The iWildCam 2019 Challenge Dataset

Sara Beery[*], Dan Morris[+], Pietro Perona[*]

California Institite of Technology[*]
1200 E California Blvd, Pasadena, CA 91125

Microsoft AI for Earth[+]
14820 NE 36th Street, Redmond, WA, 98052

Abstract

Camera Traps (or Wild Cams) enable the automatic collection of large quantities of image data. Biologists all over the world use camera traps to monitor biodiversity and population density of animal species. The computer vision community has been making strides towards automating the species classification challenge in camera traps, but as we try to expand the scope of these models 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.

1. 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*: heat- or motion-activated cameras placed in the wild (See Fig. 1 for examples).

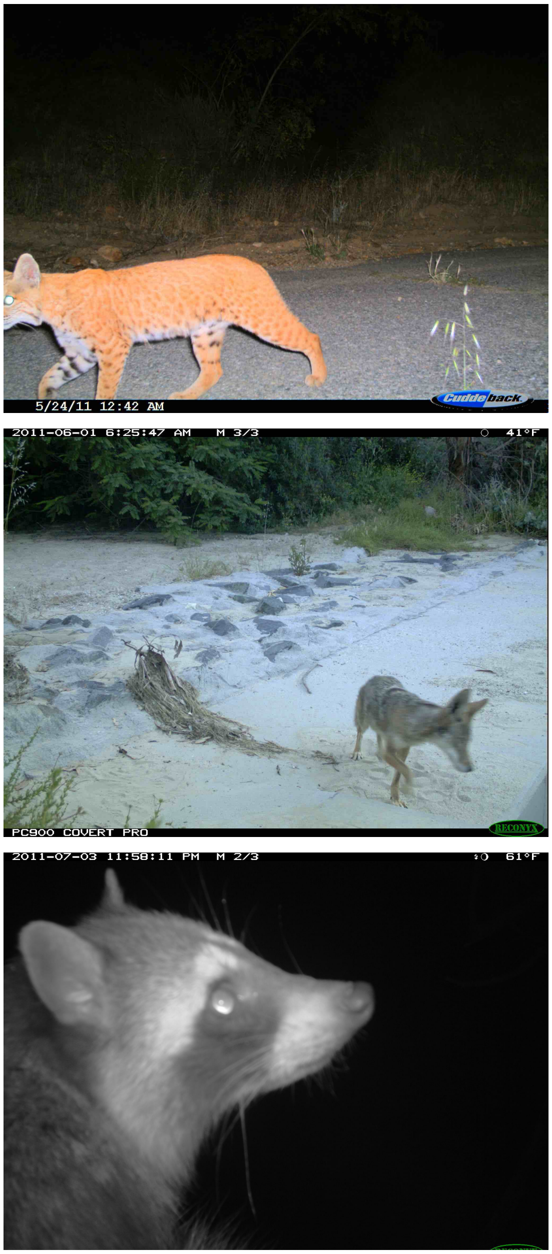

5/24/11 12:42 AM
Cuddeback
2011-06-01 6:25:47 AM M 3/3 41°F
PC900 COVERT PRO
RECONYX
2011-07-03 11:58:11 PM M 2/3 61°F

Figure 1. Examples of camera trap data.

At present, camera trap images are annotated by hand, severely limiting data scale and research productivity. Annotation of camera trap photos is time consuming and challenging. Because the images are taken automatically based on a triggered sensor, there is no guarantee that the animal will be centered, focused, well-lit, or an appropriate scale (they can be either very close or very far from the camera, each causing its own problems). See Fig. 3 for examples of these challenges. Further, up to 70% of the photos at any given location are triggered by something other than an animal, such as wind in the trees, a passing car, or a hiker.

Automating camera trap labeling is not a new challenge for the computer vision community [1, 3–10, 12–17]. However, most of the proposed solutions have used the same camera locations for both training and testing the performance of an automated system. If we wish to build systems that are trained once to detect and classify animals, and then deployed to new locations without further training, we must measure the ability of machine learning and computer vision to *generalize* to new environments. Both the 2018 [2] and 2019 iWildCam challenges focus on generalization, and encourage computer vision researchers to tackle generalization in creative, novel ways.

2. The iWildCam 2019 Dataset

The data for the 2019 challenge is curated from the Caltech Camera Traps (CCT) which was also used for the iWildCam 2018 Challenge [2], a new camera trap dataset from Idaho (IDFG), and two alternate data domains: iNaturalist and Microsoft TrapCam-AirSim.

2.1. 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. **??**. 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. 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.

2.2. 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. 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.

2.3. Additional Data Domains

2.3.1 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. 2. We allow the use of iNaturalist data from both the 2017 and 2018 iNaturalist competition datasets [11]. 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.

2.3.2 Microsoft TrapCam-AirSim

This synthetic data generator utilizes a modular natural environment within Microsoft AirSim [1**?**] 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. 2.

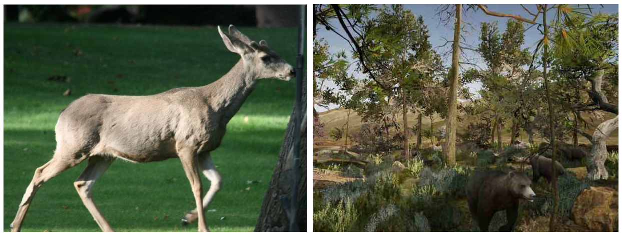
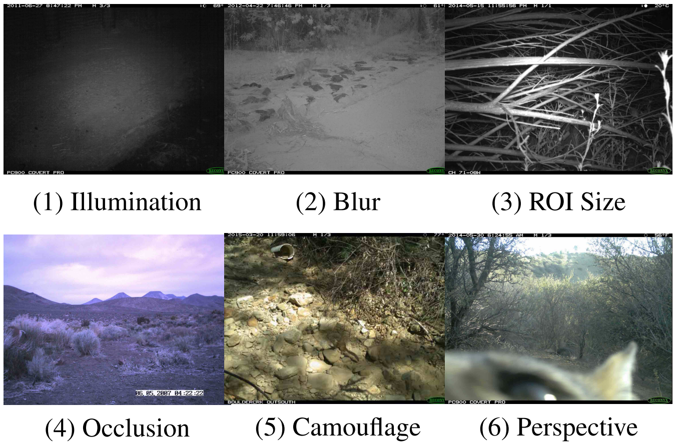


Figure 2. **Altenate domain examples.** (Left) iNaturalist, (Right) TrapCam-AirSim



(1) Illumination (2) Blur (3) ROI Size

(4) Occlusion (5) Camouflage (6) Perspective

Figure 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.

3. The iWildCam Challenge 2019

The iWildCam Challenge 2019 was conducted through Kaggle as part of FGVC6 at CVPR19. We used macro-average F1 score as our competition metric, to slightly emphasize recall over precision and to encourage more emphasis on rare classes, as opposed to rewarding high performance on common classes proportionally to their unbalanced level of occurrence.

3.1. 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.

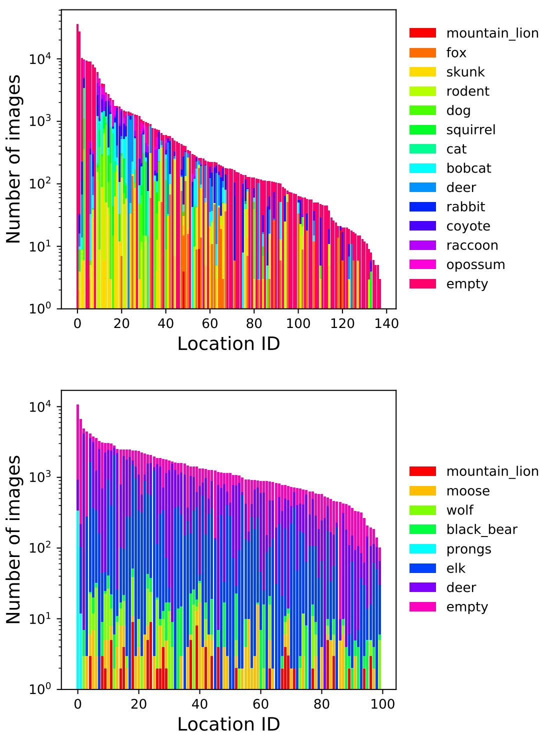


Figure 4. **Number of annotations for each location**. (Top) CCT locations, containing 14 classes. (Bottom) IDFG locations, containing images of 8 classes. The distribution of images per location is long-tailed, and each location has a different and peculiar class distribution.

3.2. 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.

4. Conclusions

Camera traps provide a unique experimental setup that allow us to explore the generalization of models while controlling for many nuisance factors. This dataset provides a test bed for studying generalization to not only new geographic regions, but also new species not seen during training. This dataset is the first designed to explicitly study generalization to a new region with a non-identical class set. The difficulty of the associated challenge demonstrates the innate inability for our current state-of-the-art methods to handle this type of generalization.

In subsequent years, we plan to extend the iWildCam Challenges by adding new regions and species worldwide. We hope to use the knowledge we gain throughout these challenges to facilitate the development of models or systems of models that can accurately provide real-time species ID in camera trap images at a global scale. Any forward progress made will have a direct impact on the scalability of biodiversity research geographically, temporally, and taxonomically.

References

[1] S. Beery, Y. Liu, D. Morris, J. Piavis, A. Kapoor, M. Meister, and P. Perona. Synthetic examples improve generalization for rare classes. *arXiv preprint arXiv:1904.05916*, 2019.

[2] S. Beery, G. van Horn, O. MacAodha, and P. Perona. The iwildcam 2018 challenge dataset. *arXiv preprint arXiv:1904.05986*, 2019.

[3] S. Beery, G. Van Horn, and P. Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[4] G. Chen, T. X. Han, Z. He, R. Kays, and T. 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.

[5] J.-H. Giraldo-Zuluaga, A. Salazar, A. Gomez, and A. Diaz-Pulido. Camera-trap images segmentation using multi-layer robust principal component analysis. *The Visual Computer*, pages 1–13, 2017.

[6] K.-H. Lin, P. Khorrami, J. Wang, M. Hasegawa-Johnson, and T. 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.

[7] A. Miguel, S. Beery, E. Flores, L. Klemesrud, and R. Bayrakcismith. Finding areas of motion in camera trap images. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 1334–1338. IEEE, 2016.

[8] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, C. Packer, and J. Clune. Automatically identifying wild animals in camera trap images with deep learning. *arXiv preprint arXiv:1703.05830*, 2017.

[9] X. Ren, T. X. Han, and Z. 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.

[10] A. Swanson, M. Kosmala, C. Lintott, R. Simpson, A. Smith, and C. Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, 2:150026, 2015.

[11] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[12] A. G. Villa, A. Salazar, and F. 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.

[13] M. J. Wilber, W. J. Scheirer, P. Leitner, B. Heflin, J. Zott, D. Reinke, D. K. Delaney, and T. 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.

[14] H. Yousif, J. Yuan, R. Kays, and Z. 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.

[15] X. Yu, J. Wang, R. Kays, P. A. Jansen, T. Wang, and T. Huang. Automated identification of animal species in camera trap images. *EURASIP Journal on Image and Video Processing*, 2013(1):52, 2013.

[16] Z. Zhang, T. X. Han, and Z. 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.

[17] Z. Zhang, Z. He, G. Cao, and W. 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.