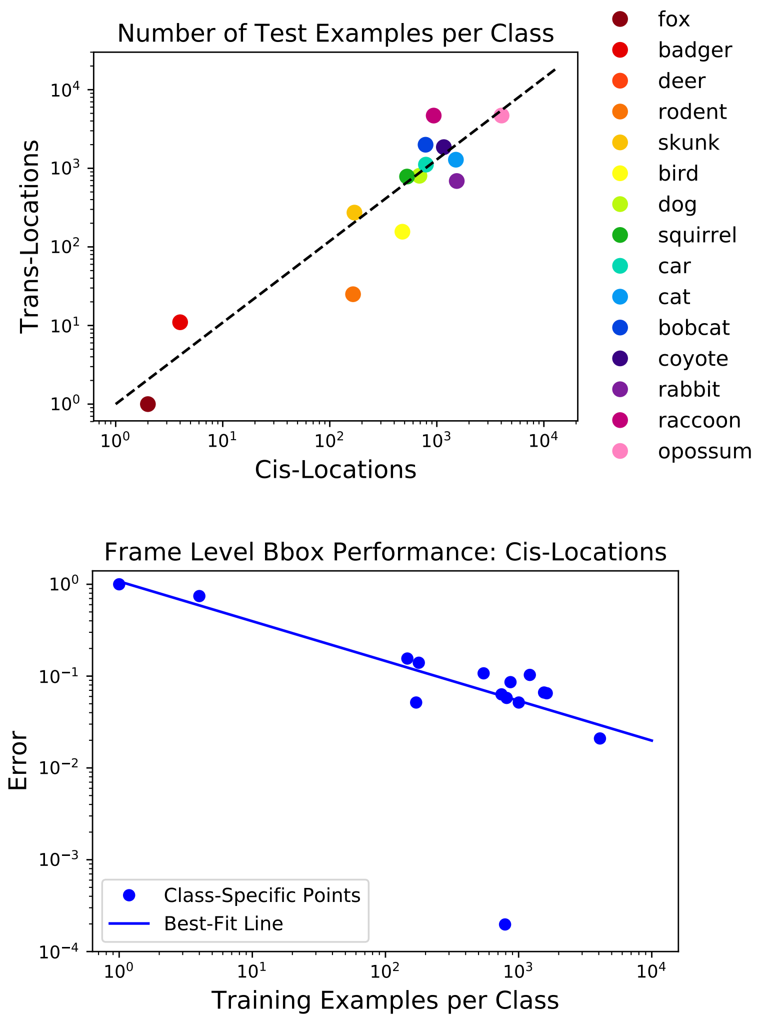


Figure 4.7: (Top) Distribution of species across the two test sets. (Bottom) An example of line fit used to generate the plots in Fig. 4.5

Analysis

There is a significantly lower generalization error in our detection experiments when not using sequences than what we observed in the classification experiments (∼ 30% error increase for detections vs ∼ 115% error increase for classification). When using sequence information, the generalization error for detections is reduced to only ∼ 5%.

Qualitatively, we found the mistakes can often be attributed to nuisance factors that make frames difficult. We see examples of all 6 nuisance factors described in Fig. 4.3 causing detection failures. The errors remaining at the sequence level occur when these nuisance factors are present in all frames of a sequence, or when the sequence only contains a single, challenging frame containing an animal. Examples

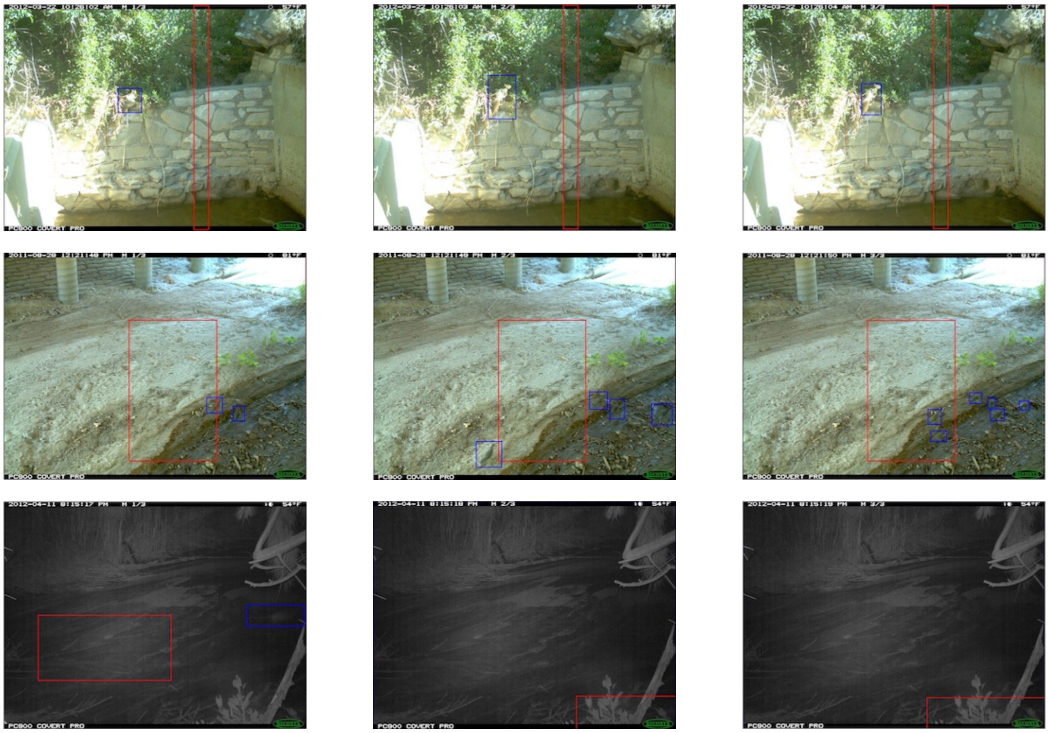


Figure 4.8: **Trans-detection failure cases at the sequence level**: Highest-confidence detection in red, ground truth in blue. In all cases the confidence of the detection was lower than 0.2. The first two sequences have small ROI, compounded with challenging lighting in the first and camouflaged birds in the second. In the third the opossum is poorly illuminated and only visible in the first frame.

of sequence-level detection failures can be seen in Fig. 4.8. The generalization gap at the frame level implies that our models are better able to deal with nuisance factors at locations seen during training.

Our experiments show that there is a small generalization gap when we use sequence information. However, overall performance has not saturated, and current state-of-the-art detectors are not achieving high precision at high recall values (1% precision at recall= 95%). So while we are encouraged by the results, there is still room for improvement. When we consider frames independently, we see that the generalization gap reappears. Admittedly this is a difficult case as it is not clear what the performance of a human would be without sequence information. However, we know that there are objects that can be detected in these frames and this dataset will challenge the next generation of detection models to accurately localize these difficult cases.

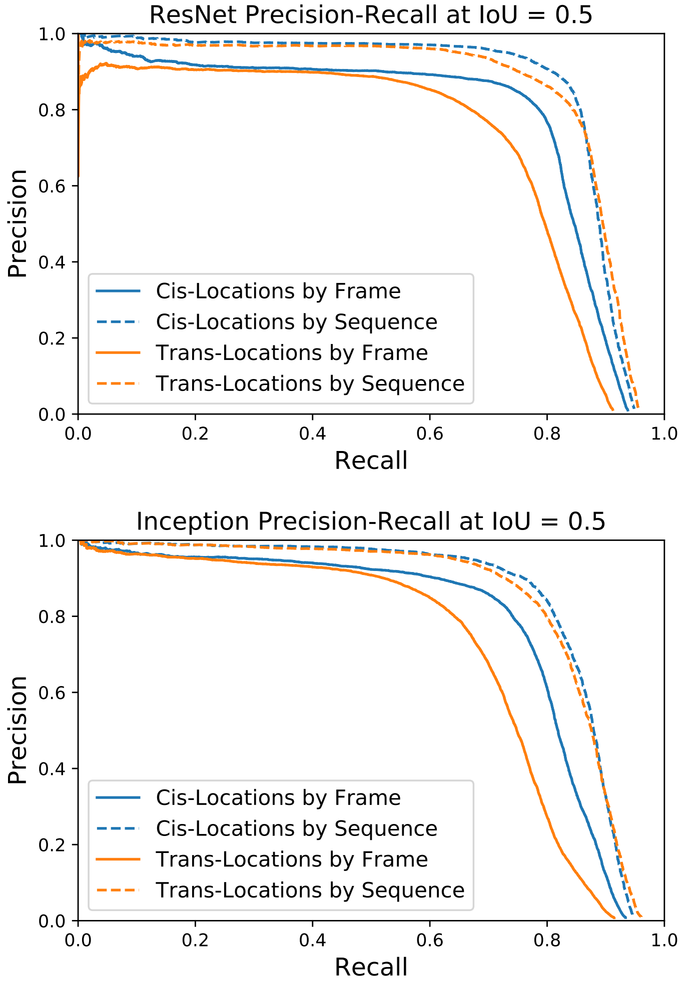


Figure 4.9: Faster-RCNN precision-recall curves at an IoU of 0.5, by frame and by sequence, using a confidence-based approach to determine which frame should represent the sequence

Table 4.2: Detection mAP at IoU=0.5 across experiments.

	Cis-Locations		Trans-Locations		Error Increase	
Sequence Information	ResNet	Inception	ResNet	Inception	ResNet	Inception
None	77.10	77.57	70.17	71.37	30%	27.6%
Most Confident	84.78	86.22	84.09	85.44	4.5%	5.6%
Oracle	94.95	95.04	92.13	93.09	55.8%	39.3%

4.6 Conclusions

The question of generalization to novel image statistics is taking center stage in visual recognition. Many indicators point to the fact that current systems are data-inefficient and do not generalize well to new scenarios. Current systems are, in essence, glorified pattern-matching machines, rather than intelligent visual learners.

Many problem domains face a generalization challenge where the test conditions are potentially highly different than what has been seen during training. Self driving cars navigating new cities, rovers exploring new planets, security cameras installed in new buildings, and assistive technologies installed in new homes are all examples where good generalization is critical for a system to be useful. However, the most popular detection and classification benchmark datasets [9, 11, 21, 25] are evaluating models on test distributions that are the same as the train distributions. Clearly it is important for models to do well on data coming from the same distribution as the train set. However, we argue that it is important to characterize the generalization behavior of these models when the test distribution deviates from the train distribution. Current datasets do not allow researchers to quantify the generalization behavior of their models.

We contribute a new dataset and evaluation protocol designed specifically to analyze the generalization behavior of classification and detection models. Our experiments reveal that there is room for significant improvement on the generalization of state-of-the-art classification models. Detection helps to improve overall classification accuracy, and we find that while detectors generalize better to new locations, there is room to improve their precision at high recall rates.

Camera traps provide a unique experimental setup that allow us to explore the generalization of models while controlling for many nuisance factors. Our current dataset is already revealing interesting behaviors of classification and detection models. There is still more information that we can learn by expanding our dataset in both data quantity and evaluation metrics. We plan to extend this dataset by adding additional locations, both from the American Southwest and from new regions. Drastic landscape and vegetation changes will allow us to investigate generalization in an even more challenging setting. Rare and novel events are frequently the most important and most challenging to detect and classify, and while our dataset already has these properties, we plan to define experimental protocols and data splits for benchmarking low-shot performance and the open-set problem of detecting and/or classifying species not seen during training.

Acknowledgements We would like to thank the USGS and NPS for providing data. This work was supported by NSFGRFP Grant No. 1745301, the views are those of the authors and do not necessarily reflect the views of the NSF. Compute time was provided by an AWS Research Grant.

References

[1] emammal: A tool for collecting, archiving, and sharing camera trapping images and data. `https://emammal.si.edu/`. Accessed: 2018-03-13.

[2] Mohammadreza Babaee, Duc Tung Dinh, and Gerhard Rigoll. A deep convolutional neural network for background subtraction. *arXiv preprint arXiv:1702.01731*, 2017.

[3] Csaba Benedek and Tamás Szirányi. A mixed markov model for change detection in aerial photos with large time differences. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[4] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer, 2012.

[5] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 1, 2017.

[6] Guobin Chen, Tony X. Han, Zhihai He, Roland Kays, and Tavis Forrester. Deep convolutional neural network based species recognition for wild animal monitoring. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 858–862. IEEE, 2014.

[7] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. *arXiv preprint arXiv:1711.11556*, 2017.

[8] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[10] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[12] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189. PMLR, 2015.

[14] Timnit Gebru, Judy Hoffman, and Li Fei-Fei. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1358–1367. IEEE, 2017.

[15] Jhony-Heriberto Giraldo-Zuluaga, Augusto Salazar, Alexander Gomez, and Angélica Diaz-Pulido. Camera-trap images segmentation using multi-layer robust principal component analysis. *The Visual Computer*, pages 1–13, 2017.

[16] Serge Belongie Grant Van Horn, Scott Laurie and Pietro Perona. Lean multiclass crowdsourcing. *Computer Vision and Pattern Recognition*, 2018.

[17] Hironori Hattori, Vishnu Naresh Boddeti, Kris Kitani, and Takeo Kanade. Learning scene-specific pedestrian detectors without real data. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3819–3827. IEEE, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[19] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[20] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*, 2017.

[21] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

[22] Neeraj Kumar, Peter N Belhumeur, Arijit Biswas, David W Jacobs, W John Kress, Ida C Lopez, and João VB Soares. Leafsnap: A computer vision system for automatic plant species identification. In *European conference on computer vision*, pages 502–516. Springer, 2012.

[23] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86 (11):2278–2324, 1998.

[24] Kai-Hsiang Lin, Pooya Khorrami, Jiangping Wang, Mark Hasegawa-Johnson, and Thomas S Huang. Foreground object detection in highly dynamic scenes using saliency. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 1125–1129. IEEE, 2014.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[26] Agnieszka Miguel, Sara Beery, Erica Flores, Loren Klemesrud, and Rana Bayrakcismith. Finding areas of motion in camera trap images. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 1334–1338. IEEE, 2016.

[27] Gregory Murphy. *The big book of concepts*. MIT press, 2004.

[28] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1447–1454. IEEE, 2006.

[29] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25): E5716–E5725, 2018.

[30] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[31] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3d models. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 1278–1286. IEEE, 2015.

[32] Jean Ponce, Tamara L Berg, Mark Everingham, David A Forsyth, Martial Hebert, Svetlana Lazebnik, Marcin Marszalek, Cordelia Schmid, Bryan C Russell, Antonio Torralba, et al. Dataset issues in object recognition. In *Toward category-level object recognition*, pages 29–48. Springer, 2006.

[33] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, page 1, 2018.

[34] Anant Raj, Vinay P Namboodiri, and Tinne Tuytelaars. Subspace alignment based domain adaptation for rcnn detector. *arXiv preprint arXiv:1507.05578*, 2015.

[35] Xiaobo Ren, Tony X. Han, and Zhihai He. Ensemble video object cut in highly dynamic scenes. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1947–1954. IEEE, 2013.

[36] Robert R Schaller. Moore's law: past, present and future. *IEEE spectrum*, 34 (6):52–59, 1997.

[37] Merrielle Spain and Pietro Perona. Some objects are more equal than others: Measuring and predicting importance. In *European Conference on Computer Vision (ECCV)*, pages 523–536. Springer, 2008.

[38] Pierre-Luc St-Charles, Guillaume-Alexandre Bilodeau, and Robert Bergevin. Subsense: A universal change detection method with local adaptive sensitivity. *IEEE Transactions on Image Processing*, 24(1):359–373, 2015.

[39] Baochen Sun and Kate Saenko. From virtual to reality: Fast adaptation of virtual object detectors to real domains. In *BMVC*, volume 1, page 3, 2014.

[40] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*, 2:150026, 2015.

[41] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[42] Kevin Tang, Vignesh Ramanathan, Li Fei-Fei, and Daphne Koller. Shifting weights: Adapting object detectors from image to video. In *Advances in Neural Information Processing Systems*, pages 638–646, 2012.

[43] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011.

[44] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

[45] Grant van Horn, Jessie Barry, Serge Belongie, and Pietro Perona. The Merlin Bird ID smartphone app (http://merlin.allaboutbirds.org/download/).

[46] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.

[47] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[48] Alexander Gomez Villa, Augusto Salazar, and Francisco Vargas. Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecological Informatics*, 41:24–32, 2017.

[49] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[50] Peter Welinder, Max Welling, and Pietro Perona. A lazy man's approach to benchmarking: Semisupervised classifier evaluation and recalibration. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3262–3269. IEEE, 2013.

[51] Kimberly Wilber, Walter J. Scheirer, Phil Leitner, Brian Heflin, James Zott, Daniel Reinke, David K. Delaney, and Terrance E. Boult. Animal recognition in the mojave desert: Vision tools for field biologists. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 206–213. IEEE, 2013.

[52] Jiaolong Xu, Sebastian Ramos, David Vázquez, and Antonio M López. Domain adaptation of deformable part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12):2367–2380, 2014.

[53] Hayder Yousif, Jianhe Yuan, Roland Kays, and Zhihai He. Fast human-animal detection from highly cluttered camera-trap images using joint background modeling and deep learning classification. In *Circuits and Systems (ISCAS), 2017 IEEE International Symposium on*, pages 1–4. IEEE, 2017.

[54] Xiaoyuan Yu, Jiangping Wang, Roland Kays, Patrick A Jansen, Tianjiang Wang, and Thomas Huang. Automated identification of animal species in camera trap images. *EURASIP Journal on Image and Video Processing*, 2013 (1):52, 2013.

[55] Yang Zhan, Kun Fu, Menglong Yan, Xian Sun, Hongqi Wang, and Xiaosong Qiu. Change detection based on deep siamese convolutional network for optical aerial images. *IEEE Geoscience and Remote Sensing Letters*, 14(10): 1845–1849, 2017.

[56] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, page 6, 2017.

[57] Zhi Zhang, Tony X. Han, and Zhihai He. Coupled ensemble graph cuts and object verification for animal segmentation from highly cluttered videos. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 2830–2834. IEEE, 2015.

[58] Zhi Zhang, Zhihai He, Guitao Cao, and Wenming Cao. Animal detection from highly cluttered natural scenes using spatiotemporal object region proposals and patch verification. *IEEE Transactions on Multimedia*, 18(10):2079–2092, 2016.

Chapter 5

IWILDCAM: BRINGING NOVEL CAMERA TRAP CHALLENGES TO THE COMPUTER VISION COMMUNITY

Sara Beery, Grant Van Horn, Oisin MacAodha, and Pietro Perona. The iWildCam 2018 challenge dataset. *The Fifth Fine-Grained Visual Categorization Workshop at CVPR*, 2018.

Sara Beery, Dan Morris, and Pietro Perona. The iWildCam 2019 challenge dataset. *The Sixth Fine-Grained Visual Categorization Workshop at CVPR*, 2019.

Sara Beery, Elijah Cole, and Arvi Gjoka. The iWildCam 2020 competition dataset. *The Seventh Fine-Grained Visual Categorization Workshop at CVPR*, 2020.

Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iWildCam 2021 competition dataset. *The Eighth Fine-Grained Visual Categorization Workshop at CVPR*, 2021.

Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the WILDS benchmark for unsupervised adaptation. *International Conference on Machine Learning*, 2022. URL `https://arxiv.org/abs/2112.05090`.

5.1 Abstract

The iWildCam competition serves as a yearly competition at the Fine-Grained Visual Categorization Workshop at the Computer Vision and Pattern Recognition Conference. It brings open, novel challenges to the computer vision community each year, helping to bridge the gap between the development of novel computer vision methods and the creation of impactful tools for the camera trap ecology community. Over the past 5 years, over 500 international teams from the computer vision and the ecology communities have taken part in the competition, increasing the visibility of these interdisciplinary and impactful problems. The iWildCam 2020 dataset was also included as a core challenge within the WILDS benchmark, the

first large-scale cross-application domain shift benchmark, as well as its extension WILDS 2.0 which introduces unlabeled data for 8 of the original WILDS datasets.

5.2 iWildCam 2018

Camera traps are a valuable tool for studying biodiversity, but research using this data is limited by the speed of human annotation. With the vast amounts of data now available it is imperative that we develop automatic solutions for annotating camera trap data in order to allow this research to scale. A promising approach is based on deep networks trained on human-annotated images [83]. iWildCam 2018 is a challenge dataset designed to explore whether such solutions generalize to novel locations, since systems that are trained once and may be deployed to operate automatically in new locations would be most useful.

Dataset

All images in the iWildCam 2018 dataset come from the American Southwest. By limiting the geographic region, the flora and fauna seen across the locations remain consistent. The current task is not to deal with entirely new regions or species, but instead to be able to recognize the same species of animals in the same region with a different camera background. In the future we plan to extend this dataset to include other regions, in order to tackle the challenges of both recognizing animals in new regions, and to the open-set problem of recognizing species of animals that have never before been seen. Examples of data from different locations can be seen in Fig. 4.2. Our dataset consists of $292,732$ images across 143 locations, each labeled as either containing an animal, or as empty. See Fig. 5.1 for the distribution of classes and images across locations. We do not filter the stream of images collected by the traps, rather this is the same data that a human biologist currently sifts through. Therefore the data is unbalanced in the number of images per location, distribution of species per location, and distribution of species overall (see Fig. 5.1). The class of each image was provided by expert biologists from the NPS and USGS. Due to different annotation styles and challenging images, we approximate that the dataset contains up to 5% annotation error.

Data Challenges

The animals in the images can be challenging to detect, even for humans. We find six main nuisance factors inherent to camera trap data (Fig. 4.3). When an image is too difficult to classify on its own, biologists will often refer to an easier image in

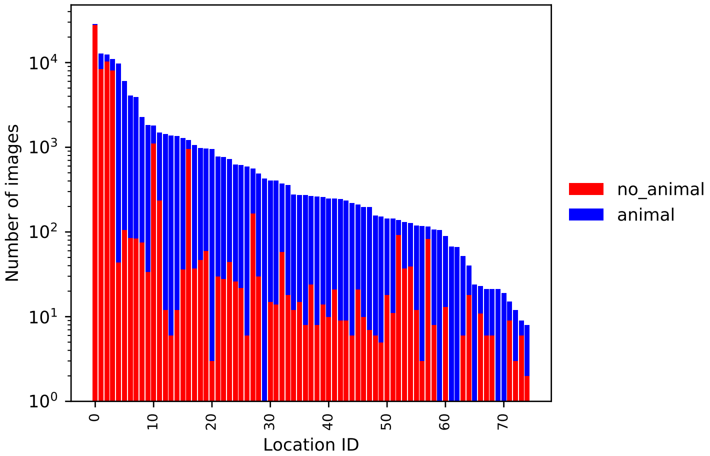


Figure 5.1: Number of annotations for each location, over the two classes. The distribution images per location is long-tailed, and each location has a different and peculiar class distribution.

the same sequence and then track motion by flipping between sequence frames in order to generate a label for each frame (e.g.is the animal still present or has it gone off the image plane?). This implies that sequence information is a valuable tool in difficult cases.

Data Split and Baseline

From our pool of 143 locations, we selected 70 locations at random to use as training data. We selected 10% of the data from our training locations and 5 random new locations to use as validation data. The remaining 68 locations are used as test data. This gives us $149,359$ training images, $17,784$ validation and $125,589$ test images.

We trained a baseline model using the InceptionV3 architecture, pretrained on ImageNet, with an initial learning rate of 0.0045, rmsprop with a momentum of 0.9, and a square input resolution of 299. We employed random cropping (containing most of the region), horizontal flipping, and random color distortion as data augmentation. This baseline achieved 74.1% accuracy on the test set.

Competition Results

The iWildCam Challenge 2018 was conducted through Kaggle as part of FGVC5 at CVPR18 and had 10 participating teams[1]. The final leaderboard from the held-out private test data can be seen in Fig. 5.2. The winning method by Stefan

[1] https://www.kaggle.com/c/iwildcam2018

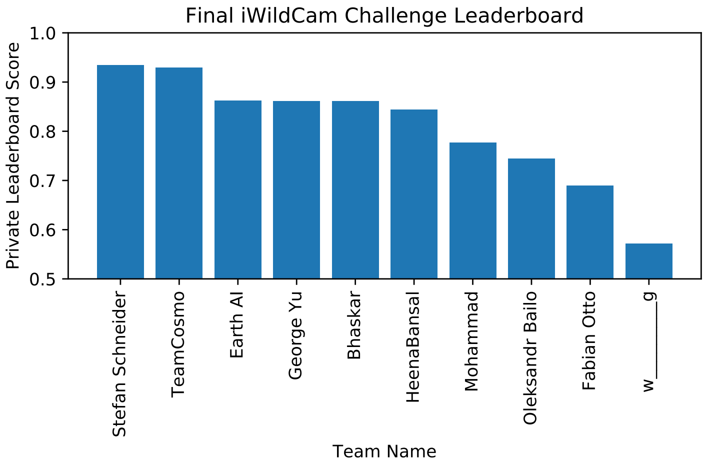


Figure 5.2: The final private leaderboard from the iWildCam Challenge 2018. These results show accuracies on the 50% held-out private test data randomly selected by Kaggle.

Schneider achieved an accuracy of 93.431%. It consisted of an ensemble of 5 models considering 5 different image sizes (50, 75, 100, 125, 150), all based on the VGG16 architecture. The models were trained from scratch using the Adam optimizer and data augmentation tools were used to randomly flip the images along the horizontal axis and add a range of blurring during training. Stefan considered a variety of models including AlexNet, GoogLeNet, DenseNet, ResNet and his own personal networks in the ensemble but found VGG16 outperformed all of them. He also considered a domain adaption model in an attempt to remove associations of location from the model but found this did not improve overall performance.

5.3 iWildCam 2019

As we try to expand the scope of computer vision models that identify species in camera traps from specific regions where we have collected training data to different areas we are faced with an interesting problem: how do you classify a species in a new region that you may not have seen in previous training data?

In order to tackle this problem, we have prepared a dataset and challenge where the training data and test data are from different regions, namely The American Southwest and the American Northwest. We use the Caltech Camera Traps dataset, collected from the American Southwest, as training data. We add a new dataset from the American Northwest, curated from data provided by the Idaho Department of Fish and Game (IDFG), as our test dataset. The test data has some class overlap with the training data, some species are found in both datasets, but there are both species seen during training that are not seen during test and vice versa. To help fill the gaps in the training species, we allow competitors to utilize transfer learning from two alternate domains: human-curated images from iNaturalist and synthetic images from Microsoft's TrapCam-AirSim simulation environment.

Dataset

The data for the 2019 challenge is curated from the Caltech Camera Traps (CCT) which was also used for the iWildCam 2018 Challenge [13], a new camera trap dataset from Idaho (IDFG), and two alternate data domains: iNaturalist and Microsoft TrapCam-AirSim.

Caltech Camera Traps

All images in this dataset, which was used for the iWildCam 2018 Challenge, come from the American Southwest. By limiting the geographic region, the flora and fauna seen across the locations remain consistent. Examples of data from different locations can be seen in Fig. 4.2. This dataset consists of 292, 732 images across 143 locations, each labeled with an animal class, or as empty. The classes represented are bobcat, opossum, coyote, raccoon, dog, cat, squirrel, rabbit, skunk, rodent, deer, fox, mountain lion, empty. We do not filter the stream of images collected by the traps, rather this is the same data that a human biologist currently sifts through. Therefore the data is unbalanced in the number of images per location, distribution of species per location, and distribution of species overall (see Fig. 5.4). The class of each image was provided by expert biologists from the NPS and USGS. Due to different annotation styles and challenging images, we approximate that the dataset contains up to 5% annotation error.

IDFG

The Idaho Department of Fish and Game provided labeled data from Idaho to use as an unseen test set, which we call IDFG. The test set contains 153,730 images from 100 locations in Idaho. It covers the classes mountain lion, moose, wolf, black bear, pronghorn, elk, deer, and empty. See Fig. 5.4 for the distribution of classes and images across locations. Similarly to CCT, we do not filter the images so the data is innately unbalanced.

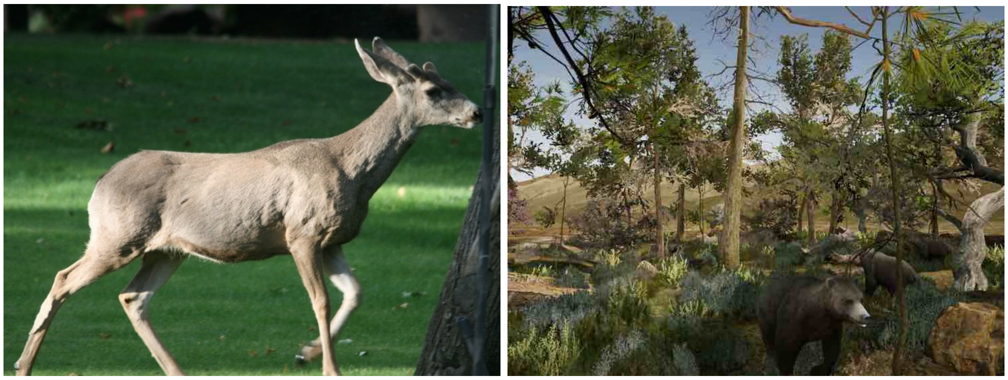


Figure 5.3: **Altenate domain examples.** (Left) iNaturalist, (Right) TrapCam-AirSim

Additional Data Domains

iNaturalist iNaturalist is a website where citizen scientists can post photos of plants and animals and work together to correctly ID the photos, an example of an iNaturalist image can be seen in Fig. 5.3. We allow the use of iNaturalist data from both the 2017 and 2018 iNaturalist competition datasets [126]. For ease of entry, we did the work to map our classes into the iNaturalist taxonomy. We also determined which mammals might be seen in Idaho using the iNaturalist API: bobcat, opossum, coyote, raccoon, dog, cat, squirrel, rabbit, skunk, rodent, deer, fox, mountain lion, moose, small mammal, elk, pronghorn, bighorn sheep, black bear, wolf, bison, and mountain goat. We curated an iNat-Idaho dataset that contains all iNat classes that might occur in Idaho, mapped into our class set in order to make adapting iNaturalist data for this challenge as simple as possible.

Microsoft TrapCam-AirSim This synthetic data generator utilizes a modular natural environment within Microsoft AirSim [16, 106] that can be randomly populated with flora and fauna. The distribution and types of animals, trees, bushes, rocks, and logs can be varied and randomly seeded to create images from a diverse set of classes and landscapes, from an open plain to a dense forest. An example of a TrapCam-AirSim image containing a bison can be seen in Fig. 5.3.

Challenge

The iWildCam Challenge 2019 was conducted through Kaggle as part of FGVC6 at CVPR19 and had 336 participating teams[2]. We used macro-average F1 score as our competition metric, to slightly emphasize recall over precision and to encourage

[2] https://www.kaggle.com/c/iwildcam-2019-fgvc6

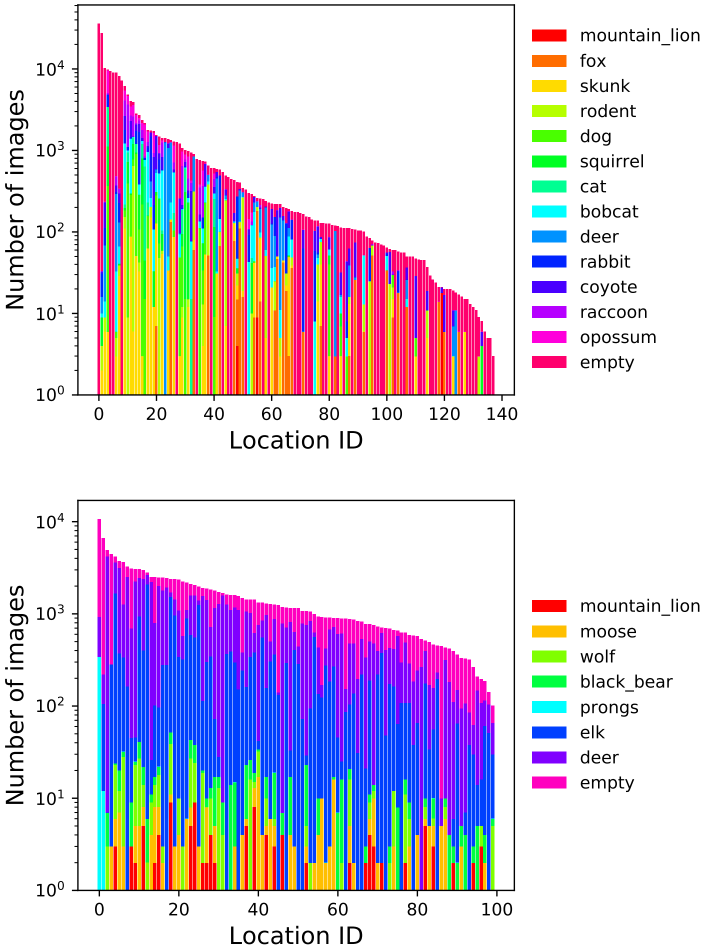


Figure 5.4: **Number of annotations for each location**. (Top) CCT locations, containing 14 classes. (Bottom) IDFG locations, containing images of 8 classes. The distribution of images per location is long-tailed, and each location has a different and peculiar class distribution.

more emphasis on rare classes, as opposed to rewarding high performance on common classes proportionally to their unbalanced level of occurrence.

Data Split and Baseline

We do not explicitly define a validation set for this challenge, instead letting competitors create their own validation set from the CCT training set and the two external data domains, iNat and TrapCam-AirSim. We use the IDFG data as our test set. Unsupervised annotation of the test set, using the provided detector or any clustering methods, is allowed. Explicit annotation of the test set is not.

We trained a simple whole-image classification baseline using the Inception-Resnet-

V2 architecture, pretrained on ImageNet and trained simultaneously on the CCT and iNat-Idaho datasets with no class rebalancing or weighting, with an initial learning rate of 0.0045, rmsprop with a momentum of 0.9, and a square input resolution of 299. We employed random cropping (containing most of the region), horizontal flipping, and random color distortion as data augmentation. This baseline achieved 0.125 macro-averaged F1 score and accuracy of 27.6% on the IDFG test set.

Camera Trap Animal Detection Model

We also provide a general animal detection model which competitors are free to use as they see fit. The model is a tensorflow Faster-RCNN model with Inception-Resnet-v2 backbone and atrous convolution. Sample code for running the detector over a folder of images can be found at https://github.com/Microsoft/CameraTraps. We have run the detector over each dataset, and provide the top 100 boxes and associated confidences for each image.

5.4 iWildCam 2020

As we try to expand the geographic scope of species identification models for camera trap data from a few regions to regions worldwide we are faced with an interesting question: how do we train models that perform well on diverse new (unseen during training) camera trap locations? Can we utilize data from other modalities, such as citizen science data and remote sensing data? In order to tackle this problem, we have prepared a challenge where the training data and test data are from different cameras spread across the globe.

The 2020 iWildCam challenge includes a new component: the use of multiple data modalities (see Fig. 5.5). An ecosystem can be monitored in a variety of ways (e.g. camera traps, citizen scientists, remote sensing) each of which has its own strengths and limitations. To facilitate the exploration of techniques for combining these complementary data streams, we provide a time series of remote sensing imagery for each camera trap location as well as curated subsets of the iNaturalist competition datasets matching the species seen in the camera trap data. It has been shown that species classification performance can be dramatically improved by using information beyond the image itself [17, 74] so we expect that participants will find creative and effective uses for this data.

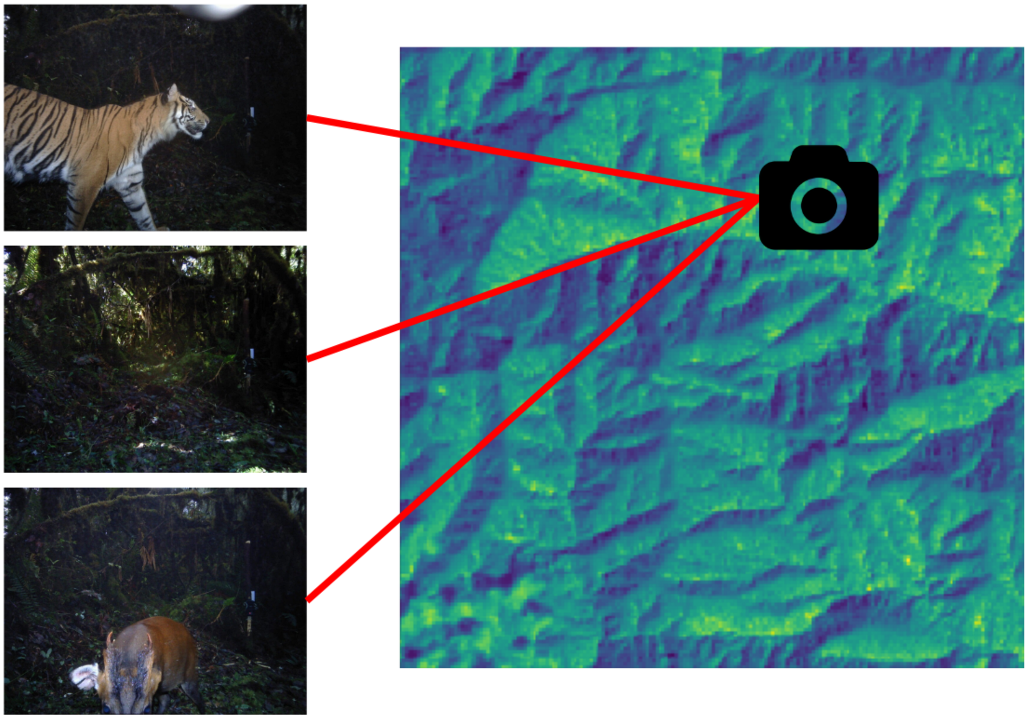


Figure 5.5: **The iWildCam 2020 dataset.** This year's dataset includes data from multiple modalities: camera traps, citizen scientists, and remote sensing. Here we can see an example of data from a camera trap paired with a visualization of the infrared channel of the paired remote sensing imagery.

Dataset

The dataset consists of three primary components: (i) camera trap images, (ii) citizen science images, and (iii) multispectral imagery for each camera location.

Camera Trap Data

The camera trap data (along with expert annotations) is provided by the Wildlife Conservation Society (WCS) [5]. We split the data by camera location, so no images from the test cameras are included in the training set to avoid overfitting to one set of backgrounds [12].

The training set contains $217,959$ images from 441 locations, and the test set contains $62,894$ images from 111 locations. These 552 locations are spread across 12 countries in different parts of the world. Each image is associated with a location ID so that images from the same location can be linked. As is typical for camera traps, approximately 50% of the total number of images are empty (this varies per location).

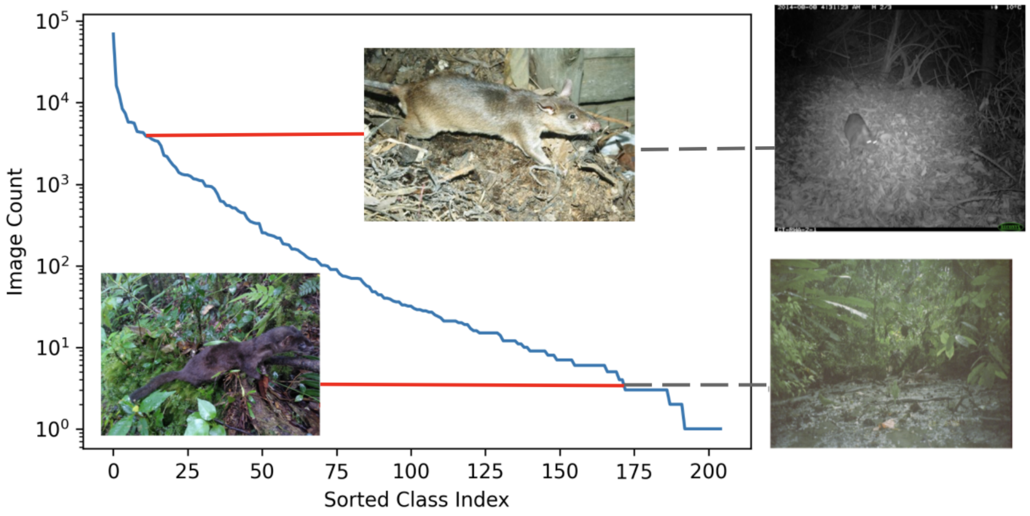


Figure 5.6: **Camera trap class distribution.** Per-class distribution of the camera trap data, which exhibits a long tail. We show examples of both a common class (the African giant pouched rat) and a rare class (the Indonesian mountain weasel). Within the plot we show images of each species, centered and focused, from iNaturalist. On the right we show images of each species within the frame of a camera trap, from WCS.

There are 276 species represented in the camera trap images. The class distribution is long-tailed, as shown in Fig. 5.6. Since we have split the data by location, some classes appear only in the training set. Any images with classes that appeared only in the test set were removed.

iNaturalist Data

iNaturalist is an online community where citizen scientists post photos of plants and animals and collaboratively identify the species [3]. To facilitate the use of iNaturalist data, we provide a mapping from our classes into the iNaturalist taxonomy.[3] We also provide the subsets of the iNaturalist 2017-2019 competition datasets [126] that correspond to species seen in the camera trap data. This data provides 13,051 additional images for training, covering 75 classes.

Though small relative to the camera trap data, the iNaturalist data has some unique characteristics. First, the class distribution is completely different (though it is still long tailed). Second, iNaturalist images are typically higher quality than the

[3] Note that for the purposes of the competition, competitors may only use iNaturalist data from the iNaturalist competition datasets.

corresponding camera trap images, providing valuable examples for hard classes. See Fig. 5.7 for a comparison between iNaturalist images and camera trap images.

Remote Sensing Data

For each camera location we provide multispectral imagery collected by the Landsat 8 satellite [125]. All data comes from the the Landsat 8 Tier 1 Surface Reflectance dataset [49] provided by Google Earth Engine [50]. This data has been been atmospherically corrected and meets certain radiometric and geometric quality standards.

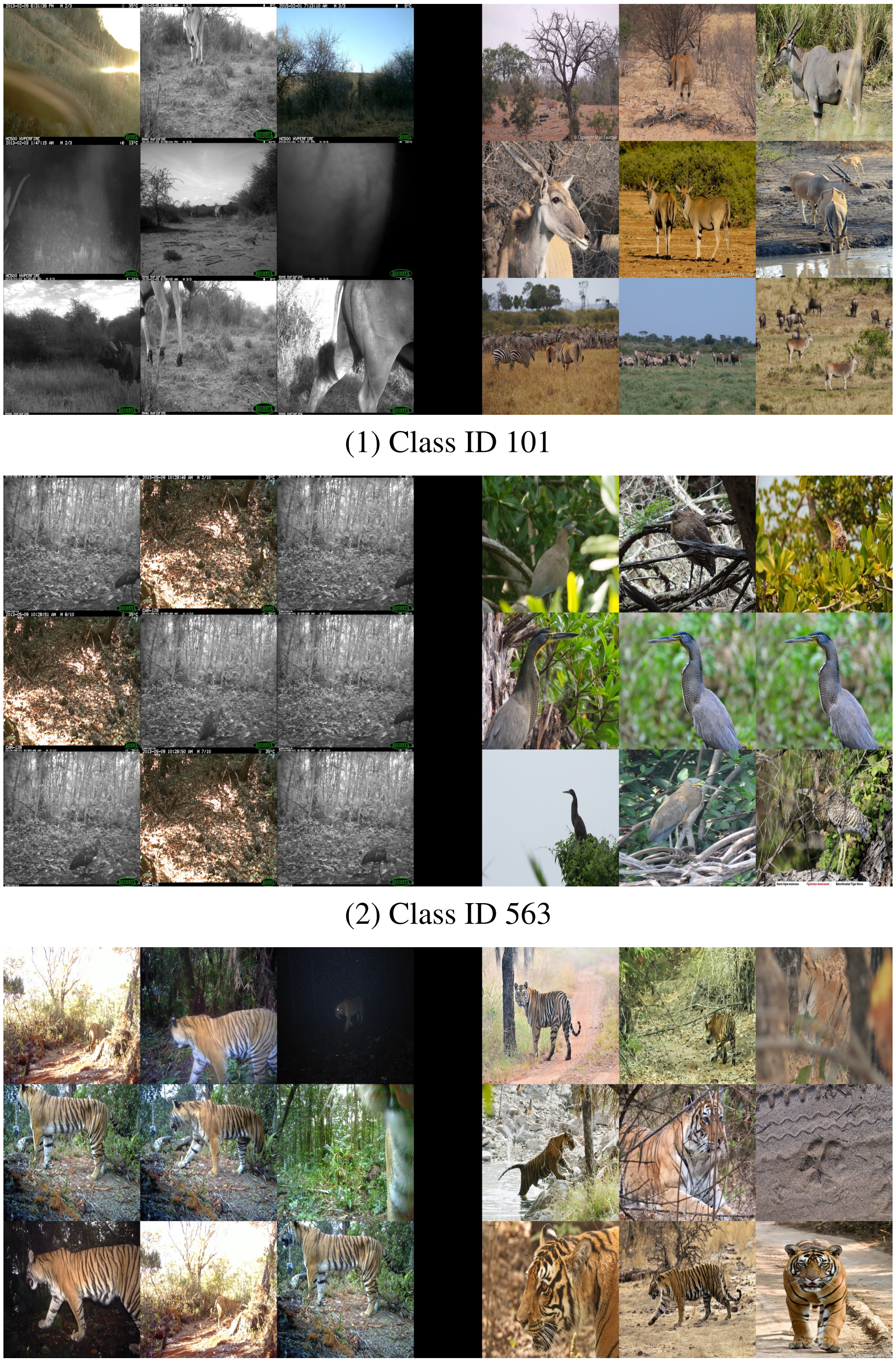
Data collection. The precise location of a camera trap is generally considered to be sensitive information, so we first obfuscate the coordinates of the camera. For each time point when imagery is available (the Landsat 8 satellite images the Earth once every 16 days), we extract a square *patch* centered at the obfuscated coordinates consisting of 9 bands of multispectral imagery and 2 bands of per-pixel metadata. Each patch covers an area of 6km × 6km. Since one Landsat 8 pixel covers an area of 30m^2, each patch is $200 \times 200 \times 11$ pixels. Note that the bit depth of Landsat 8 data is 16.

The multispectral imagery consists of 9 different bands, ordered by descending frequency / ascending wavelength. Band 1 is ultra-blue. Bands 2, 3, and 4 are traditional blue, green, and red. Band 5-9 are infrared. Note that bands 8 and 9 are from a different sensor than bands 1-7 and have been upsampled from 100m^2/pixel to 30m^2/pixel. Refer to [49] or [125] for more details.

Each patch of imagery has two corresponding *quality assessment* (QA) bands which carry per-pixel metadata. The first QA band (`pixelqa`) contains automatically generated labels for classes like `clear`, `water`, `cloud`, or `cloud shadow` which can help to interpret the pixel values. The second QA band (`radsatqa`) labels the pixels in each band for which the sensor was saturated. Cloud cover and saturated pixels are common issues in remote sensing data, and the QA bands may provide some assistance. However, they are automatically generated and cannot be trusted completely. See [49] for more details.

Baseline Results

We trained a basic image classifier as a baseline for comparison. The model is a randomly initialized Inception-v3 with input size 299 × 299, which was trained using only camera trap images. During training, images were randomly cropped



(1) Class ID 101

(2) Class ID 563

(3) Class ID 154

Figure 5.7: **Camera trap data (left) vs iNaturalist data (right).** (1) Animal is large, so camera trap image does not fully capture it. (2) Animal is small, so it makes up a small part of the camera trap images. (3) Quality is equivalent, although iNaturalist images have more camera pose and animal pose variation.

and perturbed in brightness, saturation, hue, and contrast. We used the `rmsprop` optimizer with an initial learning rate of 0.0045 and a decay factor of 0.94.

Let C be the number of classes. We trained using a class balanced loss from [32], given by

$$\mathcal{L}'(\mathbf{p}, y) = \frac{1-\beta}{1-\beta^{n_y}}\mathcal{L}(\mathbf{p}, y)$$

where $\mathbf{p} \in \mathbb{R}^C$ is the vector of predicted class probabilities (after softmax), $y \in \{1, \ldots, C\}$ is the ground truth class, $\mathcal{L}$ is the categorical cross-entropy loss, n_y is the number of samples for class y, and β is a hyperparameter which we set to 0.9.

This baseline achieved a macro-averaged F1 score of 0.62 and an accuracy of 62% on the iWildCam 2020 test set.

Challenge

The iWildCam Challenge 2020 was conducted through Kaggle as part of FGVC7 at CVPR20 and had 121 participating teams[4].

5.5 iWildCam 2021

In order to estimate the abundance of a species from camera trap data, ecologists need to know not just which species were seen, but also how many individuals of each species were seen. Object detection techniques can be used to find the number of individuals in each image. However, since camera traps collect images in motion-triggered bursts, simply adding up the number of detections over all frames is likely to lead to an incorrect estimate. Overcoming these obstacles may require incorporating spatiotemporal reasoning or individual re-identification in addition to traditional species detection and classification.

The computer vision community has been making steady progress improving automated systems for species classification and localization in camera trap images over the past decade [12, 14, 16, 17, 17, 26, 47, 77, 83, 84, 104, 120, 131, 140, 147, 152]. Classifications of species seen in a given image or sequence are used by ecologists to generate species richness models [38], species occurrence models [75] or species distribution models [42], which describe (stated simply) where in a region or around the world a species might live (or be able to live). However, these types of models do not typically describe the *abundance* (population size of a given species in an area) or *density* (how that population is spatially distributed [97]) of the species.

[4] https://www.kaggle.com/c/iwildcam-2020-fgvc7

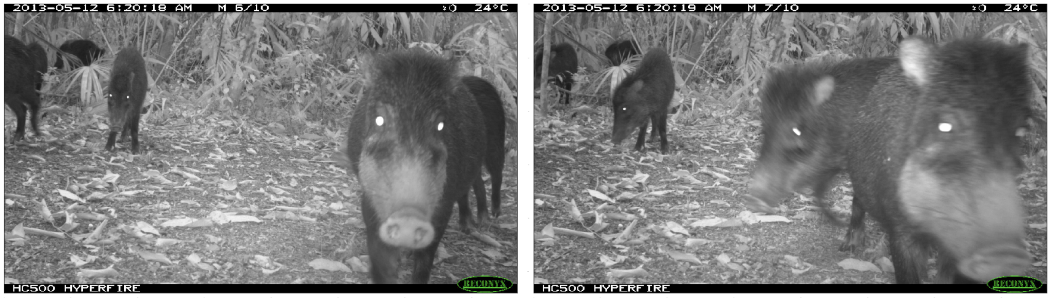


Figure 5.8: **How many pigs are there?** This year's challenge focuses on counting individuals across a sequence of camera trap images. Because the images are taken no faster than one frame per second, there are often temporal discontinuities between frames that make traditional tracking methods perform badly. However, humans are able to use a combination of spatio-temporal logic and visual re-identification to match individuals between frames.

A common method for population estimation is *mark-recapture*, which requires individual animals to be identified and recognized in future imagery [112]. Though strides are being made in visual re-identification for species with strong biometric markings such as zebras [105, 130], many species are not visually re-identifiable by humans, making data collection and analysis difficult. To address this, ecological models have been developed that estimate abundance from counts of individuals of a species captured in each camera across short time windows [80, 97]. The iWildCam 2021 competition [5] seeks to automate that counting process to enable abundance estimation to scale efficiently to large data collections, and one day to global data repositories such as Wildlife Insights [6].

Competitors will categorize and count species across short bursts of images in the test data. No count labels have been provided for the training set, in hopes that competitors will develop methods that can learn to count without explicit training labels, as most public camera trap data is not labeled with counts [1]. We provide competitors with species labels along with weakly-supervised detections [14] and instance segmentations [19] to help them to disambiguate individuals. The competition also maintains the multi-modal aspects of the iWildcam 2020 challenge [15] by providing citizen science images for the species of interest, remote sensing imagery for each camera location, and obfuscated geolocation for most cameras.

[5] iWildCam 2021 is hosted on Kaggle: https://www.kaggle.com/c/iwildcam2021-fgvc8

Dataset

The 2021 training set contains $203,314$ images from 323 locations, and the WCS test set contains $60,214$ images from 91 locations. These 414 locations are spread across 12 countries in different parts of the world. Each image is associated with a location ID so that images from the same location can be linked. In some cases, WCS biologists placed multiple cameras at the same location. We denote this with a sub-location ID, which communicates that the background and hardware of the camera at these sub-locations is different, but the physical location is the same. As is typical for camera traps, approximately 50% of the total number of images are empty (this varies per location). The iWildCam 2021 dataset is slightly smaller than the iWildCam 2020 dataset. We removed images from iWildCam 2020 that were found to be corrupted, mislabeled, or labeled with ambiguous categories like 'start'.

There are 206 species represented in the camera trap images. The class distribution is long-tailed, as shown in Fig. 5.6. Since we have split the data by location, some classes appear only in the training set. Any images with classes that appeared only in the test set were removed.

Count Labels Count labels for the test data were collected in collaboration with Centaur Labs [2]. We showed human annotators sequences of images that they could freely scroll through. Each sequence was labeled by between 3 and 30 individual annotators, with additional annotations collected for examples where annotators did not agree. Final counts were determined by majority vote, weighted by annotator performance on an expert-labeled subset. Sequences found to have multiple species were manually annotated by experts.

Obfuscated GPS Locations In order to allow competitors to try to use the geographic location of the cameras to improve their classification [74], we worked with WCS to release obfuscated GPS coordinates for most of the camera trap locations. The precise coordinates of the cameras have been obfuscated randomly to within 1 km for privacy and security reasons, and correspond to the centers of the provided remote sensing imagery. Some of the obfuscated GPS locations were not released at the request of WCS, but we can confirm that all locations without GPS are from the same country.

iNaturalist Data

iNaturalist is an online community where citizen scientists post photos of plants and animals and collaboratively identify the species [3]. Similar to iWildCam 2020, we provide a mapping from our classes into the iNaturalist taxonomy.[6] We also provide the subsets of the iNaturalist 2017-2019 competition datasets [126] that correspond to species seen in the camera trap data. This curated set provides $13,051$ additional images for training, covering 75 classes.

Though small relative to the camera trap data, the iNaturalist data has some unique characteristics. First, the class distribution is completely different (though it is still long tailed). Second, iNaturalist images are typically higher quality than the corresponding camera trap images, providing valuable examples for hard classes. See [15] for a comparison between iNaturalist images and camera trap images.

Remote Sensing Data

In addition to the raw remote sensing data for each camera location outlined in [15], this year we have provided pre-extracted ImageNet [34] features. We use an ImageNet-pretrained ResNet-50 [52] to extract features from the RGB channels of each multispectral image.

Provided Models

The MegaDetector Competitors are free to use the Microsoft AI for Earth MegaDetector [14] (a general and robust camera trap detection model)as they see fit. Megadetector V3 detects animal and human classes, while the MegaDetector V4 adds a vehicle class. Any version of the MegaDetector is allowed to be used in this competition. The models can be downloaded on the Microsoft Camera Traps GitHub repository [4]. We provide the top MegaDetector V3 boxes and associated confidences along with our WCS image metadata.

DeepMAC Along with MegaDetector box labels, we also provide a method to extract corresponding segmentation masks within each detected box. The segmentations are derived from the DeepMAC model [19]. Although DeepMAC is designed as an instance segmentation model (i.e. detection+segmentation), for this competition we provide an instance of the model which takes boxes as input from the user. Combined with the MegaDetector box labels, or a user-provided detection model, this can be used to extract a per-detection segmentation mask. We provide the DeepMAC masks associated with MegaDetector V3 boxes on Kaggle. Examples of segmentation results paired with MegaDetector V3 boxes can be seen in Fig. 5.9.

[6] Note that for the purposes of the competition, competitors may only use iNaturalist data from the 2017-2021 iNaturalist competition datasets.

The DeepMAC model was originally trained on all of COCO [70] and achieves a detection and mask mAP of 44.5 % and 39.7 % respectively.

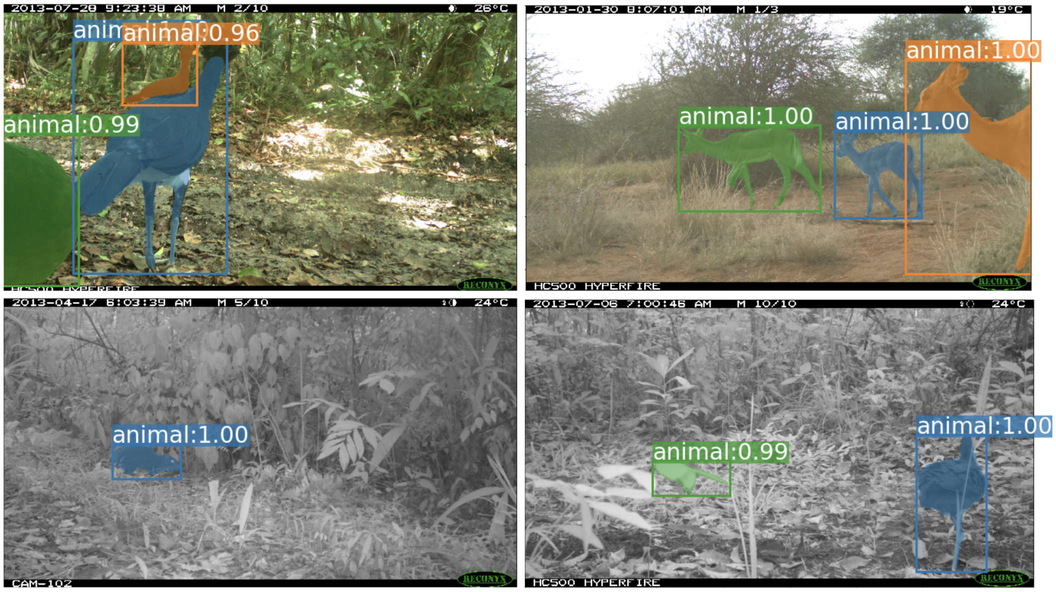

2013-07-28 9:23:38 AM M 2/10 26°C
anim
animal:0.96
animal:0.99
2013-01-30 8:07:01 AM M 1/3 19°C
animal:1.00
animal:1.00
animal:1.00
2013-04-17 6:03:39 AM M 5/10 24°C
animal:1.00
CAM-102
2013-07-06 7:00:46 AM M 10/10 24°C
animal:0.99
animal:1.00
HC500 HYPERFIRE

Figure 5.9: Segmentation results from DeepMAC, paired with MegaDetector V3 boxes. You can see in the lower right example that if the boxes are in error, the segmentation model will still provide its best guess at a segmentation (here it has segmented part of a plant that was a MegaDetector false positive).

Evaluation

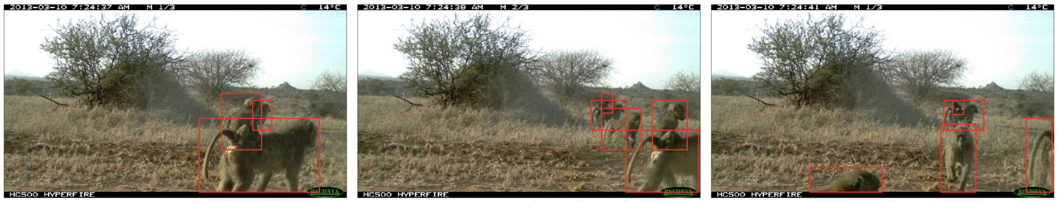


Figure 5.10: Here, the MegaDetector correctly boxed all animals and the classification model also correctly predected “baboon" as the class for all three images in the sequence. Our majority vote classification for the sequence is therefore “baboon" (correct) and our baseline model would see 5 boxes in both the second and 3rd image (the maximum number of boxes in any frame across the sequence) and predict “5 baboons". This prediction is close, but in fact there is one baboon in image 2 that is not visible in image 3, and one baboon in image 3 that is new, so the correct answer for this sequence would be “6 baboons".

Let $X \in \{0, 1, 2, ...\}^{n\times m}$ be a matrix of predictions, so each entry x_{ij} is the predicted count for species $j \in \{1, ..., m\}$ in sequence $i \in \{1, ..., n\}$. Let $Y \in \{0, 1, 2, ...\}^{n\times m}$ be the matrix of corresponding ground truth counts. Submissions will be evaluated using *mean columnwise root-mean-squared error* (MCRMSE) given by

$$\mathrm{MCRMSE}(X, Y) = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - y_{ij})^2}. \tag{5.1}$$

We selected this metric out of the options provided by Kaggle in order to capture both species identification mistakes and count mistakes as well as to ensure false predictions on empty sequences would contribute to the error. Because many sequences are empty in camera trap data and because many species are rare, the metric tends to be a small number even when the actual errors in counts are large. To convert the metric to something more interpretable, we can un-normalize the metric from MCRMSE to the *summed columnwise root summed squared error* (SCRSSE) given by

$$\begin{aligned}\mathrm{SCRSSE}(X, Y) &= m\sqrt{n}\mathrm{MCRMSE}(X, Y)\\ &= \sum_{j=1}^{m}\sqrt{\sum_{i=1}^{n}(x_{ij} - y_{ij})^2}.\end{aligned} \tag{5.2}$$

Baseline Results

We built our simple counting baselines from our iWildCam 2020 classification baseline model (see details in [15]), the iWildCam 2020 winning submission, and the provided MegaDetector V3 results. The results can be seen in Table 5.1, and the simple baselines are described below.

- **Max boxes:**. We assume that all high-confidence animal boxes (≥ 0.8) for an image are correct, and that the species in all boxes match our majority-vote classification prediction for that sequence. We take the maximum number of boxes from any image in the sequence and use that as our count. This will be a lower bound on the actual number of individuals across the sequence since it prevents double counting multiple images of the same individual. Example in Fig 5.10.
- **Sum boxes:** We assume that all high-confidence animal boxes (≥ 0.8) for each image are correct, and that the species in all boxes match our majority-vote classification prediction for that sequence. We take the sum of boxes across the sequence and use that as our count. This will be a upper bound on the actual number of individuals since individuals seen in multiple frames will be double counted.

Baseline	**MCRMSE**	**SCRSSE**
All zeros	0.03938	844.73
Max boxes A	0.05890	1263.50
Sum boxes A	0.17550	3753.49
One per predicted species A	0.04061	871.15
Max boxes B	0.03720	798.051
Sum boxes B	0.19897	4268.13
One per predicted species B	**0.03593**	**770.72**

Table 5.1: **Simple baseline results on the test set.** For the (A) set, we used the classification predictions from our naive classification baseline from the iWildCam 2020 competition [15]. For the (B) set we used the classification predictions from the iWildCam 2020 competition winning solution from Megvii Research Nanjing. We use the MegaDetector V3 boxes and a set of simple heuristics to generate counts from the species prediction.

- **One per predicted species:** We add a count of one for each unique species predicted by our image-level classification model across the sequence. This will be a lower bound on the actual number of individuals across the sequence as it just assumes that one animal was seen per species, regardless of detection results.
- **All zeros:** Just predict zero for all instances. Under our chosen metric this performs surprisingly well. This is for two reasons. First, camera trap data frequently has a small number of animals for any given species. Second, the model is double penalized if the count is correct but the species is incorrect (one penalty for missing the correct species count and one for overpredicting the incorrect species count).

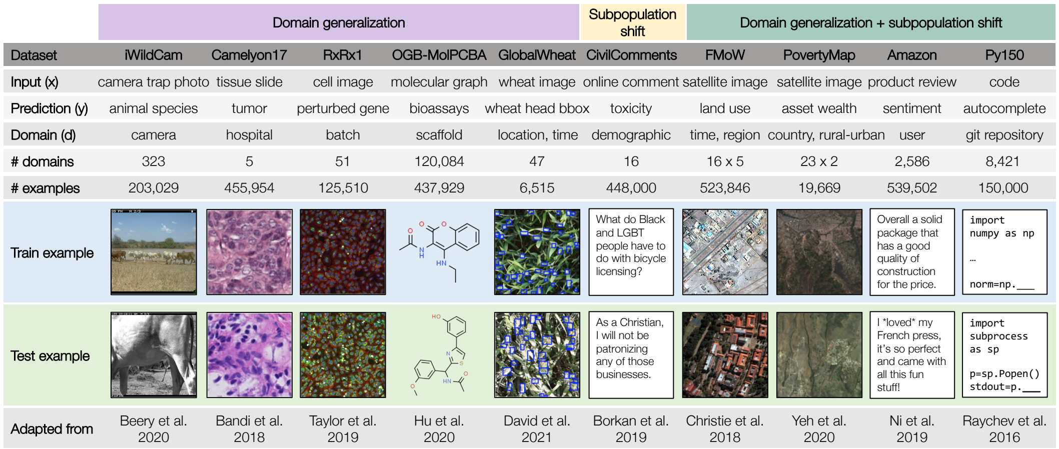
Challenge

The iWildCam Challenge 2021 was conducted through Kaggle as part of FGVC8 at CVPR21 and had 42 participating teams[7].

5.6 iWildCam in the WILDS distribution shift benchmark

We included iWildCam in Wilds, a curated benchmark of 10 datasets (see Fig. 5.11) reflecting a diverse range of distribution shifts that naturally arise in real-world applications, such as shifts across hospitals for tumor identification; across camera traps for wildlife monitoring; and across time and location in satellite imaging

[7] https://www.kaggle.com/c/iwildcam2021-fgvc8



	Domain generalization					Subpopulation shift	Domain generalization + subpopulation shift			
Dataset	iWildCam	Camelyon17	RxRx1	OGB-MolPCBA	GlobalWheat	CivilComments	FMoW	PovertyMap	Amazon	Py150
Input (x)	camera trap photo	tissue slide	cell image	molecular graph	wheat image	online comment	satellite image	satellite image	product review	code
Prediction (y)	animal species	tumor	perturbed gene	bioassays	wheat head bbox	toxicity	land use	asset wealth	sentiment	autocomplete
Domain (d)	camera	hospital	batch	scaffold	location, time	demographic	time, region	country, rural-urban	user	git repository
# domains	323	5	51	120,084	47	16	16 x 5	23 x 2	2,586	8,421
# examples	203,029	455,954	125,510	437,929	6,515	448,000	523,846	19,669	539,502	150,000

Train example						What do Black and LGBT people have to do with bicycle licensing?			Overall a solid package that has a good quality of construction for the price.	`import numpy as np`
										`…`
										`norm=np.___`

Test example						As a Christian, I will not be patronizing any of those businesses.			I *loved* my French press, it's so perfect and came with all this fun stuff!	`import subprocess as sp`
										`p=sp.Popen()`
										`stdout=p.___`

Adapted from	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Hu et al. 2020	David et al. 2021	Borkan et al. 2019	Christie et al. 2018	Yeh et al. 2020	Ni et al. 2019	Raychev et al. 2016

Figure 5.11: The WILDS benchmark contains 7 datasets across a diverse set of application areas, data modalities, and dataset sizes. Each dataset comprises data from different domains, and the benchmark is set up to evaluate models on distribution shifts across these domains.

and poverty mapping. On each dataset, we show that standard training yields substantially lower out-of-distribution than in-distribution performance. This gap remains even with models trained by existing methods for tackling distribution shifts, underscoring the need for new methods for training models that are more robust to the types of distribution shifts that arise in practice. To facilitate method development, we provide an open-source package that automates dataset loading, contains default model architectures and hyperparameters, and standardizes evaluations. **The full paper, code, and leaderboards are available at `https://wilds.stanford.edu`.**

WILDS Benchmark Overview

Distribution shifts—where the training distribution differs from the test distribution—pose significant challenges for machine learning (ML) systems deployed in the wild. In this work, we consider two common types of distribution shifts: *domain generalization* and *subpopulation shift* (5.12). Both of these shifts arise naturally in many real-world scenarios, and prior work has shown that they can substantially degrade model performance. In domain generalization, the training and test distributions comprise data from related but distinct domains, such as patients from different hospitals [148], images taken by different cameras [12], bioassays from different cell types [68], or satellite images from different countries and time periods [59]. In subpopulation shift, we consider test distributions that are subpopulations of the training distribution, with the goal of doing well even on the worst-case

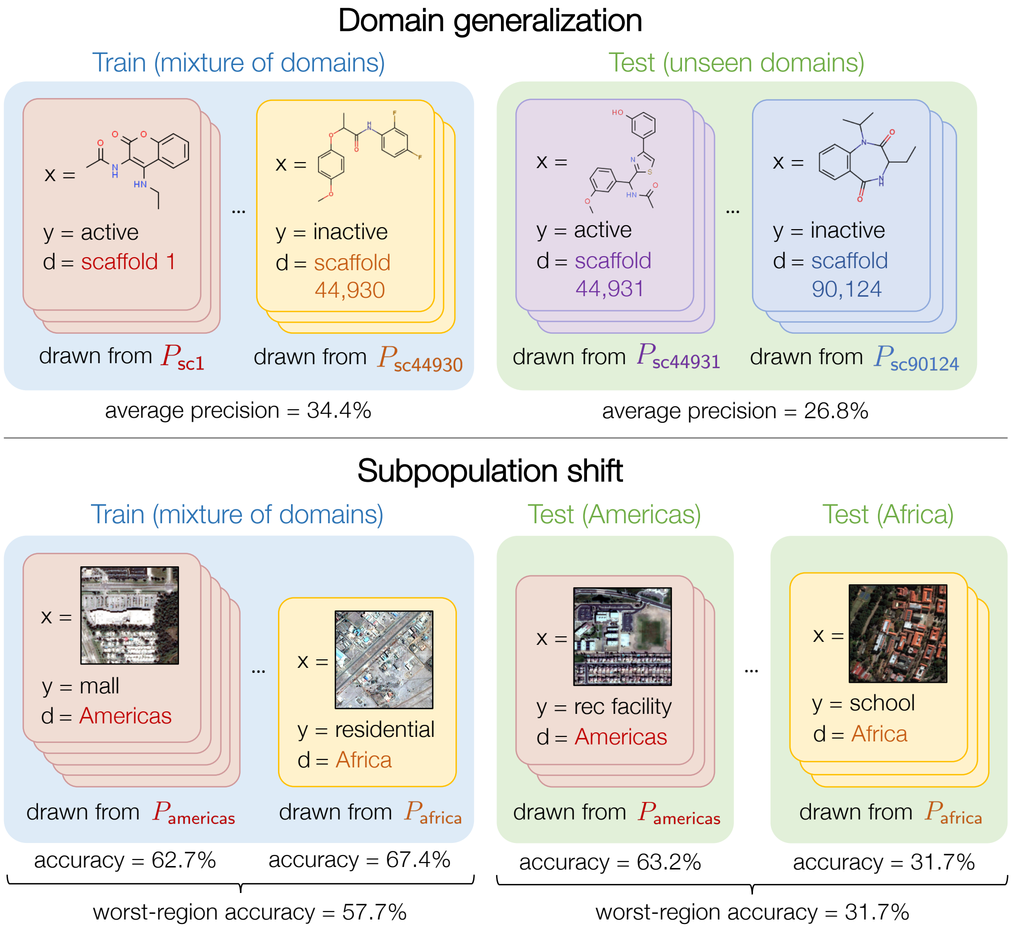


Figure 5.12: In the WILDS datasets, each data point (x, y, d) is drawn from a domain d. Each domain corresponds to a distribution P_d over data points that are similar in some way, e.g., molecules with the same scaffold structure, satellite images from the same region, or patients from the same hospital. We study two types of distribution shifts over domains. **Top:** In *domain generalization*, the training and test distributions comprise disjoint sets of domains, and the goal is to generalize to domains unseen during training, e.g., molecules with a new scaffold structure in OGB-MOLPCBA [58]. **Bottom:** In *subpopulation shift*, the training and test domains overlap, but their relative proportions differ. We typically assess models by their worst performance over test domains, each of which correspond to a subpopulation of interest, e.g., different geographical regions in FMoW-WILDS [29].

subpopulation; e.g., we might seek models that perform well on all demographic subpopulations, including minority individuals [21].

Despite their ubiquity, these real-world distribution shifts are under-represented in the datasets widely used in the ML community today [45]. Most of these datasets were instead designed for the standard i.i.d. setting, with training and test sets from the same distribution, and prior work on retrofitting them with distribution shifts has focused on shifts that are cleanly characterized but not necessarily likely to arise in real-world deployments. For instance, many recent papers have studied datasets with shifts induced by synthetic transformations, such as changing the color of MNIST digits [8], or by targeted data splits, such as generalizing from cartoons to photos [65]. Datasets like these are important testbeds that allow for systematic studies; but to develop and evaluate methods for real-world distribution shifts, it is also necessary to complement these existing datasets with ones that capture shifts in the wild.

WILDS is a curated collection of benchmark datasets with evaluation metrics and train/test splits that represent the kinds of distribution shifts that ML models face in the wild (5.14). WILDS datasets span a broad array of societally-important applications with natural distribution shifts: animal species categorization [15], tumor identification [10], bioassay prediction [58, 141], text toxicity classification [20], sentiment analysis [82], land use classification [29], and poverty mapping [146]. These 7 datasets reflect distribution shifts arising from different cameras, hospitals, molecular scaffolds, demographics, users, countries, and time periods.

WILDS builds on extensive data-collection efforts by domain experts, who are often forced to grapple with distribution shifts to make progress on problems in their applications. To design WILDS, we worked with them to identify, select, and adapt datasets that fulfilled the following criteria:

1. **Distribution shifts with performance drops.** The train/test splits reflect shifts that substantially degrade model performance, i.e., with a large gap between in-distribution and out-of-distribution performance.

2. **Real-world relevance.** The training/test splits and evaluation metrics are motivated by real-world scenarios, and chosen in conjunction with domain experts to be consistent with prior work in their corresponding applications.

3. **Potential leverage.** Distribution shift benchmarks must be non-trivial but also possible to solve, as models cannot be expected to generalize to arbitrary distribution shifts. We constructed each WILDS dataset to have training data from multiple domains, with domain annotations and other metadata available at training time. We hope that these can be used to learn robust models: e.g., for domain generalization, one could use these annotations to learn models that are invariant to domain-specific features, while for subpopulation shift, one could learn models that perform uniformly well across each subpopulation.

We chose the WILDS datasets to collectively encompass a diversity of tasks, data modalities, dataset sizes, and numbers of domains, so as to enable evaluation across a broad range of real-world distribution shifts. To make these datasets accessible, we have substantially modified most of them, e.g., to clarify the distribution shift, standardize the data splits, or preprocess the data for use in standard ML frameworks.

Datasets are significant catalysts for ML research. Likewise, benchmarks that curate and standardize datasets—e.g., the GLUE and SuperGLUE benchmarks for language understanding [133, 134] and the Open Graph Benchmark for graph ML [58]—can accelerate research by focusing community attention, easing development on multiple datasets, and enabling systematic comparisons between approaches. With WILDS, we aim to facilitate progress on handling real-world distribution shifts in a broad range of societally-important ML applications.

Problem settings

Each WILDS dataset is associated with a type of domain shift: domain generalization, subpopulation shift, or a hybrid of both (5.14). In each setting, we can view the overall data distribution as a mixture of D domains $\mathcal{D} = \{1, \ldots, D\}$. Each domain $d \in \mathcal{D}$ corresponds to a fixed data distribution P_d over (x, y, d), where x is the input, y is the prediction target, and all points sampled from P_d have domain d. We encode the domain shift by assuming that the training distribution $P^{\text{train}} = \sum_{d\in\mathcal{D}} q_d^{\text{train}} P_d$ has mixture weights q_d^{train} for each domain d, while the test distribution $P^{\text{test}} = \sum_{d\in\mathcal{D}} q_d^{\text{test}} P_d$ is a different mixture of domains with weights q_d^{test}. For convenience, we define the set of training domains as $\mathcal{D}^{\text{train}} = \{d \in \mathcal{D} \mid q_d^{\text{train}} > 0\}$, and likewise, the set of test domains as $\mathcal{D}^{\text{test}} = \{d \in \mathcal{D} \mid q_d^{\text{test}} > 0\}$.

At training time, the learning algorithm gets the domain annotations d, i.e., the training set comprises points $(x, y, d) \sim P^{\text{train}}$. At test time, the model gets either x or (x, d) drawn from P^{test}, depending on the application.

Domain generalization (5.12-Top). In domain generalization, we aim to generalize to test domains that are similar to but distinct from the training domains, i.e., the training domains $\mathcal{D}^{\text{train}}$ and test domains $\mathcal{D}^{\text{test}}$ are disjoint, with $\mathcal{D}^{\text{train}} \cap \mathcal{D}^{\text{test}} = \emptyset$. For example, in CAMELYON17-WILDS, we train on data from some hospitals and test on a different hospital. We typically seek to minimize the average error on the test distribution.

Subpopulation shift (5.12-Bottom). In subpopulation shift, we aim to perform well across a wide range of domains seen during training time. For example, in

CivilComments-wilds, the domains d represent particular demographics, some of which are a minority in the training set, and we seek high accuracy on each of these subpopulations without observing their demographic identity d at test time. Concretely, all test domains are seen during training time, with $\mathcal{D}^{\text{test}} \subseteq \mathcal{D}^{\text{train}}$, but the proportions of the domains change, with $q^{\text{test}} \neq q^{\text{train}}$. We typically seek to minimize the maximum error over all test domains.

Hybrid settings. Some settings are a combination of both domain generalization and subpopulation shift. For example, in the FMoW-wilds dataset, the inputs are satellite images and the domains correspond to the year and geographical region in which they were taken. We simultaneously consider domain generalization across time (i.e., the training set comprises images taken before a certain year, and the test set comprises images taken afterwards) and subpopulation shift across geographical regions (i.e., there are images from the same geographical regions in the training and test sets, but at different proportions).

Baseline algorithms for distribution shifts

Many algorithms have been proposed for training models that are more robust to particular distribution shifts than standard ERM models. Unlike ERM, these algorithms tend to utilize domain annotations during training, with the goal of learning a model that can generalize across domains. In this section, we evaluate several representative algorithms from prior work and show that the out-of-distribution performance drops remain.

Domain generalization baselines. Methods for domain generalization typically involve adding a penalty to the ERM objective that encourages some form of invariance across domains. We include two such methods as representatives:

- **CORAL** [117], which penalizes differences in the means and covariances of the feature distributions (i.e., the distribution of last layer activations in a neural network) for each domain. Conceptually, CORAL is similar to other methods that encourage feature representations to have the same distribution across domains [44, 67, 69, 71, 124].

- **IRM** [8], which penalizes feature distributions that have different optimal linear classifiers for each domain. This builds on earlier work on invariant predictors [90].

Other techniques for domain generalization include conditional variance regularization [54]; self-supervision [24]; and meta-learning-based approaches [9, 39, 66].

Subpopulation shift baselines. In subpopulation shift settings, our aim is to train models that perform well on all relevant subpopulations. We test the following approach:

- **Group DRO** [57, 99], which uses distributionally robust optimization to explicitly minimize the loss on the worst-case domain during training. Group DRO builds on the maximin approach developed in Meinshausen and Bühlmann [76].

Other methods for subpopulation shifts include reweighting methods based on class/-domain frequencies [32, 110]; label-distribution-aware margin losses [22]; adaptive Lipschitz regularization [23]; slice-based learning [28, 94]; style transfer across domains [48]; or other DRO algorithms that do not make use of explicit domain information and rely on, for example, unsupervised clustering [86, 114].

Subpopulation shifts are also connected to the well-studied notions of tail performance and risk-averse optimization (Chapter 6 in Shapiro et al. [107]). For example, optimizing for the worst case over all subpopulations of a certain size, regardless of domain, can guarantee a certain level of performance over the smaller set of subpopulations defined by domains [40, 41].

The iWildCam2020-wilds Benchmark in Wilds

Animal populations have declined 68% on average since 1970 [7]. To better understand and monitor wildlife biodiversity loss, ecologists commonly deploy camera traps—heat or motion-activated static cameras placed in the wild [136]—and then use ML models to process the data collected [6, 14, 84, 119, 137]. Typically, these models would be trained on photos from some existing camera traps and then used across new camera trap deployments. However, across different camera traps, there is drastic variation in illumination, camera angle, background, vegetation, color, and relative animal frequencies, which results in models generalizing poorly to new camera trap deployments [12].

We study this shift on a variant of the iWildCam 2020 dataset [15].

Setup

Problem setting. We consider the domain generalization setting, where the domains are camera traps, and we seek to learn models that generalize to photos taken from new camera deployments (5.3). The task is multi-class species classification. Concretely, the input x is a photo taken by a camera trap, the label y is one of 186 different animal species, and the domain d is an integer that identifies the camera trap that took the photo.

The training set contains 142,202 images from 245 camera traps, and the test set contains 38,943 images from 47 camera traps. A total of 324 camera traps are spread across multiple countries in different parts of the world. Each image is associated with a camera trap ID. As is typical for camera traps, approximately 50% of the total number of images are empty (this varies per location).

Data. The dataset comprises 217,609 images from 324 different camera traps spread across 12 countries in different parts of the world. The original camera trap data comes from the Wildlife Conservation Society.[8] Approximately half of the images do not contain any animal species; this corresponds to one of the 186 class labels. We split the dataset by randomly partitioning the data by camera traps:

1. **Training:** 142,202 images taken by 245 camera traps.
2. **Validation (OOD):** 20,784 images taken by 32 different camera traps.
3. **Test (OOD):** 38,943 images taken by 47 different camera traps.
4. **Validation (ID):** 7,819 images taken by the same camera traps as the training set (but distinct from the training images).
5. **Test (ID):** 7,861 images taken by the same camera traps as the training set (but distinct from the training images).

The camera traps were randomly distributed across the training, validation (OOD), and test (OOD) sets.

The original iWildCam 2020 Kaggle competition similarly split the dataset by camera trap, though the competition focused on average accuracy. We consider a smaller subset of the data here; see 5.6.

[8] `http://lila.science/datasets/wcscameratraps`

Table 5.2: Baseline results on iWildCam2020-wilds.

	Test (ID)		Test (OOD)	
Algorithm	Macro F1	Average accuracy	Macro F1	Average accuracy
ERM	**82.5** (1.3)	**96.5** (0.2)	**27.8** (1.3)	**62.9** (0.5)
CORAL	68.3 (8.8)	93.1 (2.2)	26.3 (1.4)	62.5 (1.7)
IRM	16.45 (2.8)	57.26 (6.8)	13.93 (2.3)	50.78 (3.0)
Group DRO	58.6 (2.2)	89.3 (1.8)	23.8 (0.7)	63.0 (0.8)

Evaluation. We evaluate models by their macro F1 score (i.e., we compute the F1 score for each class separately, then average those scores). We also report the average accuracy of each model across all test images, but primarily use the macro F1 score to better capture model performance on rare species. In the natural world, protected and endangered species are rare by definition, and are often the most important to accurately monitor. However, common species are much more likely to be captured in camera trap images; this imbalance can make metrics like average accuracy an inaccurate picture of model effectiveness.

Potential leverage. Though the problem is challenging for existing ML algorithms, adapting to photos from different camera traps is simple and intuitive for humans. Repeated backgrounds and habitual animals, which cause each sensor to have a unique class distribution, provide a strong implicit signal across data from any one location. We anticipate that approaches that utilize the provided camera trap annotations can learn to factor out these common features and avoid learning spurious correlations between particular backgrounds and animal species.

Baseline results

ERM results and performance drops. We trained a ResNet-50 [52] that was pretrained on ImageNet. Model performance dropped substantially and consistently going from in-distribution (ID) to out-of-distribution (OOD) camera traps (5.2), with a macro F1 score of 82.5 on the ID test set but only 27.8 on the OOD test set. Similarly, the model obtained an average accuracy of 96.5% on the ID test set but only 62.9% on the OOD test set. The large discrepancy between ID and OOD model performance suggests that there is significant room for improvement.

Additional baseline methods. We trained CORAL, IRM, and Group DRO models, treating each camera trap as a domain. However, these did not improve upon the

ERM baseline (Table 5.2). We observed that as the F1 score on the OOD validation set tends to converge more quickly than on the ID validation set, early stopping on OOD validation sometimes leads to selecting an early epoch, which in turn leads to low scores on the ID validation and test sets. For example, this occurred for one random seed out of three in the CORAL models, which partially contributes to its lower and more variable ID accuracy and F1 scores. The IRM models performed especially poorly on this dataset; we suspect that this is because the default estimator of the IRM penalty term can be negatively biased when examples are sampled without replacement from small domains, but further investigation is needed.

Discussion. Even though there is significant label imbalance, the overall label distribution is approximately the same in the ID and OOD split, suggesting that it is not primarily label shift that accounts for the performance drop. Across locations, there is drastic variation in illumination, camera angle, background, vegetation, and color. This variation, coupled with considerable differences in the distribution of animals between camera traps, likely encourages the model to overfit to specific animal species appearing in specific locations, which may account for the performance drop.

The original iWildCam 2020 competition allows users to use MegaDetector [14], which is an animal detector trained on a large set of data beyond what is provided in the training set. The MegaDetector training set is not publicly available. To facilitate more controlled experiments, we intentionally do not use MegaDetector in our baselines for iWildCam2020-wilds. We welcome leaderboard submissions that use MegaDetector, as it is useful to see how much better models can perform when they use MegaDetector or other similar animal detectors, but we will distinguish these submissions from others that only use what is provided in the training set.

Broader context

Differences across data distributions at different sensor locations is a common challenge in automated wildlife monitoring applications, including using audio sensors to monitor animals that are easier heard than seen such as primates, birds, and marine mammals [31, 111, 115], and using static sonar to count fish underwater to help maintain sustainable fishing industries [91, 103, 128]. As with camera traps, each static audio sensor has a specific species distribution as well as a sensor specific background noise signature, making generalization to new sensors challeng-

Table 5.3: The number of examples and camera traps in each split for iWildCam2020-wilds.

Split	# Examples	# Camera traps
Training	142,202	245
Validation (ID)	7,819	223
Test (ID)	7,861	224
Validation (OOD)	20,784	32
Test (OOD)	38,943	47

ing. Similarly, static sonar used to measure fish escapement have sensor-specific background reflectance based on the shape of the river bottom. Moreover, since species are distributed in a non-uniform and long-tailed fashion across the globe, it is incredibly challenging to collect sufficient samples for rare species to escape the low-data regime. Implicitly representing camera-specific distributions and background features in per-camera memory banks and extracting relevant information from these via attention has been shown to help overcome some of these challenges for static cameras [17].

More broadly, shifts in background, image illumination and viewpoint have been studied in computer vision research. First, several works have shown that object classifiers often rely on the background rather than the object to make its classification [96, 109, 142]. Second, common perturbations such as blurriness or shifts in illumination, tend to reduce performance [37, 55, 122]. Finally, shifts in rotation and viewpoint of the object has been shown to degrade performance [11].

Additional details

Data processing. We generate the data splits in three steps. First, to generate the OOD splits, we randomly split all locations into three groups: Validation (OOD), Test (OOD), and Others. Then, to generate the ID splits, we split the Others group uniformly at random into three sets: Training, Validation (ID), and Test (ID).

When doing the ID split, some locations only ended up in some of but not all of Training, Validation (ID), and Test (ID). For instance, if there were very few dates for a specific location (camera trap), it may be that no examples from that location ended up in the train split. This defeats the purpose of the ID split, which is to test performance on locations that were seen during training. We put these locations in the train split. Finally, any images in the test set with classes not present in the train set were removed.

Modifications to the original dataset. In the competition on Kaggle there is a held-out test set that we are not utilizing, as the test set is intended to be reused in a future competition and is not yet public. Instead, we constructed our own test set by splitting the Kaggle competition training data into our own splits: train, validation (ID), validation (OOD), test (ID), test (OOD).

Images are organized into sequences, but we treat each image separately. In the iWildCam 2020 competition, the top participants utilize the sequence data and also use a pretrained MegaDetector animal detection model that outputs bounding boxes over the animals. These images are cropped using the bounding boxes and then fed into a classification network. As we discuss in 5.6, we intentionally do not use MegaDetector in our experiments.

Baseline model details. We train a ResNet-50 with batch size 16 for 18 epochs, on images resized to 224 by 224. We pick hyperparameters by doing a grid search over different learning rates, 10^{-3}, 10^{-4} and 10^{-5} and different weight decay, 0, 10^{-4} and 10^{-5}. The optimizer is Adam. We pick the best hyperparameters and run 3 seeds.

When training the CORAL baseline, we use the best best learning rate and weight decay from ERM. To pick the penalty weight we do a grid search over $\{0.1, 1, 10\}$.

5.7 iWildCam in WILDS 2.0

Machine learning systems deployed in the wild are often trained on a source distribution but deployed on a different target distribution. Unlabeled data can be a powerful point of leverage for mitigating these distribution shifts, as it is frequently much more available than labeled data and can often be obtained from distributions beyond the source distribution as well. However, existing distribution shift benchmarks with unlabeled data do not reflect the breadth of scenarios that arise in real-world applications. iWildCam2020-wilds was included in the Wilds 2.0 update, which extends 8 of the 10 datasets in the Wilds benchmark of distribution shifts to include curated unlabeled data that would be realistically obtainable in deployment. These datasets span a wide range of applications (from histology to wildlife conservation), tasks (classification, regression, and detection), and modalities (photos, satellite images, microscope slides, text, molecular graphs). The update maintains consistency with the original Wilds benchmark by using identical labeled training, validation, and test sets, as well as the evaluation metrics. We systematically benchmark state-of-the-art methods that leverage unlabeled data, including domain-invariant, self-training, and self-supervised methods, and show that their success on Wilds is limited. To facilitate method development and evaluation, we provide an open-source package that automates data loading and contains all of the model architectures and methods used in this paper. Code and leaderboards are available at `https://wilds.stanford.edu`.

Overview of Wilds 2.0

Distribution shifts—when models are trained on a source distribution but deployed on a different target distribution—are frequent problems for machine learning systems in the wild [45, 61, 92]. In this update, we focus on the use of unlabeled data to mitigate these shifts. Unlabeled data is a powerful point of leverage as it is more readily available than labeled data and can often be obtained from distributions beyond the source distribution. For example, in the crop detection task in 5.13, we wish to learn a model that can extrapolate to a set of target domains (farms) [33], and while we only have labeled training examples from some source domains, we have many more unlabeled examples from the source domains, from extra domains, and even directly from the target domains.

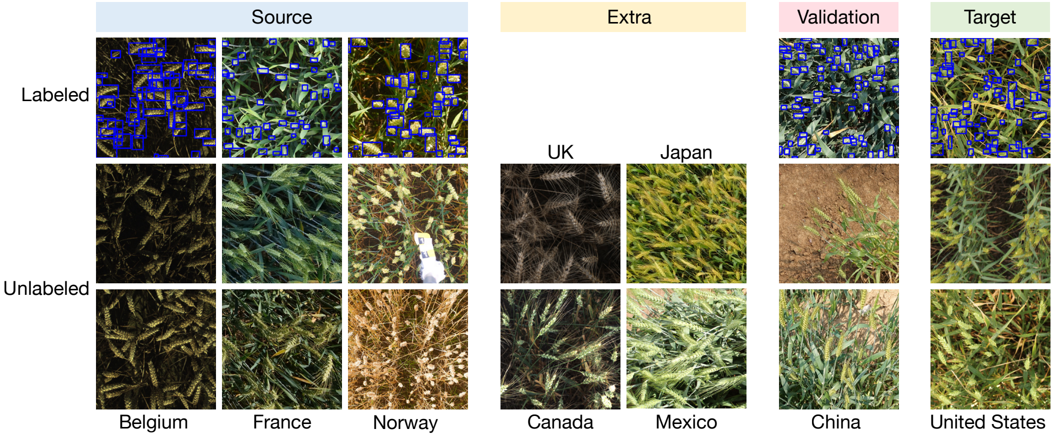


Figure 5.13: Each Wilds dataset [61] contains labeled data from the source domains (for training), validation domains (for hyperparameter selection), and target domains (for held-out evaluation). In the Wilds 2.0 update, we extend these datasets with unlabeled data from a combination of source, validation, or target domains, as well as extra domains from which there is no labeled data. The labeled data is exactly the same as in Wilds 1.0. In this figure, we illustrate the setting with the GlobalWheat-wilds dataset, where domains correspond to images acquired from different locations and at different times.

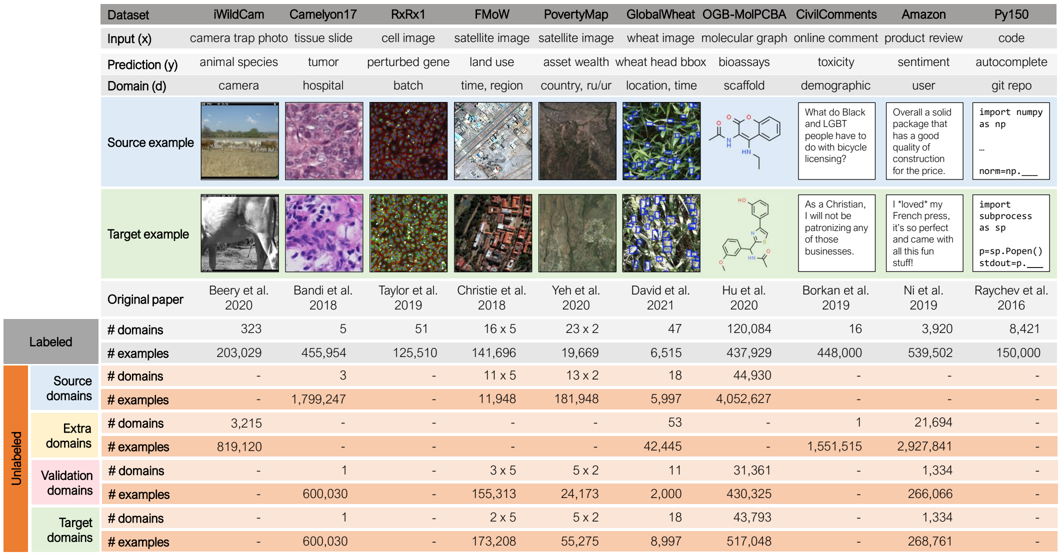
Many methods for leveraging unlabeled data have been highly successful on some types of distribution shifts [18, 150]. However, the datasets typically used for evaluating these methods do not reflect many of the realistic shifts that might occur in the wild. These evaluations tend instead to focus on shifts between photos and stylized versions like sketches [65, 89, 129] or synthetic renderings [88], or between variants of digits datasets like MNIST [62] and SVHN [81]. Unfortunately, prior work has shown that methods that work well on one type of shift need not generalize to others [36, 78, 121, 144], which raises the question of how well they would work on a wider array of realistic shifts.

In WILDS 2.0, we extend 8 of the 10 WILDS datasets[9] with curated unlabeled data acquired from the same source and target domains as the labeled data, as well as from extra domains of the same type: e.g., in the GLOBALWHEAT-WILDS dataset pictured in 5.13, we acquired unlabeled photos of wheat fields from the source and target farms as well as extra farms that were not in the original labeled dataset. In total, WILDS 2.0 adds 14.5 million unlabeled examples, expanding the number of examples for each dataset by 3–13× and **allowing us to combine the real-world relevance of WILDS with the leverage of unlabeled data**.

We developed a standardized and consistent protocol for evaluating methods that leverage the unlabeled data in WILDS 2.0. We assessed representatives from three popular categories: methods for learning domain-invariant representations [44, 117], self-training methods [63, 113, 143], and pre-training methods that rely on self-supervision [25, 35]. These methods have been successful on some types of shifts, such as going from photos to sketches, or from handwritten digits to street signs [18, 150].

Our results across the WILDS datasets are mixed: many methods did not outperform standard supervised training despite using additional unlabeled data, and the only clear successes were on two image classification datasets (CAMELYON17-WILDS and FMOW-WILDS). Successful methods relied heavily on data augmentation [25, 143], which limited their applicability to modalities where augmentation techniques are not as well developed, such as text and molecular graphs. The same methods were unsuccessful on the image regression and detection tasks, which have been relatively understudied: e.g., pseudolabel-based methods do not straightforwardly apply to regression. For the text datasets, continued language model

[9] We omitted PY150-WILDS, as code completion data is always labeled by nature of the task, and RXRX1-WILDS, as unlabeled data for that genetic perturbation task is not typically available.



	Dataset	iWildCam	Camelyon17	RxRx1	FMoW	PovertyMap	GlobalWheat	OGB-MolPCBA	CivilComments	Amazon	Py150
	Input (x)	camera trap photo	tissue slide	cell image	satellite image	satellite image	wheat image	molecular graph	online comment	product review	code
	Prediction (y)	animal species	tumor	perturbed gene	land use	asset wealth	wheat head bbox	bioassays	toxicity	sentiment	autocomplete
	Domain (d)	camera	hospital	batch	time, region	country, ru/ur	location, time	scaffold	demographic	user	git repo

	Source example								What do Black and LGBT people have to do with bicycle licensing?	Overall a solid package that has a good quality of construction for the price.	`import numpy as np`
											…
											`norm=np.___`

	Target example								As a Christian, I will not be patronizing any of those businesses.	I *loved* my French press, it's so perfect and came with all this fun stuff!	`import subprocess as sp`
											`p=sp.Popen()`
											`stdout=p.___`

	Original paper	Beery et al. 2020	Bandi et al. 2018	Taylor et al. 2019	Christie et al. 2018	Yeh et al. 2020	David et al. 2021	Hu et al. 2020	Borkan et al. 2019	Ni et al. 2019	Raychev et al. 2016
Labeled	# domains	323	5	51	16 x 5	23 x 2	47	120,084	16	3,920	8,421
	# examples	203,029	455,954	125,510	141,696	19,669	6,515	437,929	448,000	539,502	150,000
Unlabeled: Source domains	# domains	-	3	-	11 x 5	13 x 2	18	44,930	-	-	-
	# examples	-	1,799,247	-	11,948	181,948	5,997	4,052,627	-	-	-
Extra domains	# domains	3,215	-	-	-	-	53	-	1	21,694	-
	# examples	819,120	-	-	-	-	42,445	-	1,551,515	2,927,841	-
Validation domains	# domains	-	1	-	3 x 5	5 x 2	11	31,361	-	1,334	-
	# examples	-	600,030	-	155,313	24,173	2,000	430,325	-	266,066	-
Target domains	# domains	-	1	-	2 x 5	5 x 2	18	43,793	-	1,334	-
	# examples	-	600,030	-	173,208	55,275	8,997	517,048	-	268,761	-

Figure 5.14: The WILDS 2.0 update adds unlabeled data to 8 WILDS datasets. For each dataset, we kept the labeled data from WILDS and expanded the datasets by 3–13× with unlabeled data from the same underlying dataset. The type of unlabeled data (i.e., whether it comes from source, extra, validation, or target domains) depends on what is realistic and available for the application. Beyond these 8 datasets, WILDS also contains 2 datasets without unlabeled data: the PY150-WILDS code completion dataset and the RXRX1-WILDS genetic perturbation dataset. For all datasets, the labeled data and evaluation metrics are exactly the same as in WILDS 1.0. Figure adapted with permission from Koh et al. [61].

pre-training did not help, unlike in prior work [51]. These results suggest fruitful avenues for future work, such as developing data augmentation techniques for non-image modalities and more realistic hyperparameter tuning protocols.

Overall, our results underscore the importance of developing and evaluating methods for unlabeled data on a wider variety of real-world shifts than is typically studied. To this end, we have updated the open-source Python WILDS package to include unlabeled data loaders, compatible implementations of all the methods we benchmarked, and scripts to replicate all experiments in this paper. Code and public leaderboards are available at `https://wilds.stanford.edu`. By allowing developers to easily test algorithms across the variety of datasets in WILDS 2.0, we hope to accelerate the development of methods that can leverage unlabeled data to improve robustness to real-world distribution shifts.

Finally, we note that WILDS 2.0 not a separate benchmark from WILDS 1.0: the labeled data and evaluation metrics are exactly the same in WILDS 1.0 and WILDS 2.0, and future results should be reported on the overall WILDS benchmark, with a note describing what kind of unlabeled data (if any) was used. In this paper, we discuss the addition of unlabeled data and analyze the performance of methods that use the unlabeled data. For a more detailed description of the datasets, evaluation metrics, and models used, please refer to the original WILDS paper [61].

Baseline Algorithms

For our evaluation, we selected representative methods from the three categories described below. These methods exemplify current approaches to using unlabeled data to improve robustness, and they have been successful on popular domain adaptation benchmarks like DomainNet [89] and semi-supervised settings like improving ImageNet accuracy by leveraging unlabeled images from the internet [25, 143].

Domain-invariant methods. Domain-invariant methods learn feature representations that are invariant across different domains by penalizing differences between learned source and target representations [44, 71–73, 102, 117, 145, 149, 151].

For our experiments, we evaluate two classical methods:

- *Domain-Adversarial Neural Networks (DANN)* [44] penalize representations on which an auxiliary classifier can easily discriminate between source and target examples.
- *Correlation Alignment (CORAL)* [117, 118] penalizes differences between the means and covariances of the source and target feature distributions.

Self-training. Self-training methods "pseudo-label" unlabeled examples with the model's own predictions and then train on them as if they were labeled examples. These methods often also use consistency regularization, which encourages the model to make consistent predictions on augmented views of unlabeled examples [18, 113, 143]. Self-training methods have recently been successfully applied to unsupervised adaptation [18, 101, 150]. We include three representative algorithms:

- *Pseudo-Label* [63] dynamically generates pseudolabels and updates the model each batch.
- *FixMatch* [113] adds consistency regularization on top of the Pseudo-Label algorithm. Specifically, it generates pseudolabels on a weakly augmented view of the unlabeled data, and then minimizes the loss of the model's prediction on a strongly augmented view.

- *Noisy Student* [143] leverages weak and strong augmentations like FixMatch, but instead of dynamically generating pseudolabels for each batch, it alternates between a few teacher phases, where it generates pseudolabels, and student phases, where it trains to convergence on the (pseudo)labeled data.

Self-supervision. Self-supervised methods learn useful representations by training on unlabeled data via auxiliary proxy tasks. Common approaches include reconstruction tasks [35, 43, 46, 64, 132], and contrastive learning [25, 27, 53, 93], and recent work has shown that self-supervised methods can reduce dependence on spurious correlations and improve performance on domain adaptation tasks [79, 123, 135]. We use these self-supervision methods for unsupervised adaptation by first pre-training models on the unlabeled data, and then finetuning them on the labeled source data [108]. We evaluate popular self-supervised methods for vision and language:

- *SwAV* [25] is a contrastive learning algorithm that maps representations to a set of clusters and then enforces similarity between cluster assignments.

- *Masked language modeling (MLM)* [35] randomly masks some of the tokens from input text and trains the model to predict the missing tokens.

The iWildCam2020-wilds Benchmark in Wilds 2.0

The iWildCam2020-wilds dataset was adapted from the iWildCam 2020 competition dataset made up of data provided by the Wildlife Conservation Society (WCS) [15] [10]. Camera trap images are captured by motion-triggered static cameras placed in the wild to study wildlife in a non-invasive manner. Images are captured at high volumes–a single camera trap can capture 10K images in a month–and annotating these images requires species identification expertise and is time-intensive. However, there are tens of thousands of camera traps worldwide capturing images of wildlife that could be used as unlabeled training data. For example, Wildlife Insights [6] now contains almost 20M camera trap images collected across the globe, but a large proportion of that data is still unlabeled. Ideally we could capture value from those images despite the lack of available labels. We extend iWildCam2020-wilds with unlabeled data from a set of WCS camera traps entirely disjoint with the labeled dataset, representative of unlabeled data from a newly-deployed sensor network.

[10] The WCS Camera Traps Dataset can be found at `http://lila.science/datasets/wcscameratraps`

Problem setting. The task is to classify the species of animals in camera trap images. The input x is an image from a camera trap, and the domain d corresponds to the camera trap that captured the image. The target y, provided only for the labeled training images, is one of 182 classes of animals. We seek to learn models that generalize well to new camera trap deployments, so the test data comes from domains unseen during training. Additionally, we evaluate the in-distribution performance on held-out images from camera traps in the train set.

Data. The data comes from multiple camera traps around the world, all provided by the Wildlife Conservation Society (WCS). The labeled data is the same as in Koh et al. [61] and the unlabeled data comprise 819,120 images from 3215 WCS camera traps not included in iWildCam 2020:

1. **Source**: 243 camera traps.
2. **Validation (OOD):** 32 camera traps.
3. **Target (OOD):** 48 camera traps.
4. **Extra:** 3215 camera traps.

The four sets of camera traps are disjoint. The distributions of the labeled and unlabeled camera traps are very similar, except that the labeled data does not contain cameras with photos taken before LandSat 8 data was available.

Table 5.4: Data for iWildCam2020-wilds. Each domain corresponds to a different camera trap.

Split	# Domains (camera traps)	# Labeled examples	# Unlabeled examples
Source	243	129,809	0
Validation (ID)		7,314	0
Target (ID)		8,154	0
Validation (OOD)	32	14,961	0
Target (OOD)	48	42,791	0
Extra (OOD)	3215	0	819,120
Total	3538	203,029	819,120

Broader context. There are large volumes of unlabeled natural world data that have been collected in growing repositories such as iNaturalist [85], Wildlife Insights [6], and GBIF [95]. This data includes images or video collected by remote sensors or community scientists, GPS track data from an-animal devices, aerial data from drones or satellites, underwater sonar, bioacoustics, and eDNA. Methods that can harness the wealth of information in unlabeled ecological data are well-posed to make significant breakthroughs in how we think about ecological and conservation-focused research. Natural-world and ecological benchmarks that provide unlabeled data include NEWT [127], investigating efficient task learning, and Semi-Supervised iNat [116], which provides labeled data for only a subset of the taxonomic tree. Recent work has begun to adapt weakly-supervised and self-supervised approaches for these natural world settings, including probing the generality and efficacy of self-supervision [30], incorporating domain-relevant context into self-supervision [87], or leveraging weak supervision from alternative data modalities [138] or pre-trained, generic models [14, 139]. Active learning also plays a role here in seeking to adapt models efficiently to unlabeled data from novel regions with only a few targeted labels [60, 84].

Table 5.5: The in-distribution (ID) and out-of-distribution (OOD) performance of each method on iWildCam2020-wilds. Following Wilds 1.0, we ran 3–10 replicates (random seeds) for each cell. We report the standard deviation across replicates in parentheses; the standard error (of the mean) is lower by the square root of the number of replicates. Fully-labeled experiments use ground truth labels on the "unlabeled" data. We bold the highest non-fully-labeled OOD performance numbers as well as others where the standard error is within range.

	iWildCam2020-wilds (Unlabeled extra, macro F1) In-distribution	 Out-of-distribution
ERM (-data aug)	46.7 (0.6)	30.6 (1.1)
ERM	47.0 (1.4)	**32.2** (1.2)
CORAL	40.5 (1.4)	27.9 (0.4)
DANN	48.5 (2.8)	**31.9** (1.4)
Pseudo-Label	47.3 (0.4)	30.3 (0.4)
FixMatch	46.3 (0.5)	**31.0** (1.3)
Noisy Student	47.5 (0.9)	**32.1** (0.7)
SwAV	47.3 (1.4)	29.0 (2.0)
ERM (fully-labeled)	54.6 (1.5)	44.0 (2.3)

Baseline Results

5.5 shows that most methods do not improve over standard empirical risk minimization (ERM) on iWildCam2020-wilds despite access to unlabeled data and careful hyperparameter tuning. In contrast, these methods have been shown to perform well on prior unsupervised adaptation benchmarks; we verify our implementations by showing that these methods (with the exception of CORAL) outperform ERM on the *real* → *sketch* shift in DomainNet, a standard unsupervised adaptation benchmark for object classification [89, 100].

Data augmentation improved OOD performance on iWildCam2020-wilds. However, none of the benchmarked methods improved OOD performance compared to ERM even though we had access to 4× as many unlabeled images from extra domains (distinct camera traps). Note we did not have access to any images from the target domains. This was surprising, as many of these methods were originally shown to work in semi-supervised settings. One difference could be that the labeled and unlabeled examples in iWildCam2020-wilds differ more significantly (as they originate from different camera traps) than in the original FixMatch paper [113], which used i.i.d. labeled and unlabeled data, or the Noisy Student paper [143], which used ImageNet labeled data [98] and JFT unlabeled data [56].

Fully-labeled ERM models that used ground truth labels for the "unlabeled" data were available for iWildCam2020-wilds. They significantly outperformed other methods, suggesting room for improvement in how we leverage the unlabeled data.

5.8 Conclusion

The iWildCam datasets and suite of benchmarks has provided an entrypoint for many computer vision researchers into the ecological domain. The growth of interest in the competition and the additional challenges introduced over time, as well as the inclusion of iWildCam in the WILDS benchmark, has provided visibility for ecology and environmental applications as useful and impactful testbeds for computer vision research. The competitions have captured and sometimes introduced novel challenges to the computer vision community and have generated useful insight about what works and what doesn't for camera trap AI that has enabled ecology and conservation technology practitioners to prioritize their efforts when seeking to use computer vision as part of their data processing pipelines. This emphasizes the value of careful benchmark dataset curation and design, both to the computer vision community and within the application domain that the benchmark exemplifies.

Acknowledgements The iWildCam benchmarks would not have been possible without the contribution of data from the Unites States Geological Society (USGS), the National Park Service (NPS), the Wildlife Conservation Society (WCS), and iNaturalist. In particular we would like to thank Jonathan Palmer, Erin Boydston, Justin Brown, and Scott Laurie who served as our key contacts within these organizations as we curated the datasets and metadata for the various iWildCam challenges. We also would like to thank the team at Kaggle, and particularly Maggie Demkin, for supporting the iWildCam challenges as featured Kaggle competitions, and the organizing team of the Fine-Grained Visual Categorization (FGVC) Workshop at CVPR for providing a home for the competitions at CVPR and providing visibility for these challenges in the computer vision community. We want to thank all of the numerous co-authors and co-organizers of both the iWildCam competitions and the WILDS benchmark team. Finally, we would like to acknowlege the financial support provided by Microsoft AI for Earth, Google, and Centaur Labs that made these competitions possible.

References

[1] Lila.science. `http://lila.science/`. Accessed: 2019-10-22.

[2] Centaur Labs. `https://www.centaurlabs.com`.

[3] iNaturalist. `https://www.inaturalist.org/`.

[4] Microsoft Camera Traps - GitHub. `https://github.com/microsoft/CameraTraps`.

[5] Wildlife Conservation Society Camera Traps Dataset. `http://lila.science/datasets/wcscameratraps`.

[6] Jorge A. Ahumada, Eric Fegraus, Tanya Birch, Nicole Flores, Roland Kays, Timothy G. O'Brien, Jonathan Palmer, Stephanie Schuttler, Jennifer Y. Zhao, Walter Jetz, et al. Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. *Environmental Conservation*, 47(1):1–6, 2020.

[7] Rosamund Almond, Monique Grooten, and Tom Peterson. Living Planet Report 2020-Bending the curve of biodiversity loss. *World Wildlife Fund*, 2020.

[8] Martin Arjovsky, Leon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[9] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 998–1008, 2018.

[10] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

[11] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

[12] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[13] Sara Beery, Dan Morris, and Pietro Perona. The iWildCam 2019 challenge dataset. *The Sixth Fine-Grained Visual Categorization Workshop at CVPR*, 2019.

[14] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *In the Data Mining and Artificial Intelligence for Conservation Workshop at Knowledge Discovery in Databases (KDD)*, 2019. *selected to be featured at KDD Earth Day.

[15] Sara Beery, Elijah Cole, and Arvi Gjoka. The iWildCam 2020 competition dataset. *The Seventh Fine-Grained Visual Categorization Workshop at CVPR*, 2020.

[16] Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 863–873, 2020.

[17] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context R-CNN: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2020.

[18] David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. Adamatch: A unified approach to semi-supervised learning and domain adaptation. *arXiv preprint arXiv:2106.04732*, 2021.

[19] Vighnesh Birodkar, Zhichao Lu, Siyang Li, Vivek Rathod, and Jonathan Huang. The surprising impact of mask-head architecture on novel class segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7015–7025, 2021.

[20] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[21] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[22] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[23] Kaidi Cao, Yining Chen, Junwei Lu, Nikos Arechiga, Adrien Gaidon, and Tengyu Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. *arXiv preprint arXiv:2006.15766*, 2020.

[24] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2229–2238, 2019.

[25] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9912–9924, 2020.

[26] Guobin Chen, Tony X. Han, Zhihai He, Roland Kays, and Tavis Forrester. Deep convolutional neural network based species recognition for wild animal monitoring. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 858–862. IEEE, 2014.

[27] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[28] Vincent Chen, Sen Wu, Alexander J. Ratner, Jen Weng, and Christopher Ré. Slice-based learning: A programming model for residual learning in critical data slices. *Advances in neural information processing systems*, 32, 2019.

[29] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[30] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? *arXiv preprint arXiv:2105.05837*, 2021.

[31] Anne-Sophie Crunchant, David Borchers, Hjalmar Kühl, and Alex Piel. Listening and watching: Do camera traps or acoustic sensors more efficiently detect wild chimpanzees in an open habitat? *Methods in Ecology and Evolution*, 11(4):542–552, 2020.

[32] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[33] Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, Bangyou Zheng, Shouyang Liu, Norbert Kirchgessner, Goro Ishikawa, Koichi Nagasawa, Minhajul A Badhon, Curtis Pozniak, Benoit de Solan, Andreas Hund, Scott C. Chapman, Frederic Baret, Ian Stavness, and Wei Guo. Global wheat head detection (gwhd) dataset: A large and diverse dataset of high-resolution rgb-labelled images to develop and benchmark wheat head detection methods. *Plant Phenomics*, 2020, 2020.

[34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[36] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. *arXiv preprint arXiv:2007.08558*, 2020.

[37] Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pages 1–7. IEEE, 2017.

[38] Robert M. Dorazio, J. Andrew Royle, Bo Söderström, and Anders Glimskär. Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology*, 87(4):842–854, 2006.

[39] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32, 2019.

[40] John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *Annals of Statistics*, 2021.

[41] John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. *Under review*, 2, 2019.

[42] Jane Elith and John R. Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual review of Ecology, Evolution, and Systematics*, 40:677–697, 2009.

[43] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Artificial Intelligence and Statistics (AISTATS)*, pages 201–208, 2010.

[44] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[45] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[46] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[47] Jhony-Heriberto Giraldo-Zuluaga, Augusto Salazar, Alexander Gomez, and Angélica Diaz-Pulido. Camera-trap images segmentation using multi-layer robust principal component analysis. *The Visual Computer*, pages 1–13, 2017.

[48] Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*, 2020.

[49] Google Earth Engine. USGS Landsat 8 Surface Reflectance Tier 1. `https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LC08_C01_T1_SR`.

[50] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 2017. doi: 10.1016/j.rse.2017.06.031. URL `https://doi.org/10.1016/j.rse.2017.06.031`.

[51] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

[52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[53] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[54] Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *arXiv preprint arXiv:1710.11469*, 2017.

[55] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

[56] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[57] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

[58] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133, 2020.

[59] Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[60] Benjamin Kellenberger, Diego Marcos, Sylvain Lobry, and Devis Tuia. Half a percent of labels is enough: Efficient animal detection in uav imagery using deep cnns and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):9524–9533, 2019.

[61] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[62] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[63] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013.

[64] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Association for Computational Linguistics (ACL)*, 2020.

[65] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vsion*, pages 5542–5550, 2017.

[66] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[67] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5400–5409, 2018.

[68] Hongyang Li, Daniel Quang, and Yuanfang Guan. Anchor: trans-cell type prediction of transcription factor binding sites. *Genome research*, 29(2): 281–292, 2019.

[69] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *European Conference on Computer Vision (ECCV)*, pages 624–639, 2018.

[70] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[71] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105. PMLR, 2015.

[72] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217, 2017.

[73] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[74] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9596–9606, 2019.

[75] Darryl I MacKenzie, James D Nichols, J Andrew Royle, Kenneth H Pollock, Larissa L Bailey, and James E Hines. *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Elsevier, 2017.

[76] Nicolai Meinshausen and Peter Bühlmann. Maximin effects in inhomogeneous large-scale data. *Annals of Statistics*, 43, 2015.

[77] Agnieszka Miguel, Sara Beery, Erica Flores, Loren Klemesrud, and Rana Bayrakcismith. Finding areas of motion in camera trap images. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 1334–1338. IEEE, 2016.

[78] John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning (ICML)*, 2021.

[79] Samarth Mishra, Kate Saenko, and Venkatesh Saligrama. Surprisingly simple semi-supervised domain adaptation with pretraining and consistency. *arXiv preprint arXiv:2101.12727*, 2021.

[80] Anna K. Moeller, Paul M. Lukacs, and Jon S. Horne. Three novel methods to estimate abundance of unmarked animals using remote cameras. *Ecosphere*, 9(8):e02331, 2018.

[81] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[82] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 188–197, 2019.

[83] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25): E5716–E5725, 2018.

[84] Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution*, 12(1):150–161, 2021.

[85] Jill Nugent. iNaturalist. *Science Scope*, 41(7):12–13, 2018.

[86] Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019.

[87] Omiros Pantazis, Gabriel Brostow, Kate Jones, and Oisin Mac Aodha. Focus on the positives: Self-supervised learning for biodiversity monitoring. *arXiv preprint arXiv:2108.06435*, 2021.

[88] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2021–2026, 2018.

[89] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vsion*, pages 1406–1415, 2019.

[90] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 78, 2016.

[91] Kerrie A Pipal, Jeremy J Notch, Sean A Hayes, and Peter B Adams. Estimating escapement for a low-abundance steelhead population using dual-frequency identification sonar (didson). *North American Journal of Fisheries Management*, 32(5):880–893, 2012.

[92] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[93] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[94] Christopher Ré, Feng Niu, Pallavi Gudipati, and Charles Srisuwananukorn. Overton: A data system for monitoring and improving machine-learned products. *arXiv preprint arXiv:1909.05372*, 2019.

[95] Tim Robertson, Markus Döring, Robert Guralnick, David Bloom, John Wieczorek, Kyle Braak, Javier Otegui, Laura Russell, and Peter Desmet. The gbif integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. *PloS One*, 9(8):e102623, 2014.

[96] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[97] J. Marcus Rowcliffe, Roland Kays, Chris Carbone, and Patrick A. Jansen. Clarifying assumptions behind the estimation of animal density from camera trap rates. *Journal of Wildlife Management*, 2013.

[98] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[99] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

[100] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, H. Marklund, Sara Beery, E. David, I. Stavness, Wei Guo, J. Leskovec, Kate Saenko, Tatsunori B. Hashimoto, S. Levine, Chelsea Finn, and Percy Liang. Extending the WILDS benchmark for unsupervised adaptation. In *International Conference on Learning Representations (ICLR)*, 2022.

[101] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, pages 2988–2997, 2017.

[102] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

[103] Stefan Schneider and Alex Zhuang. Counting fish and dolphins in sonar images using deep learning. *arXiv preprint arXiv:2007.12808*, 2020.

[104] Stefan Schneider, Graham W. Taylor, and Stefan Kremer. Deep learning object detection methods for ecological camera trap data. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 321–328. IEEE, 2018.

[105] Stefan Schneider, Graham W. Taylor, Stefan Linquist, and Stefan C. Kremer. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 10 (4):461–470, 2019.

[106] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, pages 621–635. Springer, 2018.

[107] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

[108] Kendrick Shen, Robbie Matthew Jones, Ananya Kumar, Sang Michael Xie, and Percy Liang. How does contrastive pre-training connect disparate domains? In *NeurIPS Workshop on Distribution Shifts*, 2021.

[109] Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8218–8226, 2019.

[110] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[111] Yu Shiu, K.J. Palmer, Marie A. Roch, Erica Fleishman, Xiaobai Liu, Eva-Marie Nosal, Tyler Helble, Danielle Cholewiak, Douglas Gillespie, and Holger Klinck. Deep neural networks for automated detection of marine mammal species. *Scientific Reports*, 10(1):1–12, 2020.

[112] Scott C. Silver, Linde E.T. Ostro, Laura K. Marsh, Leonardo Maffei, Andrew J. Noss, Marcella J. Kelly, Robert B. Wallace, Humberto Gomez, and Guido Ayala. The use of camera traps for estimating jaguar panthera onca abundance and density using capture/recapture analysis. *Oryx*, 38(2):148–154, 2004.

[113] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

[114] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

[115] Dan Stowell, Michael D. Wood, Hanna Pamuła, Yannis Stylianou, and Hervé Glotin. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. *Methods in Ecology and Evolution*, 10(3): 368–380, 2019.

[116] Jong-Chyi Su and Subhransu Maji. The semi-supervised iNaturalist-aves challenge at FGVC7 workshop. *arXiv preprint arXiv:2103.06937*, 2021.

[117] Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450, 2016.

[118] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[119] Michael A. Tabak, Mohammad S. Norouzzadeh, David W. Wolfson, Steven J. Sweeney, Kurt C. VerCauteren, Nathan P. Snow, Joseph M. Halseth, Paul A. Di Salvo, Jesse S. Lewis, Michael D. White, et al. Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590, 2019.

[120] Michael A. Tabak, Mohammad S. Norouzzadeh, David W. Wolfson, Erica J. Newton, Raoul K. Boughton, Jacob S. Ivan, Eric A. Odell, Eric S. Newkirk, Reesa Y. Conrey, Jennifer Stenglein, et al. Improving the accessibility and transferability of machine learning algorithms for identification of animals in camera trap images: Mlwic2. *Ecology and Evolution*, 10(19):10374–10383, 2020.

[121] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[122] Dogancan Temel, Jinsol Lee, and Ghassan AlRegib. Cure-or: Challenging unreal and real environments for object recognition. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 137–144. IEEE, 2018.

[123] Yao-Hung Hubert Tsai, Martin Q. Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Conditional contrastive learning: Removing undesirable information in self-supervised representations. *arXiv preprint arXiv:2106.02866*, 2021.

[124] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[125] U.S. Geological Survey. Landsat 8 Imagery. `https://www.usgs.gov/land-resources/nli/landsat/landsat-8`, 2022.

[126] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[127] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. *arXiv preprint arXiv:2103.16483*, 2021.

[128] Sindre Vatnehol, Hector Peña, and Nils Olav Handegard. A method to automatically detect fish aggregations using horizontally scanning sonar. *ICES Journal of Marine Science*, 75(5):1803–1812, 2018.

[129] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

[130] Maxime Vidal, Nathan Wolf, Beth Rosenberg, Bradley P. Harris, and Alexander Mathis. Perspectives on individual animal identification from biology and computer vision. *arXiv preprint arXiv:2103.00560*, 2021.

[131] Alexander Gomez Villa, Augusto Salazar, and Francisco Vargas. Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecological Informatics*, 41:24–32, 2017.

[132] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, , and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning (ICML)*, 2008.

[133] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[134] Alex Wang, Amapreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations (ICLR)*, 2019.

[135] Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. *arXiv*, 2021.

[136] Oliver R. Wearn and Paul Glover-Kapfer. Camera-trapping for conservation: A guide to best-practices. *WWF conservation technology series*, 1(1):2019–04, 2017.

[137] Ben G. Weinstein. A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3):533–545, 2018.

[138] Ben G. Weinstein, Sergio Marconi, Stephanie Bohlman, Alina Zare, and Ethan White. Individual tree-crown detection in rgb imagery using semi-supervised deep learning neural networks. *Remote Sensing*, 11(11):1309, 2019.

[139] Ben G. Weinstein, Lindsey Gardner, Vienna Saccomanno, Ashley Steinkraus, Andrew Ortega, Kristen Brush, Glenda Yenni, Ann E. McKellar, Rowan Converse, Christopher Lippitt, et al. A general deep learning model for bird detection in high resolution airborne imagery. *bioRxiv*, 2021.

[140] Kimberly Wilber, Walter J. Scheirer, Phil Leitner, Brian Heflin, James Zott, Daniel Reinke, David K. Delaney, and Terrance E. Boult. Animal recognition in the mojave desert: Vision tools for field biologists. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 206–213. IEEE, 2013.

[141] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: A benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[142] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

[143] Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves imagenet classification. *CoRR*, abs/1911.04252, 2019. URL `http://arxiv.org/abs/1911.04252`.

[144] Sang Michael Xie, Ananya Kumar, Robbie Jones, Fereshte Khani, Tengyu Ma, and Percy Liang. In-N-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. In *International Conference on Learning Representations (ICLR)*, 2021.

[145] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *International Conference on Computer Vision (ICCV)*, pages 1426–1435, 2019.

[146] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature Communications*, 11(1):1–11, 2020.

[147] Hayder Yousif, Jianhe Yuan, Roland Kays, and Zhihai He. Fast human-animal detection from highly cluttered camera-trap images using joint background modeling and deep learning classification. In *Circuits and Systems (ISCAS), 2017 IEEE International Symposium on*, pages 1–4. IEEE, 2017.

[148] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[149] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3801–3809, 2018.

[150] Yabin Zhang, Haojian Zhang, Bin Deng, Shuai Li, Kui Jia, and Lei Zhang. Semi-supervised models are strong unsupervised domain adaptation learners. *arXiv preprint arXiv:2106.00417*, 2021.

[151] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning (ICML)*, pages 7404–7413, 2019.

[152] Zhi Zhang, Zhihai He, Guitao Cao, and Wenming Cao. Animal detection from highly cluttered natural scenes using spatiotemporal object region proposals and patch verification. *IEEE Transactions on Multimedia*, 18(10):2079–2092, 2016.

Chapter 6

SYNTHETIC EXAMPLES IMPROVE GENERALIZATION FOR RARE CLASSES

Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 863–873, 2020.

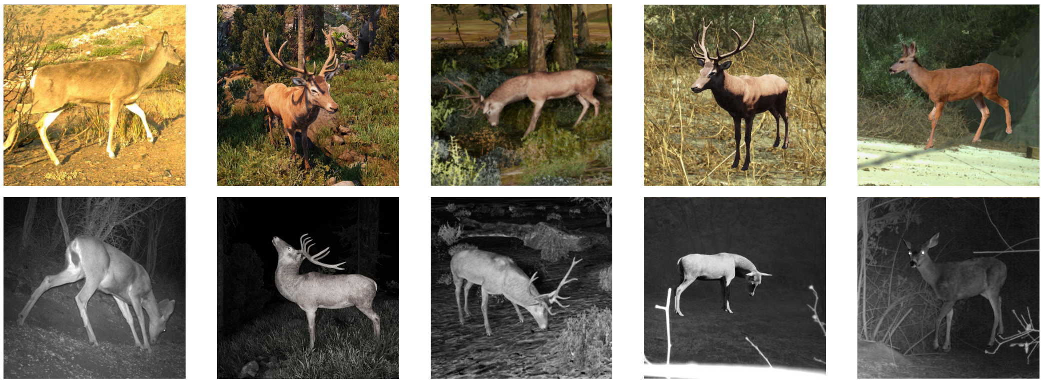
6.1 Abstract

The ability to detect and classify rare occurrences in images has important applications–for example, counting rare and endangered species when studying biodiversity, or detecting infrequent traffic scenarios that pose a danger to self-driving cars. Few-shot learning is an open problem: current computer vision systems struggle to categorize objects they have seen only rarely during training, and collecting a sufficient number of training examples of rare events is often challenging and expensive, and sometimes outright impossible. We explore in depth an approach to this problem: complementing the few available training images with ad-hoc simulated data.

Our testbed is animal species classification, which has a real-world long-tailed distribution. We present two natural world simulators, and analyze the effect of different axes of variation in simulation, such as pose, lighting, model, and simulation method, and we prescribe best practices for efficiently incorporating simulated data for real-world performance gain. Our experiments reveal that synthetic data can considerably reduce error rates for classes that are rare, that as the amount of simulated data is increased, accuracy on the target class improves, and that high variation of simulated data provides maximum performance gain.

6.2 Introduction

In recent years, computer vision researchers have made substantial progress towards automated visual recognition across a wide variety of visual domains [8, 20, 47, 51, 57, 68]. However, applications are hampered by the fact that in the real world the distribution of visual classes is long-tailed, and state-of-the-art recognition algorithms struggle to learn classes with limited data [67]. In some cases (such as recognition of rare endangered species) classifying rare occurrences correctly is



(a) Real Camera Traps **(b)** TrapCam-Unity **(c)** TrapCam-AirSim **(d)** Sim on Empty **(e)** Real on Empty

Figure 6.1: **Day and night examples for each simulation method.** We compare four different simulation methods and compare the effects of each on classification performance.

crucial. Simulated data, which is plentiful, and comes with annotation "for free," has been shown to be useful for various computer vision tasks [24, 26, 28, 32, 36, 49, 50, 53, 55, 61, 69]. However, an exploration of this approach in a long-tailed setting is still missing (see section 6.3).

As a testbed, we focus on the effect of simulated data augmentation on the real-world application of recognizing animal species in camera trap images. Camera traps are heat- or motion-activated cameras placed in the wild to monitor animal populations and behavior. The processing of camera trap images is currently limited by human review capacity; consequently, automated detection and classification of animals is a necessity for scalable biodiversity assessment. A single sighting of a rare species is of immense importance. However, training data of rare species is, by definition, scarce. This makes this domain ideal for studying methods for training detection and classification algorithms with few training examples. We utilize a technique from [8] which tests performance at camera locations both seen (cis) and unseen (trans) during training in order to explicitly study generalization (see Section 6.4 for a more detailed explanation).

We introduce two novel natural world simulators based on popular 3D game development engines for generalizable, realistic and efficient synthetic data generation. We investigate the use of simulated data as augmentation during training, and how to best combine real data for common classes with simulated data for rare classes to achieve optimal performance across the class set at test time. We consider four different data simulation methods (see Fig. 6.1) and compare the effects of each on classification performance. Finally, we analyze the effect of both increasing the number of simulated images and controlling for axes of variation to provide best practices for leveraging simulated data for real-world performance gain on rare classes.

6.3 Related work

Visual categorization datasets

Large and well-annotated public datasets allow scientists to train, analyze, and compare the performance of different methods, and have provided large performance improvements over traditional vision approaches [31, 34, 63]. The most popular datasets used for this purpose are ImageNet, COCO, PascalVOC, and OpenImages, all of which are human-curated from images scraped from the web [17, 21, 39, 43]. These datasets cover a wide set of classes across both the manufactured and natural world, and are usually designed to provide "enough" data per class to avoid the low-data regime. More recently researchers have proposed datasets that focus specifically on long-tailed distributions [8, 41, 68]. The Caltech Camera Traps dataset [8] introduced the challenge of learning from limited locations, and generalizing to new locations.

Handling imbalanced datasets

Imbalanced datasets lead to bias in algorithm performance toward well-represented classes [12]. Algorithmic solutions often use a non-uniform cost per misclassification via weighted loss [19, 29, 30]. One example, focal loss, was recently proposed to deal with the large foreground/background imbalance in detection [44].

Data solutions employ data augmentation, either by 1) over-sampling the minority classes, 2) under-sampling the majority classes, or 3) generating new examples for the minority classes. When using mini-batch gradient descent, oversampling the minority classes is similar to weighted loss. Under-sampling the majority classes is non-ideal, as this reduces information about common classes. Our paper falls into the third category: generating new training data for rare classes. Data augmentation via pre-processing, using affine and photometric transformations, is a well-established tool for improving generalization [33, 40]. Data generation and simulation have begun to be explored as data augmentation methods, see section 6.3.

Algorithmic and data solutions for imbalanced data are complementary, algorithmic advances can be used in conjunction with augmented training data.

Low-shot learning

Low-shot learning attempts to learn categories from few examples [42]. Wang and Herbert [70] do low-shot classification by regressing from small-dataset classifiers to large-dataset classifiers. Hariharan and Girshick [27] look specifically at ImageNet, using classes that are unbalanced, some with large amounts of training data, and some with little training data. Metric learning learns a representation space where distance corresponds to similarity, and uses this as a basis for low-shot solutions [14]. We consider the low-shot regime with regard to *real* data for our rare target class, but investigate the use of added synthetic data based on a human-generated articulated model of the unseen class during training instead of additional class-specific attribute labels at training and test time. This takes us outside of the traditional low-shot framework into the realm of domain transfer from simulated to real data.

Data augmentation via style transfer, generation, and simulation

Image generation via generative adversarial networks (GANs) and recurrent neural networks (RNNs), as well as style transfer and image-to-image translation have all been considered as sources for data augmentation [11, 25, 35, 45, 52, 66, 73]. These techniques require large amounts of data to generate realistic images making them non-ideal solutions for low-data regimes. Though conditional generation allows for class-specific output, the results can be difficult to interpret or control.

Graphics engines such an Unreal [1, 71] and Unity [2] leverage the expertise of human artists and complex physics models to generate photorealistic simulated images, which can be used for data augmentation. Because ground truth is known at generation, simulated data has proved particularly useful for tasks requiring detailed and expensive annotation, such as keypoints, semantic segmentations, or depth information [24, 28, 32, 49, 50, 53, 55, 61, 69]. Varol et al. [69] use synthetically-generated humans placed on top of real image backgrounds as pretraining for human pose estimation, and suggest fine-tuning a synthetically-trained model on real data. [61] uses a combination of unlabeled real data and labeled simulated data of the same class to improve real-world performance on an eye-tracking task by using GANs [24]. This method requires a large number of unlabled examples from the target class. [32, 50, 53] find that simulated data improves detection performance, and the degree of realism and variability of simulation affects the amount of improvement. They consider only small sets of non-deformable man-made objects. Richter et al. [55] showed that a segmentation model for city scenes trained with a subset of their real dataset and a large synthetic set outperforms a model trained with the full real dataset. [49] proposes a dataset and benchmark for evaluating models for unsupervised domain transfer from synthetic to real data with all-simulated training data, as opposed to simulated data only for rare classes. While this literature is encouraging, a number of questions are left unexplored. The first is a careful analysis of when simulated data is useful and, in particular, if it is useful in generalizing to new scenarios. Second, whether simulated data can be useful in highly complex and relatively unpredictable scenes such as natural scenes, as opposed to indoors and urban scenes. Third, whether it is just the synthetic objects or also the synthetic environments that contribute to learning.

Simulated datasets

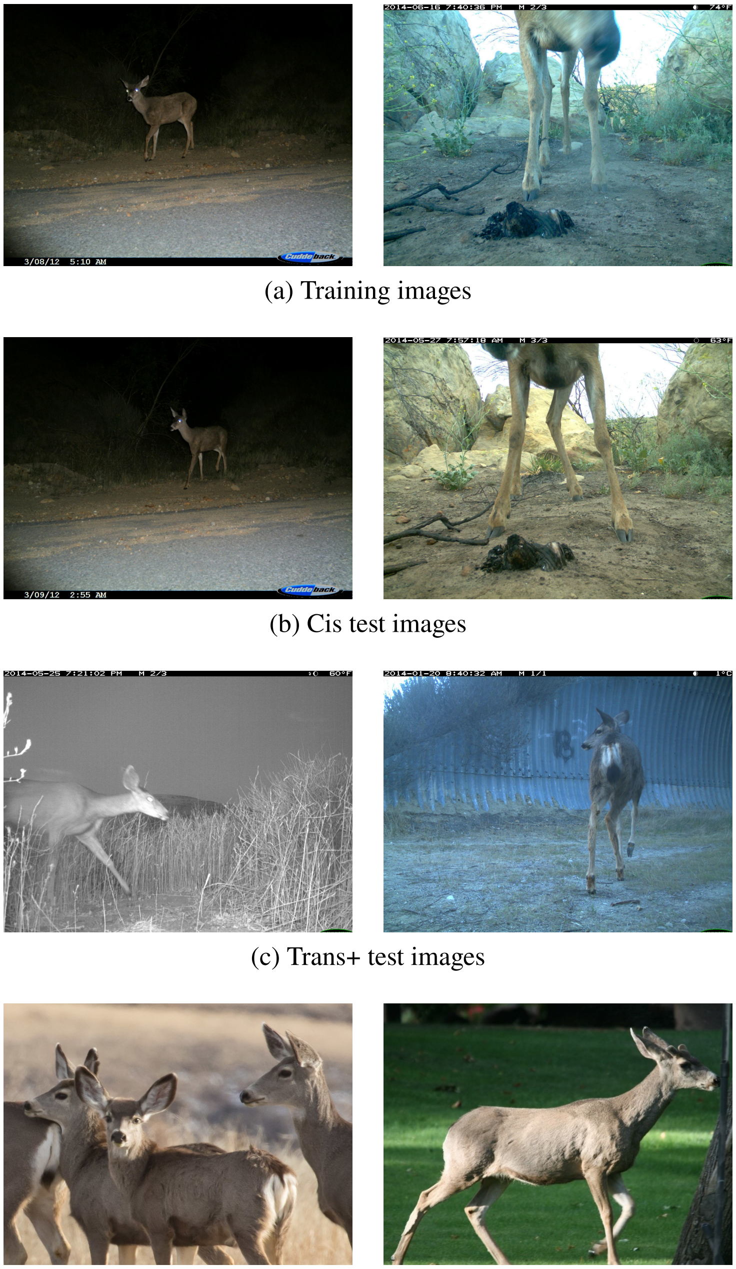
Previous efforts on synthetic dataset generation focus on non-deformable man-made objects and indoor scenes [32, 38, 53, 58, 62, 72], human pose/actions [16, 69], or urban scenes [18, 22, 26, 36, 55, 56].

Bondi et al. [10] previously released the AirSim-w data simulator within the domain of wildlife conservation, focused on creating aerial infrared imagery. The resolution and quality of the assets is designed to replicate data from 100 meters in the air, but is not realistic close-up. We contribute the first image data generators specifically for the natural world with the ability to recreate natural environments and generate near-photorealistic images of animals within the scene, including real-world nuisance factors such as challenging pose, lighting, and occlusion.

6.4 Data and simulation

Real data

Our real-world training and test data comes from the Caltech Camera Traps (CCT) dataset [8]. CCT contains $243,187$ images from 140 camera trap locations covering 30 classes of animals, curated from data provided by the United States Geological Survey and the National Park Service. We follow the CCT-20 data split laid out in [8], which was explicitly designed for in-depth generalization analysis. The split uses a subset of $57,868$ images from 20 camera locations covering 15 classes in CCT to simultaneously investigate performance on locations seen during training and generalization performance to new locations. Bounding-box annotations are provided for all images in CCT-20, whereas the rest of CCT has only class labels. In the CCT-20 data split, *cis-locations* are defined as locations seen during training and *trans-locations* as locations not seen during training (see Fig. 6.3). Nine locations are used for trans-test data, one location for trans-validation data, and data from the remaining 10 locations is split between odd and even days, with odd days as cis-test data and even days as training and cis-validation data (a 95% of data from even days for training, 5% for testing).



(a) Training images

(b) Cis test images

(c) Trans+ test images

(d) iNaturalist images

Figure 6.2: **Cis vs. Trans:** The cis-test data can be very similar to the training data: animals tend to behave similarly at a single location even across different days, so the images collected of each species are easy to memorize intra-location. The trans data has biases towards specific angles and lighting conditions that are different from those in the cis locations, and as such is very hard to learn from the training data. iNaturalist data represents a domain shift to human-curated images.

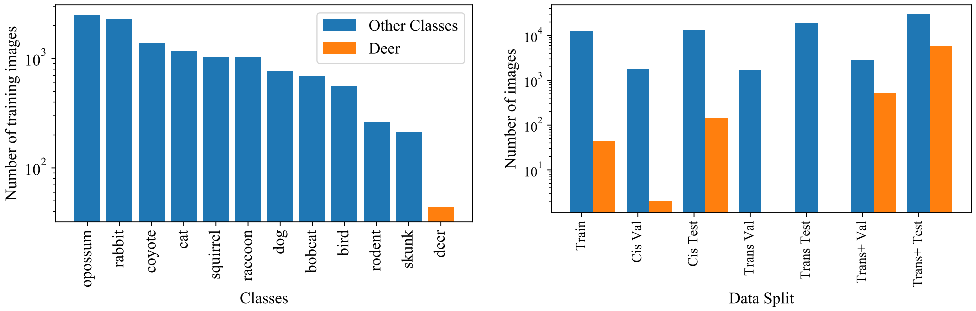


Figure 6.3: **(Top) Number of training examples for each class.** Deer are rare in the training locations from the CCT-20 data split. We focus on deer as a test species in order to investigate whether we can improve performance on a "rare" class. Since deer are not rare at other camera locations within the CCT dataset, we have enough test data to thoroughly evaluate the effect. **(Bottom) Number of examples for each data split, for deer and other classes.** In the CCT-20 data split there were no trans examples of deer. We added annotations to the trans val and test sets for an additional 16K images across 65 new locations from CCT, including 6K examples of deer. We call these augmented sets *trans+*.

To study the effect of simulated data on rare species, we focus on deer, which are rare in CCT-20, with only 44 deer examples out of the $13,553$ images in the training set (see Fig. 6.3). To focus on the performance of a single rare class, we remove the other two rare classes in CCT-20: badgers and foxes. We note that there are no deer images in the established CCT-20 trans sets. In reality, deer are far from uncommon: unlike a truly rare species, there exist sufficient images of deer in the CCT dataset outside of the CCT-20 locations to rigorously evaluate performance. To facilitate deeper investigation of generalization we have collected bounding-box annotations for an additional 16K images from CCT across 65 new locations, which we add to the trans-validation and trans-test sets to cover a wider variety of locations and classes (including deer). We call this augmented trans set *trans+* (see Fig. 6.3) and will release the annotations at publication. To further analyze generalization, we also test on data containing deer from the iNaturalist 2017 dataset [68], which represents a domain shift to human-captured and human-selected photographs. We consider *Odocoileus hemionus* (mule deer) and *Odocoileus* virginianus (white-tailed deer) images from iNaturalist, the two species of deer seen in the CCT data. In Supplementary Material we show results on an additional class, wolf.

Synthetic data

To assess generality we leverage multiple collections of woodland and animal models to create two simulation environments, which we call TrapCam-Unity and TrapCam-AirSim. Both simulation environments and source code to generate images will be provided publicly, along with the data generated for this paper. To synthesize daytime images we varied the orientation of the simulated sun in both azimuth and elevation. To create images taken at night we used a spotlight attached to the simulated camera to simulate a white-light or IR flash and qualitatively match the low color saturation of the nighttime images. To simulate animals' eyeshine (a result of the reflection of camera flash from the back of the eye), we placed small reflective balls on top of the eyes of model animals.

TrapCam-AirSim. We create a modular natural environment within Microsoft AirSim [60] that can be randomly populated with flora and fauna. The distribution and types of trees, bushes, rocks, and logs can be varied and randomly seeded to create a diverse set of landscapes, from an open plain to a dense forest. We used various off-the-shelf components such as an animal pack from Epic Studios [4] (Animals Vol 01: Forest Animals by GiM [5]), background terrain also from Unreal Marketplace [1], vegetation from SpeedTree [7], and rocks/obstructions from Megascans [6]. The actual area of the environment is small, at 50 meters, but the modularity allows many possible scenes to be built.

TrapCam-Unity. Unity 3D game development engine is a popular game development tool that offers realistic graphics, real time performance and abundant 3D assets. We take advantage of the "Book of The Dead" environment [3], a near-photorealistic, open-source forest environment published by Unity to demonstrate its high definition rendering pipeline. This off-the-shelf environment is large and rich in details, it has a diversity of subregions with significantly different statistics. We change the lighting and move throughout this large, static environment to collect data with various background scenes. We make use of 17 animated deer models from five off-the-shelf model sets, purchased from Unity Asset Store and originally developed for game development, including the GiM models used in TrapCam-AirSim. A single gaming PC (Core i7 5820K, 16GB RAM, GTX 1080Ti) generates over 300,000 full-HD images with pixel-level instance annotation per day and the throughput linearly scales to additional machines.

Simulated animals on empty images. Similar to the data generated in [69], we generate synthetic images of deer by rendering deer on top of real camera trap images containing no animals, which we call *Sim on Empty* (see Fig. 6.1). We first generate animal foreground images by randomizing the location, orientation in azimuth, pose and illumination of the deer, then paste the foreground images on top of the real empty images. A limitation is that the deer are not in realistic relationships or occlusion scenarios with the environment around them. We also note that the empty images used to construct this data come from both cis and trans locations, so Sim on Empty contains information about test-set backgrounds unavailable in the purely simulated sets. This choice is based on current camera trap literature, which first detects the presence of any animal, and then determines animal species [8, 47]. After the initial animal detection step, the empty images are known and can be utilized.

Segmented animals on empty images. We manually segment the 44 examples of deer from the training set and paste them at random on top of real empty camera trap images, which we call *Real on Empty* (see Fig. 6.1). This allows us to analyze whether the generalization challenge is related to memorizing the training deer+background or memorizing the training deer regardless of background. Similar to the Sim on Empty set, these images do not have realistic foreground/background relationships and the empty images come from both cis and trans locations.

6.5 Experiments

Beery, et al.[8] showed that detecting and localizing the presence of an “animal” (where all animals are grouped into a single class) both generalizes well to new locations and improves classification performance. We focus on classification of cropped ground-truth bounding boxes as opposed to training multi-class detectors in order to disambiguate classification and detection errors. We specifically investigate how added synthetic training data for rare classes effects model performance on both rare and common classes.

We find that the Inception-Resnet-V2 architecture [64] works best for the cropped-box classification task, based on performance comparison across architectures (see Supplementary Material). Most classification systems are pretrained on Imagenet, which contains animal classes. To ensure that our “rare” class is truly something the model is unfamiliar with, as opposed to something seen in pretraining, we pretrain our classifiers on *no-animal ImageNet*, a dataset we define by removing the “animal” subtree (all classes under synset node n00015388) from ImageNet. We use an initial learning rate of 0.0045, RMSprop with a momentum of 0.9 [65], and a square

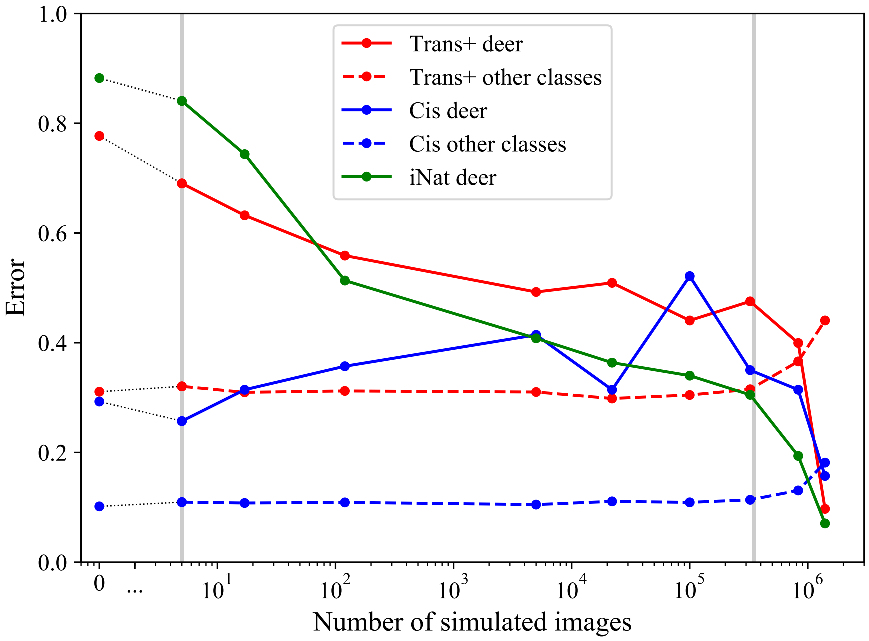
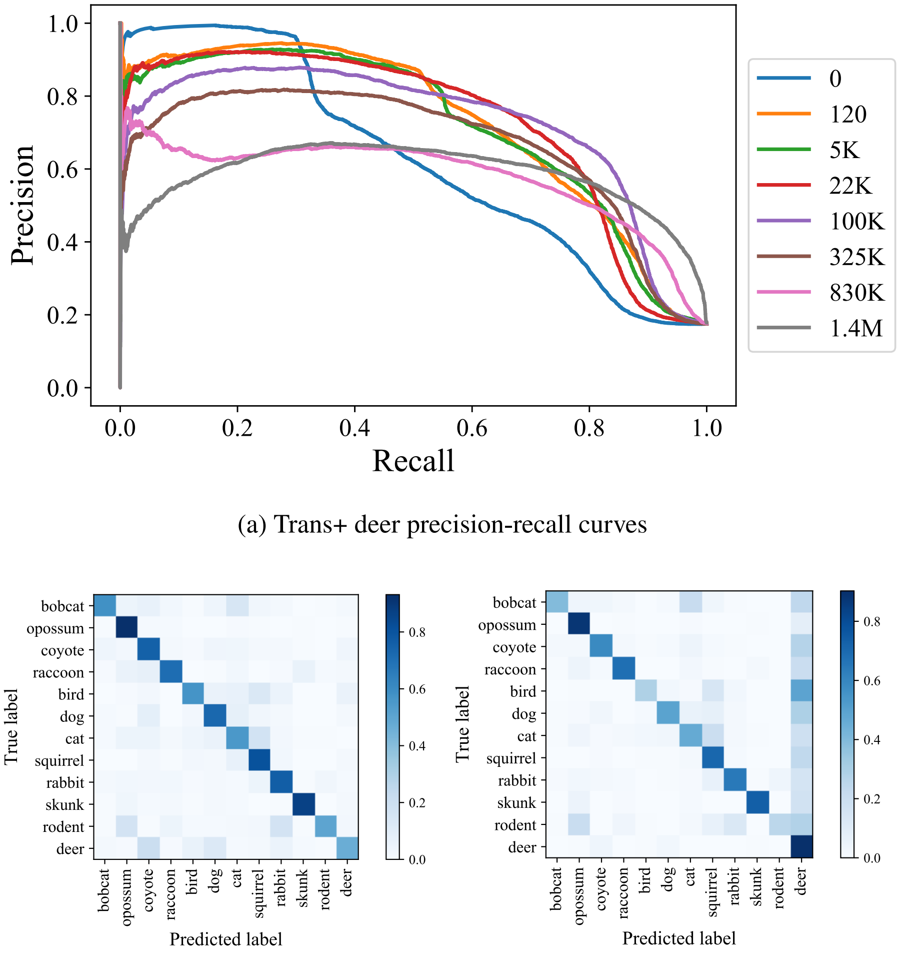


Figure 6.4: **Error as a function of number of simulated images seen during training. We divide this plot into three regions.** The leftmost region is the baseline performance with no simulated data, shown at x=0 (Note x-axis is in log scale). In the middle region, additional simulated training data increases performance on the rare class and does not harm the performance of the remaining classes (trend lines are visualized). The rightmost region, where many simulated images are added to the training set, results in a biased classifier, hurting the performance of the other classes (see Fig. 6.5 (b-c) for details). We compare the class error for "deer" and "other classes" in both the "cis" and "trans+" testing regimes. Lines marked "deer" use only the deer test images for the error computation. Lines marked "other classes" use all the images in the other classes (excluding deer) for the error computation. Error is defined as the number of incorrectly identified images divided by the number of images.

input resolution of 299. We employ random cropping (containing at least 65% of the region), horizontal flipping, color distortion, and blur as data augmentation. Model selection is performed using early stopping based on trans+ validation set performance [9].

Effect of increase in simulated data

We explore the trade-off in performance when increasing the number of simulated images, from 5 to 1.4 million, spanning 5 log units (see Fig. 6.4). Very little simulated data is needed to see a trans+ performance boost: with as few as 5 simulated images we see a 10% decrease in per-class error on trans+ deer, with



(a) Trans+ deer precision-recall curves

(b) Confusion matrix: 100K

(c) Confusion matrix: 1.4M

Figure 6.5: **(a) Trans+ PR curves for the deer class:** Note the development of a biased classifier as we add simulated training data. The baseline model (in blue) has high precision but suffers low recall. The model trained with 1.4M simulated images (in grey) has higher recall, but suffers a loss in precision. **(b-c) Evidence of a biased classifier:** Compare the deer column in the confusion matrices, the model trained with 1.4M simulated images predicts more test images as deer.

$<$ 0.5% increase in average per-class error on the other trans+ classes. As we increase the number of simulated images, trans+ performance improves: with 100K simulated images we see a 39% decrease in trans+ deer error, with $<$ 0.5% increase in error for the other trans classes. There exists some threshold ($>$ 325K) where, if passed, an increase in simulated data noticeably biases the classifier towards the deer class (see Fig. 6.5): with 1.4 million simulated images, our trans+ deer error decreases by 88%, but it comes at the cost of a 13% increase in average per-class error across the other classes. At this point there is an overwhelming class prior

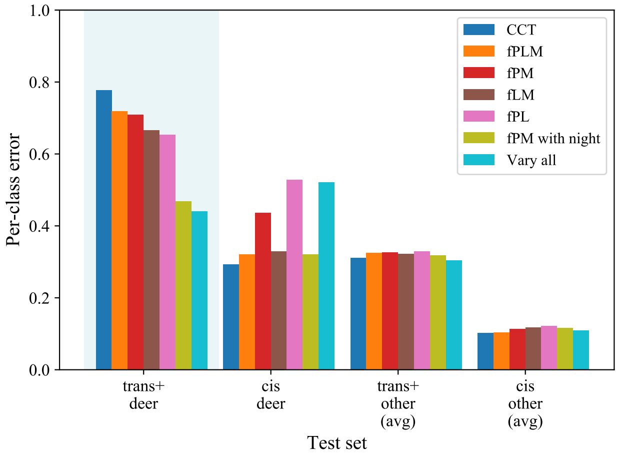


Figure 6.6: **E**rror as a function of variability of simulated images seen during training: 100K simulated deer images. Error is calculated as in Fig. 6.4, and all data is from TrapCam-Unity. Trans+ deer performance is highlighted. In the legend "CCT" means the model was trained only on the CCT-20 training set with no added simulated data. "P" means "pose," "L" means "lighting," and "M" means "model," while the prefix "f" for "fixed" denotes which of these variables were controlled for a particular experiment. For example "fPM" means the pose and the animal model were held fixed, while the lighting was allowed to vary. The variability of simulated data is extremely important, and that while all axes of variability matter, simulating nighttime images has the largest effect.

towards deer: the next-largest class at training time would be opossums with $2,514$ images, 3 orders of magnitude less.

Unsurprisingly, cis deer performance decreases with added simulated data. Although the images were taken on different days (train from even days, cis-test from odd days) the animals captured were to some extent creatures of habit. Thus, training and test images can be nearly identical from within the same locations (see Fig. 6.2). Almost all cis test deer images have at least one visually similar training image. As simulated data is added at training time, the model is forced to learn a more complex, varied representation of deer. As a result, we see cis deer performance decrease. To quantify robustness, we ran the 100K experiment three times. We found that trans+ deer error had a standard deviation of 2% and cis deer error had a standard deviation of 4%, whereas the average error across other classes had a standard deviation of 0.2% for both cis and trans.

We also investigate performance on deer images from iNaturalist [68], which are individually collected by humans and are usually relatively centered and well-focused (and therefore easier to classify) but represent a domain shift (see Fig. 6.2). Adding

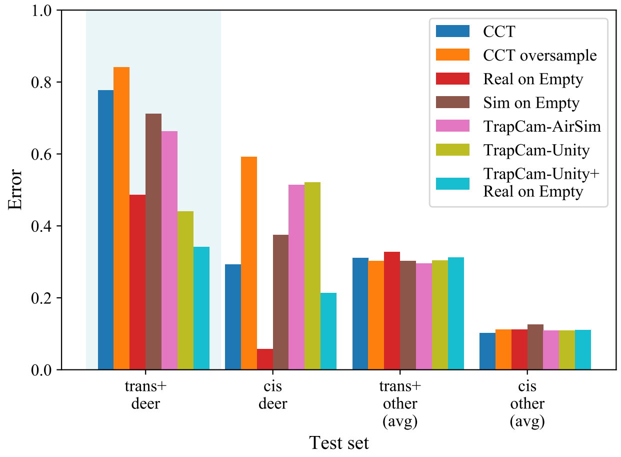


Figure 6.7: **E**rror as a function of simulated data generation method: 100K simulated deer images. Per-class error is calculated as in Fig. 6.4. Trans+ deer performance is highlighted. Oversampling decreases performance, and there is a large boost in performance from incorporating real segmented animals on different backgrounds (Real on Empty). TrapCam-Unity with everything allowed to vary (model, lighting, pose, including nighttime simulation) gives us slightly better trans+ performance, without requiring additional segmentation annotations. Combining Real on Empty with TrapCam-Unity (50K of each) gives us the best trans+ deer performance.

simulated data improves performance on the iNaturalist deer images (see Fig. 6.4), demonstrating the robustness and generality of the representation learned.

Effect of variation in simulation

In order to understand which aspects of the simulated data are most beneficial, we consider three dimensions of variation during simulation: pose, lighting, and animal model. Using the TrapCam-Unity simulator, we generate 100K daytime simulated images for each of these experiments. As a control, we create a set of data where the pose, lighting, and animal model are all fixed. We then create sets with varied pose, varied lighting, and varied animal model, each with the other variables held fixed. An additional set of data is generated varying all of the above. Unsurprisingly, widest variation results in the best trans+ deer performance. The individual axes of variation do have an effect of performance, and some are more "valuable" than others (see Fig. 6.6). There are many more dimensions of variation that could be explored, such as simulated motion blur or variation in camera perspective. For CCT data, we find adding simulated nighttime images has the largest effect on performance. We have determined that for deer 49% of training images, 53% of cis test images, and 56% of trans+ test images were captured at night, using either IR or white flash.

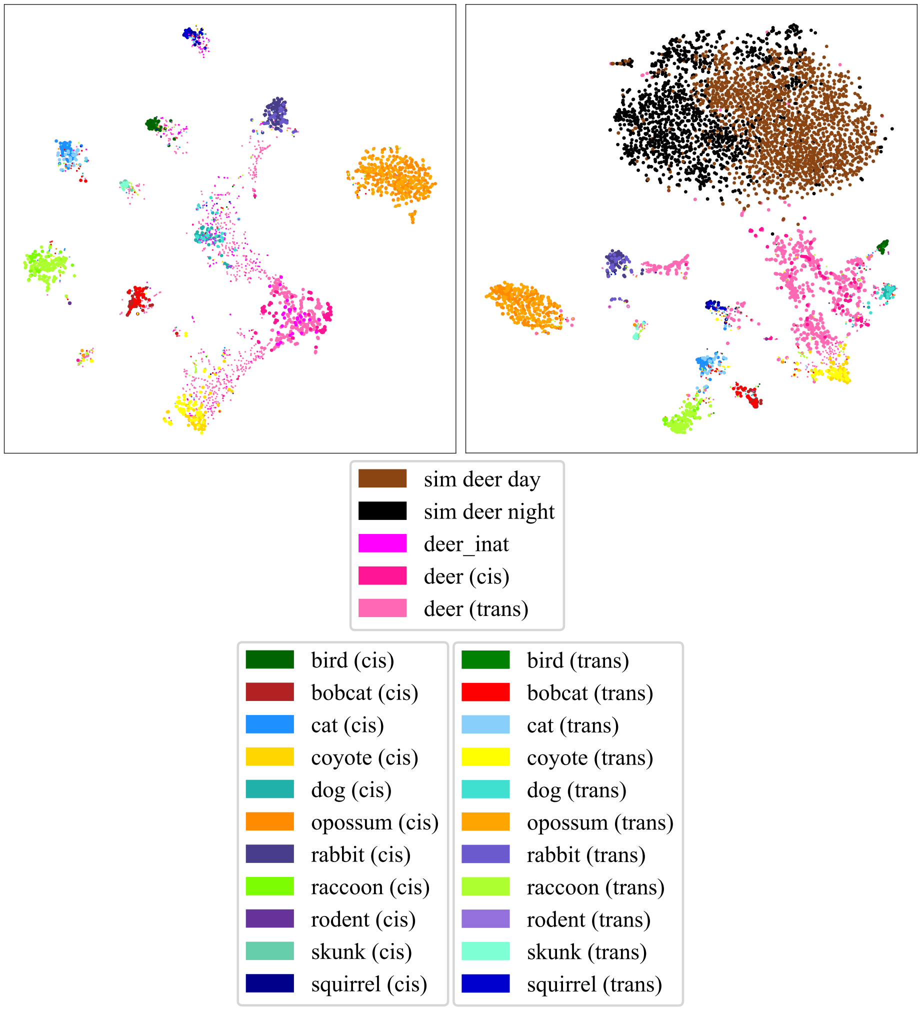


Figure 6.8: **Visualization of network activations (Left is no simulated deer, right is with 1.4M simulated deer): more deer are classified correctly as we add synthetic data, despite the synthetic data being clustered separately.** The pink points are real deer, the brown are simulated day images and the grey are simulated night images. Large markers are points that are classified correctly, while small markers are points classified incorrectly. The plots were generated by running 200-dimensional PCA over the activations at the last pre-logit layer of the network when running inference on the test sets, and then running 2-dimensional tSNE over the resulting PCA embedding.

Simulating only daytime images injects a prior towards deer being seen during the day. By training on half day and half night images we match the day/night prior for deer in the data. Not all species occur equally during the day or night, some are strictly nocturnal. Our results suggest that a good strategy is to determine the appropriate ratio of day to night images using your training set and match that ratio when adding simulated data.

Comparing simulated data generation methods

We compare performance gain from 4 methods of data synthesis, using 100K added deer images for each (see Fig. 6.7. The animal model is controlled (each simulated set uses the same GiM deer model for these experiments) for fair comparison of the efficacy of each generation method. As an additional control, we consider oversampling the rare class. This creates the same sampling prior towards deer without introducing any new information. Oversampling performs worse than just training on the unbalanced training set by causing the model to overfit the deer class to the training images. By manually segmenting out the deer in the 44 training images and randomly pasting them onto empty backgrounds we see a large improvement in performance. Cis error goes down to 6% with this method of data augmentation, which makes sense in the view of the strong similarities between the training and cis-test data (see Fig. 6.2).

Real on Empty and Sim on Empty are able to approximate both “day” and “night” imagery, a deer pasted onto a nighttime empty image is actually a reasonable approximation of an animal illuminated by a flash at night (see Fig. 6.1). They also have the additional benefit of using backgrounds from both cis and trans sets, giving them trans information not provided by the simulated datasets. TrapCam-Unity with all variability enabled is our best-performing model without requiring additional segmentation annotations. If segmentation information is available, Real on Empty combined with TrapCam-Unity (50K of each) improves both cis and trans deer performance: trans deer error decreases to 36% (a 54% decrease compared to CCT only), with $<$ 2% increase in error on trans other classes.

Visualizing the representation of data

In order to visualize how the network represents simulated data vs. real data, we use PCA and tSNE [46] to cluster the activations of the final pre-logit layer of the network. These visualizations can be seen in Fig. 6.8. Interestingly, the model learns “deer” bimodally: simulated deer are clustered almost entirely separately from real deer, with a few datapoints of each ending up in the opposite cluster. Even though those clusters overlap only slightly, the network is surprisingly able to classify more deer images correctly.

6.6 Conclusions and future work

We present two fast, realistic natural world data simulators based on popular 3D game development engines. Our simulators have 3 major advantages. **First**, they are generalizable. Thanks to the abundant 3D assets available online in the game development community, integrating a new species in a new environment from off the shelf assets is simple and fast. **Second**, not only are the graphics near-photorealistic, the pipeline also generates animals with realistic pose, animation, and interactions with the environment. **Third**, data generation is efficient. A single gaming PC generates over 300,000 full-HD images with pixel-level instance annotation per day and the throughput linearly scales to additional machines.

We explore using the simulated data to augment rare classes during training. Towards this goal, we compare multiple sources of natural-world data simulation, explicitly measure generalization via the cis-vs-trans paradigm, examine trade-offs in performance as the number of simulated images seen during training is increased, and analyze the effect of controlling for different axes of variation and data generation methods.

From our experiments we draw three main lessons. First: using synthetic data can considerably reduce error rates for classes that are rare, and with segmentation annotations we can reduce error rates even further by additionally randomly pasting segmented images of rare classes on empty background images. Second: as the amount of simulated data is increased, accuracy on the target class improves. However, with 1000x more simulated data than the common classes, we see negative effects on the performance of other classes due to the high class imbalance. Third: the variation of simulated data generated is very important, and maximum variation provides maximum performance gain.

While an increase in simulated data corresponds to an increase in target class performance, the representation of simulated data overlaps only rarely with real data (see Fig. 6.8). It remains to be studied whether embedding techniques [59], domain adaptation techniques [23, 74], or style transfer [24, 61] could be used to encourage a higher overlap in representation between the synthetic and real data, and if that overlap would lead to an increase in categorization accuracy. Additionally, the bias induced by adding large amounts of simulated data could be addressed with algorithmic solutions such as those in [15, 19, 29, 30]. We have not discussed the drawbacks related to model training with large quantities of synthetic data (epoch time, data storage, etc.). In future, we will explore merging the simulator and classifier so that highly variable synthetic data could be requested "online" without storing raw frames.

Simulation is a fast, interpretable, and controllable method of data generation that is easy to use and easy to adapt to new classes. This allows for an integrated and evolving training pipeline with new classes of interest: simulated data can be generated iteratively based on needs or gaps in performance. Our analysis suggests a general methodology when using simulated data to improve rare-class performance: 1) generate small, variable sets of simulated data (even small sets can drive improvement), 2) add these sets to training and analyze performance to determine ideal ratios and dimensions of variation, 3) take advantage of ease and speed of generation to create an abundance of data based on this ideal distribution, and determine an operating point of number of added simulated images to optimize performance between rare target class and other classes based on the project goal.

Further, the performance gains we have demonstrated, along with the data generation tools we contribute to the community, will allow biodiversity researchers focused endangered species to improve classification performance on their target species. Adding each new species to the simulation tools currently requires the assistance of a graphics artist. However, automated 3D modeling techniques, such as those proposed in [13, 37, 48, 54], might eventually become an inexpensive and practical source of data to improve few-shot learning.

The improvement we have found in rare-class categorization is encouraging, and the release of our data generation tools and the data we have generated will provide a good starting point for other researchers studying imbalanced data, simulated data augmentation, or natural-world domains.

6.7 Acknowledgements

We would like to thank the USGS and NPS for providing data. This work was supported by NSFGRFP Grant No. 1745301, the views are those of the authors and do not necessarily reflect the views of the NSF. Compute provided by Microsoft AI for Earth and AWS.

References

[1] Unreal game engine. https://www.unrealengine.com/en-US/what-is-unreal-engine-4. Accessed: 2019-02-05.

[2] Unity game engine. https://unity3d.com/. Accessed: 2019-02-05.

[3] Unity book of the dead. https://unity3d.com/book-of-the-dead. Accessed: 2019-03-21.

[4] Epic studios. http://epicstudios.com/. Accessed: 2019-03-21.

[5] Forest animals by GiM. https://www.unrealengine.com/marketplace/en-US/animals-vol-01-forest-animals. Accessed: 2019-03-21.

[6] Quixel megascans library. https://quixel.com/megascans. Accessed: 2019-03-21.

[7] Speedtree. https://store.speedtree.com/. Accessed: 2019-03-21.

[8] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[9] Yoshua Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade*, pages 437–478. Springer, 2012.

[10] Elizabeth Bondi, Debadeepta Dey, Ashish Kapoor, Jim Piavis, Shital Shah, Fei Fang, Bistra Dilkina, Robert Hannaford, Arvind Iyer, Lucas Joppa, et al. Airsim-w: A simulation environment for wildlife conservation with uavs. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, page 40. ACM, 2018.

[11] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3722–3731, 2017.

[12] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

[13] Thomas J. Cashman and Andrew W. Fitzgibbon. What shape are dolphins? Building 3D morphable models from 2D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):232–244, 2013.

[14] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1153–1162, 2016.

[15] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[16] César Roberto de Souza12, Adrien Gaidon, Yohann Cabon, and Antonio Manuel López. Procedural generation of videos to train deep action recognition networks. 2017.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[18] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[19] Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[20] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[21] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[22] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2016.

[23] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189. PMLR, 2015.

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[25] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

[26] Sanghui Han, Alex Fafard, John Kerekes, Michael Gartley, Emmett Ientilucci, Andreas Savakis, Charles Law, Jason Parhan, Matt Turek, Keith Fieldhouse, et al. Efficient generation of image chips for training deep learning algorithms. In *Automatic Target Recognition XXVII*, volume 10202, page 1020203. International Society for Optics and Photonics, 2017.

[27] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy*, 2017.

[28] Hironori Hattori, Vishnu Naresh Boddeti, Kris Kitani, and Takeo Kanade. Learning scene-specific pedestrian detectors without real data. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3819–3827. IEEE, 2015.

[29] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9):1263–1284, 2008.

[30] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328. IEEE, 2008.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[32] Stefan Hinterstoisser, Olivier Pauly, Hauke Heibel, Marek Martina, and Martin Bokeloh. An annotation saved is an annotation earned: Using fully synthetic training for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vsion workshops*, 2019.

[33] Andrew G. Howard. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*, 2013.

[34] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*, 2017.

[35] Daniel Jiwoong Im, Chris Dongjoo Kim, Hui Jiang, and Roland Memisevic. Generating images with recurrent adversarial networks. *arXiv preprint arXiv:1602.05110*, 2016.

[36] Shunping Ji, Yanyun Shen, Meng Lu, and Yongjun Zhang. Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples. *Remote Sensing*, 11(11):1343, 2019.

[37] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[38] Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: an interactive 3d environment for visual AI. *CoRR*, abs/1712.05474, 2017. URL `http://arxiv.org/abs/1712.05474`.

[39] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

[40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[41] Neeraj Kumar, Peter N Belhumeur, Arijit Biswas, David W Jacobs, W John Kress, Ida C Lopez, and João VB Soares. Leafsnap: A computer vision system for automatic plant species identification. In *European conference on computer vision*, pages 502–516. Springer, 2012.

[42] Fei-Fei Li, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (4):594–611, 2006.

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[44] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[45] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4990–4998, 2017.

[46] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[47] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25): E5716–E5725, 2018.

[48] Frederik Pahde, Mihai Puscas, Jannik Wolff, Tassilo Klein, Nicu Sebe, and Moin Nabi. Low-shot learning from imaginary 3d model. *arXiv preprint arXiv:1901.01868*, 2019.

[49] Xingchao Peng, Ben Usman, Kuniaki Saito, Neela Kaushik, Judy Hoffman, and Kate Saenko. Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. *arXiv preprint arXiv:1806.09755*, 2018.

[50] Bojan Pepik, Rodrigo Benenson, Tobias Ritschel, and Bernt Schiele. What is holding back convnets for detection? In *German Conference on Pattern Recognition*, pages 517–528. Springer, 2015.

[51] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, page 1, 2018.

[52] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[53] Param S. Rajpura, Hristo Bojinov, and Ravi S. Hegde. Object detection using deep cnns trained on synthetic images. *arXiv preprint arXiv:1706.06782*, 2017.

[54] Bernhard Reinert, Tobias Ritschel, and Hans-Peter Seidel. Animated 3d creatures from single-view video by skeletal sketching. In *Graphics Interface*, pages 133–141, 2016.

[55] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. *Lecture Notes in Computer Science*, page 102–118, 2016. ISSN 1611-3349. doi: 10.1007/978-3-319-46475-6_7. URL `http://dx.doi.org/10.1007/978-3-319-46475-6_7`.

[56] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3234–3243, 2016.

[57] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[58] Manolis Savva, Angel X. Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. MINOS: Multimodal indoor simulator for navigation in complex environments. *arXiv:1712.03931*, 2017.

[59] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[60] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and Service Robotics*, pages 621–635. Springer, 2018.

[61] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[62] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[63] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[64] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[65] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2):26–31, 2012.

[66] Toan Tran, Trung Pham, Gustavo Carneiro, Lyle Palmer, and Ian Reid. A bayesian data augmentation approach for learning deep models. In *Advances in Neural Information Processing Systems*, pages 2797–2806, 2017.

[67] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

[68] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[69] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[70] Yu-Xiong Wang and Martial Hebert. Learning to learn: Model regression networks for easy small sample learning. In *European Conference on Computer Vision*, pages 616–634. Springer, 2016.

[71] Yi Zhang Siyuan Qiao Zihao Xiao Tae Soo Kim Yizhou Wang Alan Yuille Weichao Qiu, Fangwei Zhong. Unrealcv: Virtual worlds for computer vision. *ACM Multimedia Open Source Software Competition*, 2017.

[72] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*, 2018.

[73] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[74] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018.

Chapter 7

LONG TERM TEMPORAL CONTEXT FOR PER CAMERA OBJECT DETECTION

Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context R-CNN: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2020.

7.1 Abstract

In static monitoring cameras, useful contextual information can stretch far beyond the few seconds typical video understanding models might see: subjects may exhibit similar behavior over multiple days, and background objects remain static. Due to power and storage constraints, sampling frequencies are low, often no faster than one frame per second, and sometimes are irregular due to the use of a motion trigger. In order to perform well in this setting, models must be robust to irregular sampling rates. In this paper we propose a method that leverages temporal context from the unlabeled frames of a novel camera to improve performance at that camera. Specifically, we propose an attention-based approach that allows our model, **Context R-CNN**, to index into a long term memory bank constructed on a per-camera basis and aggregate contextual features from other frames to boost object detection performance on the current frame.

We apply Context R-CNN to two settings: (1) species detection using camera traps, and (2) vehicle detection in traffic cameras, showing in both settings that Context R-CNN leads to performance gains over strong baselines. Moreover, we show that increasing the contextual time horizon leads to improved results. When applied to camera trap data from the Snapshot Serengeti dataset, Context R-CNN with context from up to a **month** of images outperforms a single-frame baseline by 17.9% mAP, and outperforms S3D (a 3d convolution based baseline) by 11.2% mAP.

7.2 Introduction

We seek to improve recognition within passive monitoring cameras, which are static and collect sparse data over long time horizons.[1] Passive monitoring deployments

[1] Models and code will be released online.

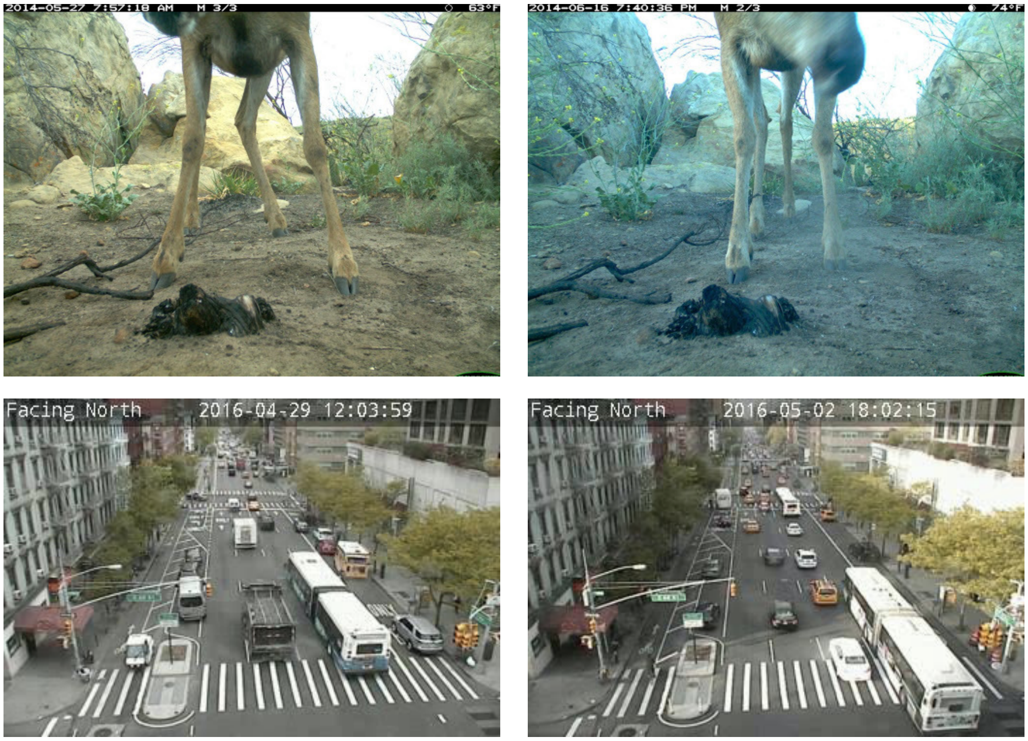

2014-05-27 7:57:18 AM M 3/3 63°F
2014-06-16 7:40:36 PM M 2/3 74°F
Facing North 2016-04-29 12:03:59
Facing North 2016-05-02 18:02:15

Figure 7.1: **Visual similarity over long time horizons.** In static cameras, there exists significantly more long term temporal consistency than in data from moving cameras. In each case above, the images were taken on separate days, yet look strikingly similar.

are ubiquitous and present unique challenges for computer vision but also offer unique opportunities that can be leveraged for improved accuracy.

For example, depending on the triggering mechanism and the camera placement, large numbers of photos at any given camera location can be empty of any objects of interest (up to 75% for some ecological camera trap datasets) [30]. Further, as the images in static passive-monitoring cameras are taken automatically (without a human photographer), there is no guarantee that the objects of interest will be centered, focused, well-lit, or an appropriate scale. We break these challenges into three categories, each of which can cause failures in single-frame detection networks:

- **Objects of interest partially observed.** Objects can be very close to the camera and occluded by the edges of the frame, partially hidden in the environment due to camouflage, or very far from the camera.

- **Poor image quality.** Objects are poorly lit, blurry, or obscured by weather conditions like snow or fog.
- **Background distractors.** When moving to a new camera location, there can exist salient background objects that cause repeated false positives.

These cases are often difficult even for humans. On the other hand, there are aspects of the passive monitoring problem domain that can give us hope–for example, subjects often exhibit similar behavior over multiple days, and background objects remain static, suggesting that it would be beneficial to provide temporal context in the form of additional frames from the same camera. Indeed we would expect humans viewing passive monitoring footage to often rewind to get better views of a difficult-to-see object.

These observations forms the intuitive basis for our model that can learn how to find and use other potentially easier examples from the same camera to help improve detection performance (see Figure 7.2). Further, like most real-world data [41], both traffic camera and camera trap data have long-tailed class distributions. By providing context for rare classes from other examples, we improve performance in the long tail as well as on common classes.

More specifically, we propose a detection architecture, *Context R-CNN*, that learns to differentiably index into a long-term memory bank while performing detection within a static camera. This architecture is flexible and is applicable even in the aforementioned low, variable framerate scenarios. At a high level, our approach can be framed as a non-parametric estimation method (like nearest neighbors) sitting on top of a high-powered parametric function (Faster R-CNN). When train and test locations are quite different, one might not expect a parametric method to generalize well [4], whereas Context R-CNN is able to leverage an unlabeled 'neighborhood' of test examples for improved generalization.

We focus on two static-camera domains:

- **Camera traps** are remote static monitoring cameras used by biologists to study animal species occurrence, populations, and behavior. Monitoring biodiversity quantitatively can help us understand the connections between species decline and pollution, exploitation, urbanization, global warming, and policy.

- **Traffic cameras** are static monitoring cameras used to monitor roadways and intersections in order to analyze traffic patterns and ensure city safety.

In both domains, the contextual signal within a single camera location is strong, and we allow the network to determine which previous images were relevant to the current frame, regardless of their distance in the temporal sequence. This is important within a static camera, as objects exhibit periodic, habitual behavior that causes them to appear days or even weeks apart. For example, an animal might follow the same trail to and from a watering hole in the morning and evening every night, or a bus following its route will return periodically throughout the day.

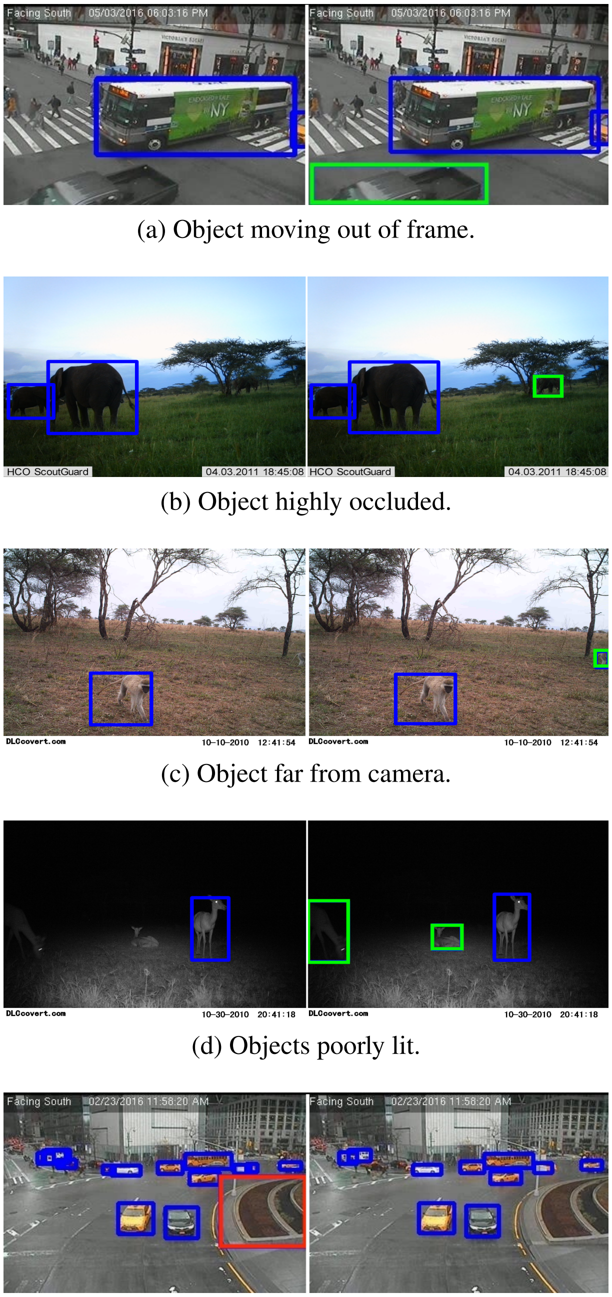
To summarize our main contributions:

- We propose *Context R-CNN*, which leverages temporal context for improving object detection regardless of frame rate or sampling irregularity. It can be thought of as a way to improve generalization to novel cameras by incorporating unlabeled images.
- We demonstrate major improvements over strong single-frame baselines; on a commonly-used camera trap dataset we improve mAP at 0.5 IoU by 17.9%.
- We show that Context R-CNN is able to leverage up to a month of temporal context which is significantly more than prior approaches.

7.3 Related Work

Single frame object detection. Driven by popular benchmarks such as COCO [23] and Open Images [21], there have been a number of advances in single frame object detection in recent years. These detection architectures include anchor-based models, both single stage (e.g., SSD [26], RetinaNet [24], Yolo [31, 32]) and two-stage (e.g., Fast/Faster R-CNN [14, 17, 33], R-FCN [10]), as well as more recent anchor-free models (e.g., CornerNet [22], CenterNet [55], FCOS [40]). Object detection methods have shown great improvements on COCO- or Imagenet-style images, but these gains do not always generalize to challenging real-world data (See Figure 7.2).

Video object detection. Single frame architectures then form the basis for video detection and spatio-temporal action localization architectures, which build upon single frame models by incorporating contextual cues from other frames in order


Facing South 05/03/2016 06:03:16 PM
Facing South 05/03/2016 06:03:16 PM
HCO ScoutGuard 04.03.2011 18:45:08
HCO ScoutGuard 04.03.2011 18:45:08
DLCcovert.com 10-10-2010 12:41:54
DLCcovert.com 10-10-2010 12:41:54
DLCcovert.com 10-30-2010 20:41:18
DLCcovert.com 10-30-2010 20:41:18
Facing South 02/23/2016 11:58:20 AM
Facing South 02/23/2016 11:58:20 AM

(a) Object moving out of frame.

(b) Object highly occluded.

(c) Object far from camera.

(d) Objects poorly lit.

(e) Background distractor.

Figure 7.2: **Static Monitoring Camera Challenges.** Images taken without a human photographer have no quality guarantees; we highlight challenges which cause mistakes in single-frame systems (left) and are fixed by our model (right). False single-frame detections are in **red**, detections missed by the single-frame model and corrected by our method are in **green**, and detections that are correct in both models are in **blue**. Note that in camera traps, the intra-image context is very powerful due to the group behavior of animal species.

to deal with more specific challenges that arise in video data including motion blur, occlusion, and rare poses. Leading methods have used pixel level flow (or flow-like concepts) to aggregate features [7, 56–58] or used correlation [13] to densely relate features at the current timestep to an adjacent timestep. Other papers have explored the use of 3d convolutions (e.g., I3D, S3D) [8, 28, 47] or recurrent networks [20, 25] to extract better temporal features. Finally, many works apply video specific postprocessing to "smooth" predictions along time, including tubelet smoothing [15] or SeqNMS [16].

Object-level attention-based temporal aggregation methods. The majority of the above video detection approaches are not well suited to our target setting of sparse, irregular frame rates. For example, flow based methods, 3d convolutions and LSTMs typically assume a dense, regular temporal sampling. And while models like LSTMs can theoretically depend on all past frames in a video, their effective temporal receptive field is typically much smaller. To address this limitation of recurrent networks, the NLP community has introduced attention-based architectures as a way to take advantage of long range dependencies in sentences [3, 12, 42]. The vision community has followed suit with attention-based architectures [27, 37, 38] that leverage longer term temporal context.

Along the same lines and most relevant to our work, there are a few recent works [11, 36, 45, 46] that rely on non-local attention mechanisms in order to aggregate information at the object level across time. For example, Wu et al [45] applied non-local attention [44] to person detections to accumulate context from pre-computed feature banks (with frozen pre-trained feature extractors). These feature banks extend the time horizon of their network up to 60s in each direction, achieving strong results on spatio-temporal action localization. We similarly use a frozen feature extractor that allows us to create extremely long term memory banks which leverage the spatial consistency of static cameras and habitual behavior of the subjects across long time horizons (up to a month). However Wu et al use a 3d convnet (I3D) for short term features which is not well-suited to our setting due to low, irregular frame rate. Instead we use a single frame model for the current frame which is more similar to [11, 36, 46] who proposed variations of this idea for video object detection achieving strong results on the Imagenet Vid dataset. In contrast to these three papers, we augment our model with an additional dedicated short term attention mechanism which we show to be effective in experiments. Uniquely, our approach also allows negative examples into memory which allows the model to learn to ignore salient false positives in empty frames due to their immobility; we find that our network is able to learn background classes (e.g., rocks, bushes) without supervision.

More generally, our paper adds to the growing evidence that this attention-based approach of temporally aggregating information at the object level is highly effective for incorporating more context in video understanding. We argue in fact that it is especially useful in our setting of sparse irregular frame samples from static cameras. Whereas a number of competing baselines like 3d convolutions and flow based techniques perform nearly as well as these attention-based models on Imagenet Vid, the same baselines are not well-suited to our setting. Thus, we see a larger performance boost from prior, non-attention-based methods to our attention-based approach.

Camera traps and other visual monitoring systems. Image classification and object detection have been increasingly explored as a tool for reducing the arduous task of classifying and counting animal species in camera trap data [4–6, 29, 30, 34, 43, 49, 50, 53]. Detection has been shown to greatly improve the generalization of these models to new camera locations [4]. It has also been shown in [4, 30, 49] that temporal information is useful. However, previous methods cannot report per-image species identifications (instead identifying a class at the burst level), cannot handle image bursts containing multiple species, and cannot provide per-image localizations and thus species counts, all of which are important to biologists.

In addition, traffic cameras, security cameras, and weather cameras on mountain passes are all frequently stationary and used to monitor places over long time scales. For traffic cameras, prior work focuses on crowd counting (e.g., counting the number of vehicles or humans in each image) [2, 9, 35, 52, 54]. Some recent works have investigated using temporal information in traffic camera datasets [48, 51], but these methods only focus on short term time horizons, and do not take advantage of long term context.

7.4 Method

Our proposed approach, Context R-CNN, builds a "memory bank" based on contextual frames and modifies a detection model to make predictions conditioned on this memory bank. In this section, we discuss (1) the rationale behind our choice of detection architecture, (2) how to represent contextual frames, and (3) how to incorporate these contextual frame features into the model to improve predictions.

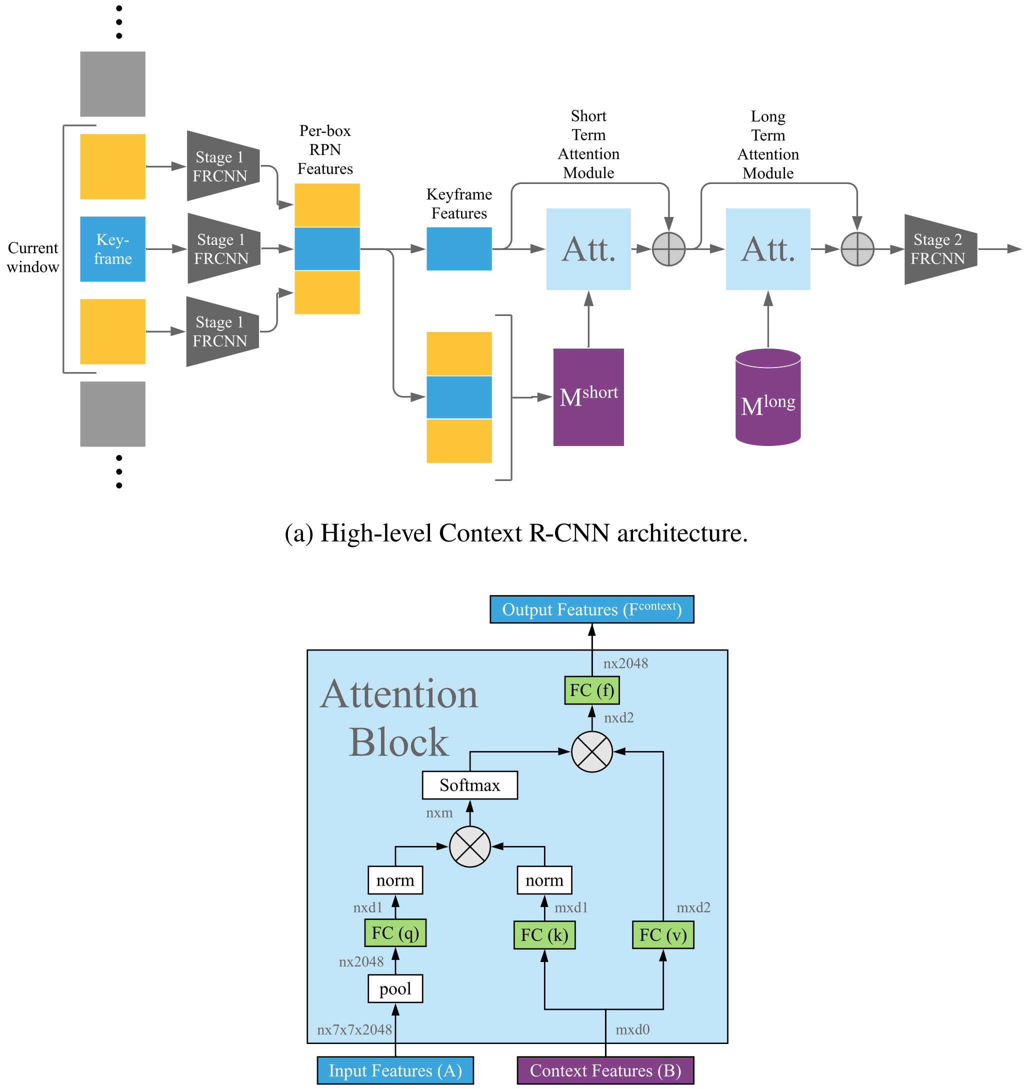
Due to our sparse, irregular input frame rates, typical temporal architectures such as 3d convnets and recurrent neural networks are not well-suited, due to a lack of inter-frame temporal consistency (there are significant changes between frames). Instead, we build Context R-CNN on top of single frame detection models. Additionally, building on our intuitions that moving objects exhibit periodic behavior and tend to appear in similar locations, we hope to inform our predictions by conditioning on instance level features from contextual frames. Because of this last requirement, we choose the Faster R-CNN architecture [33] as our base detection model as this model remains a highly competitive meta-architecture and provides clear choices for how to extract instance level features. Our method is easily applicable to any two stage detection framework.

As a brief review, Faster R-CNN proceeds in two stages. An image is first passed through a first-stage region proposal network (RPN) which, after running non-max suppression, returns a collection of class agnostic bounding box proposals. These box proposals are then passed into the second stage, which extracts instance-level features via the ROIAlign operation [18, 19] which then undergo classification and box refinement.

In Context R-CNN, the first-stage box proposals are instead routed through two attention-based modules that (differentiably) index into memory banks, allowing the model to incorporate features from contextual frames (seen by the same camera) in order to provide local and global temporal context. These attention-based modules return a contextually-informed feature vector which is then passed through the second stage of Faster R-CNN in the ordinary way. In the following section (7.4), we discuss how to represent features from context frames using a memory bank and detail our design of the attention modules. See Figure 7.3 for a diagram of our pipeline.

Building a memory bank from context features

Long Term Memory Bank (M^{long}). Given a keyframe i_t, for which we want to detect objects, we iterate over all frames from the same camera within a pre-defined time horizon $i_{t-k} : i_{t+k}$, running a frozen, pre-trained detector on each frame. We build our long term memory bank (M^{long}) from feature vectors corresponding to resulting detections. Given the limitations of hardware memory, deciding what to store in a memory bank is a critical design choice. We use three strategies to ensure that our memory bank can feasibly be stored.



(a) High-level Context R-CNN architecture.

(b) Single attention block.

Figure 7.3: **Context R-CNN Architecture.** **(a)** The high-level architecture of the model, with short term and long term attention used sequentially. Short term and long term attention are modular, and the system can operate with either or both. **(b)** We see the details of our implementation of an attention block, where n is the number of boxes proposed by the RPN for the keyframe, and m is the number of comparison features. For short term attention, m is the total number of proposed boxes across all frames in the window, shown in (a) as M^{short}. For long term attention, m is the number of features in the long term memory bank M^{long} associated with the current clip. See section 7.4 for details on how this memory bank is constructed.

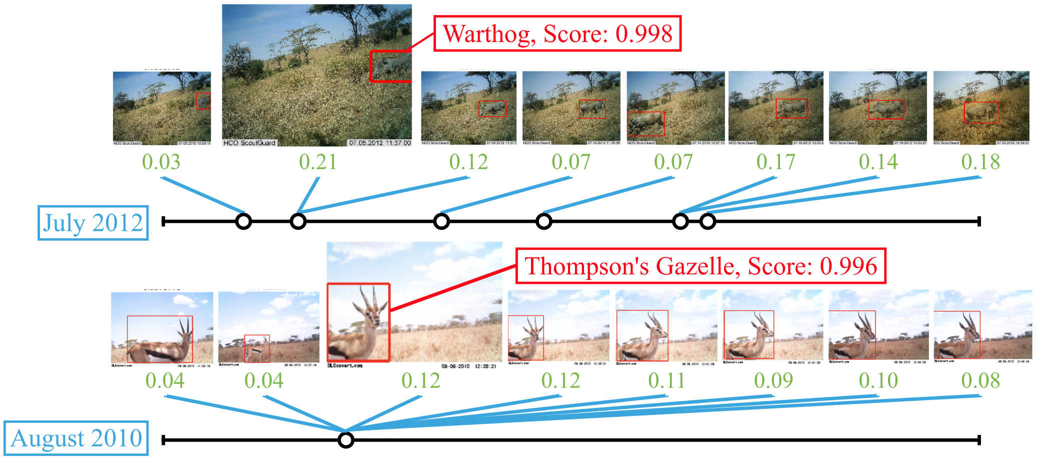


Figure 7.4: **Visualizing attention.** In each example, the keyframe is shown at a larger scale, with Context R-CNN's detection, class, and score shown in red. We consider a time horizon of one month, and show the images and boxes with highest attention weights (shown in green). The model pays attention to objects of the same class, and the distribution of attention across time can be seen in the timelines below each example. A warthogs' habitual use of a trail causes useful context to be spread out across the month, whereas a stationary gazelle results in the most useful context to be from the same day. The long term attention module is adaptive, choosing to aggregate information from whichever frames in the time horizon are most useful.

- We take instance level feature tensors after cropping proposals from the RPN and save only a spatially pooled representation of each such tensor concatenated with a spatiotemporal encoding of the datetime and box position (yielding per-box embedding vectors).

- We curate by limiting the number of proposals for which we store features–we consider multiple strategies for deciding which and how many features to save to our memory banks, see section 7.6 for more details.

- We rely on a pre-trained single frame Faster R-CNN with Resnet-101 backbone as a frozen feature extractor (which therefore need not be considered during backpropagation). In experiments we consider an extractor pretrained on COCO alone, or fine-tuned on the training set for each dataset. We find that COCO features can be used effectively but that best performance comes from a fine-tuned extractor (see Table 7.1(c)).

Together with our sparse frame rates, by using these strategies we are able to construct memory banks holding up to 8500 contextual features–in our datasets, this is sufficient to represent a month's worth of context from a camera.

Short Term Memory (M^{short}). In our experiments we show that it is helpful to include a separate mechanism for incorporating short term context features from nearby frames, using the same, trained first-stage feature extractor as for the keyframe. This is different from our long term memory from above which we build over longer time horizons with a frozen feature extractor. In contrast to long term memory, we do not curate the short term features: for small window sizes it is feasible to hold features for all box proposals in memory. We take the stacked tensor of cropped instance-level features across all frames within a small window around the current frame (typically $\leq$ 5 frames) and globally pool across the spatial dimensions (width and height). This results in a matrix of shape (# proposals per frame $*$ # frames) $\times$ (feature depth) containing a single embedding vector per box proposal (which we call our *Short Term Memory*, M^{short}), that is then passed into the short term attention block.

Attention module architecture

We define an attention block [42] which aggregates from context features keyed by input features as follows (see Figure 7.3): Let A be the tensor of input features from the current frame (which in our setting has shape $[n \times 7 \times 7 \times 2048]$, with n the number of proposals emitted by the the first-stage of Faster R-CNN). We first spatially pool A across the feature width and height dimensions, yielding A^{pool} with shape $[n \times 2048]$. Let B be the matrix of context features, which has shape $[m \times d_0]$. We set $B = M^{short}$ or M^{long}. We define $k(\cdot;\theta)$ as the *key* function, $q(\cdot;\theta)$ as the *query* function, $v(\cdot;\theta)$ as the *value* function, and $f(\cdot;\theta)$ as the final projection that returns us to the correct output feature length to add back into the input features. We use a distinct θ (θ^{long} or θ^{short}) for long term or short term attention respectively. In our experiments, k, q, v and f are all fully-connected layers, with output dimension 2048. We calculate attention weights w using standard dot-product attention:

$$w = \text{Softmax}\left((k(A^{pool};\theta) \cdot q(B;\theta)) \,/\, (T\sqrt{d})\right), \tag{7.1}$$

where $T > 0$ is the softmax temperature, w the attention weights with shape $[n \times m]$, and d the feature depth (2048).

We next construct a context feature $F^{context}$ for each box by taking a projected, weighted sum of context features:

$$F^{context} = f(w \cdot v(B;\theta);\theta), \tag{7.2}$$

where $F^{context}$ has shape $[n \times 2048]$ in our setting. Finally, we add $F^{context}$ as a per-feature-channel bias back into our original input features A.

7.5 Data

Our model is built for variable, low-frame-rate real-world systems of static cameras, and we test our methods on two such domains: camera traps and traffic cameras. Because the cameras are static, we split each dataset into separate camera locations for train and test, to ensure our model does not overfit to the validation set [4].

Camera Traps. Camera traps are usually programmed to capture an image burst of $1 - 10$ frames (taken at 1 fps) after each motion trigger, which results in data with variable, low frame rate. In this paper, we test our systems on the Snapshot Serengeti (SS) [39] and Caltech Camera Traps (CCT) [4] datasets, each of which have human-labeled ground truth bounding boxes for a subset of the data. We increase the number of bounding box labeled images for training by pairing class-agnostic detected boxes from the Microsoft AI for Earth MegaDetector [5] with image-level species labels on our training locations. SS has 10 publicly available seasons of data. We use seasons $1 - 6$, containing 225 cameras, 3.2M images, and 48 classes. CCT contains 140 cameras, 243K images, and 18 classes. Both datasets have large numbers of false motion triggers, 75% for SS and 50% for CCT; thus many images contain no animals. We split the data using the location splits proposed in [1], and evaluate on the images with human-labeled bounding boxes from the validation locations for each dataset (64K images across 45 locations for SS and 62K images across 40 locations for CCT).

Traffic Cameras. The CityCam dataset [52] contains 10 types of vehicle classes, around 60K frames and 900K annotated objects. It covers 17 cameras monitoring downtown intersections and parkways in a high-traffic city, and "clips" of data are sampled multiple times per day, across months and years. The data is diverse, covering day and nighttime, rain and snow, and high and low traffic density. We use 13 camera locations for training and 4 cameras for testing, with both parkway and downtown locations in both sets.

7.6 Experiments

We evaluate all models on held-out camera locations, using established object detection metrics: mean average precision (mAP) at 0.5 IoU and Average Recall (AR). We compare our results to a (comparable) single-frame baseline for all three datasets. We focus the majority of our experiments on a single dataset, Snapshot

	Snapshot Serengeti		Caltech Camera Traps		CityCam	
Model	mAP	AR	mAP	AR	mAP	AR
Single Frame	37.9	46.5	56.8	53.8	38.1	28.2
Context R-CNN	**55.9**	**58.3**	**76.3**	**62.3**	**42.6**	**30.2**

(a) Results across datasets

Snapshot Serengeti	mAP	AR
One minute	50.3	51.4
One hour	52.1	52.5
One day	52.5	52.9
One week	54.1	53.2
One month	**55.6**	**57.5**

(b) Time horizon

Snapshot Serengeti	mAP	AR
One box per frame	**55.6**	**57.5**
COCO features	50.3	55.8
Only positive boxes	53.9	56.2
Subsample half	52.5	56.1
Subsample quarter	50.8	55.0

(c) Selecting memory

Snapshot Serengeti	mAP	AR
Single Frame	37.9	46.5
Maj. Vote	37.8	46.4
ST Spatial	39.6	36.0
S3D	44.7	46.0
SF Attn	44.9	50.2
ST Attn	46.4	55.3
LT Attn	55.6	57.5
ST+LT Attn	**55.9**	**58.3**

(d) Comparison across models

CityCam	mAP	AR
Single Frame	38.1	28.2
Top 1 Box	40.5	29.3
Top 8 Boxes	**42.6**	**30.2**

(e) Adding boxes to M^{long}

Table 7.1: **Results.** All results are based on Faster R-CNN with a Resnet 101 backbone. We consider the Snapshot Serengeti, Caltech Camera Traps, and CityCam datasets. All mAP values employ an IoU threshold of 0.5, and AR is reported for the top prediction (AR@1).

Serengeti, investigating the effects of both short term and long term attention, the feature extractor, the long term time horizon, and the frame-wise sampling strategy for M^{long}. We further explore the addition of multiple features per frame in CityCam.

Main Results

Context R-CNN strongly outperforms the single-frame Faster RCNN with Resnet-101 baseline on both the Snapshot Serengeti (SS) and Caltech Camera Traps (CCT) datasets, and shows promising improvements on CityCam (CC) traffic camera data as well (See Table 7.1 (a)). For all experiments, unless otherwise noted, we use a fine-tuned dataset specific feature extractor for the memory bank. **We show an**

absolute mAP at 0.5 IoU improvement of 19.5% on CCT, 17.9% on SS, and 4.5% on CC. Recall improves as well, with AR@1 improving 2% on CC, 11.8% on SS, and 8.5% on CCT.

For SS, we also compare against several baselines with access to short term temporal information (Table 7.1(d)). All short term experiments use an input window of 3 frames. Our results are as follows:

- We first consider a simple majority vote **(Maj. Vote)** across the high-confidence single-frame detections within the window, and find that it does not improve over the single-frame baseline.
- We attempt to leverage the static-ness of the camera by taking a temporal-distance-weighted average of the RPN box classifier features from the key frame with the cropped RPN features from the same box locations from the surrounding frames **(ST Spatial)**, and find it outperforms the single-frame baseline by 1.9% mAP.
- **S3D** [47], a popular video object detection model, outperforms single-frame by 6.8% mAP despite being designed for consistently sampled high frame rate video.
- Since animals in camera traps occur in groups, cross-object intra-image context is valuable. An intuitive baseline is to restrict the short term attention context window (M^{short}) to the current frame **(SF Attn)**. This removes temporal context, showing how much improvement we gain from explicitly sharing information across the box proposals in a non-local way. We see that we can gain 7% mAP over a vanilla single-frame model by adding this non-local attention module.
- When we increase the short term context window to three frames, keyframe plus two adjacent, **(ST Attn)** we see an additional improvement of 1.5% mAP.
- If we consider *only* long term attention with a time horizon of one month **(LT Attn)**, we see a 9.2% mAP improvement over short term attention.
- By combining both attention modules into a single model **(ST+LT Attn)**, we see our highest performance at 55.9% mAP, and show in Figure 7.5 that we improve for all classes in the imbalanced dataset.

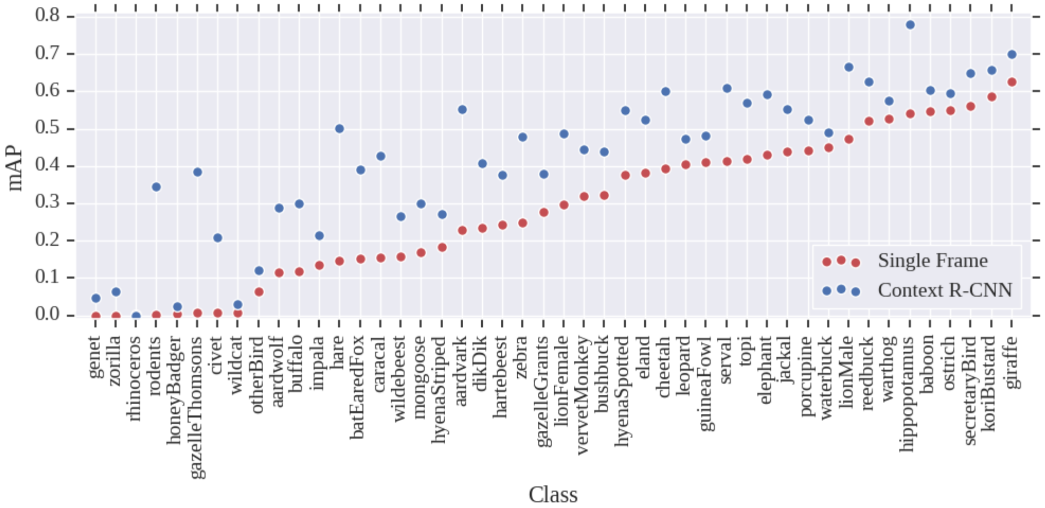
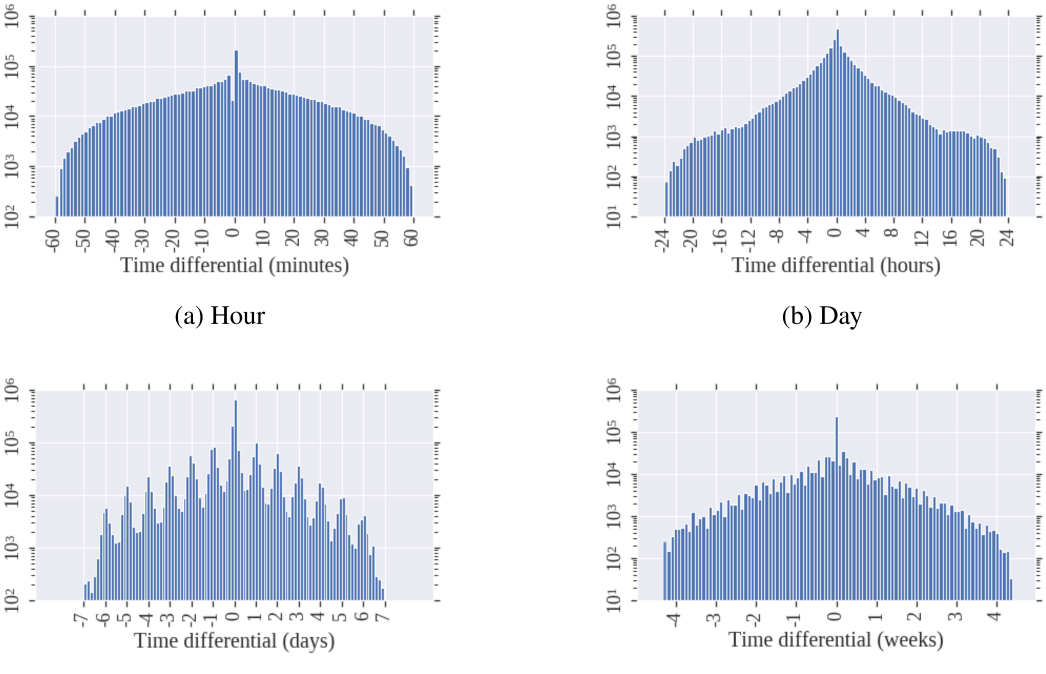


Figure 7.5: **Performance per class.** Our performance improvement is consistent across classes: we visualize SS per-species mAP from the single-frame model to our best long term and short term memory model.

Changing the Time Horizon (Table 7.1(b))

We ablate our long term only attention experiments by increasing the time horizon of M^{long}, and find that performance increases as the the time horizon increases. We see a large performance improvement over the single-frame model even when only storing a minute-worth of representations in memory. This is due to the sampling strategy, as highly-relevant bursts of images are captured for each motion trigger. The long term attention block can adaptively determine how to aggregate this information, and there is much useful context across images within a single burst. However, some cameras take only a single image at a trigger; in these cases the long term context becomes even more important. The adaptability of Context R-CNN to be trained on and improve performance across data with variable frame rates *and* with different sampling strategies (time lapse, motion trigger, heat trigger, and bursts of 1-10 images per trigger) is a valuable attribute of our system.

In Figure 7.6, we explore the time differential between the top scoring box for each image and the features it most closely attended to, using a threshold of 0.01 on the attention weight. We can see day/night periodicity in the week- and month-long plots, showing that attention is focused on objects captured at the same time of day. As the time horizon increases, the temporal diversity of the attention module increases and we see that Context R-CNN attends to what is available across the time horizon, with a tendency to focus higher on images nearby in time (see examples in Figure 7.4).



(a) Hour

(b) Day

(c) Week

(d) Month

Figure 7.6: **Attention over time.** We threshold attention weights at 0.01, and plot a histogram of time differentials from the highest-scoring object in the keyframe to the attended frames for varied long term time horizons. Note that the y-axis is in log scale. The central peak of each histogram shows the value of nearby frames, but attention covers the breadth of what is provided: namely, **if given a month worth of context, Context R-CNN will use it**. Also note a strong day/night periodicity when using a week-long or month-long memory bank.

Contextual features for constructing M^{long}.

Feature extractor (Table 7.1(c)). For Snapshot Serengeti, we consider both a feature extractor trained on COCO, and one trained on COCO and then fine-tuned on the SS training set. We find that while a month of context from a feature extractor tuned for SS achieves 5.3% higher mAP than one trained only on COCO, we are able to outperform the single-frame model by 12.4% using memory features that have never before seen a camera trap image.

Subsampling memory (Table 7.1(c)). We further ablate our long term memory by decreasing the stride at which we store representations in the memory bank, while maintaining a time horizon of one month. If we use a stride of 2, which subsamples the memory bank by half, we see a drop in performance of 3.1% mAP at 0.5. If we increase the stride to 4, we see an additional 1.7% drop. If instead of increasing the stride, we instead subsample by taking only positive examples (using an oracle to determine which images contain animals for the sake of the experiment), we find that performance still drops (explored below).

Keeping representations from empty images. In our static camera scenario, we choose to add features into our long term memory bank from all frames, both empty and non-empty. The intuition behind this decision is the existence of salient background objects in the static camera frame which do not move over time, and can be repeatedly and erroneously detected by single-frame architectures. We assume that the features from the frozen extractor are visually representative, and thus sufficient for both foreground and background representation. By saving representations of highly-salient background objects, we thus hope to allow the model to learn per-camera salient background classes and positions without supervision, and to suppress these objects in the detection output.

In Figure 7.7, we see that adding empty representations reduces the number of false positives across all confidence thresholds compared to the same model with only positive representations. We investigated the 100 highest confidence "false positives" from Context R-CNN, and found that in almost all of them (97/100), the model had correctly found and classified animals that were missed by human annotators. The Snapshot Serengeti dataset reports 5% noise in their labels [39], and looking at the high-confidence predictions of Context R-CNN on images labeled "empty" is intuitively a good way to catch these missing labels. Some of these are truly challenging, where the animal is difficult to spot and the annotator mistake is unfortunate but reasonable. Most are truly just label noise, where the existence of an animal is obvious, suggesting that our performance improvement estimates are likely conservative.

Keeping multiple representations per image (Table 7.1(e)). In Snapshot Serengeti, there are on average 1.6 objects and 1.01 classes per image across the non-empty images, and 75% of the images are empty. The majority of the images contain a single object, while a few have large herds of a single species. Given this, choosing only the top-scoring detection to add to memory makes sense, as that object is likely to be representative of the other objects in the image (*e.g.,* keeping only one zebra example from an image with a herd of zebra). In CityCam, however, on average there are 14 objects and 4 classes per frame, and only 0.3% of frames are empty. In this scenario, storing additional objects in memory is intuitively useful, to ensure that the memory bank is representative of the camera location. We investigate adding features from the top-scoring 1 and 8 detections, and find that selecting 8 objects

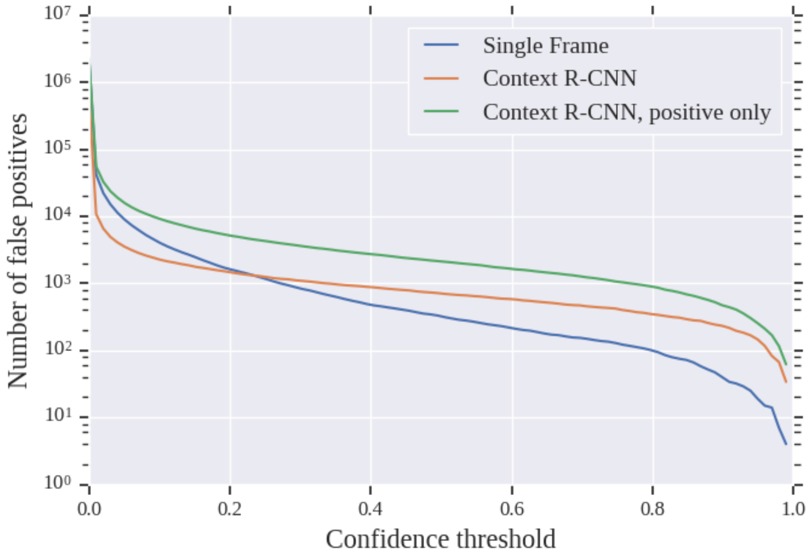


Figure 7.7: **False positives on empty images.** When adding features from empty images to the memory bank, we reduce false positives across all confidence thresholds compared to the same model without negative representations. Note that the y-axis is in log scale. The single frame model has fewer high-confidence false positives than either context model, but when given positive and negative context Context R-CNN is able to suppress low-confidence detections. By analyzing Context R-CNN's 100 most high-confidence detections on images labeled "empty" we found 97 images where the annotators missed animals.

per frame yields the best performance (see Table 7.1(e)). A logical extension of our approach would be selecting objects to store based not only on confidence, but also diversity.

Failure modes. One potential failure case of this similarity-based attention approach is the opportunity for hallucination. If one image in a test location contains something that is very strongly misclassified, that one mistake may negatively influence other detections at that camera. For example, when exploring the confident "false positives" on the Snapshot Serengeti dataset (which proved to be almost universally true detections that were missed by human annotators) the 3/100 images where Context R-CNN erroneously detected an animal were all of the same tree, highly confidently predicted to be a giraffe.

7.7 Conclusions and Future Work

In this work, we contribute a model that leverages per-camera temporal context up to a month, far beyond the time horizon of previous approaches, and show that in the static camera setting, attention-based temporal context is particularly beneficial. Our method, Context R-CNN, is general across static camera domains, improving detection performance over single-frame baselines on both camera trap and traffic camera data. Additionally, Context R-CNN is adaptive and robust to passive-monitoring sampling strategies that provide data streams with low, irregular frame rates.

It is apparent from our results that what and how much information is stored in memory is both important and domain specific. We plan to explore this in detail in the future, and hope to develop methods for curating diverse memory banks which are optimized for accuracy and size, to reduce the computational and storage overheads at training and inference time while maintaining performance gains.

7.8 Acknowlegdements

We would like to thank Pietro Perona, David Ross, Zhichao Lu, Ting Yu, Tanya Birch and the Wildlife Insights Team, Joe Marino, and Oisin MacAodha for their valuable insight. This work was supported by NSFGRFP Grant No. 1745301, the views are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] Lila.science. `http://lila.science/`. Accessed: 2019-10-22.

[2] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. pages 483–498, 2016.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[5] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *In the Data Mining and Artificial Intelligence for Conservation Workshop at Knowledge Discovery in Databases (KDD)*, 2019. *selected to be featured at KDD Earth Day.

[6] Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 863–873, 2020.

[7] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 331–346, 2018.

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[9] Antoni B. Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[10] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[11] Hanming Deng, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Object guided external memory network for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6678–6687, 2019.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3038–3046, 2017.

[14] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vsion*, pages 1440–1448, 2015.

[15] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 759–768, 2015.

[16] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465*, 2016.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[19] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*, 2017.

[20] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. Object detection in videos with tubelet proposal networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 727–735, 2017.

[21] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018.

[22] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vsion*, pages 2980–2988, 2017.

[25] Mason Liu and Menglong Zhu. Mobile video object detection with temporally-aware feature maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5686–5695, 2018.

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[28] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018.

[29] Agnieszka Miguel, Sara Beery, Erica Flores, Loren Klemesrud, and Rana Bayrakcismith. Finding areas of motion in camera trap images. In *Image*

Processing (ICIP), 2016 IEEE International Conference on, pages 1334–1338. IEEE, 2016.

[30] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25): E5716–E5725, 2018.

[31] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7263–7271, 2017.

[32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[34] Stefan Schneider, Graham W. Taylor, and Stefan Kremer. Deep learning object detection methods for ecological camera trap data. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 321–328. IEEE, 2018.

[35] Ankit Parag Shah, Jean-Bapstite Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. Cadp: A novel dataset for cctv traffic camera based accident analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–9. IEEE, 2018.

[36] Mykhailo Shvets, Wei Liu, and Alexander C Berg. Leveraging long-range temporal relationships between proposals for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9756–9764, 2019.

[37] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[38] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *arXiv preprint arXiv:1904.01766*, 2019.

[39] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*, 2:150026, 2015.

[40] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *arXiv preprint arXiv:1904.01355*, 2019.

[41] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[43] Alexander Gomez Villa, Augusto Salazar, and Francisco Vargas. Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecological Informatics*, 41:24–32, 2017.

[44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[45] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.

[46] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9217–9225, 2019.

[47] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018.

[48] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. Spatiotemporal modeling for crowd counting in videos. *Proceedings of the IEEE International Conference on Computer Vision*, pages 5151–5159, 2017.

[49] Hayder Yousif, Jianhe Yuan, Roland Kays, and Zhihai He. Fast human-animal detection from highly cluttered camera-trap images using joint background modeling and deep learning classification. In *Circuits and Systems (ISCAS), 2017 IEEE International Symposium on*, pages 1–4. IEEE, 2017.

[50] Xiaoyuan Yu, Jiangping Wang, Roland Kays, Patrick A Jansen, Tianjiang Wang, and Thomas Huang. Automated identification of animal species in

camera trap images. *EURASIP Journal on Image and Video Processing*, 2013 (1):52, 2013.

[51] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and José MF Moura. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3667–3676, 2017.

[52] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and Jose MF Moura. Understanding traffic density from large-scale web camera data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5898–5907, 2017.

[53] Zhi Zhang, Zhihai He, Guitao Cao, and Wenming Cao. Animal detection from highly cluttered natural scenes using spatiotemporal object region proposals and patch verification. *IEEE Transactions on Multimedia*, 18(10):2079–2092, 2016.

[54] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. In *Advances in Neural Information Processing Systems*, pages 8559–8570, 2018.

[55] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[56] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 408–417, 2017.

[57] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2349–2358, 2017.

[58] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. Towards high performance video object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7210–7218, 2018.

Chapter 8

THE AUTO ARBORIST DATASET: A LARGE-SCALE BENCHMARK FOR MULTIVIEW URBAN FOREST MONITORING UNDER DOMAIN SHIFT

8.1 Abstract

Generalization to novel domains is a fundamental challenge for computer vision. Near-perfect accuracy on benchmarks is common, but these models do not work as expected when deployed outside of the training distribution. To build computer vision systems that truly solve real-world problems at global scale, we need benchmarks that fully capture real-world complexity, including geographic domain shift, long-tailed distributions, and data noise.

We propose urban forest monitoring as an ideal testbed for studying and improving upon these computer vision challenges, while working towards filling a crucial environmental and societal need. Urban forests provide significant benefits to urban societies. However, planning and maintaining these forests is expensive. One particularly costly aspect of urban forest management is monitoring the existing trees in a city: e.g., detecting tree locations, species, and health. Monitoring efforts are currently based on tree censuses built by human experts, costing cities millions of dollars per census and thus collected infrequently.

Most previous investigations into automating urban forest monitoring focused on datasets from one or two cities, usually covering only common categories. To address these shortcomings, we introduce a new large-scale dataset that joins public tree censuses from 23 cities with a large collection of street level and aerial imagery. Our Auto Arborist dataset contains over 2.5M trees and 344 genera and is >2 orders of magnitude larger than the closest dataset in the literature. We introduce baseline results on our dataset across modalities as well as metrics for the detailed analysis of generalization with respect to geographic distribution shifts, vital for such a system to be deployed at-scale.

8.2 Introduction

Urban forests provide critical benefits to the over 4B people living in urban areas worldwide [106]. They filter air and water, capture stormwater runoff, sequester

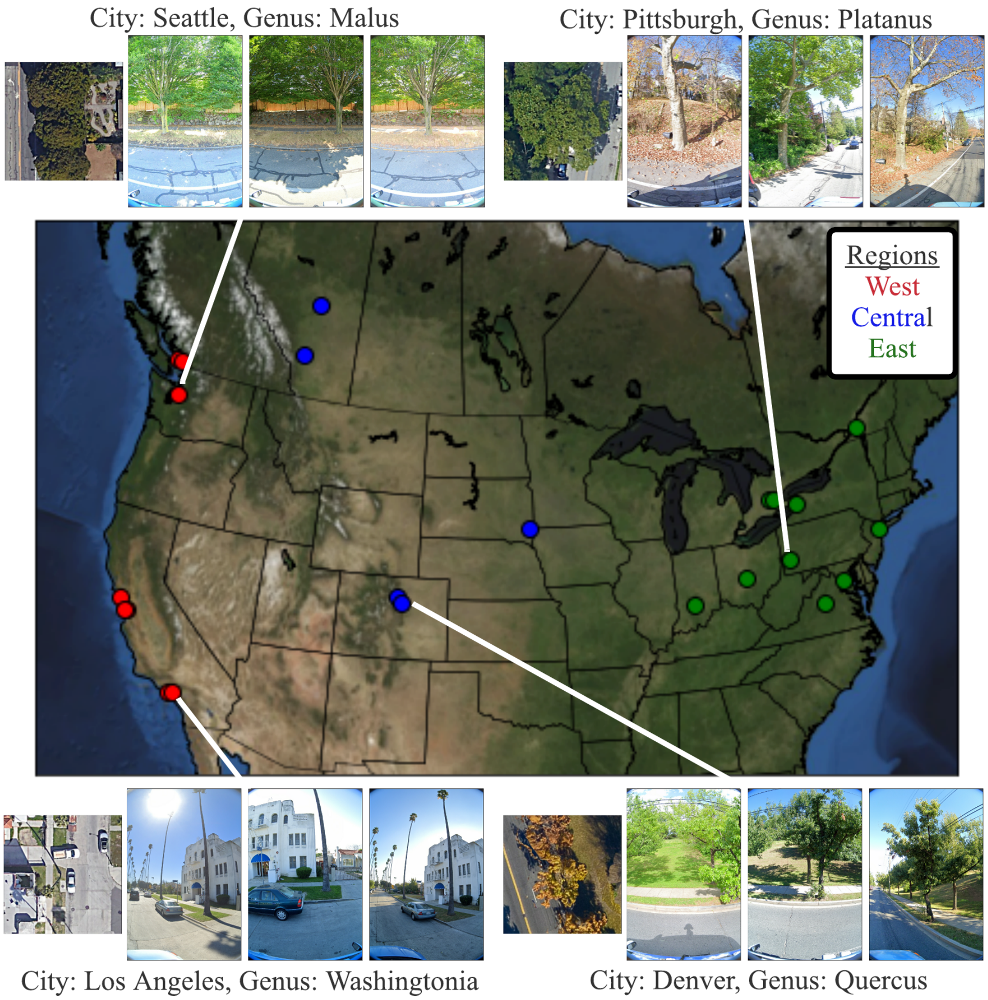

City: Seattle, Genus: Malus
City: Pittsburgh, Genus: Platanus
Regions
West
Central
East
City: Los Angeles, Genus: Washingtonia
City: Denver, Genus: Quercus

Figure 8.1: The Auto Arborist dataset covers 23 cities across North America, and contains paired aerial and multiview ground-level imagery for 2.6M trees across 344 unique genera.

Dataset	**Ground-level images**	**Aerial Images**	**Sites**	**Classes**
Registree [27]	46,321	28,678	1	40
Pasadena Re-ID [100]	6,141 (panos)	0	1	1
NEON Tree Evaluation* [131]	0	25,949 (boxes)	22	1
IDTreeS Classification* 2017 [97]	0	613 (boxes)	1	9
IDTreeS Classification* 2020 [56]	0	452 (boxes)	2	34
Auto Arborist (Ours)	6,479,077	2,637,208	23	344

Table 8.1: Comparison of our dataset to previous tree detection and identification datasets. Note that for Pasadena Re-ID, there is only one class ("tree") but the task is re-identification as opposed to categorization. The three datasets with an * are from wilderness forests, as opposed to urban forests.

atmospheric carbon dioxide, limit erosion and drought, and save energy in a variety of ways (e.g., by providing shade and thus reducing cooling costs and urban heat islands [99, 119, 138, 141]). In the US alone, urban forests cover 127M acres and produce ecosystem services valued at >$18B [105]. These forests make up the foundations of our urban ecosystems, and provide habitat for diverse urban wildlife and insect populations [44]. *Urban forest monitoring*, measuring the size, health and species distribution over time, allows us to (1) quantify ecosystem services including air quality improvement [20, 47], carbon sequestration [91, 105, 121], and benefits to public health [28, 47, 125, 125], (2) track damage from extreme weather events [8, 67, 98], and (3) target planting to improve robustness to climate change, disease and infestation [24, 64, 113, 114]. Further, lack of access to urban greenery is a key aspect of urban social inequality [56, 93, 103], including socioeconomic inequality [52, 73, 85] and racial inequality [21, 129]. Urban forest monitoring enables the quantification of this inequality and the pursuit of its improvement [22, 117].

To enable automated urban forest monitoring, we focus on the fundamental task of building a *tree census* (sometimes called a *tree inventory*). Due to their significant cost (a recent tree census in Los Angeles cost $2M and took 18 months [120]), tree censuses are typically conducted only by cities with the means and will to invest in these undertakings, and even then they are conducted rarely (e.g., once every 5-10 years). We seek to dramatically lower the cost of a tree census by using computer vision to help find, label, and monitor individual trees using a combination of street level and aerial imagery. An automated system could democratize access to urban forest monitoring, providing this valuable information to under-resourced cities that are already disproportionately affected by climate change [108].

While there have been prior works on urban tree species recognition from aerial [3, 4, 63, 78, 124, 137, 139, 140] or street level [94] imagery (or both, in a limited number of cases [27, 130]), a major limitation has been a lack of large-scale labeled datasets. To our knowledge, all prior works have focused on single or limited numbers of cities, and have included only the most common categories. We introduce the *Auto Arborist dataset*, a multiview urban tree classification dataset which, at 2.6 million trees is two orders of magnitude larger than those in prior work [27, 130] and contains 344 genera (and many more species). To build Auto Arborist, we draw on public tree censuses from 23 cities in the US and Canada and merge these public records with street level and overhead RGB imagery. As the first urban forest dataset to cover multiple cities, Auto Arborist allows for previously-impossible detailed analysis of generalization with respect to geographic distribution shifts, vital to building systems that scale. We propose a set of metrics to evaluate performance with respect to these geographic distribution shifts and show the strengths and weaknesses of typical deep learning models when applied to the Auto Arborist dataset.

Going beyond its immediate application to sustainability and conservation, Auto Arborist can serve as an important challenge benchmark for computer vision. There has been increasing interest in domain generalization, which is ubiquitous in real-world applications [13, 51, 81, 96]. For example, prior works have observed that a model for self-driving cars that can drive safely in San Diego may not work equally well in Seattle [50, 68, 69]. In terms of number of domains, objects, classes, and images, Auto Arborist presents a scale not seen in previous real-world domain generalization benchmarks [13–17, 81]; it focuses on detailed cross domain analysis, and incorporates multiple views and modalities.

To summarize, our main contributions are as follows:

- We develop a pipeline for combining public tree census data with street level and aerial imagery.
- We introduce the Auto Arborist dataset built using this pipeline — the largest dataset of its kind covering >2.6M trees, >9.1M images and 344 categories and first one of its kind to cover multiple cities (23 cities).
- We show that for Auto Arborist, geographic domain shift and the category imbalance are major factors in performance of typical deep learning models.
- We show that diverse geographic coverage is important for generalization to a novel city, and that both multiple views and multiple data modalities are beneficial.

8.3 Related work

Tree detection, localization, and taxonomic identification have been studied in multispectral aerial imagery [49, 78, 140], ground-level imagery [94, 101], and LIDAR [46, 70], with some recent approaches combining data across modalities [6, 27]. Costly high-resolution data collected from low-flying aircraft has been shown to improve performance [18, 109], but this data is not available for much of the world. Though past studies have considered tree detection and categorization, many of these have been limited by perspective (aerial vs ground-level views), geospatial coverage, and taxonomic coverage. Our work seeks to expand upon all three, providing a testbed for urban forest monitoring that is broad in scope and relies on datatypes

Region W (West)		
City	Trees	Genera
San Francisco, CA	154,698	195
San Jose, CA	225,655	201
Cupertino, CA	15,300	104
Vancouver, BC	121,249	93
Seattle, WA	150,983	142
Surrey, BC	62,251	72
Santa Monica, CA	25,381	126
Los Angeles, CA	391,788	202
Total West	1,147,305	328
Region C (Central)		
City	Trees	Genera
Boulder City, CO	29,489	65
Denver, CO	175,438	97
Calgary, AB	64,576	35
Sioux Falls, SD	13,277	
Edmonton, AB	76,164	32
Total Central	358,944	104
Region E (East)		
City	Trees	Genera
Washington, DC	152,983	71
Charlottesville, VA	1571	56
Pittsburgh, PA	23,382	79
Montreal, QC	208,097	61
New York, NY	560,069	68
Buffalo, NY	29,527	74
Kitchener, ON	21,265	26
Cambridge, ON	14,757	57
Columbus, OH	114,536	81
Bloomington, IN	4,772	53
Total East	1,130,959	102

Table 8.2: Cities by region. The holdout city for each region is in bold.

which are available across cities (aerial and street level RGB imagery) to enable the development of general models and methods which can be used off-the-shelf in novel cities.

Tree detection and localization from aerial data

There have been successful, broad-scale studies of tree density, canopy cover, and individual tree delineation from aerial data [9, 25, 38, 40, 46, 62, 66, 77, 92, 107, 109, 111, 122, 131], including tree crown detection across sites from the US National Ecological Observatory Network (NEON) [131, 134], tree canopy mapping in urban forests in cities across the US [93], and counting individual trees in Sub-Saharan Africa [25]. These methods rely on a diverse set of aerial data modalities, from low-resolution RGB or hyperspectral sattelite data to high-resolution RGB, hyperspectral, and LIDAR data collected from low-flying aircraft and UAVs [18]. However, there are still open challenges in maintaining performance of methods in novel regions [97, 132, 133], and methods must be well-validated and possibly adapted for any novel region before use. Tree crown delineation in dense forests remains a challenge, leading to several studies (e.g., in sub-Saharan Africa [25]) focusing on low-tree-density regions or trees outside forests [43, 116]. Further, there is only so much that can be understood from an aerial view alone. A large amount of the woody vegetation in a forest is hidden under the tree canopy. Understory trees have been mapped with very-high-resolution UAV-collected data [32, 60, 61, 86], but this data is rarely available. Our combined approach allows us to use available ground level imagery to see under the canopy.

Tree taxonomic identification from aerial data

Automated tree identification in aerial data from satellite or low-flying aircraft, including RGB, hyperspectral, LIDAR, or some combination thereof, is well-studied in the remote sensing community. [49] is a thorough review of species classification from remote sensing data which notes the lack of studies considering large spatial extents. Many studies focus on predicting species occurrence, presence/absence, or abundance for a limited set of species [2, 23, 29–31, 33, 34, 41]. Detecting and categorizing individual trees presents further complexity [3, 26, 35, 36, 41, 42, 48, 48, 53, 55, 59, 65, 70–72, 74, 75, 79, 80, 82, 110, 112, 115, 128], and recently deep learning approaches have been shown to outperform more traditional methods on this task [39, 57, 63, 89, 113, 127, 142]. Generalization to novel regions is a known challenge with many of the proposed methods [109]. The IDTrees challenges

[56, 97] were the first to propose a public benchmark for cross-site individual tree categorization, but provided limited labeled data (<1000 labeled trees from <=34 tree categories at 3 NEON wilderness forest plots). Further challenge arises when predicting species in an urban environment, where human intervention leads to a much higher diversity of tree species, with a much longer tail, than is seen in the wild [135]. For this reason, many studies of urban tree categorization focus only on common species [3, 4, 10, 11, 63, 78, 124, 137, 139, 140].

Tree detection and localization in the urban forest from ground-level data
Ground-level data, (e.g., from Google Street View [7], Mapillary [95], and iNaturalist [1]), have been identified as an important source of information for urban monitoring applications [19]. Automated measures of urban "greenness" and tree cover mapping from ground-level data have been proposed, with implications in social justice and public health [45, 87, 88, 118, 123]. Datasets such as Mapillary Vistas [104] and Cityscapes [37] facilitate semantic segmentation of urban categories, including vegetation, but do not provide instance-level information or fine-grained taxonomic labels. Similarly, most current computer vision studies of the urban forest focus on species-agnostic individual tree detection [76, 126] and localization [83, 94, 100–102] across multiple ground-level views of the same tree.

Tree taxonomic identification from combined aerial and ground-level data
Previous large-scale datasets that combine aerial and ground-level data, such as CVUSA [136], were designed for alternative tasks such as image geolocalization. Several methods exist for combining aerial + ground-level data, with tree identification as a key application [84, 116]. Here, ground-level data can include RGB imagery, LIDAR, and even physical measurements such as tree diameter or hyperspectral signature [54]. [130] and [27] proposed a system for identifying street trees using paired aerial and ground-level RGB imagery for urban forests and released a dataset of paired imagery for Pasadena. [6] proposed a class-agnostic tree detection method from aerial imagery and ground-level LIDAR. Recently, [5] used GNNs to map individual trees across aerial and ground-level community science imagery in forests. All of these prior works trained on a single city and could benefit from a much larger dataset such as ours.

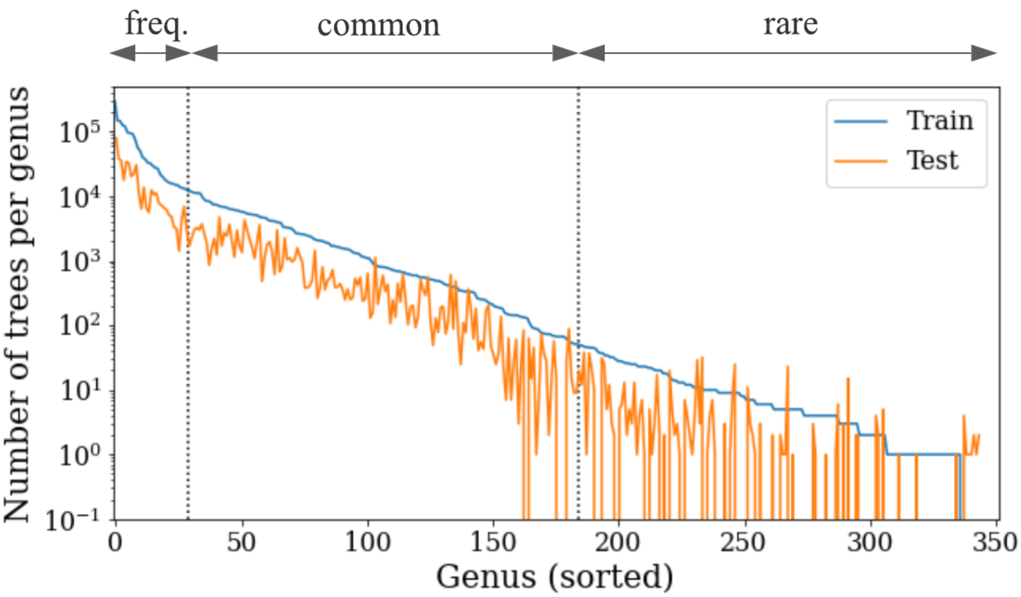


Figure 8.2: Distribution of genera in train and test, with frequent, common, and rare classes delineated.

8.4 The Auto Arborist dataset

We have generated the largest, and most geographically diverse, computer-vision-ready multi-view dataset of urban trees to date. The *Auto Arborist* dataset contains 2,637,208 trees across 23 cities. Each tree is represented by a 512 × 512 pixel aerial image where each pixel is $5cm \times 5cm$, as well as up to three 768 × 1152 pixel street level images [7] of the tree (for a total of 9,116,285 images in the dataset)[1]. To avoid taxonomic complexity arising from hybrid and sub-species when developing methods, we have chosen to focus on genus prediction (instead of species-level prediction) as our primary task and have confirmed with ecologists and city planners that a genus-level map would be highly useful as a first step. Our dataset includes 344 unique genera, with a real-world long-tailed class imbalance and unique class distribution for each city on the dataset (Figure 8.2).

Dataset curation

To curate Auto Arborist, we started from existing tree censuses which are provided by many cities online. For each tree census considered, we verified that the data contained GPS locations and genus/species labels, and was available for public use. This resulted in data from 23 cities which we then parsed into a common

[1] We are publishing all tree records (after curation/merging c.f. 8.4) and a subset of the imagery (verified to obtain consistent results to the full dataset) with personally identifiable information removed. For more information, please visit https://google.github.io/auto-arborist.

format, fixing common data entry errors (such as flipped latitude/longitude) and mapping groundtruth genus names (and their common misspellings) to a universal label map consisting of 344 categories. We also removed records with invalid genus names, such as “unidentified.” Aggregated into a single dataset, this process yielded localized records for ~5M trees.

Figure 8.1 shows a map of the 23 selected cities as well as example imagery from the dataset. We partition the cities into three separate regions for evaluation purposes (discussed further in Section 8.5). Table 8.2 summarizes the contribution from each of the cities to the Auto Arborist dataset organized by these regions. For this “v1” version of Auto Arborist, we restrict our focus to the US and Canada, with a single genus prediction task. There is room for Auto Arborist to grow in tasks and geographic area: many public tree censuses contain additional metadata (e.g., tree age, health, and trunk diameter), and there are many more cities we might include both in the US and Canada, and globally. We place our dataset in context with previously published tree classification datasets in Table 8.1, and emphasize the significantly enhanced scope in number of images across modalities, number of regions, and number of categories.

Extracting street level and aerial imagery

For each city, starting from the parsed tree census, we associate each tree census record to both street level and aerial images. For each tree in our dataset, we sample a $15m \times 15m$, 300×300 pixel RGB aerial image centered on the tree’s latitude and longitude. We consider all street level images taken within 2-10 meters of the record’s latitude and longitude, filtering out any images which do not meet all of the following criteria:

- Taken on or after Jan 1, 2018.
- Contains the base of the tree near the horizontal center of the image based on the projection of the tree’s latitude/longitude onto the image, based on estimated camera pose generated by the API.
- Contains a significant number of “tree” pixels based on a semantic segmentation model (to avoid cases when the tree has died or been removed, when possible) *and* does not contain any “person” or “bike rider” pixels based on a semantic segmentation model (to remove personally identifiable information).

After filtering, we have 2.6M tree records, each of which is associated with one aerial image and 1-3 street level images, along with a date and GPS location.

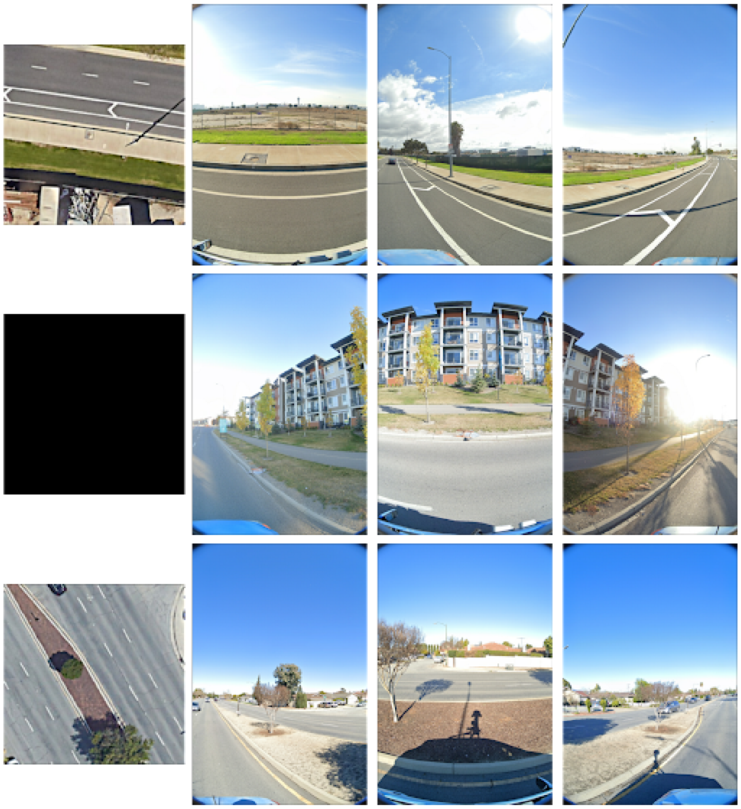


Figure 8.3: Noise in the Auto Arborist dataset includes trees that have died since the tree census was taken (top), aerial data quality, including failures causing black squares (middle), and temporal variation in deciduous trees (bottom–aerial image has leaves, but street level images are bare), which affects northern cities more than southern ones.

Challenging aspects of the Auto Arborist data

By matching street level images from existing public records rather than collecting groundtruth labels from scratch, we have been able to achieve a scale much larger than any previous datasets. As we show, scale is important for generalizing to novel cities (which is the ultimate goal). But using public records to generate data across cities also introduces a number of challenges.

Sources of noise and ambiguity. First, we address several known sources of noise and ambiguity in our dataset. See Figure 8.3 for examples of the following.

- **Label noise:** There is a known discrepancy between label accuracy of volunteer citizen scientists vs. experts (e.g., with a PhD in Ecology) [12], and there is also no ecologically-agreed-upon definition of tree vs. bush. Cities differ in their labeling protocol.

- **Presence noise:** Tree records in censuses can often be outdated. Specifically, depending on the amount of time since the data was originally collected, there is increasing possibility that trees will have been removed or have died, and new trees planted.
- **Location noise:** Different cities use different data collection protocols and different sensors, leading to discrepancy in the accuracy of the position readings (e.g., by GPS). We estimate visually that they are usually accurate within ~3 meters.
- **Image quality:** Quality of aerial imagery varies for different cities. The primary tree in a street level image can sometimes be occluded–though we try to guard against this by removing images that are too far from the tree, sometimes vehicles block the tree from view. Qualitatively, access to multiple views frequently helps mitigate occlusion issues. Finally, deciduous trees vary in appearance across seasons, with leaves turning color and then dropping in the winter.
- **Unlabeled visible trees:** Trees on private property (e.g., yards) are not labeled in public censuses, but are visible in the background. While the tree of interest is often the most prominent, the presence of trees of other genera can create classification confusion.

Distribution shift and the long tail. One of our primary challenges is to be able to do well on novel cities that were not part of the training set, but in order for a model to do so, it will have to contend with distribution shift, where the training distribution of cities differs from the novel test distribution on some new city. We remark that there are two kinds of shift that we observe in our data — what we might call "label shift," and "appearance shift." Label shift refers to when the marginal distribution $P(y)$ of labels (genera) differs from city to city even if the appearance distribution of image x conditioned on a particular label $P(x|y)$ does not change (e.g., [90]). In our setting this simply can mean that species distributions vary geographically (e.g., we tend to see Palm trees in Southern California and less in Canada), but can also come from cities having different sizes (for example, Los Angeles is much larger than Santa Monica and thus contains many more species).

Figure 8.4 visualizes the distribution shift between every pair of cities (using L_1 distance between normalized genus distributions). In some cases we can see little overlap between genera from two cities, and for cities with similar location, i.e. Denver and Boulder, we tend to see high overlap in genus distribution. However even when two cities are very similar both in size and location, it is still generally the case that one city will contain a number of genera not found in the other due to

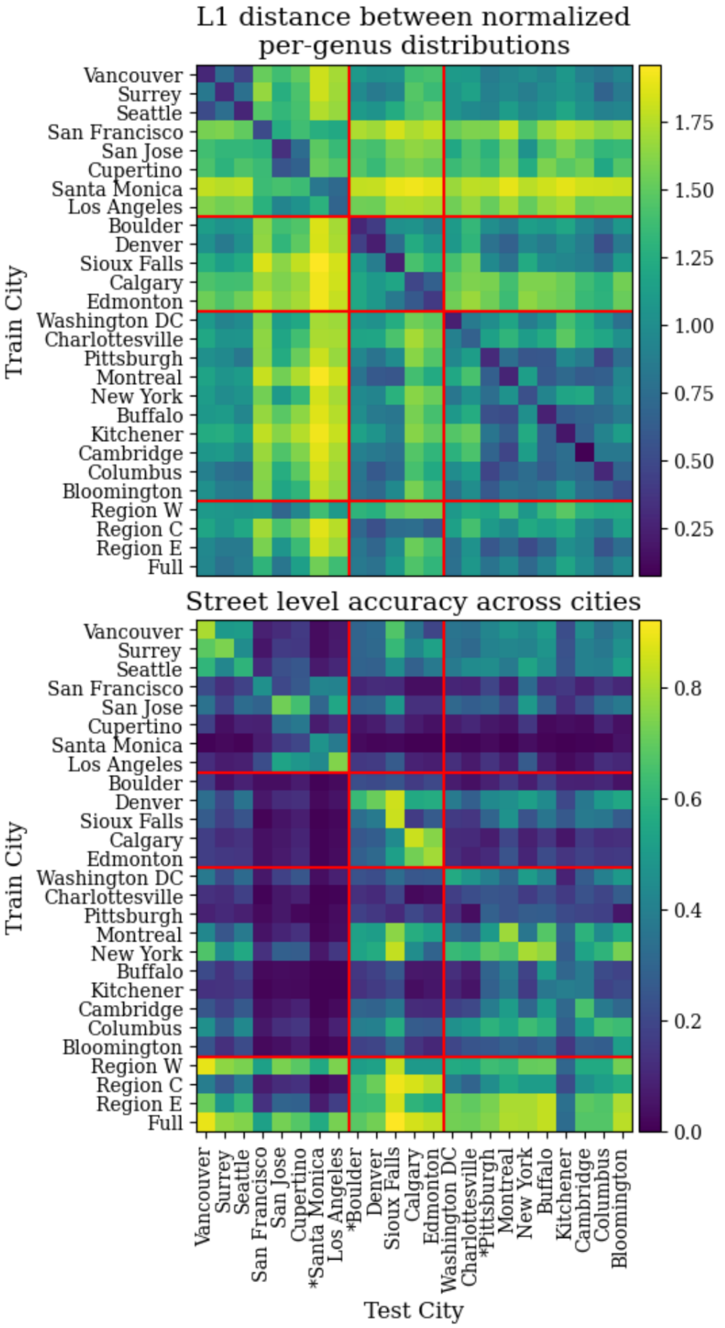


Figure 8.4: **(Top)** The distance between the distributions of training and test data for each training split and each test city (red lines represent regional boundaries). We use the L1 distance between the normalized per-class count vectors for each set as our measure of distributional distance. Because the class distribution is long-tailed and our test sets are split geographically within each city to prevent data poisoning, the train and test distribution are not identical within each city (the diagonal is non-uniform, and the matrix is not symmetric). **(Bottom)** Pairwise train/test accuracy from street level baselines.

the long tailed genus distribution. In the extreme setting of "train on one city, test on one city" we thus always have many test genera for which there are no training examples. And even in the regime of training on many cities and testing on a single holdout city, we *still* typically have classes for which there are no training examples, implying value in expanding the dataset in future.

Beyond label shift, we also see "appearance shift" — the images of a particular genus can look different depending on the city. This is partly due to different backgrounds (which can in principle be handled by masking out the background pixels, but is out of scope for this work), but it can also be due to other external factors such as weather conditions (for example, we are likely to see more leafless trees from images in Edmonton than we are to see them in LA) or even "terroir" related factors like soil composition.

8.5 Evaluation protocol

Since distribution shift is such a big factor in performance, we have chosen to set up our evaluation protocol to explicitly evaluate distribution shift based on 3 unique types of train/test splits, defined hierarchically:

1. **Per-city splits**: At the first level, we are interested in how well a city generalizes to itself. Here, each city has a defined training region and a defined test region, split geographically (usually based on latitude or longitude) to avoid overfitting on background characteristics. The test sets for each city are never used for training.
2. **Regional splits**: Next, we are interested in generalization within and across larger regions (e.g., how would we fare in cities on the East coast if we trained on West coast trees?)–for this level of evaluation, we split the cities into three regions, *Region W* (West), *Region C* (Central), and *Region E* (East) (Table 8.2). We build our regional training sets from the per-city training sets for that region. We hold out one city from each region (which we call "holdout cities") to capture performance on an in-region novel city, and also show results on all out-of-region cities.
3. **Full dataset**: For the final and largest split, we combine training data across the three regions. We maintain the same holdout cities as the regional splits for training, and test on the test sets of all cities (including the holdouts).

Evaluation metrics. Due to the long-tailed distribution of the data across genera, a pure accuracy measure is insufficient to capture performance, as it is highly biased towards frequent species. Thus, we report accuracy alongside class-averaged recall (AR), calculated as average over all classes of the proportion of correct predictions for the set of examples of that class (this is sometimes also called class-averaged accuracy). To capture performance in a more nuanced way, we also introduce an LVIS [58]-inspired breakdown of class-averaged recall for frequent ($n \geq 20,000$ examples), common ($100 \leq n < 20,000$ examples) and rare ($n < 100$ examples) subsets of our data. This results in 29 frequent, 150 common, and 165 rare genera, and we denote these metrics as FAR, CAR, and RAR respectively.

8.6 Experiments

We now demonstrate the benefits of having a multi-city, multiview dataset by training models on Auto Arborist. In this section we train separate aerial and street level baseline ResNet 101 models for each training split described in Section 8.5, including the training sets for each individual city, the regional splits, and the full dataset. Training details can be found in Appendix D.

Single city vs. regional vs. full dataset training. We begin by experimenting with single-view street level models (as the street level modality gives the most accurate results in isolation). In Figure 8.5, we compare performance on a city's test set when training on that city's training set (***city***), the aggregation of training sets from that city's region (***region***), and all available training data (***full***). Unsurprisingly, we find that more data is better–we see an average improvement of 21.3% AR across cities when going from training on a single city to the full dataset. However we note that training on a region also gives strong performance gains over training on a city itself (average improvement of 18.3% AR), and for some test cities regional training can be on par with (or even slightly better than) training on the full set.

Cross-city generalization. Next we examine cross-city generalization, where we are interested in how effective it would be to train on a certain city A if we are interested in testing on B. For this analysis, we first perform all possible *cross training* combinations, training on every train split (including per-city, regional and full) and testing each model on the test set for each city. Results for these pairwise combinations are visualized in Figure 8.4 (bottom). Here we see regional "blocks" of strong generalization, reflecting that cities generalize well to cities in the same geographical area. For example, we tend to get good performance training on one of the Pacific Northwest cities (Seattle, Vancouver, Surrey) and testing on another.

We can also see that some cities tend to generalize quite well to other cities on average whereas some cities tend to generalize poorly to other cities. Figure 8.6 shows this effect in more detail–here we use a given city as a training set and report the spread of performance when applied to other cities' test sets. In this plot, a larger gap between "self-test" (red stars) and the box implies less generalizability. Here, to remove confounding factors due to test genera not seen or rarely seen during training, we restrict computation of AR for train city A and test city B to only the "frequent" genera seen in the train split of A and the test split of B.

We observe that cities that are poor "training cities" (on the left side of Fig. 8.6) tend to be smaller and have poor performance overall, though this is not universally true (consider San Francisco). On the other hand, large cities (e.g., NYC) tend to generalize well on average. But we also see that there are no cities which generalize optimally to all others, and optimal generalization performance is only reached by training on the full dataset. Even restricting our attention to frequent, shared classes, we find that generalization ability continues to be highly correlated with label distribution similarity. In Figure 8.7, we compare AR across these shared, frequent genera with the L1 distribution distance for three cities and show they are negatively correlated–increased label distribution distance implies worse performance, even on frequently-seen classes shared between train and test cities.

Value-add of multiple views. Finally, in Table 8.3, we show that the multiview aspects of our dataset bring value. Overall, our street level models perform much better than the aerial models, generally with a difference of >20% AR and we see that using multiple views of a tree outperforms a single view. We have experimented with several techniques to combine information across street level views and aerial imagery, and find that while most of the predictive value comes from the street level imagery, there is benefit in incorporating aerial information. We combine the modalities via a simple method: average pooling the logits from multiple street level images and then combining with aerial logits via a Mixture of Experts (MoE) model:

$$f(x_{SL}, x_A) = x_{SL} \cdot \text{sigmoid}(w) + x_A \cdot (1 - \text{sigmoid}(w)). \tag{8.1}$$

where x_{SL} and x_A are street level and aerial logits, $w \in \mathbb{R}^n$ are learned parameters, and n is the number of classes.

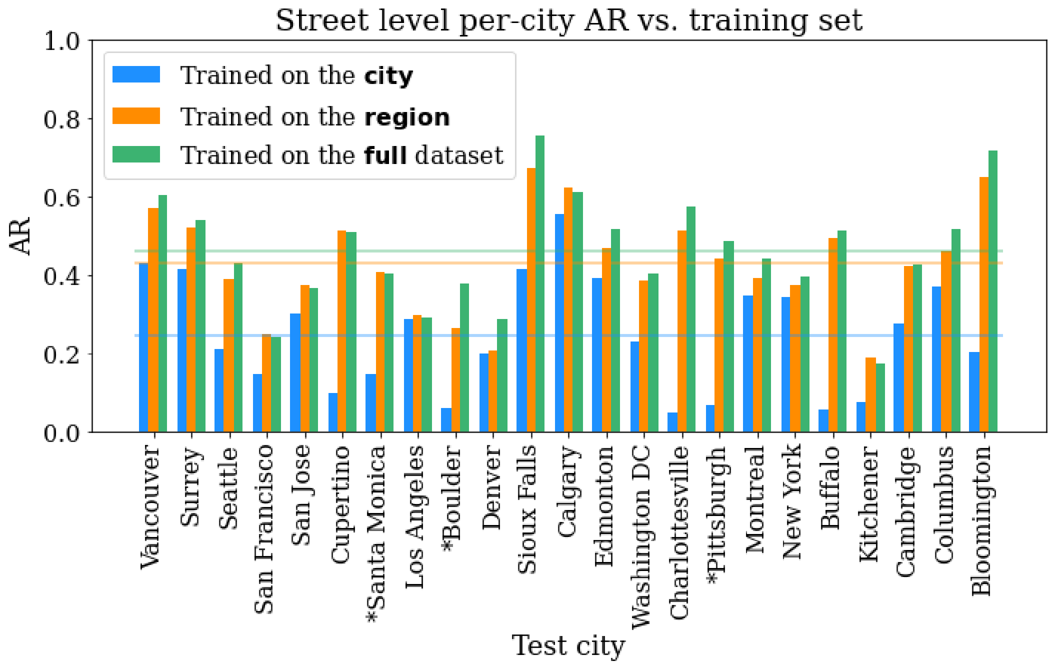


Figure 8.5: Performance growth when adding regional and continental diversity. For each city, we show test performance from a model trained on that city, trained on the respective region for that city, and trained on the full dataset. Note that performance improves on our holdout cities as well, despite the regional and full training sets not including data from those cities. Average performance from models trained on per-city, regional, and full are shown as horizontal lines.

Combining modalities in this way yields an average $\sim$ 1% boost for each regional model, compared to average-pooling logits across multiple street level views, and a 3-5% boost over predicting from a single street level image. For the full training set, we find that preserving the regional variations in learned MoE weights (w) is important–thus our best model (which achieves 49.96% AR) uses street level and aerial models trained on the full dataset but MoE weights specialized for the region to which a city belongs. We conjecture that this regional dependence is due primarily to regional variations in aerial image quality/availability. In Fig. 8.8, we visualize the per-genus weights learned by the MoE per region. Looking more closely at the MoE weights, we find that our models only assigns nonzero weights to aerial data for classes that have ≥400 training examples. Moreover, we see that we are able to rely on aerial images more in Region W compared to the other two regions.

We show results from our best model in Table 8.4, reporting accuracy and AR for the full dataset, and broken down by frequent and common genera. Notably, many cities have >80% accuracy, and Vancouver and Sioux Falls see >90%. There is

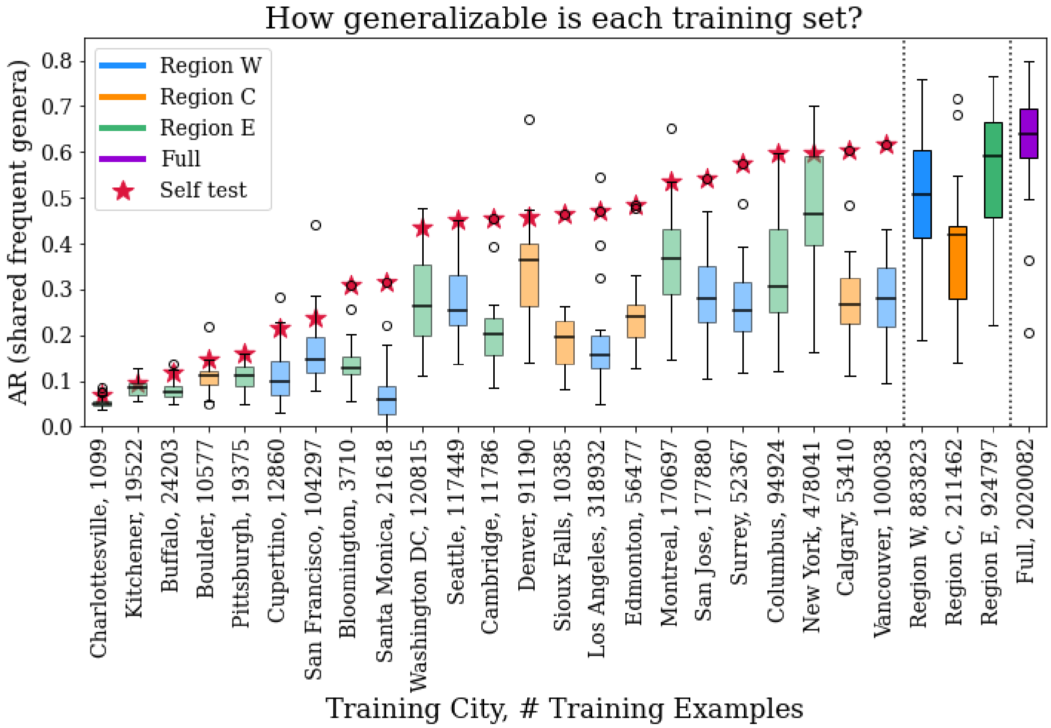


Figure 8.6: For each training set, we show the distribution of AR across the test cities, and highlight the "self test" case where a city is tested on its own test set. A larger gap between "self test" and the box implies less generalizability. Here, to remove confounding factors due to test genera not seen or rarely seen during training, we compute AR for train city A and test city B to be the average per-genus recall across the "frequent" genera seen in both train(A) and test(B).

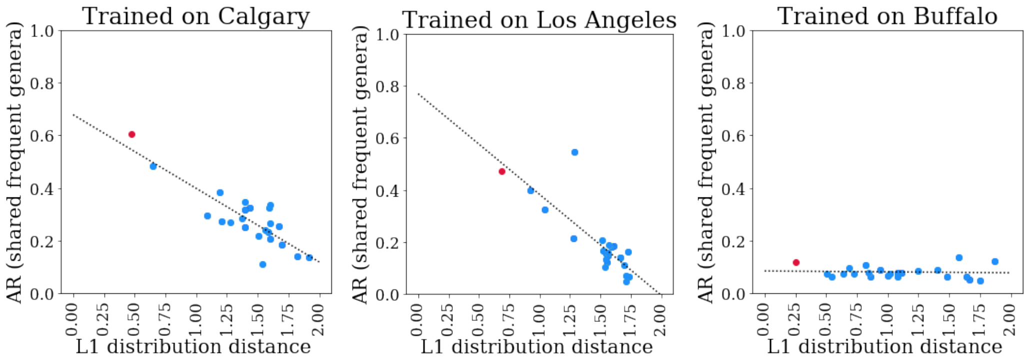


Figure 8.7: Digging further into the generalizability of a given training set, we visualize the generalization gap between AR on shared, frequent genera testing on that same city ("self-test," in red) vs other cities, and plot against the L1 distance between the genus distribution of train vs. test, as seen in Fig. 8.4. We see that frequently these are anti-correlated, but training sets for some cities (like Buffalo) struggle to perform well across the board.

Train Set	Aerial	1 SL	3 SL	A+SL
Region W	20.63	41.53	45.12	**46.07**
Region C	18.8	44.77	46.91	**47.12**
Region E	17.54	43.25	45.13	**46.21**
Full	18.7	46.13	49.0	**49.23**
Full w/ Regional MoE				**49.96**

Table 8.3: City-averaged percent AR for different regions and ensembling strategies. Street level imagery is much more informative than aerial, and combining multiple street level images gives a further boost. However, even though aerial performance on its own is quite low, we see benefit in adding the aerial imagery when making predictions. We find that while the features from the full model are more discriminative, we see best performance using full model features paired with region-specific Mixture of Experts to combine aerial and street level predictions.

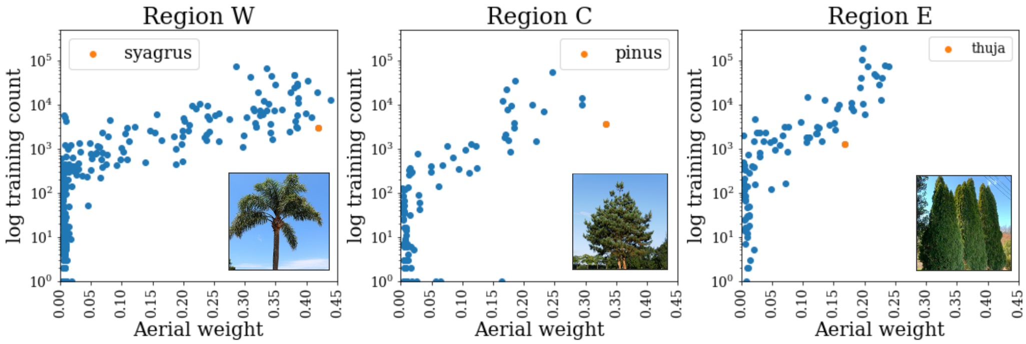


Figure 8.8: Each regional MoE learns to use aerial information only for genera with more than ∼400 training examples. Notably, the distribution of the three is quite different, and there are certain genera that are more “aerially distinctive” (we have highlighted one for each region).

still significant room to improve on AR across the board. Rare class performance was 0.0 for every city, unsurprising given most rare classes have <10 examples. This points to potential gain from low-shot and long-tail learning methods such as logit-adjustment, but we find that such methods struggle to perform well under such a high degree of imbalance (see Supplementary).

8.7 Limitations and future work

We have presented a baseline modeling approach meant to highlight the performance of a typical CNN and present simple methods for combining signals from multiple views — there is much room for improvement, particularly on rare classes. In future, to predict on cities with no past census, we would need to first localize and geocode the trees to be classified. We also hope to expand our dataset to include more cities,

City	Acc	AR	FAR	CAR
Vancouver, BC	93.28	67.51	82.76	63.35
Surrey, BC	82.35	58.96	75.82	48.80
Seattle, WA	79.68	46.55	74.65	43.08
San Francisco, CA	58.71	26.39	37.87	31.37
San Jose, CA	77.71	40.07	63.35	41.13
Cupertino, CA	74.14	56.86	65.28	55.40
Santa Monica, CA	56.26	43.29	64.93	44.09
Los Angeles, CA	76.24	32.62	52.56	35.80
Boulder City, CO	73.23	42.23	58.61	32.88
Denver, CO	76.46	29.72	57.16	22.02
Sioux Falls, SD	93.78	76.76	81.52	62.50
Calgary, AB	88.81	62.18	70.92	52.32
Edmonton, AB	87.55	56.67	62.58	43.99
Washington, DC	77.44	44.49	67.31	30.05
Charlottesville, VA	73.52	57.77	73.38	42.16
Pittsburgh, PA	78.84	54.93	71.83	43.97
Montreal, QC	85.51	49.49	64.99	39.08
New York, NY	82.54	42.77	66.38	28.16
Buffalo, NY	86.03	54.01	71.92	43.41
Kitchener, ON	33.96	17.94	21.31	4.49
Cambridge, ON	72.16	47.69	65.84	34.38
Columbus, OH	69.28	55.71	68.29	47.32
Bloomington, IN	85.50	73.52	79.82	64.46

Table 8.4: Per-city performance (%) with our best model trained on the full dataset combining aerial and multiview street level modalities. AR is class-averaged recall for each city, averaged over the test classes for that city. FAR is “Frequent” AR, CAR is “Common” AR, which serve to further disentangle the commonality of a species in the training data with its per-city performance. Holdout cities in bold.

both in North America and worldwide, and include species level predictions and additional features such as tree size and health.

Auto Arborist represents an important first step towards global-scale urban forest monitoring. This has implications for environmental justice: given that marginalized communities have less access to urban greenery, systems trained on Auto Arborist could help equitize access to urban forests by empowering quantifiable analysis and targeted replanting. However we must be responsible with our technology — to this end, we protect the privacy of residents of these urban and suburban areas by explicitly filtering out any imagery containing humans, and blur vehicle plates. Secondly, we will need efficient human-in-the-loop validation protocols before such a system could be trusted, to ensure science policy is not based on poorly-generalized ML predictions.

8.8 Seeing the forest for the trees (Conclusions)

Climate change and loss of ecological diversity are among the most pressing issues of our time. Monitoring is a first crucial step to understanding and mitigating the effects of global warming on urban forests, but many cities cannot afford regular tree censuses. Towards the goal of broad, accessible, and affordable urban forest monitoring, we have introduced the Auto Arborist dataset. This dataset is the first of its kind to expand beyond a single city and common categories: Auto Arborist contains 2.6 million trees across 23 cities, covering 344 unique genera. This dataset will enable the computer vision community to tackle urban forest monitoring at scale, and our evaluation protocols help us measure performance without data poisoning, and to evaluate generalization to novel cities.

Acknowledgements

We would like to thank Ruth Alcantara, Tanya Birch, and Dan Morris from Google AI for Nature and Society, John Quintero, Stafford Marquardt, Tim Taylor, Jacqueline Hea, Xiaoqi Yin, Puneet Lall, and Matt Manolides from Google Geo, Janel Thamkul, Karan Gill, Tom Duerig, Abhijit Kundu, David Ross, Vighnesh Birodkar and the Visual Dynamics team from Google Perception, and Pietro Perona for their support. This work was supported in part by the Resnick Sustainability Institute and was undertaken while Beery was a Student Researcher at Google.

References

[1] iNaturalist. `https://www.inaturalist.org`. Accessed: 2021-11-11.

[2] Samuel Adelabu, Onisimo Mutanga, Elhadi E. Adam, and Moses Azong Cho. Exploiting machine learning algorithms for tree species classification in a semiarid woodland using rapideye image. *Journal of Applied Remote Sensing*, 7(1):073480, 2013.

[3] Michael Alonzo, Bodo Bookhagen, and Dar A. Roberts. Urban tree species mapping using hyperspectral and lidar data fusion. *Remote Sensing of Environment*, 148:70–83, 2014.

[4] Mike Alonzo, Keely Roth, and Dar A. Roberts. Identifying Santa Barbara's urban tree species from AVIRIS imagery using canonical discriminant analysis. *Remote Sensing Letters*, 4(5):513–521, 2013.

[5] Kenza Amara, David Dao, Björn Lütjens, Dava Newman, Tom Crowther, and Ce Zhang. One forest: Towards a global species dataset by fusing remote sensing and citizen science data with graph neural networks. *Fragile Earth Workshop at KDD*, 2020.

[6] Daniel Amigo, David Sánchez Pedroche, Jesús García, and José M. Molina. Automatic individual tree detection from combination of aerial imagery, lidar and environment context. In *International Workshop on Soft Computing Models in Industrial and Environmental Applications*, pages 294–303. Springer, 2021.

[7] Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. Google street view: Capturing the world at street level. *Computer*, 43(6):32–38, 2010.

[8] Emeka C. Anyanwu and Ian Kanu. The role of urban forest in the protection of human environmental health in geographically-prone unpredictable hostile weather conditions. *International Journal of Environmental Science & Technology*, 3(2):197–201, 2006.

[9] Mélaine Aubry-Kientz, Anthony Laybros, Ben Weinstein, James GC Ball, Toby Jackson, David Coomes, and Grégoire Vincent. Multisensor data fusion for improved segmentation of individual tree crowns in dense tropical forests. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3927–3936, 2021.

[10] Bulut Aygüneş, Selim Aksoy, and Ramazan Gökberk Cinbiş. Weakly supervised deep convolutional networks for fine-grained object recognition in multispectral images. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 1478–1481. IEEE, 2019.

[11] Bulut Aygunes, Ramazan Gokberk Cinbis, and Selim Aksoy. Weakly supervised instance attention for multisource fine-grained object recognition with an application to tree species classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176:262–274, 2021.

[12] Nick Bancks, Eric A. North, and Gary R. Johnson. An analysis of agreement between volunteer-and researcher-collected urban tree inventory data. 2018.

[13] Sara Beery, Grant Van Horn, Oisin MacAodha, and Pietro Perona. The iWildCam 2018 challenge dataset. *The Fifth Fine-Grained Visual Categorization Workshop at CVPR*, 2018.

[14] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[15] Sara Beery, Dan Morris, and Pietro Perona. The iWildCam 2019 challenge dataset. *The Sixth Fine-Grained Visual Categorization Workshop at CVPR*, 2019.

[16] Sara Beery, Elijah Cole, and Arvi Gjoka. The iWildCam 2020 competition dataset. *The Seventh Fine-Grained Visual Categorization Workshop at CVPR*, 2020.

[17] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iWildCam 2021 competition dataset. *The Eighth Fine-Grained Visual Categorization Workshop at CVPR*, 2021.

[18] Even Bergseng, Hans Ole Ørka, Erik Næsset, and Terje Gobakken. Assessing forest inventory information obtained from different inventory approaches and remote sensing data sources. *Annals of Forest Science*, 72(1):33–45, 2015.

[19] Filip Biljecki and Koichi Ito. Street view imagery in urban analytics and gis: A review. *Landscape and Urban Planning*, 215:104217, 2021.

[20] John Blum. Contribution of ecosystem services to air quality and climate change mitigation policies: the case of urban forests in barcelona, spain. In *Urban Forests*, pages 21–54. Apple Academic Press, 2017.

[21] Bob Bolin, Sara Grineski, and Timothy Collins. The geography of despair: Environmental racism and the making of south phoenix, arizona, usa. *Human Ecology Review*, pages 156–168, 2005.

[22] Christopher G. Boone, Geoffrey L. Buckley, J. Morgan Grove, and Chona Sister. Parks and people: An environmental justice inquiry in Baltimore, Maryland. *Annals of the Association of American Geographers*, 99(4):767–787, 2009.

[23] Mirco Boschetti, Luigi Boschetti, Simone Oliveri, Luigi Casati, and Ian Canova. Tree species mapping with airborne hyper-spectral mivis data: the ticino park study case. *International Journal of Remote Sensing*, 28(6): 1251–1261, 2007.

[24] Leslie Brandt, Abigail Derby Lewis, Robert Fahey, Lydia Scott, Lindsay Darling, and Chris Swanston. A framework for adapting urban forests to climate change. *Environmental Science & Policy*, 66:393–402, 2016.

[25] Martin Brandt, Compton J. Tucker, Ankit Kariryaa, Kjeld Rasmussen, Christin Abel, Jennifer Small, Jerome Chave, Laura Vang Rasmussen, Pierre Hiernaux, Abdoul Aziz Diouf, et al. An unexpectedly large count of trees in the west african sahara and sahel. *Nature*, 587(7832):78–82, 2020.

[26] Tomas Brandtberg. Classifying individual tree species under leaf-off and leaf-on conditions using airborne lidar. *ISPRS Journal of Photogrammetry and Remote Sensing*, 61(5):325–340, 2007.

[27] Steve Branson, Jan Dirk Wegner, David Hall, Nico Lang, Konrad Schindler, and Pietro Perona. From Google Maps to a fine-grained catalog of street trees. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135:13–30,

Jan 2018. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2017.11.008. URL `http://dx.doi.org/10.1016/j.isprsjprs.2017.11.008`.

[28] Gregory N. Bratman, Christopher B. Anderson, Marc G. Berman, Bobby Cochran, Sjerp De Vries, Jon Flanders, Carl Folke, Howard Frumkin, James J. Gross, Terry Hartig, et al. Nature and mental health: An ecosystem service perspective. *Science Advances*, 5(7):eaax0903, 2019.

[29] Dick J. Brus, Geerten M. Hengeveld, Dennis Walvoort, Paul W. Goedhart, Nanny A.H. Heidema, Gurt-Jan Nabuurs, and Karl Gunia. Statistical mapping of tree species over europe. *European Journal of Forest Research*, 131(1): 145–157, 2012.

[30] Alexandre Carleer and Eléonore Wolff. Exploitation of very high resolution satellite data for tree species identification. *Photogrammetric Engineering & Remote Sensing*, 70(1):135–140, 2004.

[31] Dominic Chambers, Catherine Périé, Nicolas Casajus, and Sylvie de Blois. Challenges in modelling the abundance of 105 tree species in eastern north america using climate, edaphic, and topographic variables. *Forest Ecology and Management*, 291:20–29, 2013.

[32] Francesco Chianucci, Andrea Cutini, Piermaria Corona, and Nicola Puletti. Estimation of leaf area index in understory deciduous trees using digital photography. *Agricultural and Forest Meteorology*, 198:259–264, 2014.

[33] Moses Azong Cho, Pravesh Debba, Renaud Mathieu, Laven Naidoo, J.A.N. Van Aardt, and Gregory P. Asner. Improving discrimination of savanna tree species through a multiple-endmember spectral angle mapper approach: Canopy-level analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 48(11):4133–4142, 2010.

[34] Moses Azong Cho, Renaud Mathieu, Gregory P. Asner, Laven Naidoo, J.A.N. Van Aardt, Abel Ramoelo, Pravesh Debba, Konrad Wessels, Russell Main, Izak P.J. Smit, et al. Mapping tree species composition in south african savannas using an integrated airborne spectral and lidar system. *Remote Sensing of Environment*, 125:214–226, 2012.

[35] Matthew L. Clark and Dar A. Roberts. Species-level differences in hyperspectral metrics among tropical rainforest trees as determined by a tree-based classifier. *Remote Sensing*, 4(6):1820–1855, 2012.

[36] Matthew L. Clark, Dar A. Roberts, and David B. Clark. Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. *Remote Sensing of Environment*, 96(3-4):375–398, 2005.

[37] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele.

The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[38] Thomas W. Crowther, Henry B. Glick, Kristofer R. Covey, Charlie Bettigole, Daniel S. Maynard, Stephen M. Thomas, Jeffrey R. Smith, Gregor Hintler, Marlyse C. Duguid, Giuseppe Amatulli, et al. Mapping tree density at a global scale. *Nature*, 525(7568):201–205, 2015.

[39] María Culman, Andrés C Rodríguez, Jan Dirk Wegner, Stephanie Delalieux, and Ben Somers. Deep learning for sub-pixel palm tree classification using spaceborne sentinel-2 imagery. In *Remote Sensing for Agriculture, Ecosystems, and Hydrology XXIII*, volume 11856, pages 45–50. SPIE, 2021.

[40] Darius S. Culvenor. A spatial clustering approach to automated tree crown delineation. In *Proceedings of the forum on automated interpretation of high spatial resolution digital imagery for forestry, 1999*, pages 67–88. Canadian Forest Service, 1999.

[41] Michele Dalponte, Lorenzo Bruzzone, and Damiano Gianelle. Tree species classification in the southern alps based on the fusion of very high geometrical resolution multispectral/hyperspectral images and lidar data. *Remote Sensing of Environment*, 123:258–270, 2012.

[42] Michele Dalponte, Hans Ole Ørka, Liviu Theodor Ene, Terje Gobakken, and Erik Næsset. Tree crown delineation and tree species classification in boreal forests using hyperspectral and als data. *Remote Sensing of Environment*, 140:306–317, 2014.

[43] Hubert de Foresta, Eduardo Somarriba, August Temu, Désirée Boulanger, Hélène Feuilly, Michelle Gauthier, and D Taylor. Towards the assessment of trees outside forests: A thematic report prepared in the framework of the global forest resources assessment 2010. *World*, 7, 2020.

[44] Steve H. Dreistadt, Donald L. Dahlsten, and Gordon W. Frankie. Urban forests and insect ecology. *BioScience*, 40(3):192–198, 1990.

[45] Fábio Duarte and Carlo Ratti. What big data tell us about trees and the sky in the cities. In *Humanizing Digital Reality*, pages 59–62. Springer, 2018.

[46] Laura Duncanson and Ralph Dubayah. Monitoring individual tree-based change with airborne lidar. *Ecology and Evolution*, 8(10):5079–5089, 2018.

[47] Theodore S. Eisenman, Galina Churkina, Sunit P. Jariwala, Prashant Kumar, Gina S. Lovasi, Diane E. Pataki, Kate R. Weinberger, and Thomas H. Whitlow. Urban trees, air quality, and asthma: An interdisciplinary review. *Landscape and urban planning*, 187:47–59, 2019.

[48] Robin Engler, Lars T. Waser, Niklaus E. Zimmermann, Marcus Schaub, Savvas Berdos, Christian Ginzler, and Achilleas Psomas. Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. *Forest Ecology and Management*, 310:64–73, 2013.

[49] Fabian Ewald Fassnacht, Hooman Latifi, Krzysztof Stereńczak, Aneta Modzelewska, Michael Lefsky, Lars T. Waser, Christoph Straub, and Aniruddha Ghosh. Review of studies on tree species classification from remotely sensed data. *Remote Sensing of Environment*, 186:64–87, 2016.

[50] Angelos Filos, Panagiotis Tigkas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2020.

[51] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[52] Ed Gerrish and Shannon Lea Watkins. The relationship between urban forests and income: A meta-analysis. *Landscape and Urban Planning*, 170:293–308, 2018.

[53] Azadeh Ghiyamat, Helmi Zulhaidi M Shafri, Ghafour Amouzad Mahdiraji, Abdul Rashid M Shariff, and Shattri Mansor. Hyperspectral discrimination of tree species with different classifications using single-and multiple-endmember. *International Journal of Applied Earth Observation and Geoinformation*, 23:177–191, 2013.

[54] Peng Gong, Ruiliang Pu, and Bin Yu. Conifer species recognition: An exploratory analysis of in situ hyperspectral data. *Remote Sensing of Environment*, 62(2):189–200, 1997.

[55] François A Gougeon. Comparison of possible multispectral classification schemes for tree crowns individually delineatedon high spatial resolution meis images. *Canadian Journal of Remote Sensing*, 21(1):1–9, 1995.

[56] Sarah J. Graves, Sergio Marconi, Dylan Stewart, Ira Harmon, Ben G. Weinstein, Yuzi Kanazawa, Victoria M. Scholl, Maxwell B. Joseph, Joseph McClinchy, Like Browne, et al. Data science competition for cross-site delineation and classification of individual trees from airborne remote sensing data. *bioRxiv*, 2021.

[57] Haiyan Guan, Yongtao Yu, Zheng Ji, Jonathan Li, and Qi Zhang. Deep learning-based tree classification using mobile lidar data. *Remote Sensing Letters*, 6(11):864–873, 2015.

[58] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019.

[59] Arto Haara and Mika Haarala. Tree species classification using semi-automatic delineation of trees on aerial images. *Scandinavian Journal of Forest Research*, 17(6):556–565, 2002.

[60] Hamid Hamraz, Marco A Contreras, and Jun Zhang. Forest understory trees can be segmented accurately within sufficiently dense airborne laser scanning point clouds. *Scientific Reports*, 7(1):1–9, 2017.

[61] Hamid Hamraz, Nathan B Jacobs, Marco A Contreras, and Chase H Clark. Deep learning for conifer/deciduous classification of airborne lidar 3d point clouds representing individual trees. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158:219–230, 2019.

[62] Matthew C Hansen, Peter V Potapov, Rebecca Moore, Matt Hancher, Svetlana A Turubanova, Alexandra Tyukavina, David Thau, Stephen V Stehman, Scott J Goetz, Thomas R Loveland, et al. High-resolution global maps of 21st-century forest cover change. *science*, 342(6160):850–853, 2013.

[63] Sean Hartling, Vasit Sagan, Paheding Sidike, Maitiniyazi Maimaitijiang, and Joshua Carron. Urban tree species classification using a worldview-2/3 and lidar data fusion approach and deep learning. *Sensors*, 19(6):1284, 2019.

[64] Richard J Hauer, Ian S Hanou, and David Sivyer. Planning for active management of future invasive pests affecting urban forests: the ecological and economic effects of varying dutch elm disease management practices for street trees in milwaukee, wi usa. *Urban Ecosystems*, 23(5):1005–1022, 2020.

[65] Ville Heikkinen, Timo Tokola, Jussi Parkkinen, Ilkka Korpela, and Timo Jaaskelainen. Simulated multispectral imagery for tree species classification using support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 48(3):1355–1364, 2009.

[66] Johannes Heinzel and Barbara Koch. Investigating multiple data sources for tree species classification in temperate forest and use for single tree delineation. *International Journal of Applied Earth Observation and Geoinformation*, 18:101–110, 2012.

[67] Daniel Henstra. Toward the climate-resilient city: Extreme weather and urban climate adaptation policies in two canadian provinces. *Journal of Comparative Policy Analysis: Research and Practice*, 14(2):175–194, 2012.

[68] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[69] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/hoffman18a.html`.

[70] Markus Hollaus, Werner Mücke, Bernhard Höfle, Wouter Dorigo, Norbert Pfeifer, Wolfgang Wagner, Christoph Bauerhansl, and Bruno Regner. Tree species classification based on full-waveform airborne laser scanning data. *Proceedings of SILVILASER*, pages 54–62, 2009.

[71] Johan Holmgren and Åsa Persson. Identifying species of individual trees using airborne laser scanner.

[72] Johan Holmgren, Åsa Persson, and Ulf Söderman. Species identification of individual trees by combining high resolution lidar data with multi-spectral images. *International Journal of Remote Sensing*, 29(5):1537–1552, 2008.

[73] Diane Hope, Corinna Gries, Weixing Zhu, William F. Fagan, Charles L. Redman, Nancy B. Grimm, Amy L. Nelson, Chris Martin, and Ann Kinzig. Socioeconomics drive urban plant diversity. *Proceedings of the National Academy of Sciences*, 100(15):8788–8792, 2003.

[74] Aarne Hovi, Lauri Korhonen, Jari Vauhkonen, and Ilkke Korpela. Lidar waveform features for tree species classification and their sensitivity to tree- and acquisition related parameters. *Remote Sensing of Environment*, 173: 224–237, 2016.

[75] Markus Immitzer, Clement Atzberger, and Tatjana Koukal. Tree species classification with random forest using very high spatial resolution 8-band worldview-2 satellite data. *Remote Sensing*, 4(9):2661–2693, 2012.

[76] Kenta Itakura and Fumiki Hosoi. Automatic tree detection from three-dimensional images reconstructed from 360 spherical camera using yolo v2. *Remote Sensing*, 12(6):988, 2020.

[77] Marek K. Jakubowski, Wenkai Li, Qinghua Guo, and Maggi Kelly. Delineating individual trees from lidar data: A comparison of vector-and raster-based segmentation approaches. *Remote Sensing*, 5(9):4163–4186, 2013.

[78] Ryan R. Jensen, Perry J. Hardin, and Andrew J. Hardin. Classification of urban tree species using hyperspectral imagery. *Geocarto International*, 27 (5):443–458, 2012.

[79] Thomas Key, Timothy A. Warner, James B. McGraw, and Mary Ann Fajvan. A comparison of multispectral and multitemporal information in high spatial

resolution imagery for classification of individual tree species in a temperate hardwood forest. *Remote Sensing of Environment*, 75(1):100–112, 2001.

[80] Sooyoung Kim. *Individual tree species identification using LIDAR-derived crown structures and intensity data*. University of Washington, 2008.

[81] Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[82] Ilkka Korpela, Lauri Mehtätalo, Lauri Markelin, Anne Seppänen, Annika Kangas, et al. Tree species identification in aerial image data using directional reflectance signatures. *Silva Fenn*, 48(3):1–20, 2014.

[83] Daniel Laumer, Nico Lang, Natalie van Doorn, Oisin Mac Aodha, Pietro Perona, and Jan Dirk Wegner. Geocoding of trees from street addresses and street-level images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:125–136, 2020.

[84] Sébastien Lefèvre, Devis Tuia, Jan Dirk Wegner, Timothée Produit, and Ahmed Samy Nassar. Toward seamless multiview scene analysis from satellite to street level. *Proceedings of the IEEE*, 105(10):1884–1899, 2017.

[85] Misha Leong, Robert R. Dunn, and Michelle D. Trautwein. Biodiversity and socioeconomics in the city: A review of the luxury effect. *Biology Letters*, 14(5):20180082, 2018.

[86] Linyuan Li, Jun Chen, Xihan Mu, Weihua Li, Guangjian Yan, Donghui Xie, and Wuming Zhang. Quantifying understory and overstory vegetation cover using uav-based rgb imagery in forest plantation. *Remote Sensing*, 12(2):298, 2020.

[87] Xiaojiang Li. Examining the spatial distribution and temporal change of the green view index in new york city using google street view images and deep learning. *Environment and Planning B: Urban Analytics and City Science*, 48(7):2039–2054, 2021.

[88] Xiaojiang Li, Chuanrong Zhang, Weidong Li, Robert Ricard, Qingyan Meng, and Weixing Zhang. Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry & Urban Greening*, 14(3):675–685, 2015.

[89] Wenzhi Liao, Frieke Van Coillie, Lianru Gao, Liwei Li, Bing Zhang, and Jocelyn Chanussot. Deep learning for fusion of apex hyperspectral and full-waveform lidar remote sensing data for tree species mapping. *IEEE Access*, 6:68716–68729, 2018.

[90] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130. PMLR, 2018.

[91] Changfu Liu and Xiaoma Li. Carbon storage and sequestration by urban forests in shenyang, china. *Urban Forestry & Urban Greening*, 11(2):121–128, 2012.

[92] Dexter H. Locke, Shawn M. Landry, J. Morgan Grove, and Rinku Roy Chowdhury. What's scale got to do with it? Models for urban tree canopy. *Journal of Urban Ecology*, 2(1), 2016.

[93] Dexter H. Locke, Billy Hall, J. Morgan Grove, Steward T.A. Pickett, Laura A. Ogden, Carissa Aoki, Christopher G. Boone, and Jarlath P.M. O'Neil-Dunne. Residential housing segregation and urban tree canopy in 37 US cities. *npj Urban Sustainability*, 1(1):1–9, 2021.

[94] Stefanie Lumnitz, Tahia Devisscher, Jerome R. Mayaud, Valentina Radic, Nicholas C. Coops, and Verena C. Griess. Mapping trees along urban street networks with deep learning and street-level imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:144–157, 2021.

[95] Dawei Ma, Hongchao Fan, Wenwen Li, and Xuan Ding. The state of mapillary: An exploratory analysis. *ISPRS International Journal of Geo-Information*, 9(1):10, 2020.

[96] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9596–9606, 2019.

[97] Sergio Marconi, Sarah J. Graves, Dihong Gong, Morteza Shahriari Nia, Marion Le Bras, Bonnie J. Dorr, Peter Fontana, Justin Gearhart, Craig Greenberg, Dave J. Harris, et al. A data science challenge for converting airborne remote sensing data into ecological information. *PeerJ*, 6:e5843, 2019.

[98] Kendra Marshman. The eye of the storm: extreme weather events and sustainable urban forest management. *Dalhousie Journal of Interdisciplinary Management*, 14, 2018.

[99] Robert I. McDonald, Timm Kroeger, Ping Zhang, and Perrine Hamel. The value of us urban tree cover for reducing heat-related health impacts and electricity consumption. *Ecosystems*, 23(1):137–150, 2020.

[100] Ahmed Samy Nassar, Sébastien Lefèvre, and Jan Dirk Wegner. Simultaneous multi-view instance detection with learned geometric soft-constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6559–6568, 2019.

[101] Ahmed Samy Nassar, Stefano D'aronco, Sébastien Lefèvre, and Jan D Wegner. Geograph: Graph-based multi-view object detection with geometric cues end-to-end. In *European Conference on Computer Vision*, pages 488–504. Springer, 2020.

[102] Ahmed Samy Nassar, Sébastien Lefèvre, and Jan Dirk Wegner. Multi-view instance matching with learned geometric soft-constraints. *ISPRS International Journal of Geo-Information*, 9(11):687, 2020.

[103] Lorien Nesbitt, Michael J. Meitner, Cynthia Girling, Stephen R.J. Sheppard, and Yuhao Lu. Who has access to urban vegetation? A spatial analysis of distributional green equity in 10 us cities. *Landscape and Urban Planning*, 181:51–79, 2019.

[104] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vsion*, pages 4990–4999, 2017.

[105] David J. Nowak and Eric J. Greenfield. US urban forest statistics, values, and projections. *Journal of Forestry*, 116(2):164–177, 2018.

[106] UN. Department of Economic and Social Affairs. 2018 revision of world urbanization prospects, 2018.

[107] Jarlath O'Neil-Dunne, Sean MacFaden, and Anna Royar. A versatile, production-oriented approach to high-resolution tree-canopy mapping in urban and suburban landscapes using geobia and data fusion. *Remote Sensing*, 6(12):12837–12865, 2014.

[108] Camilo Ordóñez and Peter N. Duinker. Assessing the vulnerability of urban forests to climate change. *Environmental Reviews*, 22(3):311–321, 2014.

[109] Alin-Ionuț Pleșoianu, Mihai-Sorin Stupariu, Ionuț Șandric, Ileana Pătru-Stupariu, and Lucian Drăguț. Individual tree-crown detection and species classification in very high-resolution remote sensing imagery using a deep learning ensemble model. *Remote Sensing*, 12(15):2426, 2020.

[110] Eetu Puttonen, Paula Litkey, and Juha Hyyppä. Individual tree species classification by illuminated—shaded area separation. *Remote Sensing*, 2(1): 19–35, 2010.

[111] Ilya Sherstyuk Sergio Parra Rebekah Loving, Arushi Agarwal. A network fusion model pipeline for multi-modal, deep learning for tree crown detection. 2020.

[112] Walter G. Rhode. Multisectral sensing of forest tree species. *Photogrammetric Engineering*, 38(12), 1972.

[113] Anastasiia Safonova, Siham Tabik, Domingo Alcaraz-Segura, Alexey Rubtsov, Yuriy Maglinets, and Francisco Herrera. Detection of fir trees (abies sibirica) damaged by the bark beetle in unmanned aerial vehicle images with deep learning. *Remote Sensing*, 11(6):643, 2019.

[114] Frank S. Santamour Jr. Trees for urban planting: diversity uniformity, and common sense. *C. Elevitch, The Overstory Book: Cultivating connections with trees*, pages 396–399, 2004.

[115] Lauren Sayn-Wittgenstein. Recognition of tree species on aerial photographs. *Rapport d'Information (SCF - Ottawa)*, 1978.

[116] Sebastian Schnell, Christoph Kleinn, and Göran Ståhl. Monitoring trees outside forests: A review. *Environmental monitoring and assessment*, 187 (9):1–17, 2015.

[117] Kirsten Schwarz, Michail Fragkias, Christopher G. Boone, Weiqi Zhou, Melissa McHale, J. Morgan Grove, Jarlath O'Neil-Dunne, Joseph P. McFadden, Geoffrey L. Buckley, Dan Childers, et al. Trees grow on money: Urban tree canopy cover and environmental justice. *PloS One*, 10(4):e0122051, 2015.

[118] Ian Seiferling, Nikhil Naik, Carlo Ratti, and Raphäel Proulx. Green streets-quantifying and mapping urban trees with street-level imagery and computer vision. *Landscape and Urban Planning*, 165:93–101, 2017.

[119] Kurt Shickman. Cool policies for cool cities: Best practices for mitigating urban heat islands in North American cities. *Global Cool Cities Alliance*, 2014.

[120] Greg Spotts. LA begins massive street tree census. `https://postcarboncity.wordpress.com/2019/10/28/la-begins-massive-street-tree-census/`. Accessed: 2010-09-30.

[121] Nathan L. Stephenson, Adrian J. Das, Richard Condit, Sabrina E. Russo, Patrick J. Baker, Noelle G. Beckman, David A. Coomes, Emily R. Lines, et al. Rate of tree carbon accumulation increases continuously with tree size. *Nature*, 507(7490):90–93, 2014.

[122] Dylan Stewart, Alina Zare, Sergio Marconi, Ben G. Weinstein, Ethan P. White, Sarah J. Graves, Stephanie Ann Bohlman, and Aditya Singh. Randcrowns: A quantitative metric for imprecisely labeled tree crown delineation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.

[123] Esra Suel, John W Polak, James E Bennett, and Majid Ezzati. Measuring social, environmental and health inequalities using deep learning and street imagery. *Scientific Reports*, 9(1):1–10, 2019.

[124] Gencer Sumbul, Ramazan Gokberk Cinbis, and Selim Aksoy. Fine-grained object recognition and zero-shot learning in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2):770–779, 2017.

[125] Mark S. Taylor, Benedict W. Wheeler, Mathew P. White, Theodoros Economou, and Nicholas J. Osborne. Research note: Urban street tree density and antidepressant prescription rates—a cross-sectional study in london, uk. *Landscape and Urban Planning*, 136:174–179, 2015.

[126] Andrew Thirlwell and Ognjen Arandjelović. Big data driven detection of trees in suburban scenes using visual spectrum eye level photography. *Sensors*, 20 (11):3051, 2020.

[127] Dong Tianyang, Zhang Jian, Gao Sibin, Shen Ying, and Fan Jing. Single-tree detection in high-resolution remote-sensing images based on a cascade neural network. *ISPRS International Journal of Geo-Information*, 7(9):367, 2018.

[128] Lars T. Waser, Christian Ginzler, Meinrad Kuechler, Emmanuel Baltsavias, and Lorenz Hurni. Semi-automatic classification of tree species in different forest ecosystems by spectral and geometric variables derived from airborne digital sensor (ads40) and rc30 data. *Remote Sensing of Environment*, 115 (1):76–85, 2011.

[129] Shannon Lea Watkins and Ed Gerrish. The relationship between urban forests and race: A meta-analysis. *Journal of Environmental Management*, 209:152–168, 2018.

[130] Jan D. Wegner, Steve Branson, David Hall, Konrad Schindler, and Pietro Perona. Cataloging public objects using aerial and street-level images — urban trees. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6014–6023, 2016. doi: 10.1109/CVPR.2016.647.

[131] Ben G. Weinstein, Sergio Marconi, Stephanie Bohlman, Alina Zare, and Ethan White. Individual tree-crown detection in rgb imagery using semi-supervised deep learning neural networks. *Remote Sensing*, 11(11):1309, 2019.

[132] Ben G. Weinstein, Sergio Marconi, Stephanie A. Bohlman, Alina Zare, and Ethan P. White. Cross-site learning in deep learning rgb tree crown detection. *Ecological Informatics*, 56:101061, 2020.

[133] Ben G. Weinstein, Sarah J. Graves, Sergio Marconi, Aditya Singh, Alina Zare, Dylan Stewart, Stephanie A. Bohlman, and Ethan P. White. A benchmark dataset for canopy crown detection and delineation in co-registered airborne rgb, lidar and hyperspectral imagery from the national ecological observation network. *PLoS Computational Biology*, 17(7):e1009180, 2021.

[134] Ben G. Weinstein, Sergio Marconi, Stephanie A. Bohlman, Alina Zare, Aditya Singh, Sarah J. Graves, and Ethan P. White. A remote sensing derived data set of 100 million individual tree crowns for the national ecological observatory network. *Elife*, 10:e62922, 2021.

[135] Roy Welch. Spatial resolution requirements for urban studies. *International Journal of Remote Sensing*, 3(2):139–146, 1982.

[136] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–9, 2015. doi: 10.1109/ICCV.2015.451. Acceptance rate: 30.3%.

[137] Qin Xiao, Susan L. Ustin, and E. Gregory McPherson. Using aviris data and multiple-masking techniques to map urban forest tree species. *International Journal of Remote Sensing*, 25(24):5637–5654, 2004.

[138] Qiuyan Yu, Wenjie Ji, Ruiliang Pu, Shawn Landry, Michael Acheampong, Jarlath O'Neil-Dunne, Zhibin Ren, and Shakhawat Hosen Tanim. A preliminary exploration of the cooling effect of tree shade in urban landscapes. *International Journal of Applied Earth Observation and Geoinformation*, 92: 102161, 2020.

[139] Caiyun Zhang and Fang Qiu. Mapping individual tree species in an urban forest using airborne lidar data and hyperspectral imagery. *Photogrammetric Engineering & Remote Sensing*, 78(10):1079–1087, 2012.

[140] Kongwen Zhang and Baoxin Hu. Individual urban tree species classification using very high spatial resolution airborne multi-spectral imagery using longitudinal profiles. *Remote Sensing*, 4(6):1741–1757, 2012.

[141] Carly D. Ziter, Eric J. Pedersen, Christopher J. Kucharik, and Monica G. Turner. Scale-dependent interactions between tree canopy cover and impervious surfaces reduce daytime urban heat during summer. *Proceedings of the National Academy of Sciences*, 116(15):7575–7580, 2019.

[142] Xinhuai Zou, Ming Cheng, Cheng Wang, Yan Xia, and Jonathan Li. Tree classification in complex forest point clouds based on deep learning. *IEEE Geoscience and Remote Sensing Letters*, 14(12):2360–2364, 2017.

Chapter 9

ELEPHANTBOOK

Peter Kulits, Jake Wall, Anka Bedetti, Michelle Henley, and Sara Beery. Elephantbook: A semi-automated human-in-the-loop system for elephant re-identification. In *ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 88–98, 2021.

9.1 Abstract

African elephants are vital to their ecosystems, but their populations are threatened by a rise in human-elephant conflict and poaching. Monitoring population dynamics is essential in conservation efforts; however, tracking elephants is a difficult task, usually relying on the invasive and sometimes dangerous placement of GPS collars. Although there have been many recent successes in the use of computer vision techniques for automated identification of other species, identification of elephants is extremely difficult and typically requires expertise as well as familiarity with elephants in the population. We have built and deployed a web-based platform and database for human-in-the-loop re-identification of elephants combining manual attribute labeling and state-of-the-art computer vision algorithms, known as ElephantBook. Our system is currently in use at the Mara Elephant Project, helping monitor the protected and at-risk population of elephants in the Greater Maasai Mara ecosystem. ElephantBook makes elephant re-identification usable by non-experts and scalable for use by multiple conservation NGOs.

9.2 Introduction

Reliable wildlife population monitoring is critical for effective conservation and species management. Accurate measurement of wildlife density and distribution across landscapes provides insight into trends and ecological processes such as population growth, fecundity, survival, mortality, and density-dependent regulation. A range of measurement techniques have been developed which include aerial surveys, camera trap networks, ground survey techniques, and individual-based re-identification (e.g., spatially explicit mark-recapture [56]). Individual-based recognition techniques can also be used in behavioral studies and human-wildlife conflict cases. The emergence of computational systems based on image algorithms has

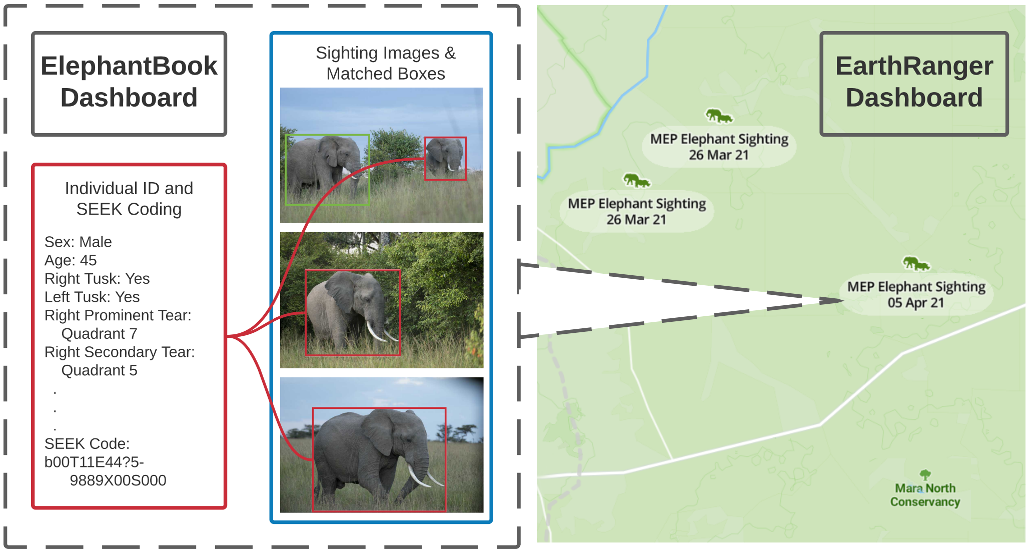

ElephantBook Dashboard
Individual ID and SEEK Coding
Sex: Male
Age: 45
Right Tusk: Yes
Left Tusk: Yes
Right Prominent Tear: Quadrant 7
Right Secondary Tear: Quadrant 5
.
.
.
SEEK Code: b00T11E44?5-9889X00S000
Sighting Images & Matched Boxes
EarthRanger Dashboard
MEP Elephant Sighting 26 Mar 21
MEP Elephant Sighting 26 Mar 21
MEP Elephant Sighting 05 Apr 21
Mara North Conservancy

Figure 9.1: ElephantBook: a system for human-in-the-loop elephant re-identification. Our system can be linked to the EarthRanger conservation land management platform [3], and it helps humans efficiently monitor elephant populations and locations from elephant sightings in the wild.

recently made traction enabling re-identification of certain species (e.g., whales, sharks, zebras, seals, lynx, and sea turtles) that present distinct morphology or patterns (e.g., contours, spots, or stripes) that facilitate visual separability among individuals [16]. However, many species are cryptic and difficult to observe, difficult even for experts to distinguish, or currently lack sufficient training data for application of computer-vision approaches.

Vital to their ecosystems, African elephants are especially important to monitor closely; they are considered ecosystem engineers who have the capacity to shape the environments in which they live, and their population density and distribution can impart multiple cascading effects on ecosystems, biodiversity, and tourism-based economies [26, 47, 53]. Both species of African elephants are threatened: the savanna elephant (*L. africana*) is endangered, and the forest elephant (*L. cyclotis*) was recently listed as critically endangered by the IUCN Red List [4]. Some populations have suffered as much as 62% population loss in recent years [43] with the ivory trade and associated poaching being the main drivers of their decline. Characterizing elephant population demographics across their range is therefore essential to conservation of the species.

Ecologists have recently attempted to create a general re-identification method that can be used by non-experts. The best known of these methods is System for

Elephant Ear-pattern Knowledge (SEEK) coding, developed by Elephants Alive [8], which uses manual attribute labels such as sex and the presence/absence of tusks to improve the accuracy and efficiency of re-identification. The Mara Elephant Project, in collaboration with the California Institute of Technology and Elephants Alive, has developed a semi-automated ensemble visual-recognition system using photographs taken by rangers and research field teams along with manual SEEK attribute labeling. ElephantBook is a novel online software solution with the goal of making elephant re-identification accessible by non-experts and scalable to multiple conservation NGOs.

9.3 Background

The Greater Mara Ecosystem

The Greater Mara Ecosystem (GME) in Kenya is a critical ecosystem given its biodiversity, large wildlife populations, and rich cultural history. It forms the northern extent of the annual migration of 2.2 million wildebeest, zebra, and gazelle from the Serengeti, and it is the most-visited tourist destination in Kenya. The most recent census results estimate there are 2,493 elephants in the GME [61]. Elephants typically live in family units consisting of related females and their offspring. Adult male elephants roam alone or in bachelor herds after they've reached an age of sexual maturity. Despite its status as one of the most beautiful and important wildlife areas in the world, the GME faces significant conservation threats: 374 elephants have been illegally killed since 2012, and there has been a 60% increase in recorded incidents of human-elephant conflict since 2017 (Mara Elephant Project unpublished data). The expansion of agriculture, infrastructure, and human populations is infringing into current elephant ranging areas and severing movement corridors. 50% of elephant range now falls outside of protected areas [70].

The Mara Elephant Project

The Mara Elephant Project (MEP), established in 2011, protects savanna elephants and works to conserve the greater Maasai Mara ecosystem (GME) in Kenya. MEP, in conjunction with the Kenyan Government, has deployed ranger teams to follow the locations of elephant groups fitted with real-time GPS tracking collars, which has led to the arrest of 373 poachers, the seizure of 1,676.5 kg of ivory, and the identification of core movement patterns of approximately 500 elephants [5]. MEP also frequently dispatches rangers to help mitigate conflicts involving "crop-raiding" elephants. Identifying which individual elephants are involved in crop-raiding is important because raiders are typically repeat offenders. Ongoing field monitoring, data analysis, and conservation efforts are needed to ensure the long-term survival of elephants and the overall GME.

Elephants Alive

Elephants Alive is a South Africa-based non-profit organization that operates across the Greater Limpopo Transfrontier Conservation Area and the southern part of Mozambique. Although officially registered with the Kruger National Park in 2003, Elephants Alive draws on data collected over a quarter of a century. Its work contributes to the long-term survival of African elephants through a greater understanding of the complex relationship between elephants and the ecosystems they occupy and by identifying science-based solutions that enable elephants and people to coexist.

EarthRanger

Vulcan's EarthRanger [3] is a real-time system for conservation-related data aggregation, storage, visualization, and dissemination [69]. It includes tracking data from wildlife, rangers, and vehicles, and it records "Events," which range from human-elephant conflict to poaching to illegal logging. Events are reported from the field using a mobile application called Cybertracker; these reports include the time, the location, and information specific to each event type. The Mara Elephant Project, along with many other NGOs in Sub-Saharan Africa, now uses EarthRanger daily to record elephant Group Sightings, including information about group size and composition. However, EarthRanger does not currently support or have any type of interface for individual-based elephant re-identification.

Human expert elephant re-identification

Elephant re-identification is a difficult task, and ecologists may spend thousands of hours over their careers cataloging and charting features that can be used to distinguish elephants. These approaches are often heavily subjective and based on the interpretation and skill of the observer, making the process difficult to replicate across multiple observers or elephant populations. Quantitative approaches are needed to reliably re-identify elephants without dependence on the one or two experts typically available within an organization. One of the most successful existing methods of differentiating between elephants relies on comparison of the elephants' ears including notches, tears, holes, and other identifiable patterns. Several orga-

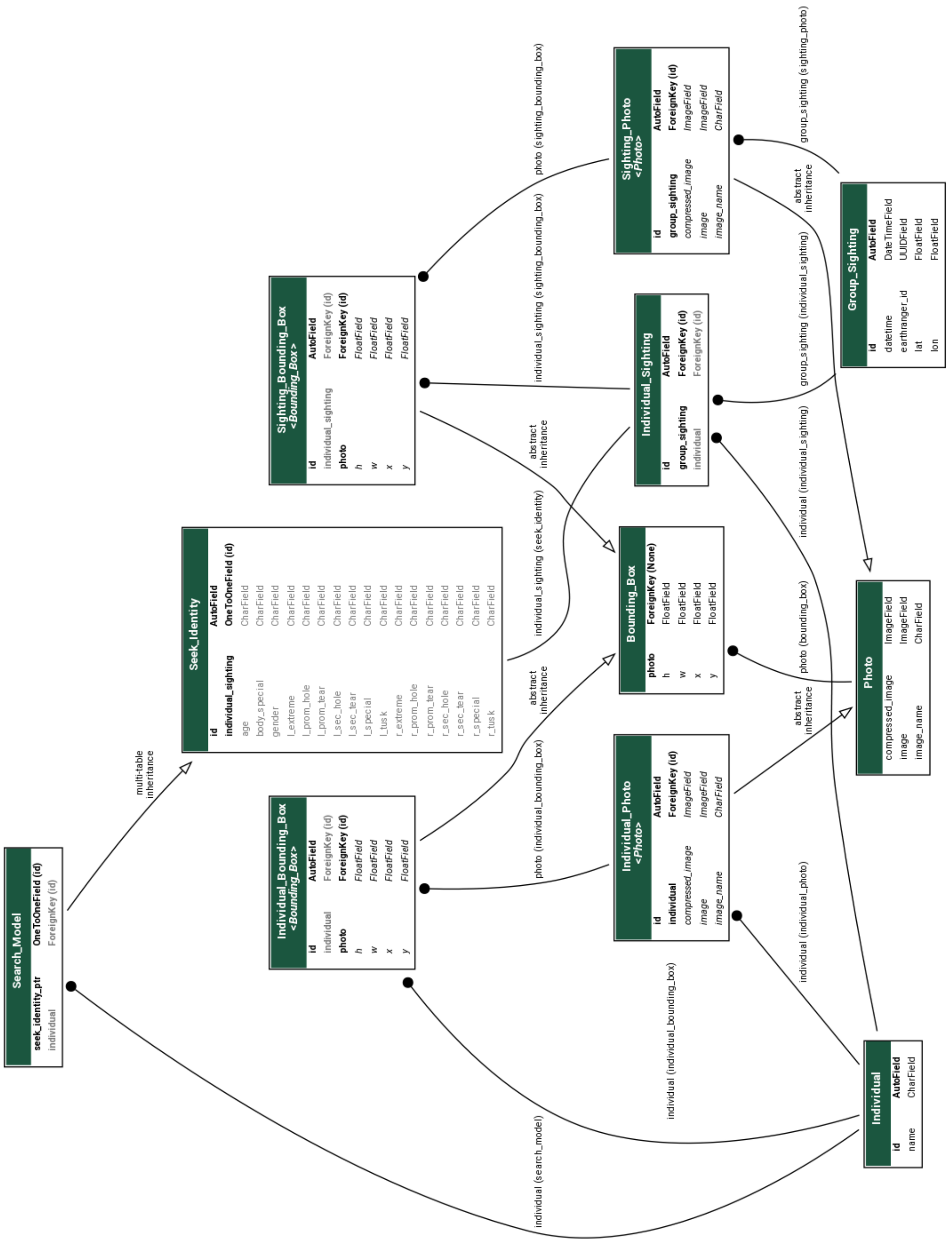


Figure 9.2: Database schema of major models in ElephantBook

nizations (e.g., Save the Elephants, Elephants Alive, Elephant Voices, Amboseli Elephant Trust) use this expert-based approach for elephant identification.

Elephants Alive developed SEEK [8], which involves a comprehensive identification dataset comprised of photos, drawings, and codes of elephant ear patterns that were collected over 25 years (since 1996). The identification system has been refined over time to exclude observer bias and accelerate the photographic identification process. We believe SEEK is the least subjective or expert-reliant elephant re-identification system in-use by any organization to date.

Automated animal re-identification

The most commonly studied re-identification problems in computer vision focus on humans, with popular benchmarks and vast literature for human facial re-identification [38, 73, 74]. There have been many recent successes in computer vision for automated species identification, in both camera trap data [9–14, 48, 57, 63] and human-captured community science data [21, 22, 42, 65–67]. Automated re-identification of *individual animals* using computer vision is an increasingly popular topic, with publications and workshops on the subject at major computer vision conferences [1]. There are several excellent reviews of computer vision for animal re-ID [54, 58, 68].

One of the main, and significant, differences between animal re-identification and other fine-grained categorization tasks is that populations are not fixed, making re-ID an open-set categorization problem [75]. You must be able to recognize if and when an individual does not already exist in your database. The set of individuals might also be quite large: even for the relatively small global population of Grevy's zebra, your full set of identities would be 8,000 individuals [49, 50].

The earliest proposed semi-automated re-identification systems go back as far as 1990, with works on whale re-identification based on human-annotated attribute similarity [46]. The next big breakthroughs in the field relied on traditional feature-engineered computer vision techniques for pattern matching (including SIFT-based feature matching) [7, 16, 18, 29, 30, 34, 41, 45, 64] and numerical representations of unique contours [6, 31, 32]. Animal re-identification, like most of computer vision, has seen significant advances with the onset of deep learning, including several neural-network-based approaches [17, 19, 20, 25, 28, 35–37, 60]. The field has recently explored metric-learning-based methods [24, 59, 75], inspired by the success of these methods for human re-identification [38, 73, 74]. Metric-learning methods are also more robust to open-set categorization, as they are similarity-based and require only a single example of an individual with which to compare, as opposed to the tens or hundreds of examples needed by data-hungry CNNs. Another common tactic for handling the open-set and data-scarce nature of re-identification is hybridizing deep networks for notable part localization with previous pattern or contour feature-based matching methods which do not require large amounts of training data per individual [62, 72].

Automated elephant re-identification

In 2010, Dabarera and Rodrigo proposed an image-based algorithm to identify individual elephants based on full-frontal facial images [23]. Korschens et al. proposed a matching algorithm based on human-labeled whole-head annotations, including the elephant's ears and tusks, where present [36]. In an extension, they released a dataset, ELPephants, and demonstrated good results on a closed set of individuals with localized feature extraction using deep nets and SVM-based feature discrimination [35]. Recent methods of robustly differentiating between elephant images in an open-set population rely on finding and matching the contours of the ear (Figure 9.9), similar to many human-expert re-identification methods like SEEK. Multi-curve matching algorithms based on human-annotated contours of elephant ears were proposed by Ardovini et al. [6] and Weideman et al. [71]. Weideman's CurvRank algorithm was originally designed for re-identification of whale flukes and dorsal fins. Recently, Weideman et al. [72] proposed an extension of CurvRank that is capable of automatically extracting matchable contours from images, and report strong results matching contours of elephant ears.

9.4 ElephantBook

Our solution, which we call ElephantBook, by default integrates with EarthRanger through its REST API to consume Group Sightings recorded by field teams. ElephantBook can also be reconfigured for use without EarthRanger if needed. It is web-based and built primarily with the Django Python package [2]. This configuration allows our system to be both lightweight and easily reconfigurable.

Human-in-the-loop re-identification pipeline

Data Collection in the Field

Rangers at the Mara Elephant Project routinely survey the Maasai Mara in search of elephants. Rangers record the time and location of every elephant sighting and submit it to EarthRanger. When possible, rangers photograph each elephant from multiple angles. If no photographs are taken, the event is resolved in EarthRanger, and no re-identification occurs.

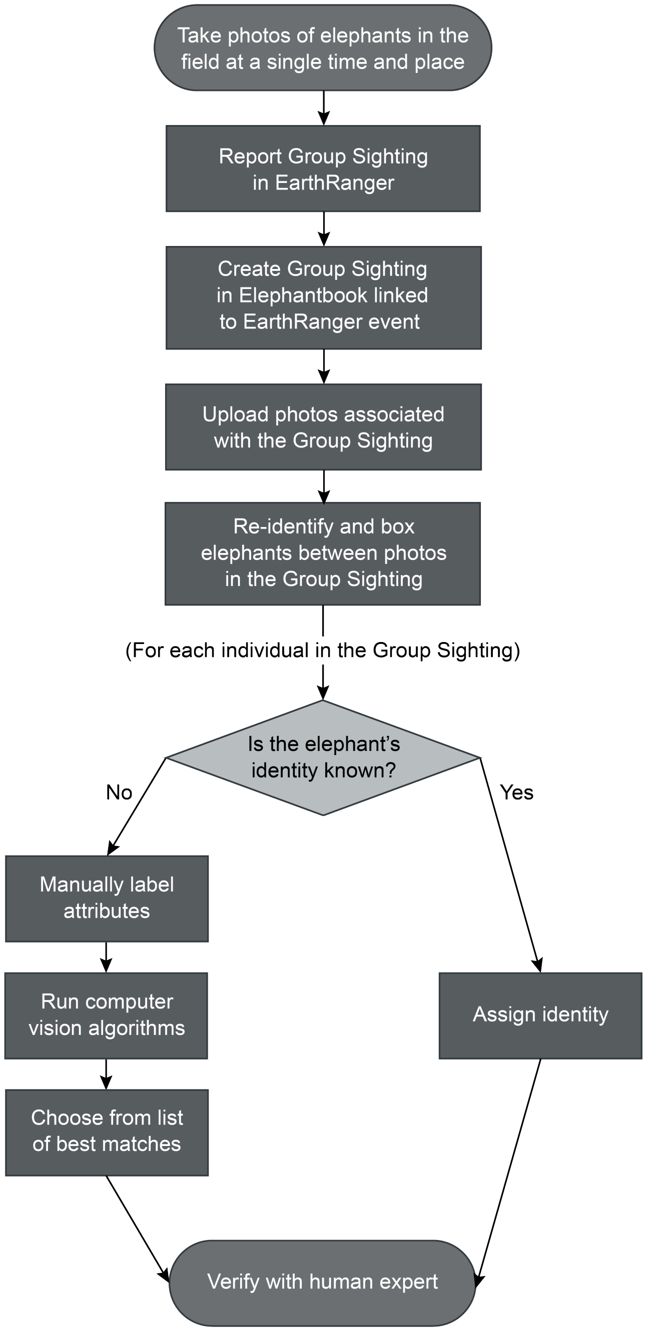

Take photos of elephants in the field at a single time and place
Report Group Sighting in EarthRanger
Create Group Sighting in Elephantbook linked to EarthRanger event
Upload photos associated with the Group Sighting
Re-identify and box elephants between photos in the Group Sighting
(For each individual in the Group Sighting)
Is the elephant's identity known?
No
Yes
Manually label attributes
Run computer vision algorithms
Choose from list of best matches
Assign identity
Verify with human expert

Figure 9.3: Workflow of elephant re-identification.

Adding a Group Sighting to ElephantBook via EarthRanger

ElephantBook pulls a list of active elephant sighting events from EarthRanger. Users select the appropriate EarthRanger event and create a corresponding ElephantBook "Group Sighting." A Group Sighting is one or more elephants spotted at the same time and place. This step is usually performed after returning from the field.

Uploading Photos

All photos taken at the same time and place of the Group Sighting are uploaded to ElephantBook. However, only photos labeled with boxes (in the next step) are used for re-identification.

Boxing Elephants

Once all photos are uploaded, elephants in each photo are boxed with an image annotation tool. While the human annotator likely will not know the name of each individual elephant in the Group Sighting photos, the annotator should be able to differentiate between elephants and identify the same elephant across multiple images. If it is impossible to tell elephants apart in a single instance, matching over a period of months is unlikely. Boxes are marked with numbers unique to each specific elephant within the Group Sighting. This identification marking reduces the number of matches we must make from the sum of the number of elephants in all photos to the number of actual elephants sighted. See Section 9.5 for more details.

Human Attribute Labeling

An "Individual Sighting" is created for each elephant identified in the previous step. An Individual Sighting is an elephant encounter at a single time and place, and it is always connected to a parent Group Sighting. Manual attribute-labeling is performed for each Individual Sighting. We use the recently-developed SEEK coding system [8]. See Section 9.6 for more details.

Computer Vision

Confidence-producing computer vision matching algorithms are run to identify potential matches. See Section 9.7 for more details.

Matching

Manual attributes are combined with the output from the computer vision algorithms to provide a list of the most likely previously identified elephants. See Section 9.8 for more details.

9.5 Bounding box annotation

We customized an open-source online bounding box annotation tool from the Visipedia project [15]. Because annotators need to match individual elephants across photos taken at a sighting, a second pane was added to allow annotators to compare multiple photos at once.

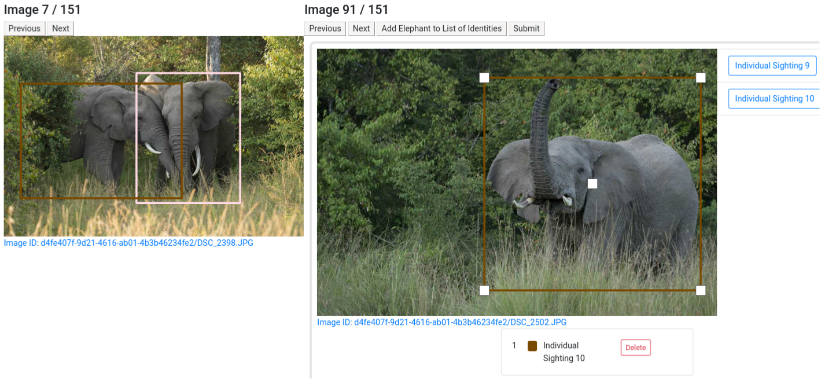


Image 7 / 151
Previous
Next
Image ID: d4fe407f-9d21-4616-ab01-4b3b46234fe2/DSC_2398.JPG
Image 91 / 151
Previous
Next
Add Elephant to List of Identities
Submit
Individual Sighting 9
Individual Sighting 10
Image ID: d4fe407f-9d21-4616-ab01-4b3b46234fe2/DSC_2502.JPG
1
Individual Sighting 10
Delete

Figure 9.4: Bounding box annotation interface.

9.6 SEEK

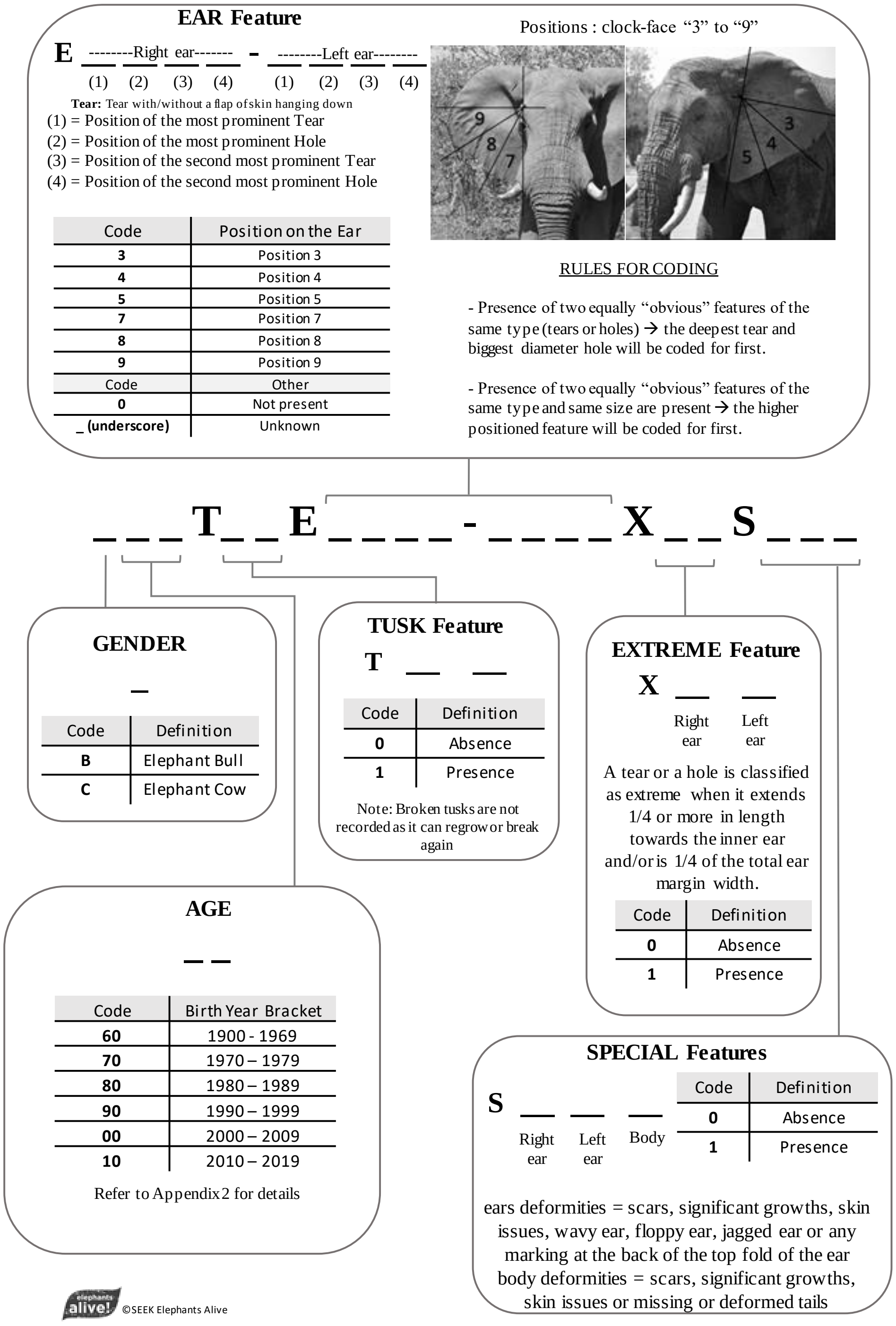
In SEEK, each elephant is assigned a unique descriptive code which is used to narrow the set of potential matches that must be considered by human experts. The code begins with the elephant’s sex and age, followed by the presence or absence of tusks (Figure 9.5). The code further defines the type and position of prominent and secondary tears and holes found on the right and left ears. Finally, it notes the presence of any extreme features on the ears and body, such as a missing tail.

9.7 Computer vision

Elephant ear localization

To allow CurvRank to focus on the ear, the localization of which is key to extracting accurate ear contours, we trained a simple elephant-ear detector.

The ELPephants dataset from the Elephant Listening Project [35]–consisting of images of African forest elephants visiting the Dzanga bai clearing in the Dzanga-Ndoki National Park in the Central African Republic–was used for training and validation. After removing duplicates, the dataset consisted of 1935 images of 276 unique individuals. The dataset is provided with the identities of the elephants, but each image was manually annotated for bounding-boxes of left and right ears. Only ears where the contours were fully visible were annotated. Annotations were made



EAR Feature

E __ __ __ __ - __ __ __ __
--------Right ear------- --------Left ear--------
(1) (2) (3) (4) (1) (2) (3) (4)

Tear: Tear with/without a flap of skin hanging down

(1) = Position of the most prominent Tear
(2) = Position of the most prominent Hole
(3) = Position of the second most prominent Tear
(4) = Position of the second most prominent Hole

Code	Position on the Ear
3	Position 3
4	Position 4
5	Position 5
7	Position 7
8	Position 8
9	Position 9
Code	Other
0	Not present
_ (underscore)	Unknown

Positions : clock-face "3" to "9"

RULES FOR CODING

- Presence of two equally "obvious" features of the same type (tears or holes) → the deepest tear and biggest diameter hole will be coded for first.

- Presence of two equally "obvious" features of the same type and same size are present → the higher positioned feature will be coded for first.

GENDER

_

Code	Definition
B	Elephant Bull
C	Elephant Cow

TUSK Feature

T __ __

Code	Definition
0	Absence
1	Presence

Note: Broken tusks are not recorded as it can regrow or break again

EXTREME Feature

X __ __
Right ear Left ear

A tear or a hole is classified as extreme when it extends 1/4 or more in length towards the inner ear and/or is 1/4 of the total ear margin width.

Code	Definition
0	Absence
1	Presence

AGE

_ _

Code	Birth Year Bracket
60	1900 - 1969
70	1970 – 1979
80	1980 – 1989
90	1990 – 1999
00	2000 – 2009
10	2010 – 2019

Refer to Appendix 2 for details

SPECIAL Features

S __ __ __
Right ear Left ear Body

Code	Definition
0	Absence
1	Presence

ears deformities = scars, significant growths, skin issues, wavy ear, floppy ear, jagged ear or any marking at the back of the top fold of the ear
body deformities = scars, significant growths, skin issues or missing or deformed tails

Figure 9.5: The SEEK coding system.

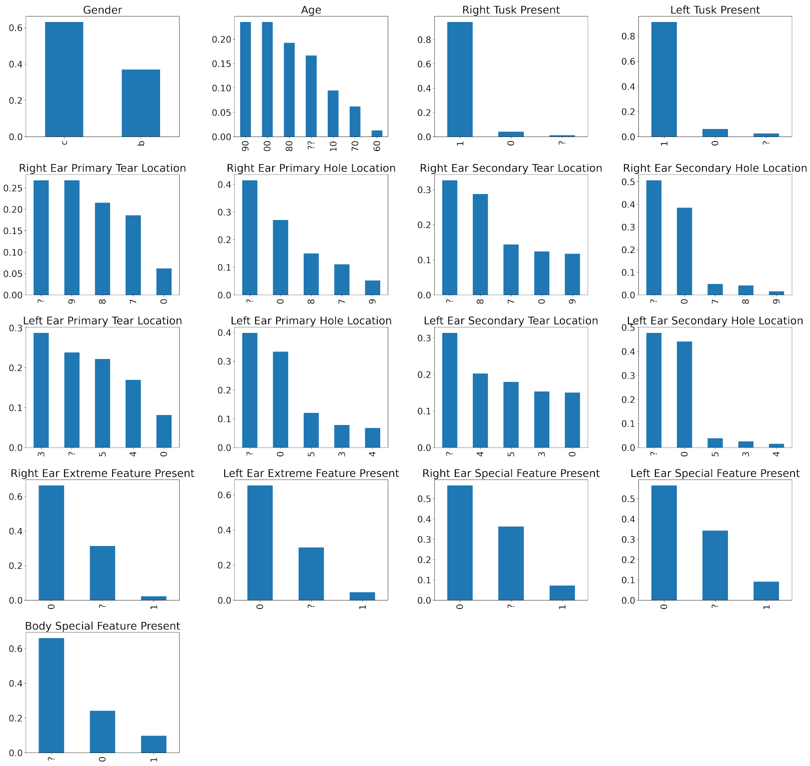


Figure 9.6: Frequency of SEEK attributes.

on 910 left ears and 1,045 right ears (Figure 9.8). Two-hundred randomly sampled images were reserved for object detection validation.

We trained a Faster R-CNN object detection model [55] with a ResNet-50 backbone [27] and added Feature Pyramid Networks (FPN) [40] in Pytorch [51]. Beginning with a model checkpoint pretrained on the Microsoft COCO dataset [39], we trained our model on 1,735 images to detect and categorize left and right ears. Our detector achieves a Mean Average Precision (mAP) [39] of 95% on our held-out test dataset of 200 randomly-selected images.

Matching ear contours with CurvRank

After extracting ear images from out ear detection model, we use CurvRank [72] to filter possible matches.

CurvRank was initially developed to recognize individual cetaceans based on contours of flukes and dorsal fins [71]. As elephant ears are also a strong re-identifiable feature and are delineated by a contoured edge, applying CurvRank to elephant re-identification was an intuitive next step, and the authors determined the transfer-

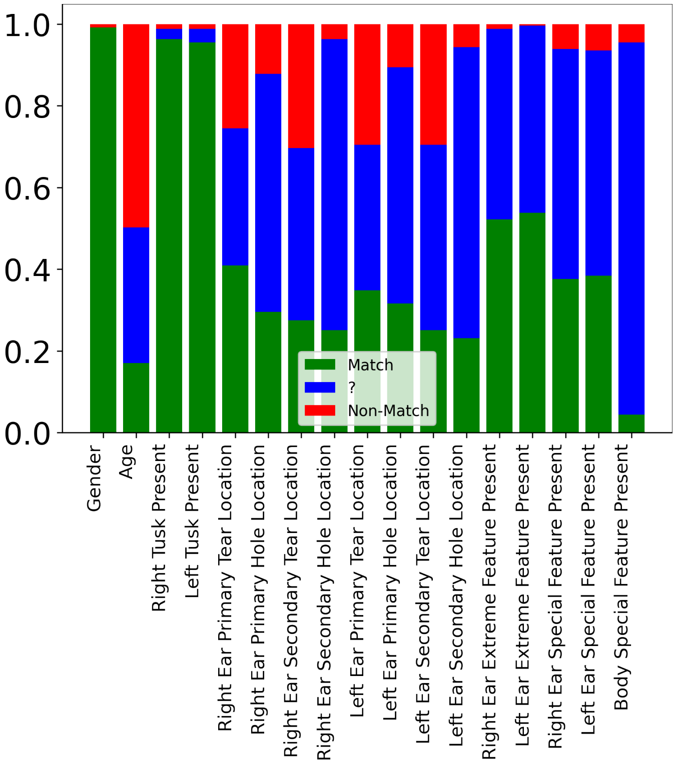


Figure 9.7: Agreement of annotators by attribute for pairs of SEEK codes on the same individual.

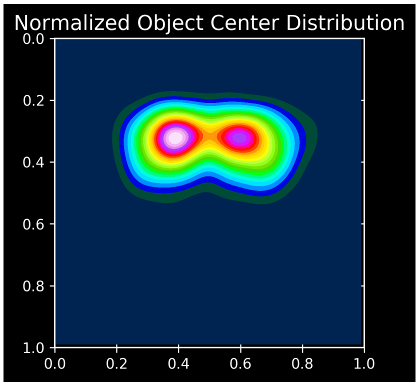


Figure 9.8: We visualize the center of the ground truth annotated boxes across our training set, and see that there is a strong bias in the imagery towards ears being in the upper center of the image, with modes slightly to the right and left.

ability of the matching algorithm from cetaceans to elephants by analyzing results on hand-drawn contours.

Recently the CurvRank authors proposed a deep-learning based algorithm to automatically extract the contours used as input to their matching algorithm [72].

They evaluated this automated approach on humpback whales and African savanna elephants, with impressive results. Their method relies on two fully convolutional neural networks for curve extraction, one coarse-grained and one fine-grained (CG-FCNN and FG-FCNN). Annotators initially traced the identifying contour in cropped images with a broad line, using a single brushstroke, to produce coarse-grained training data for the CG-FCNN. In the second step, the FG-FCNN is trained to predict for each pixel in the (ear or fluke) image the probability that it would be covered by the coarse brush stroke, producing a probability image at the same resolution as the initial image. By using the coarse, easily extracted training data to train the FG-FCNN, tedious manual effort is avoided. These pixel-level probability maps guide the third step: a shortest path contour extraction algorithm.

Once the contour is extracted, it is represented as an ordered sequence of (x, y) coordinate pairs. Then CurvRank builds an integral curvature by sliding multiple disks of increasing radius along the contour [72]. For each scale, every point is represented as the ratio of the areas of the disk for that scale on either side of the contour [72]. Feature keypoints are defined at local extrema of the integral curvature representation [33, 72], and feature descriptors are extracted from the regions between all pairs of keypoints. This set of feature descriptors forms a densely sampled, overlapping representation of the entire individual contour across multiple scales. Match similarity is determined and possible matching individuals ranked via the local naive Bayes nearest neighbors (LNBNN) algorithm [44].

The method reported a top-1 matching accuracy of 84% for high-quality, high-resolution images of elephant ears on the closed set of 132 individuals on which the model is also trained [72]. The authors remark that elephant ear recognition is more difficult than whale fluke identification, due to challenges of contour extraction against more-highly textured image backgrounds and because the identifying information is more localized and subtle.

9.8 Matching

To allow rangers to efficiently identify an individual from the large set of previously encountered elephants, rangers are presented with a ranked list of possible matches to visually examine. These matches are computed with a score function that is a linear combination of manual attribute differences and computer vision matching confidence. Each SEEK attribute of the new Individual Sighting is compared to the SEEK code of all known Individuals. For each attribute in the codes, the distance

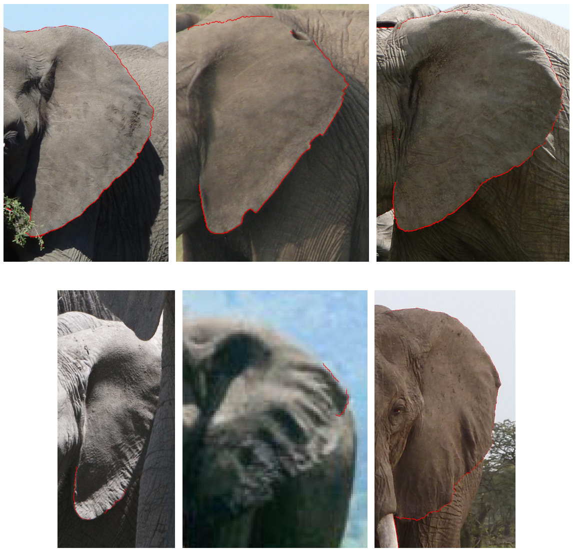


Figure 9.9: Here we show successful (top) and failed (bottom) CurvRank examples. CurvRank is highly successful in high-quality, high-resultion imagery, but performance drops off in lower-resolution or blurry data as the edge of the ear is harder to distinguish.

is zero if the attributes match, one if they differ, and 0.6 if either of them contain a wildcard character. Additionally, the weight of the age component of the distance is set to 0.4 because of the known difficulty in accurately aging elephants (Figure 9.7). The mean of these differences is taken. The weighting parameters were learned separately on a training set of codes to optimize matching accuracy.

CurvRank produces an unbounded matching score between the new Individual Sighting and all Individuals. A greater score indicates greater likelihood of a true match. CurvRank scores are subtracted from the SEEK score and multiplied by 0.1. The parameter weight of 0.1 was learned in a training set of SEEK codes and CurvRank contours.

Evaluating matching accuracy

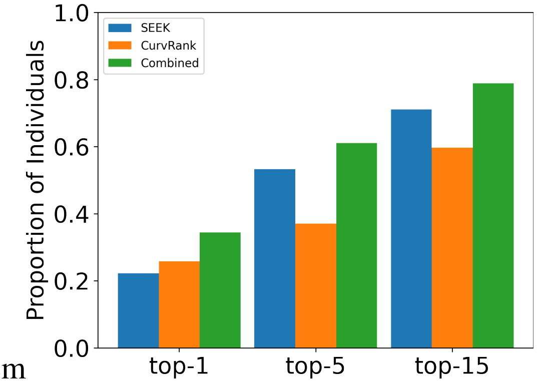
To evaluate the robustness of SEEK, CurvRank, and our proposed combination of the two, we trained a non-expert team of seven college undergraduates to perform SEEK labeling, and we collected labels from two to three students annotator for a set of Individual Sightings from the Elephant Voices collected by Joyce Poole [52]. In total, we have three annotations for 75 Individual Sightings and two annotations for 26 Individual Sightings.

We held out individuals that had at least two SEEK code annotations and two right-ear CurvRank contours. There are 45 individuals with a pair of SEEK codes and 33 individuals with a triplet of codes. Comparisons of top-k matching performance for SEEK, CurvRank, and our ensembled approach can be seen in Figure 9.12. We observe that matching performed with SEEK codes generally outperforms that of CurvRank alone, but that a combined approach is able to leverage the best of both, leading to more accurate matching. Using our combined system, with only two previous sightings of an individual in our database, we are able to match to the correct individual within the top 15 for 92.9% of sightings, and within the top 5 for 66.7%, helping rangers reduce the time needed to find the correct matched individual in the database. As Mara Elephant Project continues to collect and label Individual Sightings we will continue to analyze and hopefully improve matching performance. We expect additional sightings to improve accuracy, as it presents more potential sightings per individual to match with correctly. However, as we collect additional sightings we will also be increasing the number of individuals in the database, making the matching task more nuanced and potentially more challenging.

9.9 Mara Elephant Project initial deployment

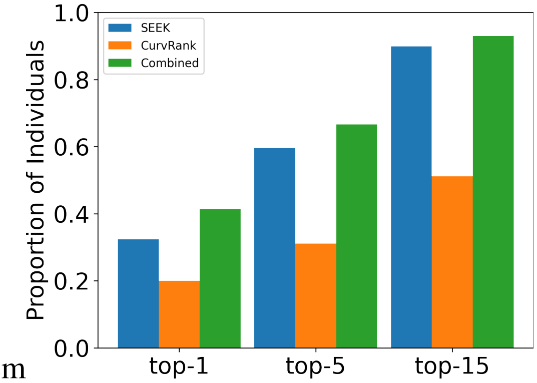
The Mara Elephant Project began using ElephantBook in January 2021 after a six-month prototyping period and so far has logged 140 Group Sightings and 251 Individual Sightings and has ingested and boxed 10,462 images of elephants. Beginning in March 2021, the organization has hired and trained a full-time team of four research assistants for collecting elephant sightings in the field, processing photos, and developing SEEK codes for individual elephants. Initial training on both field methodology for cataloging elephant Group Sightings and in the use of ElephantBook and SEEK labeling took one week. MEP's goal is to characterize and document the majority of the Mara's 2500 individuals. Extension of the ElephantBook system with partner organizations in Tanzania would further enable documenting the greater, connected elephant population stretching south into the Serengeti and consisting of >7000 individuals.

Initial experience using ElephantBook is that it is an intuitive system that mimics a typical re-identification workflow. Optimizations for low-bandwidth connections, such as compression of photos before viewing them, but also keeping original full-resolution versions available for detailed scrutiny by a SEEK coder, have greatly improved the user experience. Boxing individuals has been relatively straightforward even for novice users. Accurately labeling SEEK codes is perhaps the



8cm

Figure 9.10: One code per individual



8cm

Figure 9.11: Two codes per individual

Figure 9.12: Comparing matching accuracy for our SEEK-based matching algorithm, CurvRank, and our hybrid SEEK-CurvRank aggregated approach. We see that SEEK and CurvRank are complementary, with the combined approach outperforming either method on its own for tests with both one and two database example for each individual.

most challenging component of the ElephantBook system, particularly the correct estimation of age category which requires considerable expertise and, to a lesser degree, the determination of sex.

9.10 Conclusions and future work

We have built a robust semi-automated system for human-in-the-loop elephant re-identification, and we have deployed our system on the ground in the Greater Mara Ecosystem. This system allows the Mara Elephant Project to track a much larger population of elephants over time, as they will no longer need to collar an elephant to track its movements. The system is a needed tool to assist in their vital elephant conservation efforts. As we move forward, we will expand ElephantBook to additional parks, including the Grumeti Game Reserve in Serengeti National Park in

Tanzania and Greater Limpopo Transfrontier Conservation Area in South Africa and Southern Mozambique.

In the coming months, we will continue to collect new elephant sightings and refine our matching system to further reduce the human effort needed for re-identification. We plan to investigate automating SEEK coding and integrating additional computer vision methodology into our system, building learned representations of individual elephants beyond their ear contours. The data collected will also allow us to further analyze how these elephant features change over time and allow us to conduct deeper analysis of our current system on an expanding set of known elephants.

Acknowledgements

We would like to thank the entire team at the Mara Elephant Project for their efforts in deploying this system. This work was supported, through funding, data storage, and computing resources, by Microsoft AI for Earth, the Caltech Resnick Sustainability Institute, and NSFGRFP Grant No. 1745301, the views are those of the authors and do not necessarily reflect the views of these organizations.

References

[1] Deep learning methods and applications for animal re-identification workshop at WACV 2020. https://sites.google.com/corp/view/wacv2020animalreid/.

[2] The Web framework for perfectionists with deadlines | Django. 2021. URL https://www.djangoproject.com/.

[3] Vulcan EarthRanger: A Domain Awareness System. 2021. URL https://earthranger.com/.

[4] The IUCN Red List of Threatened Species. *IUCN Red List of Threatened Species*, 2021. URL https://www.iucnredlist.org/en.

[5] Mara Elephant Project: Increased security. 2021. URL https://maraelephantproject.org/results/increased-security/.

[6] Alessandro Ardovini, Luigi Cinque, and Enver Sangineto. Identifying elephant photos by multi-curve matching. *Pattern Recognition*, 41(6):1867–1877, 2008.

[7] Zaven Arzoumanian, Jason Holmberg, and Brad Norman. An astronomical pattern-matching algorithm for computer-aided identification of whale sharks rhincodon typus. *Journal of Applied Ecology*, 42(6):999–1011, 2005.

[8] Anka Bedetti, Cathy Greyling, Barry Paul, Jennifer Blondeau, Amy Clark, Hannah Malin, Jackie Horne, Ronny Makukule, Jessica Wilmot, Tammy Eggeling, et al. System for elephant ear-pattern knowledge (seek) to identify individual african elephants. *Pachyderm*, 61:63–77, 2020.

[9] Sara Beery, Grant Van Horn, Oisin MacAodha, and Pietro Perona. The iWildCam 2018 challenge dataset. *The Fifth Fine-Grained Visual Categorization Workshop at CVPR*, 2018.

[10] Sara Beery, Dan Morris, and Pietro Perona. The iWildCam 2019 challenge dataset. *The Sixth Fine-Grained Visual Categorization Workshop at CVPR*, 2019.

[11] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *In the Data Mining and Artificial Intelligence for Conservation Workshop at Knowledge Discovery in Databases (KDD)*, 2019. *selected to be featured at KDD Earth Day.

[12] Sara Beery, Elijah Cole, and Arvi Gjoka. The iWildCam 2020 competition dataset. *The Seventh Fine-Grained Visual Categorization Workshop at CVPR*, 2020.

[13] Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 863–873, 2020.

[14] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context R-CNN: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2020.

[15] Serge Belongie and Pietro Perona. Visipedia circa 2015. *Pattern Recognition Letters*, 72:15–24, 2016. doi: 10.1016/j.patrec.2015.11.023.

[16] Tanya Y. Berger-Wolf, Daniel I. Rubenstein, Charles V. Stewart, Jason A. Holmberg, Jason Parham, Sreejith Menon, Jonathan P. Crall, Jon Van Oast, Emre Kiciman, and Lucas Joppa. Wildbook: Crowdsourcing, computer vision, and data science for conservation. *CoRR*, abs/1710.08880, 2017. URL `http://arxiv.org/abs/1710.08880`.

[17] Clemens-Alexander Brust, Tilo Burghardt, Milou Groenenberg, Christoph Kading, Hjalmar S Kuhl, Marie L Manguette, and Joachim Denzler. Towards automated visual monitoring of individual gorillas in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2820–2830, 2017.

[18] Tilo Burghardt, Neill Campbell, Peter J Barham, Innes C Cuthill, and Richard Sherley. A fully automated computer vision system for the biometric identification of african penguins (spheniscus demersus) on robben island. In *6th International Penguin Conference (IPC07)*. University of Tasmania, Australia, 2007.

[19] Steven J.B. Carter, Ian P. Bell, Jessica J. Miller, and Peter P. Gash. Automated marine turtle photograph identification using artificial neural networks, with application to green turtles. *Journal of Experimental Marine Biology and Ecology*, 452:105–110, 2014.

[20] Gullal Singh Cheema and Saket Anand. Automatic detection and recognition of individuals in patterned species. *Lecture Notes in Computer Science*, page 27–38, 2017. ISSN 1611-3349. doi: 10.1007/978-3-319-71273-4_3. URL `http://dx.doi.org/10.1007/978-3-319-71273-4_3`.

[21] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4109–4118, 2018.

[22] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[23] Ranga Dabarera and Ranga Rodrigo. Vision Based Elephant Recognition for Management and Conservation. *Fifth International Conference on Information and Automation for Sustainability*, pages 163–166, 2010. doi: 10.1109/ICIAFS.2010.5715653.

[24] Debayan Deb, Susan Wiper, Sixue Gong, Yichun Shi, Cori Tymoszek, Alison Fletcher, and Anil K. Jain. Face recognition: Primates in the wild. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10. IEEE, 2018.

[25] Alexander Freytag, Erik Rodner, Marcel Simon, Alexander Loos, Hjalmar S Kühl, and Joachim Denzler. Chimpanzee faces in the wild: Log-euclidean cnns for predicting identities and attributes of primates. In *German Conference on Pattern Recognition*, pages 51–63. Springer, 2016.

[26] Gary Haynes. Elephants (and extinct relatives) as earth-movers and ecosystem engineers. *Geomorphology*, 157:99–107, 2012.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[28] Qi He, Qijun Zhao, Ning Liu, Peng Chen, Zhihe Zhang, and Rong Hou. Distinguishing individual red pandas from their faces. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 714–724. Springer, 2019.

[29] Lex Hiby and Phil Lovell. Computer aided matching of natural markings: A prototype system for grey seals. *Report of the International Whaling Commission*, 12:57–61, 1990.

[30] Lex Hiby, Phil Lovell, Narendra Patil, N Samba Kumar, Arjun M Gopalaswamy, and K Ullas Karanth. A tiger cannot change its stripes: Using a three-dimensional model to match images of living tigers and tiger skins. *Biology letters*, 5(3):383–386, 2009.

[31] Gilbert R. Hillman, Bernd Wursig, Glenn A. Gailey, Nasser Kehtarnavaz, et al. Computer-assisted photo-identification of individual marine vertebrates: A multi-species system. *Aquatic Mammals*, 29(1):117–123, 2003.

[32] Ruben Huele and Helias Udo de Haes. Identification of individual sperm whales by wavelet transform of the trailing edge of the flukes. *Marine Mammal Science*, 14(1):143–145, 1998.

[33] Benjamin Hughes and Tilo Burghardt. Automated visual fin identification of individual great white sharks. *International Journal of Computer Vision*, 122 (3):542–557, 2017.

[34] Marcella J. Kelly. Computer-aided photograph matching in studies using individual identification: An example from serengeti cheetahs. *Journal of Mammalogy*, 82(2):440–449, 2001.

[35] Matthias Korschens and Joachim Denzler. Elpephants: A fine-grained dataset for elephant re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[36] Matthias Körschens, Björn Barz, and Joachim Denzler. Towards Automatic Identification of Elephants in the Wild. *arXiv:1812.04418 [cs]*, 2018. URL `http://arxiv.org/abs/1812.04418`. arXiv: 1812.04418.

[37] Zhangyong Li, Chao Ge, Siwan Shen, and Xinwei Li. Cow individual identification based on convolutional neural network. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, pages 1–5, 2018.

[38] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2197–2206, 2015.

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[40] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[41] Alexander Loos and Andreas Ernst. An automated chimpanzee identification system using face detection and recognition. *EURASIP Journal on Image and Video Processing*, 2013(1):1–17, 2013.

[42] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9596–9606, 2019.

[43] Fiona Maisels, Samantha Strindberg, Stephen Blake, George Wittemyer, John Hart, Elizabeth A. Williamson, Rostand Aba'a, Gaspard Abitsi, Ruffin D. Ambahe, Fidal Amsini, Parfait C. Bakabana, Thurston Cleveland Hicks, Rosine E. Bayogo, Martha Bechem, Rene L. Beyers, Anicet N. Bezangoye, Patrick Boundja, Nicolas Bout, Marc Ella Akou, Lambert Bene Bene, Bernard Fosso, Elizabeth Greengrass, Falk Grossmann, Clement Ikamba-Nkulu, Omari Ilambu, Bila Isia Inogwabini, Fortune Iyenguet, Franck Kiminou, Max Kokangoye, Deo Kujirakwinja, Stephanie Latour, Innocent Liengola, Quevain Mackaya, Jacob Madidi, Bola Madzoke, Calixte Makoumbou, Guy Aim Malanda, Richard Malonga, Olivier Mbani, Valentin A. Mbendzo, Edgar Ambassa, Albert Ekinde, Yves Mihindou, Bethan J. Morgan, Prosper Motsaba, Gabin Moukala, Anselme Mounguengui, Brice S. Mowawa, Christian Ndzai, Stuart Nixon, Pele Nkumu, Fabian Nzolani, Lilian Pintea, Andrew Plumptre, Hugo Rainey, Bruno Bokoto de Semboli, Adeline Serckx, Emma Stokes, Andrea Turkalo, Hilde Vanleeuwe, Ashley Vosper, and Ymke Warren. Devastating Decline of Forest Elephants in Central Africa. *PLoS ONE*, 8, 2013. ISSN 19326203. doi: 10.1371/journal.pone.0059469.

[44] Sancho McCann and David G Lowe. Local naive bayes nearest neighbor for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3650–3656. IEEE, 2012.

[45] Vincent Miele, Gaspard Dussert, Bruno Spataro, Simon Chamaillé-Jammes, Dominique Allainé, and Christophe Bonenfant. Revisiting giraffe photo-identification using deep learning and network analysis. *bioRxiv*, 2020.

[46] Sally A. Mizroch, Judith A. Beard, and Macgill Lynde. Computer assisted photo-identification of humpback whales. *Report of the International Whaling Commission*, 12:63–70, 1990.

[47] Robin Naidoo, Brendan Fisher, Andrea Manica, and Andrew Balmford. Estimating economic losses to tourism in africa from the illegal killing of elephants. *Nature Communications*, 7(1):1–9, 2016.

[48] Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution*, 12 (1):150–161, 2021.

[49] Joy Omulupi Ontita, Jennifer Weston, George Anyona, Geoffrey Chege, David Kimiti, Kaia Tombak, Andrew Gersick, and Nancy Rubenstein. The state of kenya's grevy's zebras and reticulated giraffes. *Results of the Great Grevy's Rally*, 2018.

[50] Jason Parham, Jonathan Crall, Charles Stewart, Tanya Berger-Wolf, and Daniel I. Rubenstein. Animal population censusing at scale with citizen science and photographic identification. In *AAAI Spring Symposium-Technical Report*, 2017.

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[52] Joyce Poole. How to identify african elephants. URL `https://elephantvoices.org/multimedia-resources/how-to-identify-african-elephants.html`.

[53] Robert M. Pringle. Elephants as agents of habitat creation for small vertebrates at the patch scale. *Ecology*, 89(1):26–33, 2008.

[54] Prashanth C. Ravoor and T.S.B. Sudarshan. Deep learning methods for multi-species animal re-identification and tracking–a survey. *Computer Science Review*, 38:100289, 2020.

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[56] J. Andrew Royle, Richard B. Chandler, Rahel Sollmann, and Beth Gardner. *Spatial Capture-Recapture*. Academic Press, August 2013. ISBN 978-0-12-407152-0.

[57] Stefan Schneider, Graham W. Taylor, and Stefan Kremer. Deep learning object detection methods for ecological camera trap data. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 321–328. IEEE, 2018.

[58] Stefan Schneider, Graham W. Taylor, Stefan Linquist, and Stefan C. Kremer. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 10 (4):461–470, 2019.

[59] Stefan Schneider, Graham W. Taylor, and Stefan C. Kremer. Similarity learning networks for animal individual re-identification-beyond the capabilities of a human observer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 44–52, 2020.

[60] Daniel Schofield, Arsha Nagrani, Andrew Zisserman, Misato Hayashi, Tetsuro Matsuzawa, Dora Biro, and Susana Carvalho. Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances*, 2019.

[61] Kenya WildlIfe Service. Aerial total count of elephants, buffaloes and giraffes in the Masai Mara ecosystem (may 2017) - Kenya Wildlife Service. URL `http://www.kws.go.ke/content/aerial-total-count-elephants-buffaloes-and-giraffes-masai-mara-ecosytem-may-2017-0`.

[62] Ankita Shukla, Gullal Sigh Cheema, Pei Gao, Suguru Onda, Divyam Anshumaan, Saket Anand, Ryan Farrell, et al. A hybrid approach to tiger re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[63] Michael A. Tabak, Mohammad S. Norouzzadeh, David W. Wolfson, Erica J. Newton, Raoul K. Boughton, Jacob S. Ivan, Eric A. Odell, Eric S. Newkirk, Reesa Y. Conrey, Jennifer Stenglein, et al. Improving the accessibility and transferability of machine learning algorithms for identification of animals in camera trap images: Mlwic2. *Ecology and Evolution*, 10(19):10374–10383, 2020.

[64] Christopher Town, Andrea Marshall, and Nutthaporn Sethasathien. Manta matcher: Automated photographic identification of manta rays using keypoint features. *Ecology and Evolution*, 3(7):1902–1914, 2013.

[65] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604, 2015.

[66] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[67] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. *arXiv preprint arXiv:2103.16483*, 2021.

[68] Maxime Vidal, Nathan Wolf, Beth Rosenberg, Bradley P. Harris, and Alexander Mathis. Perspectives on individual animal identification from biology and computer vision. *arXiv preprint arXiv:2103.00560*, 2021.

[69] Jake Wall, George Wittemyer, Brian Klinkenberg, and Iain Douglas-Hamilton. Novel Opportunities for Wildlife Conservation and Research with Real-Time Monitoring. *Ecological Applications*, 24(4):593–601, 2014.

[70] Jake Wall, George Wittemyer, Brian Klinkenberg, Valerie LeMay, Stephen Blake, Samantha Strindberg, Michelle Henley, Fritz Vollrath, Fiona Maisels, Jelle Ferwerda, and Iain Douglas-Hamilton. Human footprint and protected areas shape elephant range across Africa. *Current Biology*, April 2021. ISSN 0960-9822. doi: 10.1016/j.cub.2021.03.042. URL https://www.sciencedirect.com/science/article/pii/S096098222100381X.

[71] Hendrik J. Weideman, Zachary M. Jablons, Jason Holmberg, Kiirsten Flynn, John Calambokidis, Reny B. Tyson, Jason B. Allen, Randall S. Wells, Krista Hupman, Kim Urian, and Charles V. Stewart. Integral Curvature Representation and Matching Algorithms for Identification of Dolphins and Whales. *arXiv:1708.07785 [cs]*, 2017. URL http://arxiv.org/abs/1708.07785. arXiv: 1708.07785.

[72] Hendrik J. Weideman, Charles V. Stewart, Jason R. Parham, Jason Holmberg, Kiirsten Flynn, John Calambokidis, D. Barry Paul, Anka Bedetti, Michelle Henley, Jerenimo Lepirei, and Frank G. Pope. Extracting identifying contours for African elephants and humpback whales using a learned appearance model. *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1276–1285, 2020.

[73] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *European conference on computer vision*, pages 1–16. Springer, 2014.

[74] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition*, pages 34–39. IEEE, 2014.

[75] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.

Chapter 10

PITFALLS AND RISKS

Sara Beery* and Elizabeth Bondi*. Can poachers find animals from public camera trap images? *CV for Animals Workshop at CVPR*, 2021.

Sara Beery* and Ellie Warren*. The Promise and Pitfalls of Machine Learning for Conservation. *WILDLABS Series on Technical Difficulties*, 2021.

Machine learning is often touted as conservation technology's silver bullet, a tool that will make conservation work easier, faster, and more effective. But those in conservation who work with machine learning can tell you from experience that it's far from a magic solution, and in fact, all the hype surrounding machine learning's potential makes its failures both surprising and frustrating. When machine learning tools fail to deliver consistent results, i.e. when a model achieves very high accuracy on a prototype dataset but doesn't work in the field, the cause is often that the prototype data wasn't representative of the end use case. This means that when the model "fails" it's really being asked to do something significantly outside the scope of what it has been trained to do. And because many conservation practitioners don't yet understand exactly what machine learning is capable of, they're more likely to buy into hype and sky-high expectations, resulting in a feedback loop that leads them to expect near-perfect performance, and then feel disappointed by the inevitable letdown.

One real challenge is that data curated for ecology tends to be project-specific, covering limited geographic areas or taxonomic groups, and collected from project-specific sensors. To build a one-size-fits-all machine learning model, you would need to collect a dataset that covers all possible use cases–which in a changing world is essentially impossible. So how does that relate to our expectations for machine learning, built around promises like 99% accuracy? In almost every paper that promises those kinds of results for the ecological community at large, the data the authors have trained their model on will not support their broad claims. Beginners expect that training a ML model is the challenging part, but really, the careful curation of diverse and representative training and evaluation data for a new task is equally if not more challenging than training the model. It is possible to attain deceptively high accuracy in a highly controlled setting–fixed sensors, time periods, sensor placement strategies, etc. The problem is that a model trained on data from one highly controlled setting is unlikely to work well in other settings, making it "fail" when used on e.g., data collected by other ecologists. Often the literature does not test their trained models in a way that shows how the models will work for other potential users. Exacerbating this, media hype makes it seem like 99% accuracy for one project means 99% accuracy for everyone, with no additional effort. That disconnect between expectations and reality leads to the perception that anything less than almost perfect results in actual practice is a failure an unusable, when in fact, a slightly-less-perfect machine learning model can still save you a lot of time and effort. When we evaluate models with generalizability as our goal, testing on data that is representative of the types of extrapolations to novel settings the model is likely face during use, it enables us to learn which models work well enough to be deployed and used off-the-shelf and which models need to be retrained for each new setting.

Another challenge is lack of clarity among conservation practitioners about the common metrics for measuring machine learning's success. Metrics like accuracy can be misleading depending on the data in question. For example, if a model achieves 90% classification accuracy while predicting whether an image is of a dog or of a mountain lion, that seems exceptional. But if the data used to measure that performance is 90% dogs, the model could get that accuracy by predicting 'dog' for every single image. When data is imbalanced, sometimes a class-averaged metric is more interpretable, because it better captures how well the model is doing across all of the possible classes. However, optimizing for class-averaged metrics frequently result in models that make more errors on common species, and if a common species is 90% of the data that can lead to a lot of errors. We often recommend users break down performance across classes, and if applicable across sensor deployments, seasons, and regions, to better understand what the model is doing for their use case and to decide which images to trust the results for, and which images to send for additional human review.

Beyond the prototyping data for training and evaluation, the choice of evaluation metrics, and the model itself, there are additional tools needed to effectively deploy ML models for conservation, including systems to get the data from the field to the models and tools for visualizing, correcting, and analyzing the results. Models like MegaDetector [2], which we trained to detects humans, vehicles, and animals (here "animal" is a single, generic class) in camera trap data, save users time combing through massive amounts of empty photos. A significant amount of effort has gone into code that makes it easy to try MegaDetector on data from new projects and to make the results of the model easy to interact with, analyze, and interpret. This code includes interactive Google CoLab Notebooks, a batch processing API to run large amounts of data through the model in parallel on the cloud, and even tools for filtering out possible repeat detections for a given camera, which sometimes occur when the model is confused by an object in the background of the frame like a rock.

When considering whether and which machine learning tools will successfully meet your needs, it is important to consider your priorities, resources, and the risk associated with errors for your study. When MegaDetector first became available to users, there was a lot of uncertainty among users about the types of possible error and how to put MegaDetector into practice most effectively and reliably for their specific camera network and target species. The model was built to work as well as possible off the shelf, anywhere in the world and for any animal taxa. That said, every user has different needs and requirements depending on their study, so each user has to weigh their own pros and cons when it comes to how they use the model. No two user's needs are the same. For example, if an ecologist is seeking to monitor invasive rodents on islands, any sighting of a rodent is highly significant, so missed detections are very high risk. For this use case, you'd use a lower confidence threshold on model detections, which reduces the risk of missing a rodent but requires human effort to filter through a larger number of false positives. With any trained machine learning model, you can pick an operating threshold that will trade off between high recall, avoiding missed detections but resulting in more potential false positives to analyze, and high precision, where your predicted results are more accurate, but you may have a higher risk of missing something important. Knowing which one of those options will lead to the right tradeoff between human processing effort and risk for a given study should be rooted in the user's expertise and knowledge of the specific application in question and the resources available.

Those without previous experience working through meeting machine learning's intrinsic challenges can be frustrated by the fact that off-the-shelf tools don't work perfectly and cannot fully automate data processing. In contrast, expectations for use should shift to include investing time to carefully analyze how well any off-the-shelf tool works for the data of interest and learning how to fine-tune an existing model if project-specific predictions are needed (e.g., species, gender, age, or behavior).

We should adjust our expectations around the use of ML for conservation from ML completely replacing human data processing to ML providing assistance to reduce the human effort needed to process newly collected data. For example, MegaDetector user Beth Gardener at the University of Washington said, “We had a big image processing party last week [...] Because of the MegaDetector; 6 of us processed over 100,000 images in one day. That would have taken weeks or months before.” Despite not completely removing the need for human eyes on the data, this is a notable speedup in processing, and a resounding success for conservation AI.

Further, expectations of use should include doing continual quality control of a model for new seasons or new deployments, and understanding that models often must be iteratively re-trained to handle new data as conditions change. ML systems for conservation should be treated as continually changing and adapting with human help as opposed to assuming model training and evaluation are a one-time thing.

One example that demonstrates the need to practice diligence and good quality control habits comes from Wildlife Insights, a machine learning platform that seeks to tackle the challenge of robust, global species identification by curating diverse camera trap data from around the world, and simultaneously provides users with a powerful platform for data management and analytics. An issue arose when our team started to analyze the performance of a new model version before its release. The new version included a large amount of training data from a new projects, and to everyone’s surprise, for no apparent reason, the model began frequently predicting the presence of domestic cats. Lots and lots of cats, on images that were clearly deer or dogs or cows, species the model had handled very well in the past.

In what seemed like a sudden catastrophic failure, we were getting detections of cats in what felt like almost every photo from certain camera trap projects. To understand what went wrong, we must go back to a potential issue that our team thought had been thoroughly investigated. Every camera trap brand has its own logo or watermark on photos. We’d previously wondered if all those different logos would bias our machine learning models or impact performance. But in earlier testing, we found that different logos in different locations in the image frame didn’t seem to throw off generalization or make a difference in our results.

The mysterious abundance of cats turned out to be the result of a large project from an urban area that captured lots of cats and used only Bushnell cameras. By some chance, most other projects already in WI were using other camera trap models, so the algorithm learned the “easy” association that an orange Bushnell logo meant the image contained a cat. The team was reminded the hard way how machine learning models will always take the easy way out, and memorize spurious correlations in the data if possible. This model was so good at everything else, but was making these weird cat errors that just didn't make sense! It took us a while to figure out what had gone wrong. We might've realized that the Bushnell logo was causing the problem sooner if we hadn't already invested time into analyzing whether logos caused issues in previous versions of the model. Because we'd already invested that time and energy, it was easier for us to overlook it because it wasn't on our radar anymore as a potential problem. But that's a good example of accepting that just because your model doesn't have a problem with something now, it doesn't mean it'll never have a problem. Don't trust the results of any model without corroborating them; otherwise, you won't recognize those problems when they do pop up. Because the Wildlife Insights team caught the error and were able to determine the cause, we were able to retrain their model after cropping out logos and verify that it fixed the issue. The new model version no longer has a love affair with cats.

With more established and familiar types of conservation technology, like camera traps themselves, the idea of failure may be easier to digest. After all, hardware can malfunction, especially when exposed to the elements and unpredictable wildlife. But with all machine learning's hype as the future of conservation tech, our own expectations may be setting machine learning up to fail. And that's unfair. Machine learning will very likely play a huge role in conservation's future, particularly as tools like MegaDetector and Wildlife Insights make it more and more user-friendly. But like we've come to accept mishaps with hardware, we need to accept machine learning's realities and current limitations in order to realize its full eventual potential.

Equally important is recognizing that machine learning's current capabilities are not its ultimate destination. This technology, like all technologies, will only improve and become more accurate and accessible over time. And with increased accessibility, we see a bright future full of promise for machine learning. Ecologists and conservationists will need to develop their intuition for and critical analysis of machine learning, and that comes with use and experience. We must to break down knowledge barriers to make machine learning more accessible for ecologists to use practically. By giving people the skills to experiment with machine learning, we can open the door to innovative ideas, and new, exciting human-AI solutions for conservation and sustainability challenges.

10.1 A case study in the risks of publishing ecological data: can poachers find animals from public camera trap images?

To protect the location of camera trap data containing sensitive, high-target species, many ecologists randomly obfuscate the latitude and longitude of the camera when publishing their data. For example, they may publish a random location within a 1km radius of the true camera location for each camera in their network. In this paper, we investigate the robustness of geo-obfuscation for maintaining camera trap location privacy, and show via a case study that a few simple, intuitive heuristics and publicly available satellite rasters can be used to reduce the area likely to contain the camera by 87% (assuming random obfuscation within 1km), demonstrating that geo-obfuscation may be less effective than previously believed.

Introduction

Monitoring biodiversity quantitatively can help us understand the connections between species decline and pollution, exploitation, urbanization, global warming, and conservation policy. Researchers study the effect of these factors on wild animal populations by monitoring changes in species diversity, population density, and behavioral patterns using camera traps. Camera traps are placed at specific, often hard-to-reach locations in the wild, and capture images when there is movement. Recently, there has been a large effort in the biology community to open-source camera trap data collections to facilitate reproducibility and provide verification (since mistakes can cause overestimates [7]), as well as promote global-scale scientific analysis. By open-sourcing the images - not just metadata - collected across organizations, scientists studying a specific taxa can pool resources and leverage bycatch (images of species that were not the target of the original study, but are still scientifically valuable) from other camera trap networks. They will also be able to study animal behavior. A great deal of camera trap images are publicly available from all over the world, including via websites hosted by Microsoft AI for Earth and University of Wyoming [1] and Google [11].

However, as mentioned on the Wildlife Insights FAQ page [12], “Won’t Wildlife Insights images reveal the locations of endangered species to poachers?” They answer that “While Wildlife Insights is committed to open data sharing, we recognize that revealing the location for certain species may increase their risk of threat. To protect the location of sensitive species, Wildlife Insights will obfuscate, or blur, the location information of all deployments made available for public download[1]

[1] Public downloads are not yet available in Wildlife Insights.

so that the exact location of a deployment containing sensitive species cannot be determined from the data. Practices to obfuscate the location information associated with sensitive species may be updated from time to time with feedback from the community." Community science (also known as citizen science) initiatives have also obfuscated locations to protect endangered species, such as eBird and iNaturalist, as there have been cases where community science and/or other open source data has informed poaching [6].

While obfuscating locations is encouraging, it is not clear whether it is sufficient to prevent geolocalization, or whether it is also necessary to blur or otherwise obfuscate portions of the images themselves. For example, for images in cities, geolocalization is typically based on recognizable landmarks and geometries, such as relationships between buildings and roads, heights of buildings, and strong architectural features or signage. The more recognizable a feature (such as a famous landmark or horizon), the easier an image is to geolocate. If you see an image containing the Chrysler building, it's easy to know that you're most likely in NYC. If you can see the outline of Mount Rainier, you're most likely in or near Seattle. An image of the exterior of a nondescript chain hotel or stretch of highway might be more difficult to place. We believe both human intuition and automated methods such as [10] take advantage of these features. The same might be said of camera trap imagery: if your image contains a noticeable landmark (for example a set of large rocky outcroppings or a body of water), it might be easier to estimate its location. In contrast, an image of dense undergrowth is only as potentially geolocalizable as your ability to recognize and model the distributions of its visible flora and fauna, and your ability to estimate a latitudinal band based on the timings of sunrise and sunset. In order to make these data publicly available, it is imperative that we understand whether these camera trap images can reveal locations, and if so, how to prevent locations from being revealed.

In this case study, we investigate how "obfuscated" these camera locations are, both with existing off-the-shelf geolocalization models and with a human-in-the-loop algorithmic approach we define as a proof-of-concept. We show that while existing models struggle to accurately locate camera trap images, a systematic method of filtering targeted to a specific conservation area using publicly available satellite rasters can be used to find specific candidate areas that are quite accurate, rendering the geo-obfuscation ineffective and indicating that the answer to our titular question is yes.

Related Work

It has been shown in prior work that geolocation can be determined from images. For example, PlaNet [10], an open-source deep learning approach pairing ground-level views and satellite data (see examples in Fig. 10.2), can predict a set of plausible locations of any image, including nature scenes. Most of these methods require multiple images for better performance, which are readily available for camera trap image collections. Other approaches focus on identifying objects in the scene, and using those identities to predict the locations [13]. Preventing locations from being revealed from images has also been considered in previous work, should this be necessary for camera trap images [14]. In particular, [14] applies to general image collections, such as those that might be posted to social media by users, and strategically deletes images until the location is ambiguous. However, all camera trap images taken from a single camera will have the same background, making it difficult to strategically remove images to reduce geolocalizability in a set of camera trap images. Camera trap images may also have very specific local landmarks, such as a well-known rock or tree, known to those familiar with an area but potentially hidden from generic deep learning methods.

Case study with Mpala Research Center

Our goal with this case study is to simply prove that geolocation is possible from camera trap imagery and metadata, indicating that sensitive animal locations could be vulnerable. We focus on Mpala Research Center and explore both an off-the-shelf deep learning method for geolocalization, as well as a human-in-the-loop method.

Mpala camera traps

The network of cameras we selected for our proof-of-concept is located at Mpala Research Center in Laikipia, Kenya. These 100 camera traps were initially placed as part of the 2020 Great Grevy's Rally, and have been continually collecting data over the past year. They capture a variety of habitats and backgrounds, including open savanna, two types of forested area, changes in elevation, and sites with visible horizon and without. You can see the diversity of landscapes across Mpala in the map in Fig. 10.1.

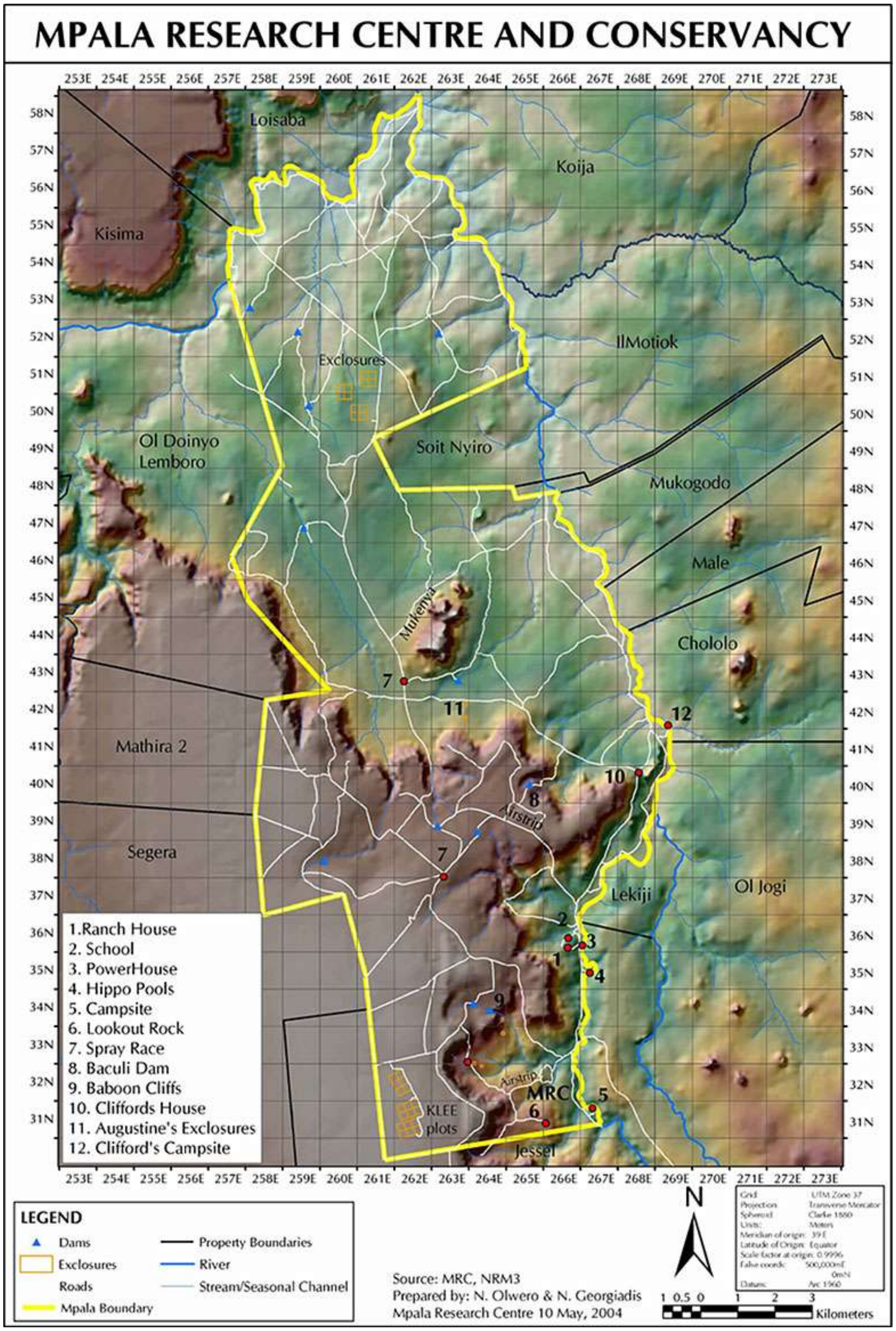
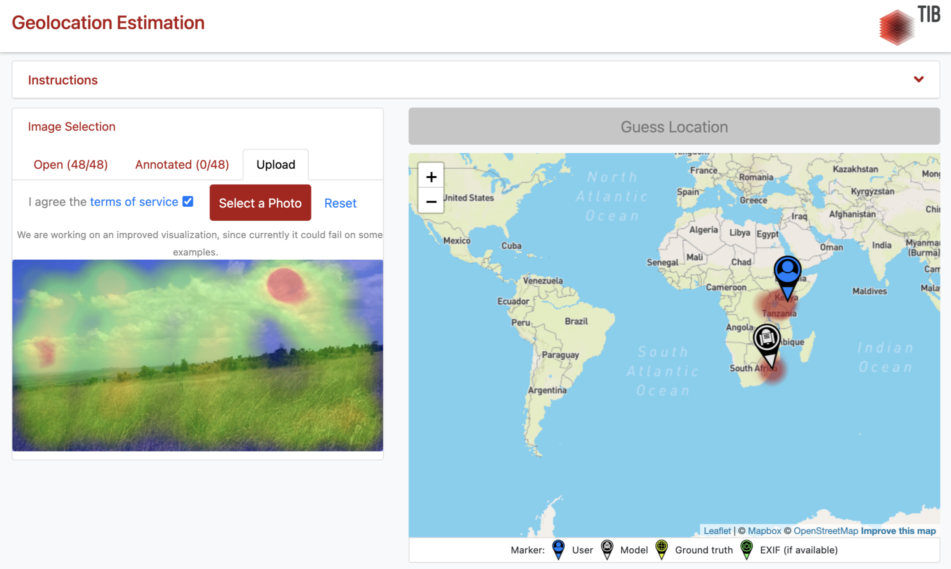


Figure 10.1: Map of Mpala Research Centre in Laikipia, Kenya, where the camera traps we studied were located.



We are working on an improved visualization, since currently it could fail on some examples.

Figure 10.2: PlaNet results on an image from a Mpala camera trap. The blue marker is an approximation of the ground truth location, the white marker is the model prediction.

Off-the-shelf results

We tried the existing PlaNet model on examples of camera trap data from Mpala, with results in Fig. 10.2. Given images from Mpala, PlaNet predicted large potential location areas covering Kenya, Tanzania, and South Africa. This may be due to the large amount of animal safari-based ecotourism in these countries. These areas are much larger than the potential randomness in location prescribed by most geo-obfuscation policies, rendering the off-the-shelf model unhelpful in further localizing the cameras. However, given camera trap-specific training data, similar deep learning-based methods may prove significantly more accurate. It is also interesting to note that the model seems to focus attention on the sky, which may imply that we should minimize the visibility of landmarks and horizons.

Satellite Rasters

We primarily utilized two sources of imagery from Google Earth Engine, specifically, (i) elevation data [8], which were collected in 2000 with a native resolution of 30 m and ranging from about 1600-1800m in our region, and (ii) Sentinel data [5], specifically the red, green, and blue bands, which were collected in 2020 at a native resolution of 10 m. We downloaded each through Google Earth Engine's platform at 10 m resolution (the minimum of the two) over the same area to cover

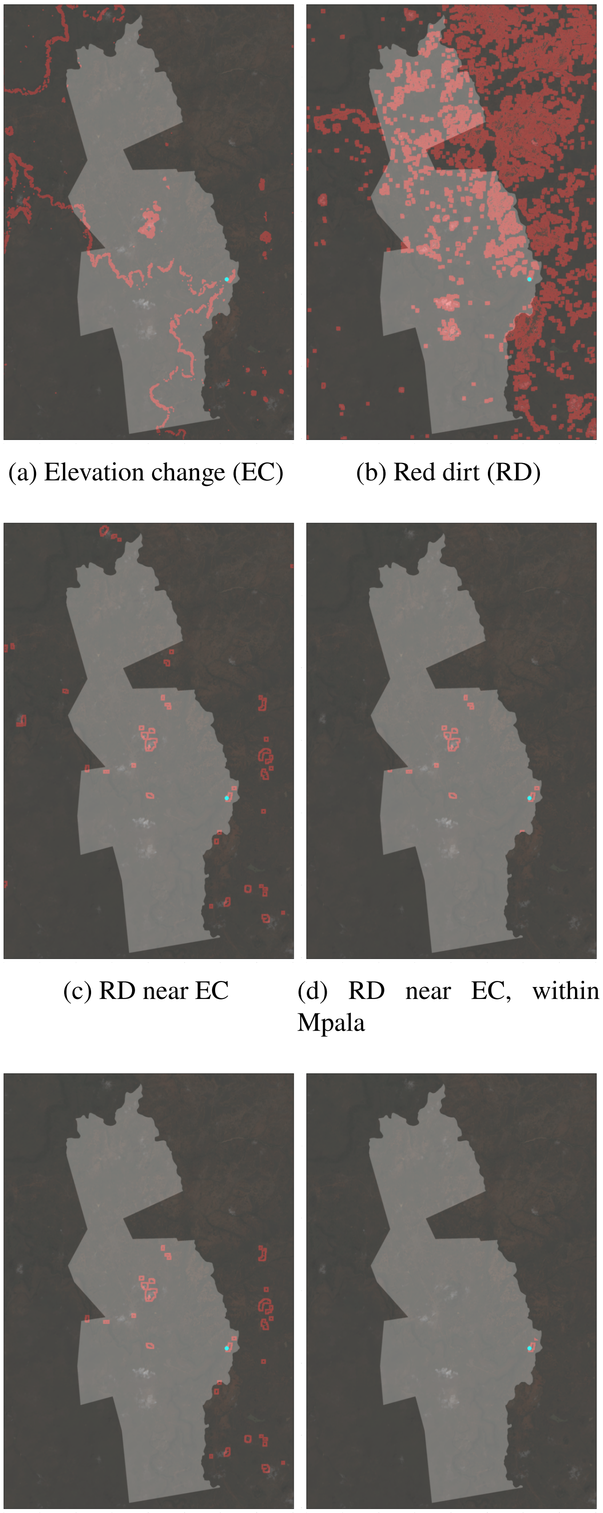
Mpala Research Centre, and all 100 camera traps. We then stacked these to form a multi-layer GeoTIFF.

We also considered using a landcover map from Google Earth Engine [3], but we found that the classes did not have a great deal of distinction or resolution over our particular area of interest. We also note that if you know what part of the world you are in, and approximate sunrise and sunset directions, then it is possible to guess the approximate camera facing from just a few images (see Fig. 10.4). Methods for automatically determining sun direction [9] and camera position [4] based on shadows have been investigated in the computer vision literature, and could be used to scale up facing estimation for a large set of cameras. These and other data could certainly be included in the future to further improve geolocation results.

Human-in-the-loop geolocalization

We manually sampled one location from the camera traps to attempt to geolocate. We chose this location because it seemed to have recognizable features, for example, red-tinted soil and a large rock nearby (see examples in Fig. 10.4), as we discussed in Sections 10.1 and 10.1. We decided to look for exactly these traits based on our observations. First, we computed the gradient of the elevation band and filtered for steep elevation change. We next searched for areas that were primarily red by thresholding the red band of the Sentinel data. We knew from our image that the camera trap was not in the red area itself, so we used morphological operators to select a small area surrounding red areas. In particular, we first did a closing operation to fill in gaps between small areas of the mask, then dilated this twice: first to represent a minimum distance away from the red area, then to represent a maximum distance away from the red area. We then subtracted the minimum dilation to get a “donut” shape around the red areas. We needed to do this for at least one of the two features in order to observe the areas of overlap (i.e., a red area nearby a rock).

Once we found these two areas, we simply carried out an AND operation between the two masks. The remaining mask represented our candidate locations for the *camera trap view*. However, the camera is placed at some distance to view this scene. Therefore, we carried out the same operation as when searching for “near red” areas. In particular, we estimated the distance of the camera from the landmarks in the image, and set the minimum and maximum dilation distances accordingly. We also adjusted the dilation distances to account for the fact that the camera might



(a) Elevation change (EC)

(b) Red dirt (RD)

(c) RD near EC

(d) RD near EC, within Mpala

(e) RD near EC, 10km obfuscation

(f) RD near EC, 1km obfuscation

Figure 10.3: Human-in-the-loop geolocation filtering. In each example, the park boundary has been overlaid for context, and the camera location is represented by a small blue dot. The red portions of the image are “potential locations” for the camera.

Filter type	**Area (km^2)**
RD near EC	6.8301
RD near EC, within Mpala	2.0688
RD near EC, 10km obfuscation	5.1373
RD near EC, 1km obfuscation	0.2641

Table 10.1: Remaining area (in square kilometers) needed to search on foot to find the camera location, using our "red dirt near elevation change" human-in-the-loop filtering method and varying levels of geo-obfuscation.

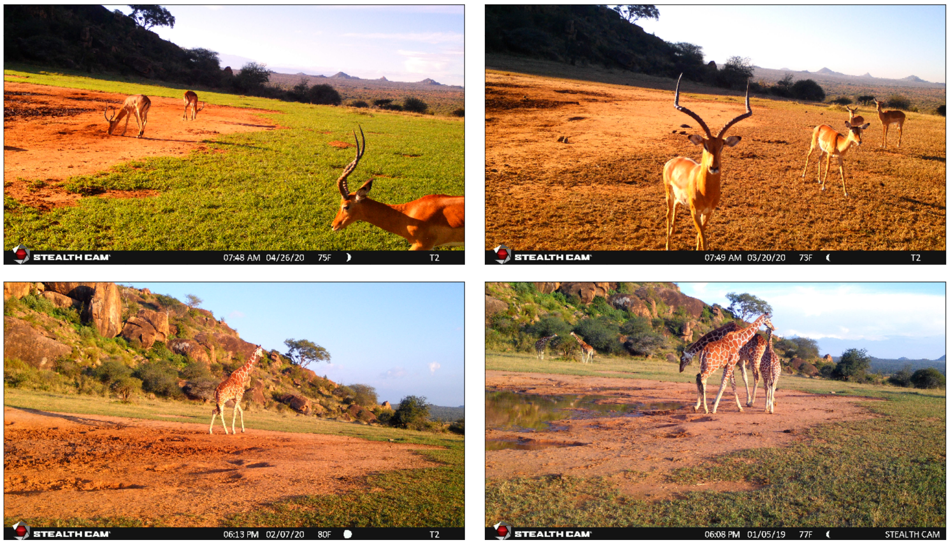


Figure 10.4: Images taken at sunrise (above) and sunset (below) imply that this camera is facing south by southeast.

be located diagonally from the area of interest, and the kernels used for dilation are pixel- rather than distance-based, resulting in differing growth distance with each dilation diagonally vs. horizontally and vertically.

This provided us with final candidate locations for the camera trap. We therefore calculated the area of these locations by simply computing the final number of candidate pixels, and then multiplying by the area of these pixels, which is $10m * 10m = 100m^2$. This gives us the final "searchable area," which we report in Table 10.1, row 1.

Geo-obfuscation

Mpala Research Centre is about 200 sq km in area, and the result from Table 10.1, row 1 contains a slightly larger region due to the rectangular image encompassing the park. To synthesize the case where we release imagery and don't provide coordinates but do provide the park name, we restrict our final candidate locations by the boundaries of the park exactly. Similarly, we repeat the calculation as though we were provided the coordinate geo-obfuscated by 10km and 1km. Our full results can be found in Table 10.1. Providing the park name and using these simple image processing techniques can narrow the search space from 200 sq km to 2 sq km, and if the provided coordinates are known to be obfuscated by 1km this can narrow the search space to 0.26 sq km. For reference, 0.005 sq km is the area of an American football field, meaning 0.26 sq km is about 52 football fields. While this is still large, we believe that it would be possible to traverse this already, and likely further refine the predictions from satellite imagery with more sophisticated methods, including estimating camera facing and/or landcover, which could cut the search area in half. We emphasize that this reduction in search space was largely due to the presence of features in the image, especially the rock and soil landmarks. Again, this implies that we should further investigate avoiding or hiding such landmarks in camera trap imagery to protect the geolocation.

Case Study Discussion

Using a very simple set of operations on human-generated heuristics based on publicly-available satellite rasters, we have shown that it is possible to drastically reduce the potential areas in which a camera may have been placed, meaning that poachers could theoretically find animals from public camera trap images. Based on our findings, one simple way to restrict the potential geolocalizability of your camera trap data could be to consciously place cameras in positions where the horizon and/or landmarks are not visible. In future work, we hope to further analyze the performance of human-in-the-loop methods and investigate fully-automated methods for geolocalization based on deep learning to better understand how to protect sensitive species while promoting scientific understanding. We therefore bring this new challenge to the computer vision community: Can we analyze which types of features in a camera trap view lead to easier geolocalization? And if so, can we adversarially remove the localizeable features to preserve privacy without removing vital ecological information?

10.2 Conclusions

While Machine learning and Computer Vision can provide significant benefits, cases like this one show that automated solutions can also provide risks to endangered species. It is vital that end users of these models be provided with the tools to evaluate the risks of using the models in real-world settings in a nuanced way. Similarly, though de-siloing ecological data can significantly improve the scale at which we are able to monitor species, we must be careful that the data we publish doesn't contain information, explicitly or implicitly, that could be used for harm.

Acknowledgements We would like to thank the entire team at Mpala Research Centre, and specifically Dan Rubenstein and the Laikipia Zebra Project Team led by Rosemary Warungu for their help managing the camera traps used to collect the data used in this case study. We would like to thank the team at WILDLABS, and specifically Ellie Warren, for their contributions to the broader discussion of "failure" in conservation technology. This work was supported, through funding, data storage, and computing resources, by Google AI for Nature and Society, Wildlife Insights, the Caltech Resnick Sustainability Institute, and NSFGRFP Grant No. 1745301, the views are those of the authors and do not necessarily reflect the views of these organizations.

References

[1] Lila.science. http://lila.science/. Accessed: 2019-10-22.

[2] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *In the Data Mining and Artificial Intelligence for Conservation Workshop at Knowledge Discovery in Databases (KDD)*, 2019. *selected to be featured at KDD Earth Day.

[3] Marcel Buchhorn, Myroslava Lesiv, Nandin-Erdene Tsendbazar, Martin Herold, Luc Bertels, and Bruno Smets. Copernicus global land cover layers—collection 2. *Remote Sensing*, 12(6):1044, 2020.

[4] Darren Caulfield and Kenneth Dawson-Howe. Direction of camera based on shadows. In *Proceedings of the Irish Machine Vision and Image Processing Conference*, pages 216–223. Citeseer, 2004.

[5] European Union/ESA/Copernicus. Sentinel-2 MSI: MultiSpectral Instrument, Level-2A. https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR.

[6] April Glaser. Online privacy is for the birds. hcn.org/articles/climate-desk-online-privacy-is-for-the-birds, 2021.

[7] Örjan Johansson, Gustaf Samelius, Ewa Wikberg, Guillaume Chapron, Charudutt Mishra, and Matthew Low. Identification errors in camera-trap

studies result in systematic population overestimation. *Scientific Reports*, 10 (1):1–10, 2020.

[8] NASA / USGS / JPL-Caltech. NASADEM: NASA NASADEM Digital Elevation 30m. URL `https://developers.google.com/earth-engine/datasets/catalog/NASA_NASADEM_HGT_001`.

[9] Scott Wehrwein, Kavita Bala, and Noah Snavely. Shadow detection and sun direction in photo collections. In *2015 International Conference on 3D Vision*, pages 460–468. IEEE, 2015.

[10] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, pages 37–55. Springer, 2016.

[11] Wildlife Insights. Bringing Cutting-Edge Technology to Wildlife Conservation. `https://www.wildlifeinsights.org`, 2022.

[12] Wildlife Insights. FAQ. `https://www.wildlifeinsights.org`, 2022.

[13] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015.

[14] Jinghan Yang, Ayan Chakrabarti, and Yevgeniy Vorobeychik. Protecting geolocation privacy of photo collections. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 524–531, 2020.

Chapter 11

CONCLUSIONS AND FUTURE DIRECTIONS

We require a real-time, adaptive, global-scale earth observation system that unites efforts across research groups in order to provide the information necessary to prioritize sustainability and conservation efforts and optimize our allocation of resources. The development of such systems requires collaborative, interdisciplinary approaches that translate diverse sources of raw information into accessible scientific insight. In this thesis, I, in collaboration with my coauthors, have contributed curated benchmark datasets that help bridge the gap between CV&ML researchers and impactful environmental challenges, novel computer vision methods that address challenges raised by these benchmarks, and robust human-AI systems that have seen widespread use in the ecological community.

I see four significant open challenges in the pursuit of global-scale environmental monitoring, and expand upon each below. These fundamental and unsolved challenges are (1) making effective use of all available modalities of data, (2) incorporating expert knowledge systematically, (3) ensuring these systems are equitable and ethical, and (4) expanding the capacity of interdisciplinary knowledge needed to work effectively on these problems.

Learning from everything: reasoning across non-homogeneous data

Data is increasingly accessible in large volumes, collected from multitudes of diverse sensors and platforms. These modalities are complementary: no one data collection method can capture the entire picture. Valuable information is captured in everything from text-based historical records to social media posts to satellite imagery. There have been amazing recent successes in multimodal CV, particularly with video+audio and images+text. However, these methods barely scratch the surface of what is possible, focusing primarily on highly correlated pairs of modalities. There has also been extensive work on multimodal data fusion in domains like diagnostic medical imagery and land cover prediction, focusing on accurate co-registration of spatial data and generating interpretably fused imagery. I seek to expand the scope, building methods that reason about data across modalities despite non-homogeneous structure and vastly inconsistent spatial and temporal scales of sampling.

Ecosystem monitoring across modalities at global scale is an exciting and important testbed for extracting scientific insight from diverse, non-homogeneous data. We have built the foundation of a multi-year research program in this space–collaborating with researchers at Google, we are undertaking the first large-scale study of tree species categorization in urban forests. Our study covers 23 cities across North America and over 2.5M trees so far, as outlined in Chapter 8. We are combining satellite imagery from Google Earth Engine with on-the-ground data from Google Street View, iNaturalist, and local tree censuses, and are developing novel methods which use cross-modal agreement over time as self-supervision to efficiently adapt to unseen cities. One of the largest challenges we must face is how to efficiently sample data for human verification (data-efficient evaluation), particularly under long-tailed distributions where rare classes are the most scientifically important. We hope to collaborate with local ecologists and community scientists to investigate a combination of self-supervision, anomaly detection, and active learning to enact efficient validation and model adaptation via community science "bio-blitzes." In the future we hope to expand our methods to wild forests, using data from the National Ecological Observatory Network (NEON) and Wildlife Insights.

Incorporating knowledge systematically into learning

There is a considerable amount of domain-specific knowledge and theory, which is almost completely ignored by current CV methodology in pursuit of "pure" data-centric approaches. This causes real harm: biases in data are propagated through to systems without a priori understanding of domain-specific risk. Further, results from black-box models are uninterpretable, making these systematic errors difficult to catch without domain experts carefully probing models with known high-risk corner cases.

General methods provide opportunity for outsize impact when designed carefully, keeping in mind the diverse set of potential use cases and risks [2]. We need to understand the tradeoffs between generality and domain specificity in our methods in order to develop rigorous ways to think about designing impactful end-to-end solutions–computer vision systems that are general purpose but optimal for each stakeholder. We have begun to study this tradeoff in large-scale systems where each user has a unique and specific goal for their ecological study, such as Wildlife Insights and our ongoing project on sustainable fisheries management [4]. We will investigate methods that can efficiently and systematically capture domain expertise as a model is training, such as risk-aware CV and data programming, or at inference time via active model adaptation.

Equitable, ethical technology

Conservation and sustainability are time-sensitive global challenges: to make rapid progress we must build and deploy solutions that are accessible to any stakeholder (academic research groups, policymakers, on-the-ground conservationists, etc.) that would benefit from, and would like to use, the technology. I witnessed firsthand the gap between cloud-based methods and user need when I deployed a network of 100 camera traps in Kenya. The fastest, most cost-effective way to extract information from the raw imagery is to mail terabytes of data to the US, where it can be quickly uploaded to the cloud and analyzed at-scale with our CV models. The cost of computation, data storage, and data movement make many automated solutions inaccessible to researchers and practitioners outside of wealthy countries like the US and Europe. I seek to make CV more equitable and deployable by developing methods which (1) increase efficiency of training, inference and data storage, and (2) incorporate and expand federated learning to enable models to learn from multiple organizations without data needing to be centralized, vital in cases where data privacy must be preserved (as we show in [1]) and where data movement is resource-constrained as bandwidth is limited (e.g., in the remote ocean).

Building interdisciplinary knowledge capacity

Another bottleneck to the widespread use of CV for conservation is access to the knowledge and skills needed to build, train, and deploy AI-based solutions for the environment. There are many more potential uses for and applications of CV in ecology then their are trained experts who are able to undertake these challenges. I am passionate about democratizing access to the powerful tools and technical skills found in in CV/ML, in order to empower conservation practitioners and ecological researchers to build their own CV systems to address their own research questions or protected area management needs. I have developed the curriculum for a three week summer school [1] designed to teach applied computer vision to senior graduate-level ecologists, where each student will bring their own ecological question and relevant data and leave the course with a prototype system to process the data using state of the art computer vision methodologies.

[1] cv4ecology.caltech.edu

We have secured funding for the school for the first three years (2021-2023) from the Resnick Sustainability Institute, Microsoft AI for Earth, and Amazon AWS. One of the main goals of this summer school is teaching the thought process behind computer vision. Beyond fundamental concepts, we teach the intuition behind how a computer vision researcher might (1) define a specific ecological problem within the framework of ML, (2) build a dataset, (3) select model architectures, and (4) evaluate performance–teaching how we think, instead of what we know [3].

References

[1] Sara Beery* and Elizabeth Bondi*. Can poachers find animals from public camera trap images? *CV for Animals Workshop at CVPR*, 2021.

[2] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *In the Data Mining and Artificial Intelligence for Conservation Workshop at Knowledge Discovery in Databases (KDD)*, 2019. *selected to be featured at KDD Earth Day.

[3] Talanquer et al. Let's teach how we think instead of what we know. *Chemistry Education Research and Practice*, 2010.

[4] Justin Kay and Matt Merrifield. The fishnet open images database: A dataset for fish detection and fine-grained categorization in fisheries. *arXiv preprint arXiv:2106.09178*, 2021.

Appendix A

RECOGNITION IN TERRA INCOGNITA: SUPPLEMENTARY MATERIAL

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

A.1 Additional experiments

Varying the amount of training data per location

We chose to use a small number of locations as this is a key variable of the generalization problem. In the limit, we would study the behavior of models trained on a single location with "unlimited" training data. We did not have access to such a dataset, and therefore used 10 training locations in order to have a sufficient number of training examples. To verify whether 10 training locations would yield significantly different results than 1 training location, we ran our bounding box experiments with a quarter, half, and all of the images available per training location, and saw trans test accuracies of 80.6%, 83.0%, and 83.4% respectively. This implies that *increasing the number of images per location would not solve the generalization problem.*

Varying the number of training locations

As an additional control, we experimented with varying the number of training locations (see Fig. A.1(Left)), and find that trans performance is stable as the number of training locations is increased beyond 2. Thus, we are confident that our dataset is adequate to measure generalization ability. We expect the generalization gap to narrow with $N >> 10$, but as the number of training locations increases the focus of the experiment shifts. We want to provide a test bed to specifically study generalization when provided with few training locations.

Varying the validation location

To analyze the effect of the validation split, we repeated our experiments with 2 other validation locations (see Fig. A.1(Right)). We find that test performance

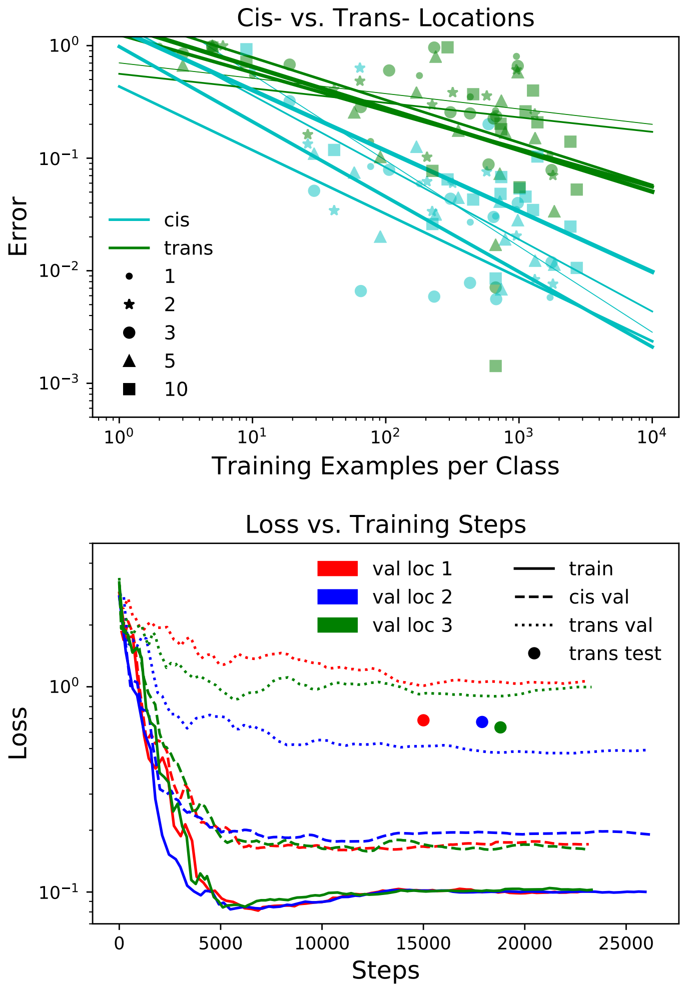


Figure A.1: **Generalization metrics are robust to N. locations and to validation.** Both plots are based on bounding box classification. **(Left)** Error per class vs. number of training examples (best-fit line width denotes number of training locations in $1, 2, 3, 5, 10$). Trans performance is stable for N locations with $2 < N \leq 10$. We chose 10 training locations to study generalization behaviors while providing maximal data for experimentation. **(Right)** Loss curves using different locations for trans-validation. The test loss for the selected model for each validation set remains stable, implying that the choice of validation location does not greatly impact trans test performance.

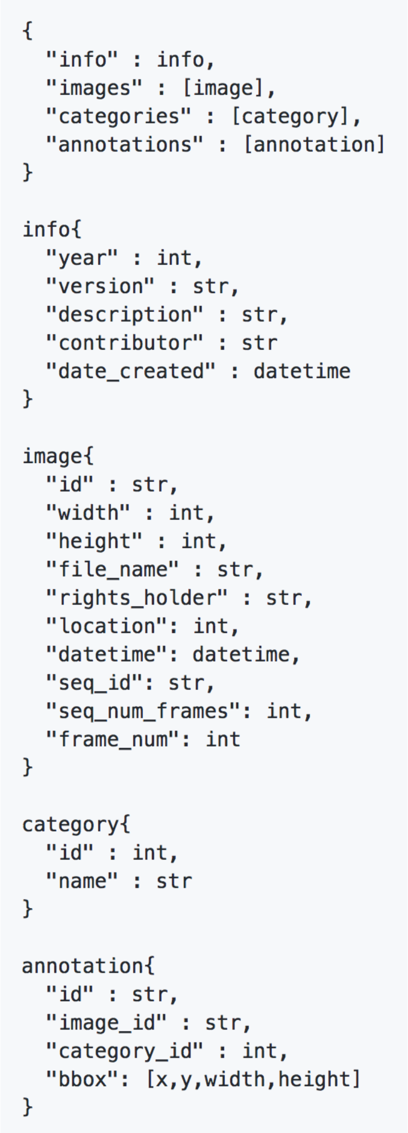
is relatively stable regardless of the validation split. Fig. A.1(Right) also shows training and validation curves for the three different validation experiments.

A.2 Data format

We chose to use an adapted version of the JSON format used by the COCO dataset with additional camera trap-specific fields, which we call COCO-CameraTraps. The format can be seen in Fig. A.2.

We added several fields for each image in order to specify camera-trap specific information. These fields include a location id, a sequence id, the number of frames in that sequence, and the frame number of the individual image. Note that not all cameras take sequences of images at a single trigger, so for some images the number of frames in the associated sequence will be one.

All data can be accessed at `https://beerys.github.io/CaltechCameraTraps/`.



```
{
  "info" : info,
  "images" : [image],
  "categories" : [category],
  "annotations" : [annotation]
}

info{
  "year" : int,
  "version" : str,
  "description" : str,
  "contributor" : str
  "date_created" : datetime
}

image{
  "id" : str,
  "width" : int,
  "height" : int,
  "file_name" : str,
  "rights_holder" : str,
  "location": int,
  "datetime": datetime,
  "seq_id": str,
  "seq_num_frames": int,
  "frame_num": int
}

category{
  "id" : int,
  "name" : str
}

annotation{
  "id" : str,
  "image_id" : str,
  "category_id" : int,
  "bbox": [x,y,width,height]
}
```

Figure A.2: COCO-CameraTraps data format

Appendix B

SYTHETIC EXAMPLES IMPROVE GENERALIZATION IN RARE CLASSES: SUPPLEMENTARY MATERIAL

Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 863–873, 2020.

B.1 Architecture selection

To select a single classification architecture to use across our experiments, we trained three classifiers: ResNet-101 V2, Inception V3, and Inception-ResNet V2. All three classifiers were pretrained on *no-animal ImageNet* then trained on the Caltech Camera Traps (CCT) training set (described in the main paper, Section 3.1) with no added simulated images. We found that Inception-ResNet V2 performed best on deer in cis and trans scenarios (see Table B.1), so we decided to use Inception-ResNet V2 as the base architecture for all further experiments.

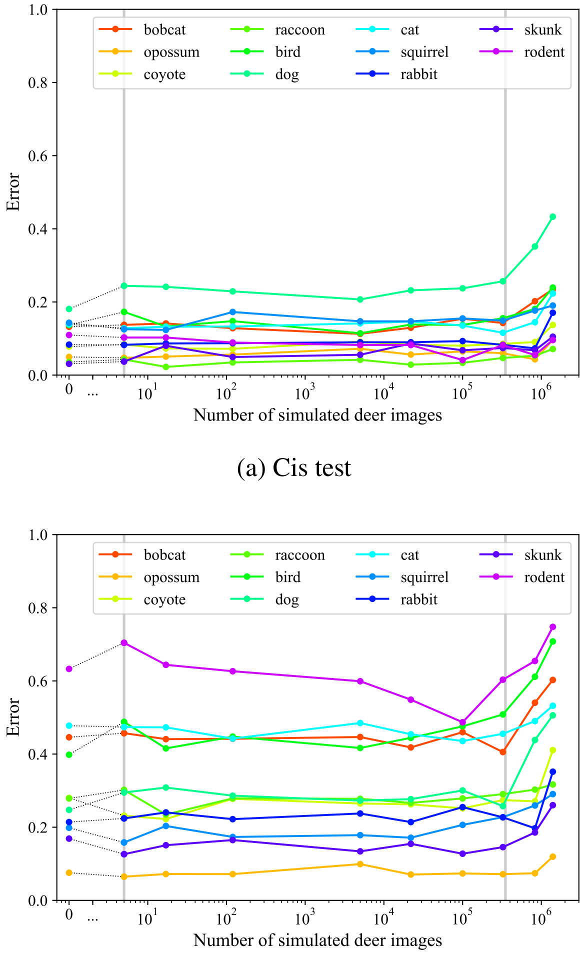
Table B.1: Error for different architectures. Error is defined as the number of incorrectly identified images divided by the number of images for each test set, where “Deer" contains only deer images and “Other" contains all non-deer images.

	Cis Test		Trans+ Test	
Architecture	Deer	Other	Deer	Other
Resnet 101 V2	47.86	11.18	88.63	**29.76**
Inception V3	50.00	11.74	81.73	32.74
Inception Resnet V2	**29.28**	**10.17**	**77.69**	31.07

B.2 Additional analysis

Per-class analysis of the effect of adding simulated deer images

By averaging over the performance of the non-deer classes in Figure 5 in the main paper, we have not changed the overall trend. The performance on each non-deer class stays reasonably constant until the number of added deer images goes above 325K.



(a) Cis test

(b) Trans test

Figure B.1: **Per-class performance on non-deer classes when adding simulated deer images.** The trends seen in Figure 5 in the main paper when averaging across classes hold for each individual class. Performance stays relatively constant until the number of added simulated deer images starts to bias the classifier, above 325K added images.

Analyzing the value of real images

We find that our simulated data is sufficient to learn to recognize some deer even without real examples, though the real examples give a large boost in performance. The performance breakdown can be seen in Table B.2. These results are promising for both researchers studying zero-shot learning and biologists studying highly endangered species: it is possible to learn a species with no real training data. This avenue remains open for further study.

Comparing night and day performance

We further analyze the effect of day and night simulation by comparing three experiments: one trained with only simulated daytime images, one trained with only simulated nighttime images, and one trained with half day and half night (see Fig B.2). We find that the models trained on only day and only night perform similarly on trans deer, and that the 50/50 split performs best on trans deer (highlighted region in Fig B.2). Training on day or night alone gives us a 20% performance boost on trans deer, while training on both gives us a 40% performance boost. This suggests that the day and night simulated images help the classifier in complementary ways: day helps with day images and night helps with night images. Performance on other classes is not strongly effected. Cis performance is quite noisy, and performs best with no added simulated data, see Fig. 2 in the main paper for further analysis.

Table B.2: Error with and without the 44 real deer examples when adding 100K simulated deer images. Error is computed as in Table B.1.

	Cis Test		Trans+ Test	
Real Training Data	Deer	Other	Deer	Other
CCT train w/o deer	94.29	18.64	68.56	34.42
CCT train w/ deer	52.14	10.91	44.05	30.47
% decrease from real deer	44.7	41.5	35.7	11.5

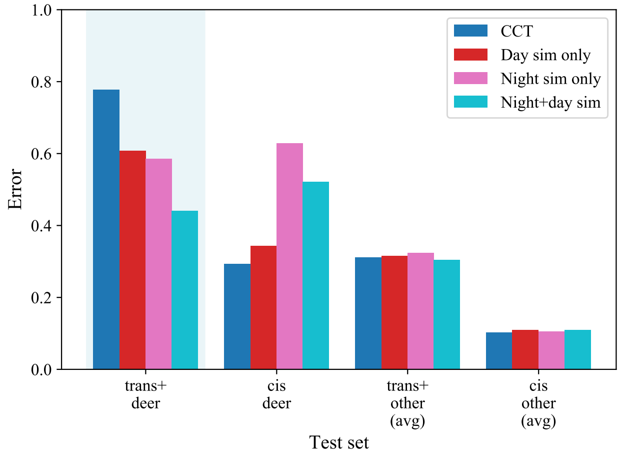


Figure B.2: Error as a function of day or night simulated images: 100K simulated deer images. Error is calculated as in Fig. 4 in the main paper. Trans+ deer performance is highlighted. Models trained on added night- or day-only simulated data perform better on trans deer than CCT alone, but the best trans deer performance comes from the 50/50 day/night split of added simulated data.

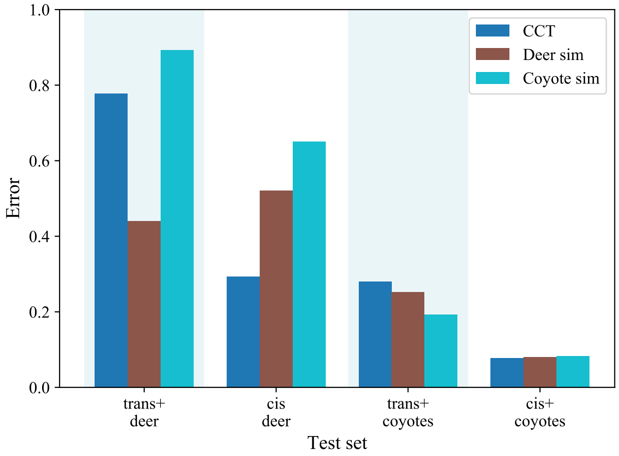
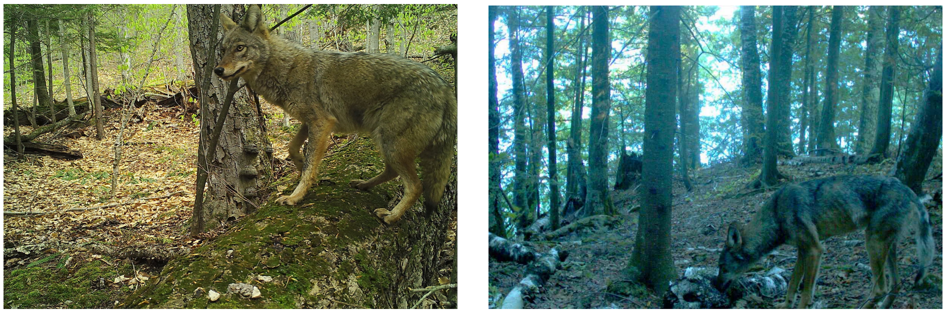


Figure B.3: **E**rror as a function of deer or coyote simulated images: 100K simulated images. Error is calculated as in Fig. 4 in the main paper. Trans+ deer and coyote performance are highlighted.



(a) Coyote [5]

(b) Wolf [16]

Figure B.4: **W**olves and coyotes are visually similar.

Investigating the effect of adding simulated data for a common class

In order to investigate how added simulated data might effect a common class, as opposed to a rare one, we created “coyote" simulated data with TrapCam-Unity, using rendered models of wolves as a proxy for coyotes. Off-the-shelf, high-quality wolf models were more widely available, and wolves and coyotes are visually very similar (see Fig.B.4). This is a coarse-grained experiment, and it remains to be seen what would happen if simulated data from two visually similar classes (e.g.wolves and coyotes) was added at the same time.

We find that adding simulated “coyote" data improves trans+ coyote performance slightly, while cis coyote performance remains the same. Unsurprisingly, for the deer class (which has few training examples) adding a large amount of simulated coyote data harms both cis and trans+ deer performance.

B.3 Creating sim and real on empty data

Alternative to the full synthetic methods of data generation with AirSim and Unity, we generated synthetic images by overlaying either simulated deer or real cropped deer on real empty background images from the CCT dataset (see Fig. B.5).

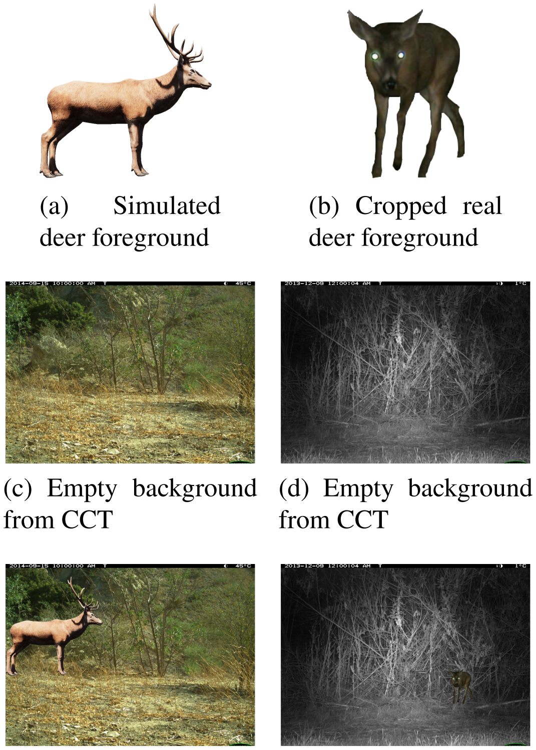
For the *Sim on Empty* dataset generation, we posed either a stag or a doe deer from the GiM model set in front of a simulated camera in Unity. We randomized the animation, orientation in azimuth (0-360 degrees), position, direction of light orientation in azimuth (0-360 degrees), and elevation (20-90 degrees).

For the *Real on Empty* dataset, we manually segmented and cropped out the 44 instances of deer from the CCT training set. Then we pasted the cropped deer foreground images on top of empty camera trap images in random locations. It is worth noting that we use real empty background to investigate the effect of real versus sim foreground deer, it is possible in future work to combine either type of foreground with sim background images.

B.4 TrapCam-AirSim details

It took time and thought to derive the overall requirements for the AirSim TrapCam environment. With a sizable number of potential biomes globally, we narrowed the scope of what we intended to build to a SW United States environment similar to what is seen in the CCT data. Eventually we settled on a sub-alpine woodland scene that is readily found across most of the Western/ Southwest US. A major requirement and challenge was how to get the most data out of a relatively small, but detailed, area - this was key to the project without expanding the size of the area of interest. The overall intent was to leverage Microsoft AirSim's computer vision mode to move a pre-configured camera around the scene, providing varied background.

We used various off-the-shelf components such as an animal pack from Epic Studios [6] (Animals Vol 01: Forest Animals by GiM [7]), background terrain from Unreal Marketplace [2], vegetation from SpeedTree [14], and rocks/obstructions from Megascans [11]. In other AirSim environments, the general scenery is fairly static with exception of particle effects (snow/rain/dust/etc). For this effort we wanted a method to vary the background, to replicate a variety of terrains within a single environment (see Fig.B.6). The actual area of the environment is small, at 50 meters long, but the modularity allows many possible scenes to be constructed. The randomization was designed to facilitate artists by allowing them to make a list of different objects to randomize from. Those objects are prioritized based on their



(a) Simulated deer foreground

(b) Cropped real deer foreground

(c) Empty background from CCT

(d) Empty background from CCT

(e) Sim on empty overlay

(f) Real on empty overlay

Figure B.5: Sim and Real on Empty Generation. (a),(c),(e) demonstrate the process of overlaying a simulated deer on top of an empty background image from the CCT dataset. (b),(d),(f) show the process of overlaying a cropped real deer on top of an empty background image from the CCT dataset.

order on the list. The BiomeTerrain class generates them by tracing random areas across the field based on a global seed. If there's space available it spawns the desired object. There are a number of object types available in TrapCam-AirSim; animal type, rocks, logs, grasses, shrubs, trees, and each type can be varied by density and distribution. Additionally, we provide 9 GiM animal models: deer (doe/stag), wolf, fox, rat, spider, bear, raccoon, and buffalo. The doe model was created by removing the antlers from the stag model with Maya [10], a common modeling tool. All animal objects were assigned segmentation IDs for efficient ground truth extraction.

We created a simple UI to vary parameters, along with a command line API for parameter configuration. The UI was constructed with Unreal Motion Graphics (UMG) Widgets and allows for future flexibility for modifications, DPI resolutions and platforms. The main core functionalities were created with C++ for better performance as a parent class for data-only blueprints, which allows the technical

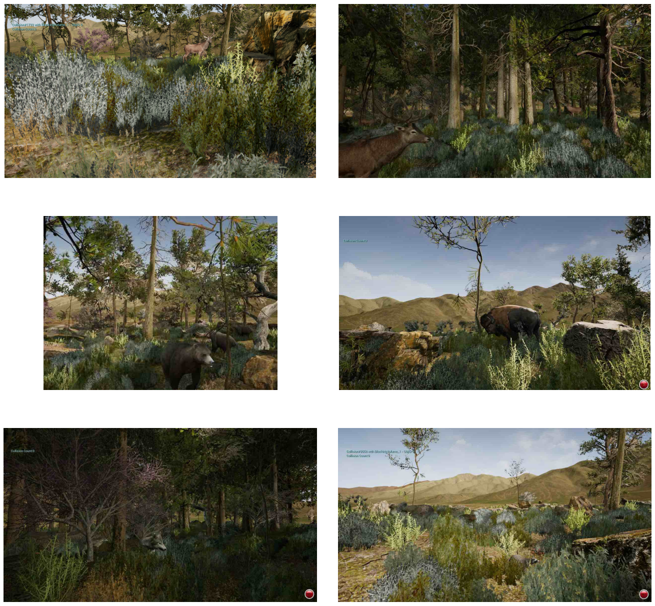


Figure B.6: **T**rapCam-AirSim environment. The TrapCam-Airsim envionment was designed to be modular and randomizeable, which allows a variety of biomes to be synthesized within a limited simulated area.

artists to easily swap assets for different environments without re-compiling the C++ code.

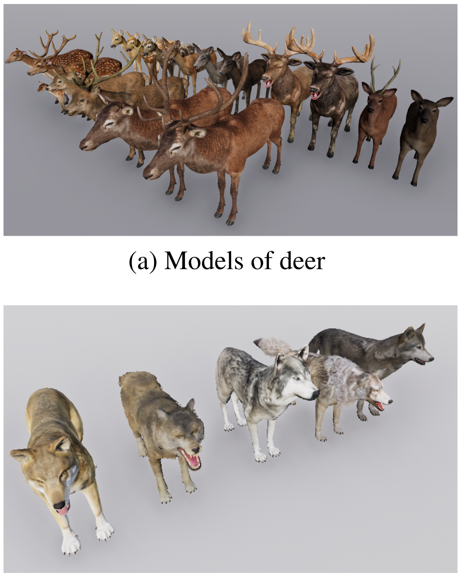
We started the requirements and scoping in mid-August 2018 with a go-ahead approximately September 6th, and produced a working prototype two weeks later, with continued development and refining through mid-October. A second phase late in the year modified the camera system to include flash capability, and animals were updated to provide eye-shine, and the UI was modified to include variability for that eye-shine.

B.5 TrapCam-Unity details

Simulation

The overall goal of our simulation is to take advantage of off-the-shelf components crafted for game development as much as possible so that we minimize manual labor and make the method more scalable and generalizeable. Specifically, we used off-the-shelf animal models and environment.

The “Book of The Dead" environment [4] we use is published for free by Unity. As shown in Fig.B.8, the near-photorealistic environment simulates a large patch of forest in a valley with volumetric grass, a variety of high definition trees, logs, and



(a) Models of deer

(b) Models of wolves

Figure B.7: **M**odels of deer and wolves. In TrapCam-Unity, we used 17 different models of deer from 5 different artists and 5 models of wolves from 5 different artists. We used the wolf models as proxies for coyotes (see Section B.2). Model details are available in Section B.5.

bushes, as well as rocks and terrain. The environment is a irregular area of roughly 20,000 m^2. It runs on a desktop PC in real time and enables us to generate large amounts of images efficiently.

To create daytime images we varied the orientation of the simulated sun in both azimuth and elevation. To create images taken at night we created a spotlight attached to the simulated camera to simulate a white-light or IR flash and qualitatively match the low color saturation of the night time images. To simulate animals' eyeshine (a result of the reflection of camera flash from the tapetum lucidum), we placed small reflective balls on top of the eyes of model animals (see Fig.B.9).

For deer simulation, we used 17 animated deer models from 5 publishers on Unity (GiM[8], 4toon[1], Protofactor[12], Red Deer[13], Janpec[9]). For coyote simulation, we used 5 models from 5 publishers (GiM[8], 4toon[1], Protofactor[12], Janpec[9], WDallgraphics[15]). We created the GiM doe model by removing the antlers of the GiM stag model with Blender[3]. For each of the animated models, we included an animation controller that contains several animation clips ranging from commonly seen behavior episodes like walking and eating, to rare occurrences like attacking and sleeping. During dataset generation, we randomly picked a clip

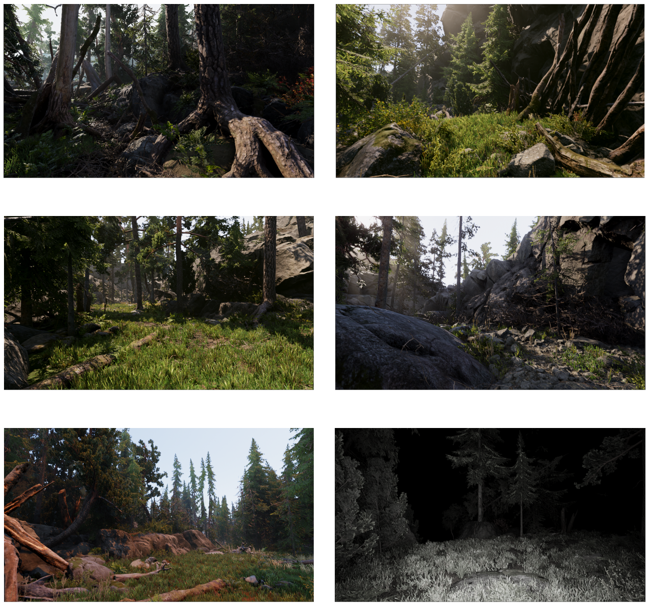


Figure B.8: **T**rapCam-Unity environment. The Book of The Dead environment is a large natural environment with diverse sub regions.

for each instance of animals and freeze it at a random time point, then we move the cameras around to sample a static scene with animals and environment.

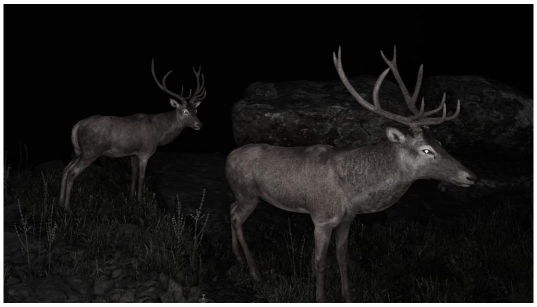


Figure B.9: **E**xample of eyeshine simulation.

We had 300 seed locations and randomly placed animals in the vicinity of a subset of the seed locations. This process was repeated multiple times to simulate animals in random locations within the environment. A similar random placement process was used to determine the locations of the cameras. All images generated are in full HD resolution (1980 x 1080).

For ground truth generation, we turned off the lighting and rendered each instance of the animal in a unique color by replacing the original animal shader with an unlit shader. We then used customized python scripts to extract animal bounding boxes by extracting pixels with these unique colors.

Scalability and Generalizability

All synthetic examples in this study are generated with off the shelf environments and models. We use our simulators to generate deer images for the sake of this study, but the simulators each currently include up to 30 simulation-ready species.

A large number of high quality assets already exist online in the game development community. For example, Unity Asset Store alone has 1382 items under the Animal category. There are also many environments available online, like the “A Boy and His Kite" environment for Unreal. Despite the abundance of readily made animal models and environments, it might still remain challenging if the species-environment combination is not covered by existing assets as the 3D assets need to be created by artists first. However, recent work in automating 3D model generation [17–20], might reduce the need for hand-crafted assets in the future.

References

[1] 4toon studio. `https://assetstore.unity.com/publishers/3695`. Accessed: 2019-03-27.

[2] Unreal game engine. `https://www.unrealengine.com/en-US/what-is-unreal-engine-4`. Accessed: 2019-02-05.

[3] Blender. `https://www.blender.org/`. Accessed: 2019-03-28.

[4] Book of the dead environment. `https://assetstore.unity.com/packages/essentials/tutorial-projects/book-of-the-dead-environment-121175`. Accessed: 2019-03-27.

[5] Coyote in a camera trap. `https://www.inaturalist.org/photos/7738216`. Accessed: 2019-03-28.

[6] Epic studios. `http://epicstudios.com/`. Accessed: 2019-03-21.

[7] Forest animals by GiM. `https://www.unrealengine.com/marketplace/en-US/animals-vol-01-forest-animals`, . Accessed: 2019-03-21.

[8] GiM studio. `https://assetstore.unity.com/publishers/18347`, . Accessed: 2019-03-27.

[9] Janpec. `https://assetstore.unity.com/publishers/1066`. Accessed: 2019-03-27.

[10] Maya. `https://www.autodesk.com/products/maya/overview`. Accessed: 2019-03-28.

[11] Quixel megascans library. https://quixel.com/megascans. Accessed: 2019-03-21.

[12] Protofactor inc. https://assetstore.unity.com/publishers/265. Accessed: 2019-03-27.

[13] Red deer studio. https://assetstore.unity.com/publishers/12623. Accessed: 2019-03-27.

[14] Speedtree. https://store.speedtree.com/. Accessed: 2019-03-21.

[15] Wdallgraphics studio. https://assetstore.unity.com/publishers/5060. Accessed: 2019-03-28.

[16] Wolf in a camera trap. https://3c1703fe8d.site.internapcdn.net/newman/csz/news/800/2018/cameratrapst.jpg. Accessed: 2019-03-28.

[17] Thomas J. Cashman and Andrew W. Fitzgibbon. What shape are dolphins? Building 3D morphable models from 2D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):232–244, 2013.

[18] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[19] Frederik Pahde, Mihai Puscas, Jannik Wolff, Tassilo Klein, Nicu Sebe, and Moin Nabi. Low-shot learning from imaginary 3d model. *arXiv preprint arXiv:1901.01868*, 2019.

[20] Bernhard Reinert, Tobias Ritschel, and Hans-Peter Seidel. Animated 3d creatures from single-view video by skeletal sketching. In *Graphics Interface*, pages 133–141, 2016.

Appendix C

LONG TERM TEMPORAL CONTEXT FOR PER-CAMERA OBJECT DETECTION: SUPPLEMENTARY MATERIAL

Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context R-CNN: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13075–13085, 2020.

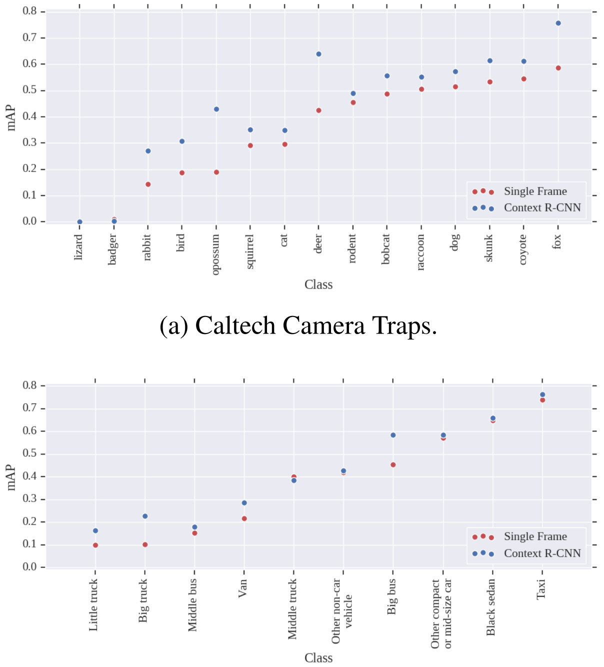
C.1 Implementation Details

We implemented our attention modules within the Tensorflow Object Detection API open-source Faster-RCNN architecture with Resnet 101 backbone [2]. Faster-RCNN optimization and model parameters are not changed between the single-frame baseline and our experiments, and we ensure robust single-frame baselines via hyperparameter sweeps. We train on Google TPUs (v3) [3] using MomentumSGD with weight decay 0.0004 and momentum 0.9. We construct each batch using 32 clips, drawing four frames for each clip spaced 1 frame apart and resizing to 640×640. Batches are placed on 8 TPU cores, colocating frames from the same clip. We augment with random flipping, ensuring that the memory banks are flipped to match the current frames to preserve spatial consistency. All our experiments use a softmax temperature of $T = .01$ for the attention mechanism, which we found in early experiments to outperform .1 and 1.

C.2 Dataset Statistics and Per-Class Performance

Each of the real-world datasets (Snapshot Serengeti, Caltech Camera Traps, and CityCam) has a long-tailed distribution of classes, which can be seen in Figure C.3. Dealing with imbalanced data is a known challenge across machine learning disciplines [1, 5], with rare classes (classes not well-represented during training) frequently proving difficult to recognize. Recognizing categories from only a few training examples is an open area of research, often referred to as “low-shot learning” or “few-shot learning.”

In Figure 5 in the main text, we demonstrate that the per-class performance universally improves for Snapshot Serengeti (SS). In Figure C.1, we show the per-class performance for Caltech Camera Traps (CCT). and CityCam (CC). Performance on



(a) Caltech Camera Traps.

(b) CityCam.

Figure C.1: **Performance per class.** Performance comparison from single-frame to our memory-based model. Note this reports mAP for each class averaged across IoU thresholds, as popularized by the COCO challenge [4].

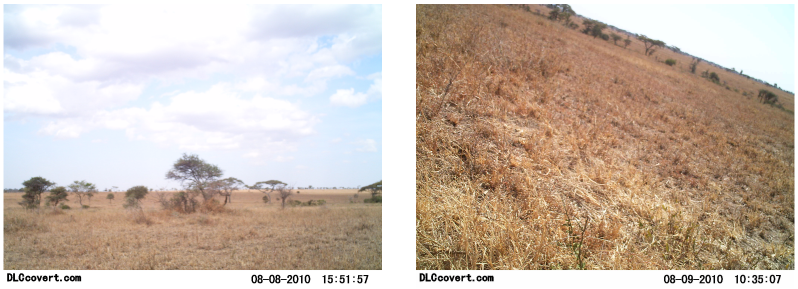
CCT improves for all classes from the single frame model. We see that for one class in CC, "Middle Truck," our method performs slightly worse; however, this class is relatively ambiguous, as the concept of "middle" size is not well-defined.

C.3 Spatiotemporal Encodings

We normalize the spatial and temporal information for each object we include in the contextual memory bank. In order to do so, we choose to use a single float between 0 and 1 to represent each of: year, month, day, hour, minute, x center coordinate, y center coordinate, object width, and object height.

We normalize each element as follows:

- **Year**: We select a reasonable window of possible years covered by our data, 1990-2030. We normalize the year within that window, representing the year in question as $\frac{year-1990}{2030-1990}$.
- **Month**: We normalize the month of the year by 12 months, *i.e.* $\frac{month}{12}$.



(a) Before.

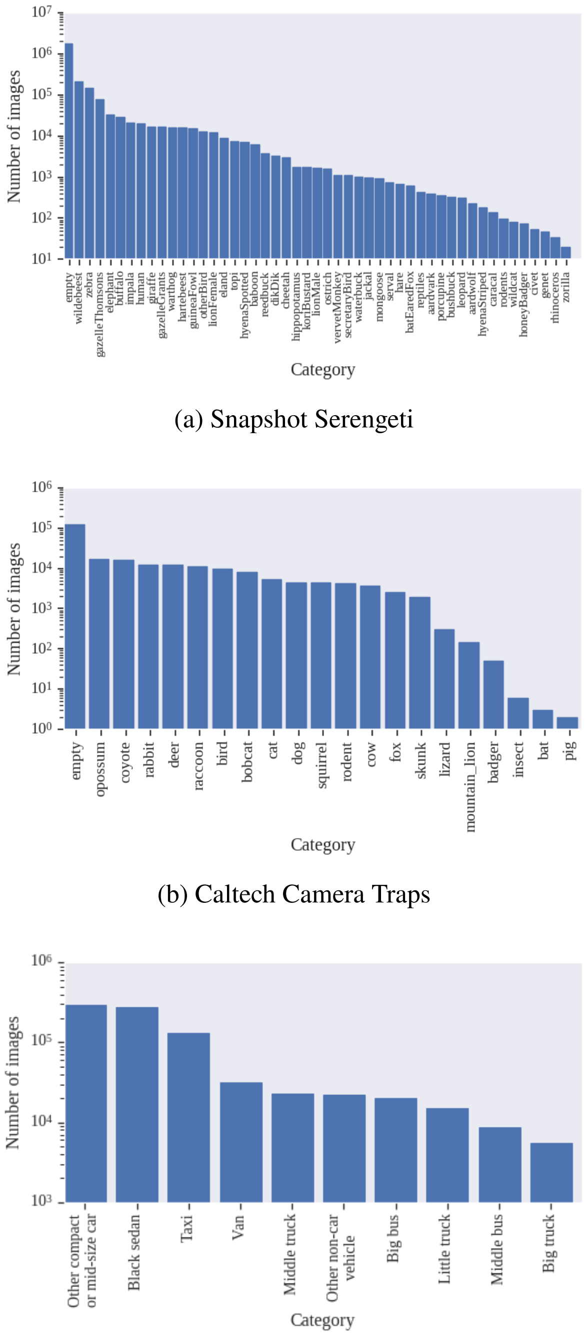
(b) After.

Figure C.2: **Our system is robust to a static camera being accidentally shifted.** Before and after example of a camera that had been bumped by an animal. The images are from the same camera. The first image was taken August 8, 2010, the next August 9, 2010. We find that the system can still utilize contextual information across a camera shift.

- **Day**: We normalize the day of the month by 31 days for simplicity, regardless of how many days there are in the month in question, *i.e.* $\frac{day}{31}$.
- **Hour**: We normalize the hour of the day by 24 hours, *i.e.* $\frac{hour}{24}$.
- **Minute**: We normalize the minute of the hour by 60 minutes, *i.e.* $\frac{minute}{60}$.
- **X Center Coordinate**: We normalize the x coordinate pixel location by the width of the image in pixels, *i.e.* $\frac{x\_center\_location\ (pixels)}{image\_width\ (pixels)}$.
- **Y Center Coordinate**: We normalize the y coordinate pixel location by the height of the image in pixels, *i.e.* $\frac{y\_center\_location\ (pixels)}{image\_height\ (pixels)}$.
- **Width of Object**: We normalize the object width in pixels by the width of the image in pixels, *i.e.* $\frac{object\_width\ (pixels)}{image\_width\ (pixels)}$.
- **Height of Object**: We normalize the object height in pixels by the height of the image in pixels, *i.e.* $\frac{object\_height\ (pixels)}{image\_height\ (pixels)}$.

C.4 Camera Movement

Our system has no hard requirements about the camera being static, instead we leverage the fact that it is static implicitly through our memory bank to provide appropriate and relevant context. We find that our system is robust to static cameras that get moved, unlike traditional background modeling approaches. In Snapshot Serengeti in particular, the animals have a tendency to rub against the camera posts



(a) Snapshot Serengeti

(b) Caltech Camera Traps

(c) CityCam

Figure C.3: **Imbalanced class distributions.** Images per category for each of the three datasets. Note the y-axis is in log scale.

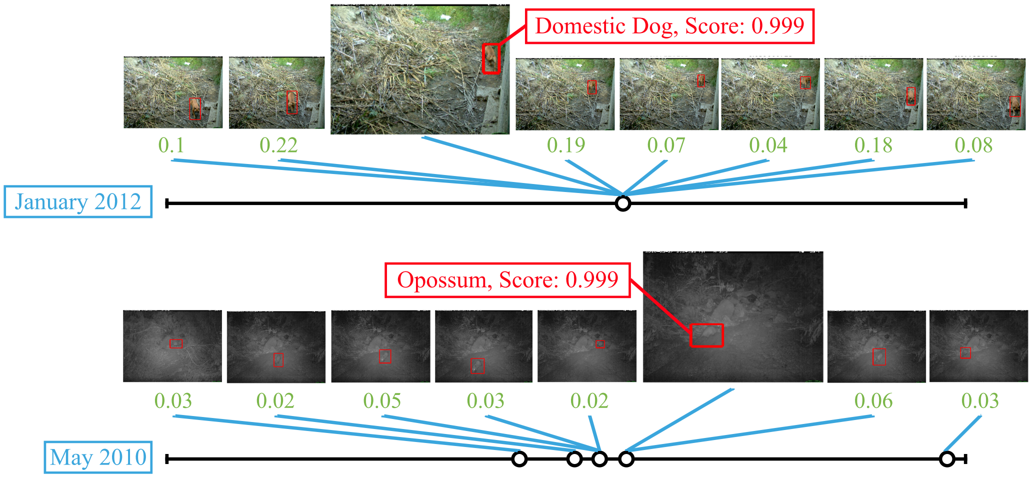


Figure C.4: **Visualizing attention.** In each example, the keyframe is shown at a larger scale, with Context R-CNN's detection, class, and score shown in red. We consider a time horizon of one month, and show the images and boxes with highest attention weights (shown in green). The model pays attention to objects of the same class, and the distribution of attention across time can be seen in the timelines below each example.

and cause camera shifts over time. Figure C.2 shows a "before and after" example of a camera being bumped or moved.

C.5 Attention Visualization

In Figure 4 in the main text, we visualize attention over time for two examples from Snapshot Serengeti. In Figure C.4 we show examples from Caltech Camera Traps. Similarly to the visualizations of attention on SS, we see that attention is adaptive to the most relevant information, paying attention across time as needed. The model consistently learns to attend to objects of the same class.

In Figure C.5, we visualize how Context R-CNN learns to learn and attend to unlabeled background classes, namely rocks and bushes. Remember that these exact camera locations were never seen during training, so the model has learned to use temporal context to determine when to ignore these salient background classes. It learns to cluster background objects of a certain type, for example bushes, across the frames at a given location. Note that these attended background objects are not always the same instance of the class, which makes sense as background classes may maintain visual similarity within a scene even if they aren't the exact same instance of that type. Species of plants or types of rock are often geographically clustered.

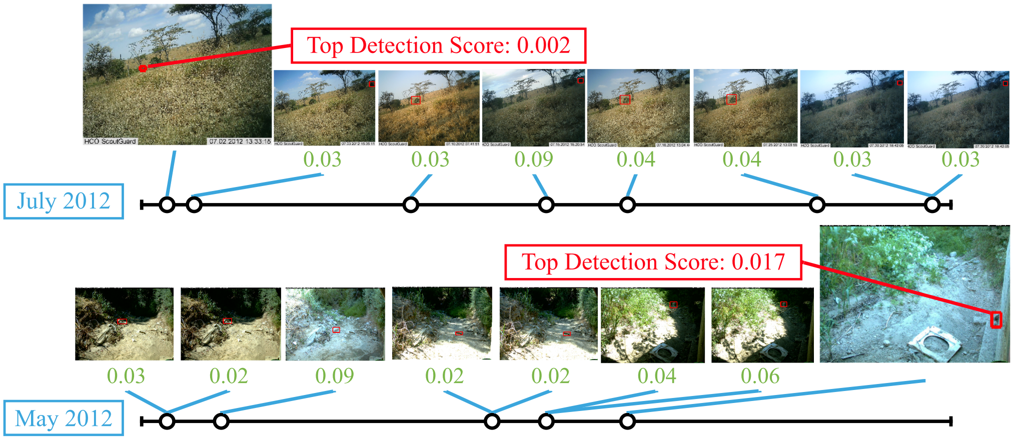


Figure C.5: **Visualizing attention on background classes.** In each example, the keyframe is shown at a larger scale, with Context R-CNN's detection, class, and score shown in red. We consider a time horizon of one month, and show the images and boxes with highest attention weights (shown in green). The first example is from SS, it shows a detected bush (an unlabeled, background class), and shows that Context R-CNN attends to the same bush over time, as well as *different* bushes in the frame. In the second example, from CCT, we see a similar situation with the background class "rock."

References

[1] Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, and Pietro Perona. Synthetic examples improve generalization for rare classes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 863–873, 2020.

[2] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *IEEE CVPR*, 2017.

[3] Sameer Kumar, Victor Bitorff, Dehao Chen, Chiachen Chou, Blake Hechtman, HyoukJoong Lee, Naveen Kumar, Peter Mattson, Shibo Wang, Tao Wang, et al. Scale mlperf-0.6 models on google tpu-v3 pods. *arXiv preprint arXiv:1909.09756*, 2019.

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[5] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

Appendix D

AUTO ARBORIST: SUPPLEMENTARY MATERIAL

D.1 Model training details

Each baseline model was trained with a batch size of 128 on 32 TPU cores on full-resolution images (input size 512x512 for aerial, 1152x768 for street level). Train time data augmentation included random horizontal flipping and minimal random cropping (at least 80% of the image maintained after the crop) for both data types. For simplicity, we used a linear warmup for one epoch, then trained for an additional 4 epochs at a 0.01 learning rate, then a final epoch at a 0.001 learning rate. We anticipate that training for additional epochs would improve all models, but that the relative performance would be maintained.

D.2 Further analysis

Accuracy vs. distribution distance In Supplementary Figure D.1 we show the pairwise accuracy vs distribution distance for all train/test pairs, with markersize denoting the number of training examples. We note the correlation between accuracy and distribution distance, as well as note that the same accuracy level can be achieved between train and test pairs with less training data if the distributions are more similar (the markers tend to get larger as you move from left to right on any horizontal line).

Dirichlet constant	AR	FAR	CAR	RAR
baseline	41.66	58.97	39.77	0.0
1	9.22	1.40	13.28	6.39
10	25.71	12.46	33.1	**17.88**
100	39.46	36.52	45.11	6.85
1K	**43.13**	49.49	**46.09**	0.58
10K	42.83	54.34	43.94	0.0
100K	41.98	**60.48**	40.04	0.02

Table D.1: Logit-adjusted loss ablation study, training on Region W with Santa Monica holdout. Metrics are averaged across the test sets of all 8 cities in the region (see Table 2 in the main paper).

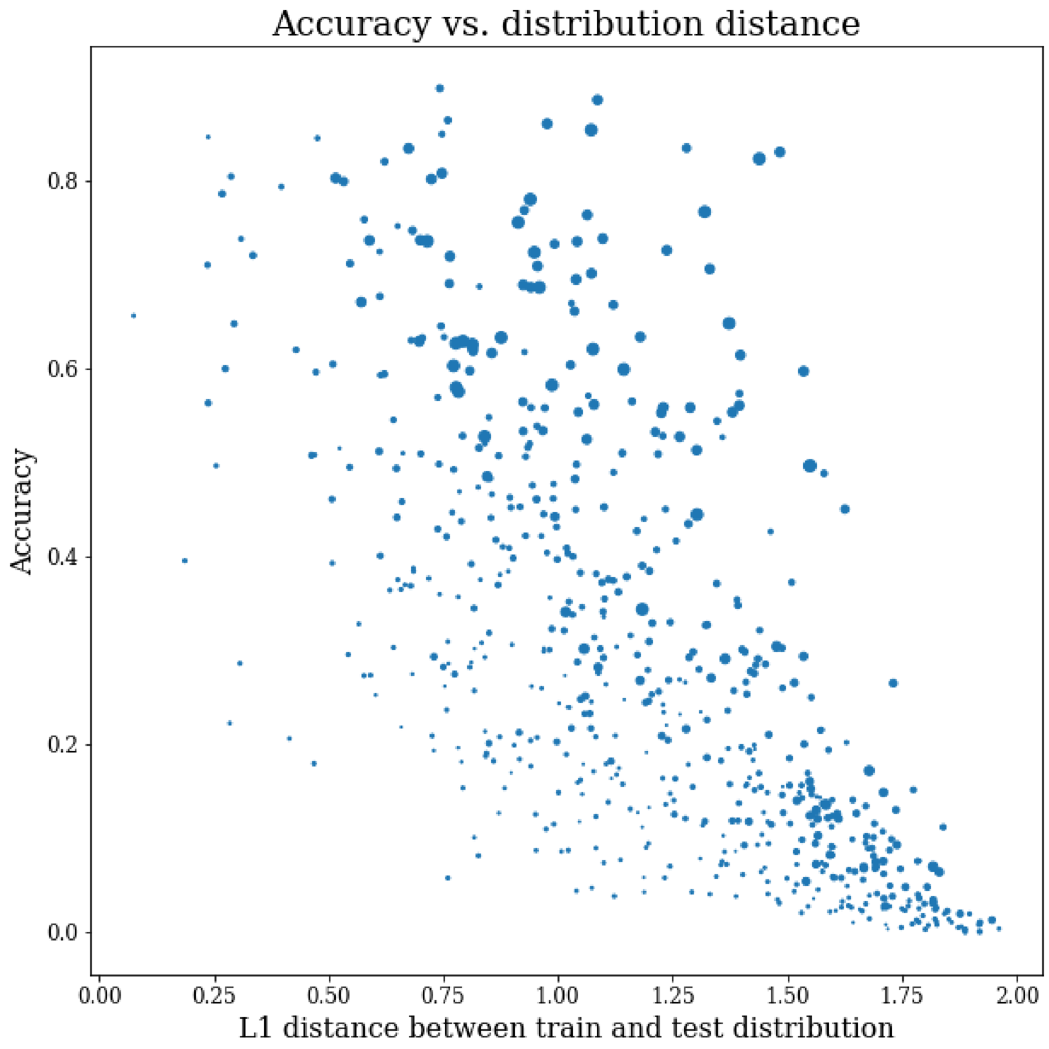


Figure D.1: There is a strong correlation between distribution similarity and performance, notably models can achieve the same accuracy with significantly less training data if their distributions are similar.

Handling the data imbalance We consider the logit-adjustment method for training on imbalanced data proposed in [1]. Eqn. D.1 defines the logit adjusted softmax cross-entropy loss we used in our experiments:

$$l(y, f(x)) = -\log \frac{e^{f_y(x)+\log \pi_y}}{\sum_{y' \in [L]} e^{f'_y(x)+\log \pi'_y}}, \tag{D.1}$$

where $[L]$ is the set of all genera. We construct our logit adjustment term π based on the per-genus counts in the training set, and expand the logit-adjusted loss to include a Dirichlet smoothing term to regularize the highly long-tailed nature of our data.

$$\pi_y = (\text{count}(y) + c)^{-1} \text{ for } y \in [L] \tag{D.2}$$

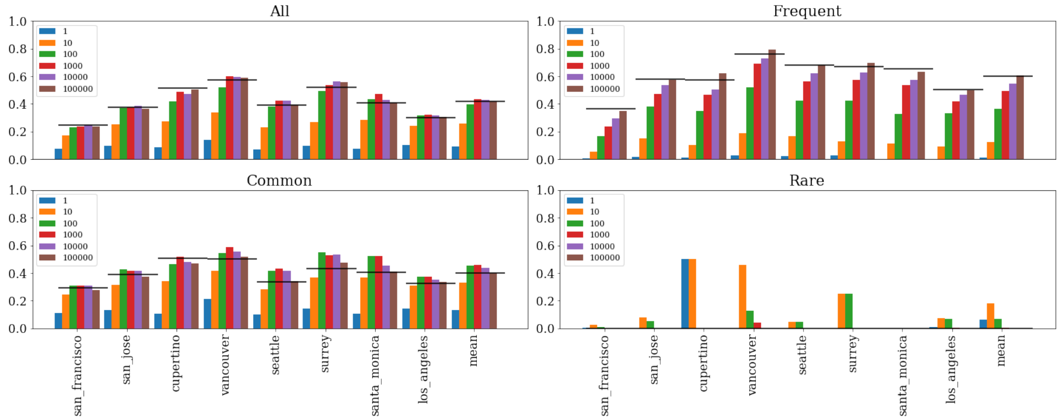


Figure D.2: Loss balancing results per-city as well as averages across the Region W cities, for different values of our Dirichlet smoothing constant c. You can see the explicit tradeoff between frequent and common classes, with rare classes remaining a significant challenge. Baseline results are shown as a black line.

where c is our Dirichlet smoothing constant. We experiment on Region W and find that while $c = 1000$ gives us a 2% boost in city-averaged AR, this is not a clear win, as it improves city-averaged AR-C (by 6.32%) but decreases city-averaged AR-F (by 9.48%). See Table D.1 for ablation results on Region W, different parameters perform better for different rarity subsets and different cities (see Fig. D.2 for additional detail).

D.3 Data sources and licenses

The public tree censuses for the 23 cities in our dataset are linked in Supplementary Table D.2, along with the licensing information for each. We visualize the data per-city in Figure D.3, showing the coordinates of each tree as well as coloring the most common genera to show regional shifts in distribution.

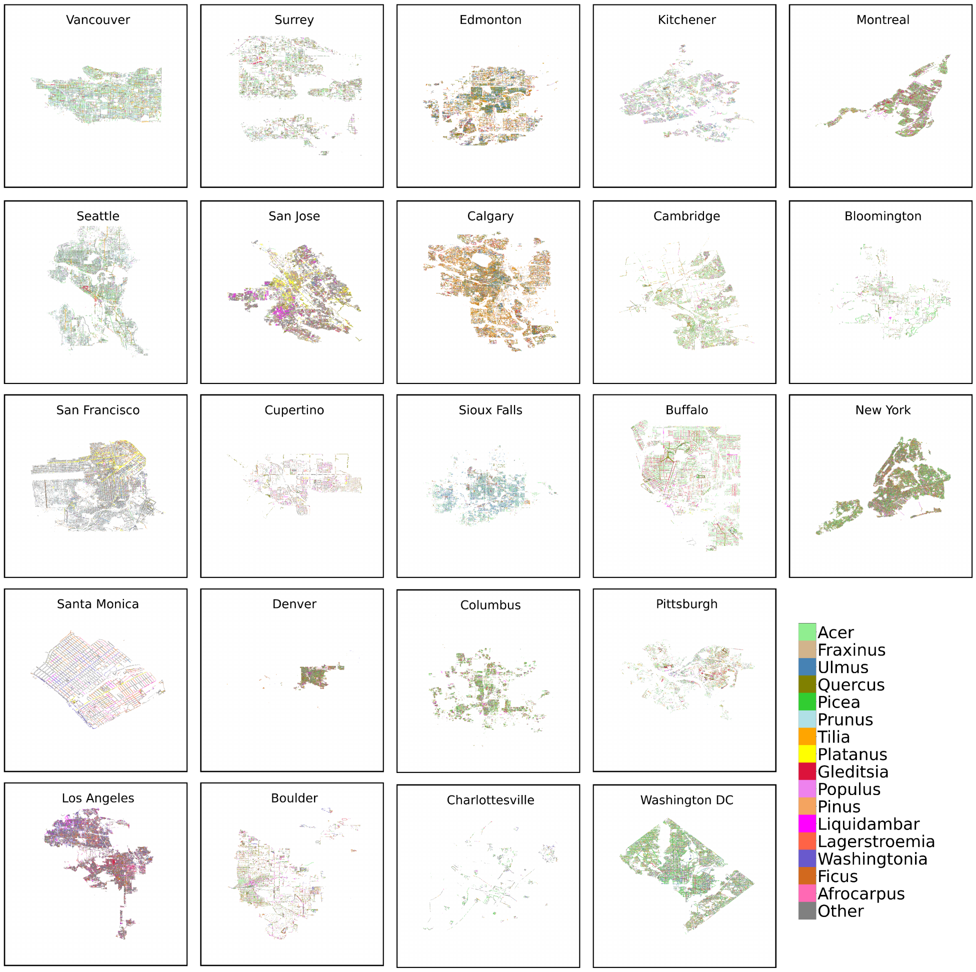


Figure D.3: Visualizing tree locations across all 23 cities. Note how the most common species (denoted by color) shifts across different geographic areas.

City	Data Source	License
Vancouver, BC	https://opendata.vancouver.ca/explore/dataset/street-trees/information/?disjunctive.species_name&disjunctive.common_name&disjunctive.height_range_id	https://opendata.vancouver.ca/pages/licence/
Surrey, BC	https://data.surrey.ca/dataset/park-specimen-trees	https://data.surrey.ca/pages/open-government-licence-surrey
Seattle, WA	https://www.seattle.gov/transportation/projects-and-programs/programs/trees-and-landscaping-program/seattle-tree-inventory-map	https://opendatacommons.org/licenses/pddl/summary/
San Francisco, CA	https://data.sfgov.org/City-Infrastructure/Street-Tree-List/tkzw-k3nq	https://opendatacommons.org/licenses/pddl/1-0/
San Jose, CA	https://gisdata-csj.opendata.arcgis.com/datasets/7db16e012fe8402db45074cd260c8f4e_510	https://gisdata-csj.opendata.arcgis.com/datasets/7db16e012fe8402db45074cd260c8f4e_510/explore
Cupertino, CA	https://gis-cupertino.opendata.arcgis.com/datasets/caa50a924b7d4b5ba8e8a4cbfd0d7f13	https://gis-cupertino.opendata.arcgis.com/datasets/Cupertino::trees-2/about
Santa Monica, CA	https://data.smgov.net/Public-Assets/Trees-Inventory/w8ue-6cnd	https://opendatacommons.org/licenses/by/1-0/
Los Angeles, CA	https://geohub.lacity.org/datasets/266c6255b1fc4ae8b8f100d8696e1fa4_0	Usage was approved by StreetsLA
Boulder City, CO	https://boulder.maps.arcgis.com/apps/opsdashboard/index.html#/328aac5a588840c99edee239672f7ca2	https://creativecommons.org/publicdomain/zero/1.0/
Denver, CO	https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-tree-inventory	https://creativecommons.org/licenses/by/3.0/
Sioux Falls, SD	https://opendata.arcgis.com/datasets/c880d62ae5fb4652b1f8e6cbca244107_10.csv	https://creativecommons.org/licenses/by/3.0/
Calgary, AB	https://data.smgov.net/Public-Assets/Trees-Inventory/w8ue-6cnd	https://data.calgary.ca/d/Open-Data-Terms/u45n-7awa
Edmonton, AB	https://data.edmonton.ca/Environmental-Services/Trees-Map/udbt-eiax	https://data.edmonton.ca/stories/s/City-of-Edmonton-Open-Data-Terms-of-Use/msh8-if28/

Table D.2: Data location and licensing information for the public tree censuses used to curate our dataset.

City	Data Source	License
Charlottesville, VA	https://hub.arcgis.com/datasets/charlottesville::tree-inventory-point/explore?location=38.038850%2C-78.483500%2C13.89	https://creativecommons.org/licenses/by/4.0/
Pittsburgh, PA	https://data.wprdc.org/dataset/city-trees	https://creativecommons.org/licenses/by/4.0/
Montreal, QC	https://donnees.montreal.ca/ville-de-montreal/arbres	https://donnees.montreal.ca/licence-d-utilisation
New York, NY	https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/pi5s-9p35	https://opendata.cityofnewyork.us/overview/
Buffalo, NY	https://data.buffalony.gov/Quality-of-Life/Tree-Inventory/n4ni-uuec/data	https://creativecommons.org/share-your-work/public-domain/cc0/
Kitchener, ON	https://data.waterloo.ca/datasets/KitchenerGIS::tree-inventory/about	https://data.waterloo.ca/datasets/KitchenerGIS::tree-inventory/about
Cambridge, ON	https://data.waterloo.ca/datasets/cityofcambridge::street-trees/explore	https://maps.cambridge.ca/images/opendata/Open%20data%20licence.pdf
Columbus, OH	https://opendata.columbus.gov/datasets/public-owned-trees/explore?location=39.974897%2C-82.996371%2C14.00&showTable=true	https://creativecommons.org/share-your-work/public-domain/cc0/
Bloomington, IN	https://data.bloomington.in.gov/dataset/public-tree-inventory	https://opendefinition.org/od/2.1/en/

Table D.3: (Part 2) Data location and licensing information for the public tree censuses used to curate our dataset.

References

[1] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.