An Animal Detection Pipeline for Identification

Jason Parham*, Charles Stewart
Rensselaer Polytechnic Institute
{parhaj,stewart}@rpi.edu

Jonathan Crall
Kitware, Inc.
jon.crall@kitware.com

Daniel Rubenstein
Princeton University
dir@princeton.edu

Jason Holmberg
Wild Me
jason@wildme.org

Tanya Berger-Wolf
University of Illinois-Chicago
tanyabw@uic.edu

Abstract

This paper proposes a 5-component detection pipeline for use in a computer vision-based animal recognition system. The end result of our proposed pipeline is a collection of novel annotations of interest (AoI) with species and viewpoint labels. These AoIs, for example, could be fed as the focused input data into an appearance-based animal identification system. The goal of our method is to increase the reliability and automation of animal censusing studies and to provide better ecological information to conservationists. Our method is able to achieve a localization mAP of 81.67%*, a species and viewpoint annotation classification accuracy of* 94.28% *and* 87.11%*, respectively, and an AoI accuracy of* 72.75% *across 6 animal species of interest. We also introduce the Wildlife Image and Localization Dataset (WILD), which contains* $5{,}784$ *images and* $12{,}007$ *labeled annotations across 28 classification species and a variety of challenging, real-world detection scenarios.*

1. Introduction

Computer vision-based methods are being used increasingly as tools to assist wild animal object recognition. The ability to identify individual animals from images enables population surveys through sight-resight statistics and forms the basis for demographic studies. The pipeline of processing for animal recognition includes several stages, starting with the detection of animals in images and ending with identification decisions. By making all stages of this pipeline more reliable and automated, animal identification studies can be increased in spatial and temporal resolution, provide better conservation statistics, and – importantly – allow citizens without specialized training to participate in engaging census data collection events [1, 5, 23, 26].

*Corresponding author.

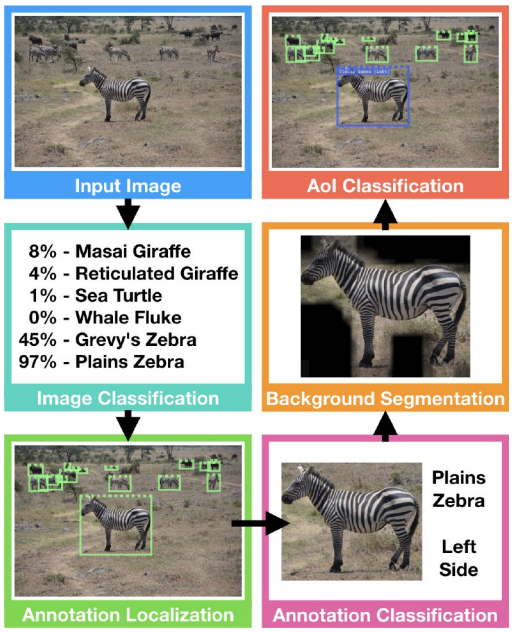


Figure 1. An overview of our detection pipeline: 1) image classification provides a score for the species that exist in the image, 2) annotation localization places bounding boxes over the animals, 3) annotation classification adds species and viewpoint labels to each annotation, 4) annotation background segmentation computes a species-specific foreground-background mask, and 5) AoI classification predicts the focus of the image.

The goal of this paper is to present a series of algorithm components for the first stage of an animal recognition pipeline, namely *detection*. Detection includes the obvious steps of finding animals in images, determining their species, and placing bounding boxes around them, creating what we refer to as an *annotation*. But, the problem is more complex than this, especially when large data volumes gathered by non-specialists are considered: there may be multiple (or no) animals from several different species

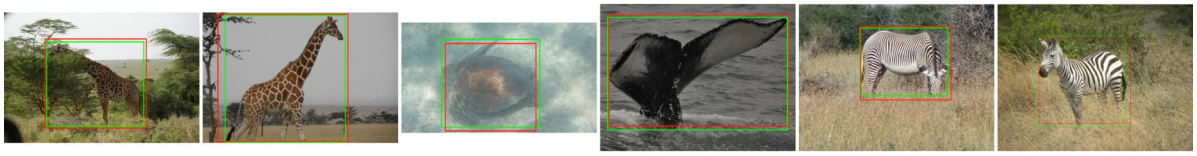


Figure 2. Example annotation localization predictions on single-sighting exemplar images for each of the 6 species of interest. The green boxes designate ground-truth bounding box coordinates and the red boxes represent the annotation localization bounding box predictions. Since we also perform annotation classification, we treat these bounding boxes more like salient object predictions.

in an image; some annotations may have poor quality while others may show only parts of an animal due to self occlusion, or occlusion by other animals or vegetation; and the animals may be seen from a range of viewpoints and poses, only some of which show identifiable information.

In response to these challenges we propose a five stage detection process (Figure 1), each of which is a separate deep convolutional neural network (DCNN): 1) whole-image classification to select the images that indeed show the species or species of interest, 2) bounding-box localization to form the annotations, 3) annotation classification to determine the species and viewpoint, 4) coarse annotation segmentation to narrow the pixel-level focus of the identification algorithm, and 5) a classifier to select in each image what we define as "annotations of interest", a novel concept introduced here. An annotation of interest (or AoI) is the primary subject(s) of a picture from the perspective of identification. For example, in Figure 1, one zebra is the clear subject of the image, despite the presence of many other zebras. Furthermore, if this zebra were not in the picture, there would be no AoIs. This allows identification to work with far fewer annotations, especially when animals can not be reliably photographed in isolation, and focuses the computations on the intended subject(s)s of each image.

Along with our proposed detection pipeline, we also present a new detection dataset called WILD (Wildlife Images and Localizations Dataset). The purpose of WILD is to provide a more realistic set of *real-world* animal sightings, with scenarios not commonly in public datasets. The species that are catalogued in WILD are 1) Masai giraffe (*Giraffa camelopardalis tippelskirchi*), 2) reticulated giraffe (*Giraffa reticulata*), 3) sea turtle (*Chelonia mydas* and *Eretmochelys imbricata*), 4) humpback whale fluke (above-water flukes of *Megaptera novaeangliae*), 5) Grevy's zebra (*Equus grevyi*), and 6) plains zebra (*Equus quagga*).

The rest of the paper is organized as follows. Section 2 outlines related work on animal censusing, object detection, and identification algorithms. Section 3 describes each components in the detection pipeline. Section 4 the describes WILD dataset and presents the results of the detection pipeline running on that dataset. Section 5 concludes with a brief discussion and suggestions for future work. Due to space limitations and to focus on the main concepts, we mostly provide summaries of the primary DCNN architecture, saving details for the more novel components.

2. Related Work

The components of the detector pipeline are modeled on a variety of deep learning architectures. The whole-image and annotation classifiers are most related to the Overfeat DCNN by Sermanet et al. [25]. The localization network is modeled off of the You Only Look Once (YOLO, version 1) detector by Redmon et al. [21] whereas the background segmentation network is a patch-based variant of the FCNN (Fully Convolutional Neural Network) discussed by Long et al. in [18]. The AoI classifier concept is novel, but its DCNN architecture has structural similarities to [25] and shares inspiration to objectives in deep saliency object detection [4, 14, 17] and attention networks [10, 27].

Our pipeline is designed to be used as a black-box component within a larger individual animal identification pipeline. Various frameworks [2, 9, 20] in the conservation literature have included computer vision components.

3. Methods

In this section we will describe the various components of our pipeline. All models (except for the annotation localizer) are trained using Lasagne [7] and Theano [3] on a single NVIDIA TITAN X GPU with 12GB. The annotation localizer is trained using a Python wrapper around the original open-source implementation[1] by Redmon et al.

3.1. Image Classification

The purpose of the image classifier is to predict the existence of species of interest within an image. Unlike the original ILSVRC classification challenge that offered only a dominant whole-image class with 1-class and 5-class testing modes, we often need to classify images containing multiple animal sightings of possibly more than one species. Therefore, we structure the classifier to predict a multi-label, multi-class vector where the corresponding index for a species is set to 1 if at least one animal of that species exists in the image and 0 otherwise. The network

[1] https://github.com/pjreddie/darknet

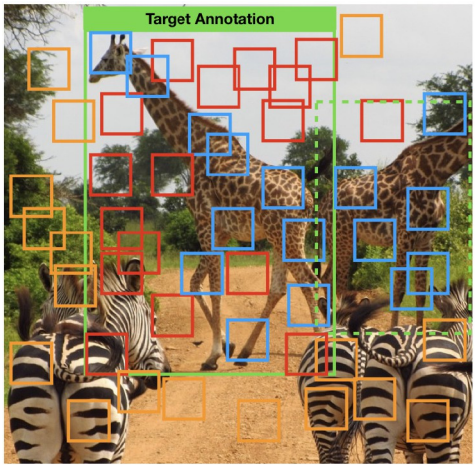


Figure 3. An illustration of the background segmentation patch sampling (using giraffes) and the utility of a cleaning procedure. The target giraffe (green, solid) has a collection of labeled positive patches (blue and red) and negative patches (orange) that are sampled outside the bounding box. The blue patches are *true* positives whereas the red patches are incorrectly-labeled *true* negatives. The goal of the cleaning procedure is to automatically convert all red boxes into orange boxes. Best viewed in color.

takes as input a 192×192 pixel image that is reduced to a $5 \times 5 \times 128$ feature vector via convolutional and max pooling layers. The network then adds a 256-unit dense layer, followed by a feature pooling layer, a dropout [11] layer ($p = 0.5$), and another 256-unit dense layer. The final dense layer has 6 output values, one for each of the species of interest, which are activated by a sigmoid function. The model's weights are optimized using a binary cross-entropy loss, applied independently for each output class.

The image classifier can be thought of as a fast, high-pass content filter to prevent irrelevant images from being processed further along the pipeline. One example of where is useful is in processing raw images taken by a camera trap. It can be common for images collected by a motion-triggered camera trap to have a high ratio of false positive images, which do not contain any sightings of a species of interest. These images are irrelevant to identification and should be filtered out as distractions to reduce overall processing time.

3.2. Annotation Localization

The annotation localization network design is based on the You Only Look Once (YOLO, version 1) network by Redmon et al. [19, 21]. The YOLO architecture is a variant of single-shot detectors (e.g. SSD[17]), which directly outputs a fixed-length regression output for a given fixed-sized input image with no need for a separate region proposal network (RPN). Refer to [21] for in-depth implementation details. The network's goal is to perform bounding box regression and species classification around all objects of interest, the result being a collection of image sub-regions that can be cropped into a candidate list of object annotations. The YOLO network takes 448 pixel by 448 pixel images as input and predicts a 7 by 7 classification grid with 2 anchor boxes per cell. The output of the network is therefore 98 bounding box coordinates along with an object score for each of the 6 species classes on each bounding box.

The predicted bounding boxes by the annotation localization network have associated species label classifications. Since we are performing annotation classification anyway, we essentially treat these localizations as salient object detections. Instead, we opt to use the output of the annotation classification network as the final annotation species, as detailed below in Section 3.3. The motivation for using an annotation localizer is clear: we want processing on relevant input sub-regions, which eliminates distracting pixel information and reduces identification computation. Examples of exemplar object localizations for each of the 6 species can be viewed in Figure 2.

3.3. Annotation Classification

The annotation classification network architecture is very similar to the image classification component except that it performs a standard single-label, multi-class classification. We intentionally train a separate set of weights for the convolutional feature extractors in each component. This obviously increases redundancy but allows for specialized filters for each task. By keeping each task semantically compartmentalized, we also achieve the advantage of being able to optimize each component independently without needing to validate the performance impact of a unified feature extraction across the entire pipeline. The goal for the annotation classification network is to provide corrected species and additional viewpoint classifications. The network takes as input smaller, 128 pixel by 128 pixel subregions that represent the resampled annotation proposals. Input images are reduced to a $5 \times 5 \times 256$ feature vector for classification via convolutional, max pooling, and batch normalization [12] (BN) layers. The network then adds a 512-unit dense layer, followed by a feature pooling layer, a Dropout layer ($p = 0.5$), and another 512-unit dense classification layer. We combine the species and viewpoints into paired classification combinations for the last dense layer of the network (activated by softmax). The model's weights are optimized using a categorical cross-entropy loss.

Alongside the species labels, we also classify the annotations based on the viewpoint of the animal, relative to the camera. The viewpoints for zebras and giraffes are labeled with one of 8 discretized yaw locations around the animal, from the set {`front`, `front-right`, `right`, `back-right`, `back`, `back-left`, `left`, `front-left`}. Sea turtles are commonly captured from above and sometimes from below, so we constrain their

allowed viewpoints to the set of 6 viewpoints {`front`, `right`, `back`, `left`, `top`, `bottom`}. Whale flukes also have a similar restriction where they are label from a set of 4 {`top`, `bottom`, `right`, `left`}, with the most common being `bottom` when the angled fluke is viewed above water and typically looking towards the rear of the animal. The species and viewpoints pairs are combined into 42 distinct combinations to create the set of available classification labels for training. The label pairing used by the annotation classifier does cause an inherent class imbalance, but achieving balanced viewpoints across all species in a real-world, unstructured setting is an impractical task.

The primary task of annotation classification is to correctly label the annotation's species *and* the correct viewpoint together. Poor scoring bounding boxes do not continue in the pipeline. The fallback task of this network is to perform species classification; therefore, any incorrect predictions by the annotation classifier are preferred to happen within the same species (i.e. incorrect viewpoint classification, but correct species classification).

3.4. Annotation Background Segmentation

The annotation background segmentation neural networks are a distinct type of architecture called a Fully Convolutional Neural Network (FCNN). Our usage of the FCNN architecture is unique from deep learning segmentation architectures (like [18, 22]) in that we do not require fully-segmented ground-truth for training. Instead, we structure background segmentation as a binary classification problem where species-specific body patches are classified against background negative patches. The goal of this detection component is to produce a species-specific background mask, which can be used to eliminate or otherwise down-weight distracting non-animal pixel information. As such, we can train the networks on patches of 48×48 pixels but, during forward inference, the networks transparently scale up and adapt to arbitrarily-sized inputs for computing binary classification *maps*. A key insight to the architecture is that during training the input images are reduced via convolutional and max pooling layers to a 1 pixel by 1 pixel patch with 128 channels. During inference, the network's output is expected to increase to $W \times H \times 128$, where W and H are down-sampled resolutions of the original input size. We then use Dropout ($p = 0.4$) and a single softmax Network-in-Network [15] layer to engineer the correct number of classification outputs while retaining a fully convolutional structure. Importantly, the last layer's softmax activation is applied along the channel dimension, across spatial dimensions.

To create the training patch data, a target annotation is selected and its corresponding image is resampled such that the annotation has a width of 300 pixels. Then, random patch locations and scales (within $\pm 80\%$) are sampled uniformly across the image with positive patches being centered inside the annotation (or an annotation of the same species) and negative patches centered outside all annotations for that species. Positive patch exemplars are thus species-specific and are meant to cover sub-regions within an animal body whereas negative patches are meant to be a representative sampling of background foliage, terrain, other animals of a different species, etc.

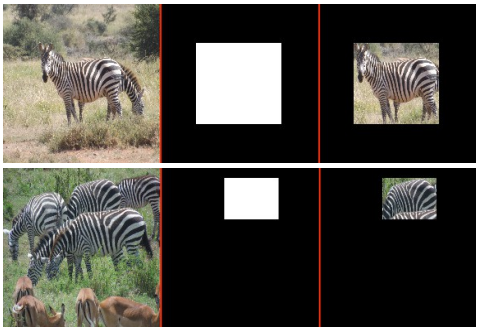


Figure 4. Example input to the AoI classifier. The positive AoI training example (top row) is comprised of the resampled RGB image (left) and the annotation segmentation mask (middle). The right-most column depicts their combined representation. As shown in the negative example (bottom row), the masked annotation is of an occluded, background animal and is not an AoI.

Our proposed positive patch sampling scheme can be problematic, however. The bounding box localizations of giraffes, for example, generally have large amounts of negative space around the neck and around the legs (Figure 3). When positive patches are sampled from within giraffe bounding boxes, a certain number of patches are incorrectly labeled as positive where they actually contain only negative, background pixel information. To help correct for this dataset label noise, we employ the use of an automated cleaning procedure during training. At the start of training, the network is given the original labels and asked to perform unaltered binary classification. When the learning rate is decreased (and only after the model achieves an overall accuracy $\geq 90\%$), we run the currently learned model on the training and validation data to find potentially incorrect labels. Any label that has a $\geq 95\%$ prediction of belonging to the opposite ground-truth label is automatically "cleaned" and its binary label flipped. We have found that the cleaning procedure helps training smoothness and improves the qualitative performance of the results.

3.5. Annotation of Interest (AoI) Classification

The novel task for the AoI classification network to solve is to predict an *a posteriori* decision concerning the composition of an image: "why did the photographer take this picture?" This is not a question of artistic composition. It is instead motivated by the goal of processing only the most

identifiable animals in a given image. Therefore, to answer the question of "why", the first task is to understand an image's semantic composition with relation to all captured animals in the scene. We construct the ground-truth AoI labels by marking the individual annotation(s) that are the interest, subject, or focus of each image. More concretely, an AoI should have most or all of the following properties:

- is of a distinguishable individual animal (i.e. free-standing, well-lit, clearly visible, etc.),
- is relatively large and has decent resolution,
- is commonly located near the center of the image, and
- is in focus and not blurred

Conversely, an annotation should not considered an AoI if it has one or more of the following opposite properties:

- is a part of an overlapping herd or group of animals
- is relatively small and/or contains few pixels
- is out of focus or is otherwise blurry
- is located around the edges of the image
- is occluded by other animals or objects by area $\geq 25\%$
- is off the edge of the frame of the image by area $\geq 25\%$

These properties demand that the annotation not be reviewed in isolation (i.e. by only viewing its cropped sub-region); the decision that an annotation is an AoI must be made by weighing the entire context of the image and against any other annotations it coexists with. Further, because these conditions are fairly strict, there are rarely more than one or two AoIs in a particular image, and many images have no AoIs. The overarching motivation for AoI classification is to prioritize further processing on only the most *identifiable* annotations. While the concept of identifiability is algorithm-dependent, we structure AoI as a generalized, easy-to-determine proxy.

The AoI classifier has a very similar convolutional and dense layer structure to the image classifier, except for three differences: 1) it takes as input a 4-channel input image, comprised of a red, blue, and green color channels stacked with a fourth annotation bounding box mask, 2) the output layer (with a softmax activation function) has only two outputs for simple binary classification, and 3.) the network weights are optimized using categorical cross-entropy loss. Examples of a positive and a negative training input images can be viewed in Figure 4. The end result of the AoI classifier is eliminating the need to perform identification processing on background and partially-visible animals, which cause identification confusion and drastically increases the need for a *human-in-the-loop* reviewer.

4. Results

This section describes the WILD data and presents experimental results for the detection pipeline. All software to train the various algorithms, perform inference on new imagery, and evaluation on WILD is open-source.[2]

[2] wildme.org

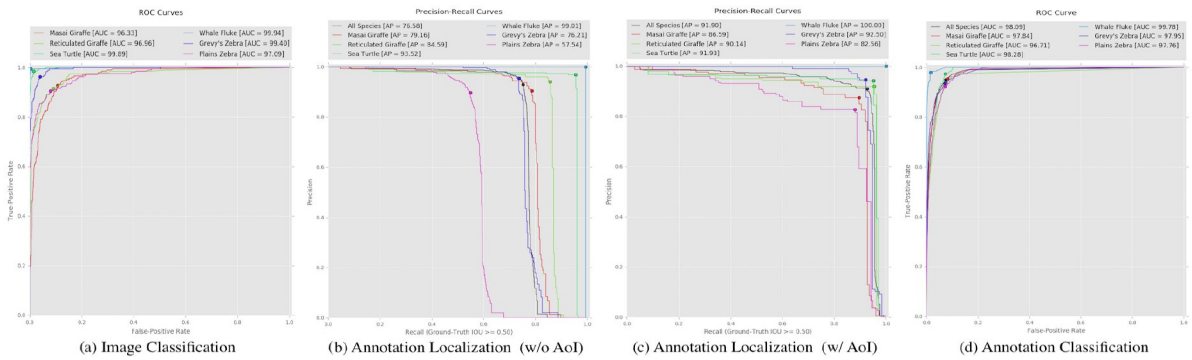
Species	Images	Annots.	AoIs
Masai Giraffe	1,000	1,468	611
Reticulated Giraffe	1,000	1,301	595
Sea Turtle	1,000	1,002	567
Whale Fluke	1,000	1,006	595
Grevy's Zebra	1,000	2,173	669
Plains Zebra	1,000	2,921	561
TOTAL	**5,784**	**9,871**	**3,598**

Table 1. WILD: a breakdown of the number of images, annotations, and AoIs per species. The total number of images is less than 6,000 because some species share sightings within the same image, specifically between zebras and giraffes, which demonstrates the need for a multi-labeled image classifier. There are also an additional 2,136 annotations in this dataset of miscellaneous categories (`car`, `boat`, `bird`, etc.) that are ignored in this paper.

4.1. WILD

We created a new ground-truthed dataset for the tasks presented in this paper, called WILD; WILD is comprised of photographs taken by biologists, wildlife rangers, citizen scientists [13], and conservationists, and captures detection scenarios that are uncommon in publicly-available computer vision datasets like PASCAL [8], ILSVRC [24], and COCO [16]. These datasets make little distinction between living animal sightings vs. abstract representations of an animal (e.g. a stuffed zebra animal toy, fondant zebras on birthday cakes). These abstractions distract from our task of detecting real-world sightings of animals in the wild and are inappropriate for the follow-on task of individual identification. In WILD all of the images in the dataset were taken *in situ* by on-the-ground photographers. Another example is that zebras and giraffes tend to form groups and stand closely together, creating sightings with frequent bounding box overlap, occlusion, and cross-species co-location. Finally, WILD contains closely relative species of giraffes and of zebras that must be distinguished.

We gathered a dataset of 5,784 images and hand-annotated 12,007 bounding box localizations across 28 classes. The 6 species of interest that are the focus of this paper represent 9,871 annotations in the dataset. A breakdown of the number of images and annotations that contain each species can be viewed in Table 1. We assigned cropped annotation sub-regions to human reviewers for labeling the species and dominant viewpoint of the animal. We then tasked reviewers to pick the annotation(s) in each image for AoI classification, the guidelines for which can be reviewed in Section 3.5. In summary, a total of 3,602 annotations were marked as AoIs. The dataset was then partitioned into two sets: training (4,623 images) and testing (1,161 images) through an $80/20\%$ stratified split based on the number of annotations in each image. This results in a total of 7,841 annotations for training and 2,030 for testing.



(a) Image Classification (b) Annotation Localization (w/o AoI) (c) Annotation Localization (w/ AoI) (d) Annotation Classification

Figure 5. Performance curves. The image classifier ROC curve (a) achieves a minimum area-under-the-curve (AUC) of 96.33%. The annotation localizer precision-recall curves (b) reports an unfiltered mean average-precision (mAP) of 81.67% across all 6 species with an Intersection-over-Union (IoU) threshold of 50%. The drastic drop in performance of the plains zebra species can be contributed to the high number of background – likely small-sized – annotations for this species; focusing (b) on just AoIs (c) increases mAP to 90.62%. The annotation classifier ROC curve (d) achieves an AUC of 98.09% for all species, which are independently averaged across all of their associated viewpoints. The point that is plotted on each of the curves indicate the operating point which is closest to a perfect prediction.

We distribute[3] the dataset in the ubiquitous PASCAL VOC format, with additional metadata attributes to mark viewpoints and AoI flags. All results are reported on held-out validation data on the WILD dataset.

4.2. Image Classification

The image classifier does a good job at correctly predicting species existence within an image, as shown in Figure 5 (a). The worst-performing species (Masai giraffe) still achieves a ROC area-under-the-curve (AUC) of 96.33% and the best-performing species (whale fluke) has an almost-perfect 99.94% AUC. The mean AUC across all species is 98.27%. With appropriate operating points selected independently for each species (indicated by the colored dots on each curve), the overall image classifier accuracy is 64.77%. When the image classifier is applied to the test data, there are no images that result are incorrectly suppressed.

4.3. Annotation Localization

The annotation localization model has a spread of accuracies across the different species, mostly based on that specific species's level of sighting difficulty. For example, the whale fluke and sea turtle localizations achieve 99.01% and 93.52% average-precision (AP), respectively. This makes intuitive sense: a mostly rigid body part sighted against a stark background of the sea, ocean floor, or sky will be easier to localize compared to a compact herd of overlapping, occluded, and varying animals. As displayed in Figure 5 (b), this difference in difficulty can be seen noticeably in the relatively poor performance of the plains zebra localizations at only 57.54%. By referencing Table 1 we can see that the ratio of AoIs to annotations is lower at 19.21% for plains zebras compared to the average of 42.18%. Furthermore, the ratio of annotations per image is also the highest at 2.921, compared to the average of 1.645. Nevertheless, the localizer achieves a mAP of 81.67% across all species.

We further analyze the performance of the localizer when only annotations marked as AoIs are considered, in Figure 5 (c). As detailed in Section 3.5, AoIs should be clearly distinguishable, relatively large, and free of major occlusions. As shown in [19], these are the major causes of error for animal localization. The annotation localization performance improves to 90.62% mAP and improves recall on all species, with the most marked improvement being the plains zebra localizations. This indicates that most of the localization errors are from background, small, occluded, or otherwise unidentifiable animals.

4.4. Annotation Classification

The annotation classifier has two goals: 1) species classification only and 2) species *and* viewpoint combined classification. While the network is optimized to learn the combined classifications, it can be used to predict accurate species-only labels for annotations. As seen in Figure 5 (d), the species-specific ROC curves achieve at least 96.71% AUC across all species. The species ROC operating curves in this figure are calculated by taking an average over the associated ROC curves for its respective viewpoints.

The effect of these two annotation classification goals can be visualized in Figure 6. The white squares indicate species and the inter-species classification accuracy is 94.28%; therefore, any values outside of the squares are incorrect species classifications. We can see that the majority of the inter-species classification error (only 5.72%) is between the two sub-genus species of giraffes and some additional error between the two zebra classes. This makes in-

[3] http://lev.cs.rpi.edu/public/datasets/wild.tar.gz (1.4GB)

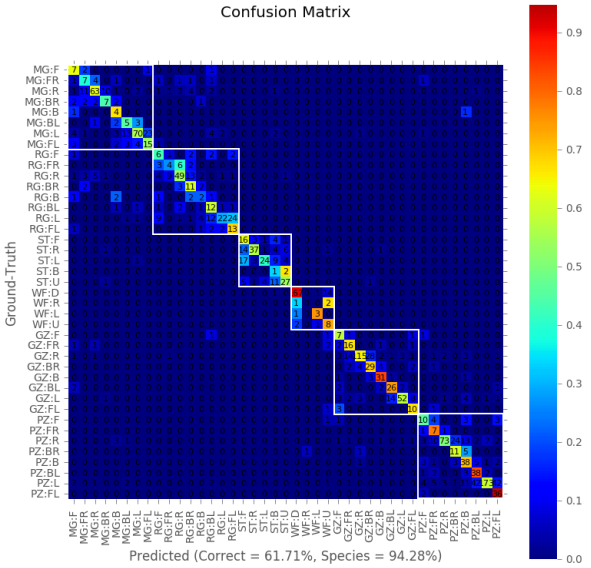


Figure 6. The classification confusion matrix for the annotation classifier, marked with abbreviated {species:viewpoint}. The white boxes represent the separate species classes.

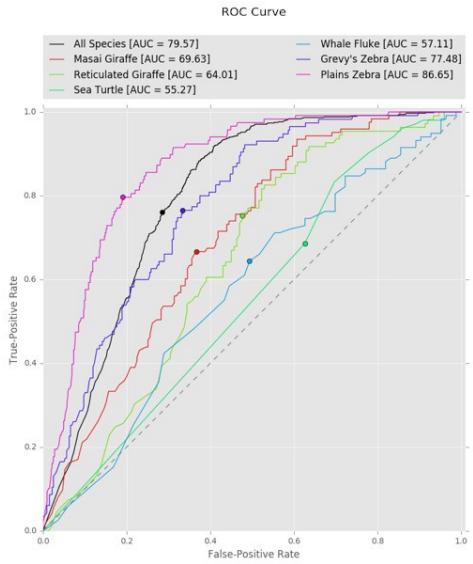


Figure 7. The AoI classifier ROC curves. The best AoI classifications were achieved by the plains zebra species, mostly due to the lower AoI to annotations ratio.

tuitive sense as the species look fairly similar and can have subtle differences between their appearances. It is worth noting here that the whale fluke and sea turtle species have almost no inter-species confusion, which is supported by their ROC AUC values of 99.78% and 98.28%, respectively. Out of the species errors that are made, 62.93% are due to incorrect sub-genus giraffe and zebra classifications. In summary, the overall genus (zebras vs. giraffes vs. whale flukes vs. sea turtles) classification accuracy is 96.40%.

The overall accuracy of species and viewpoint combination classifications is 61.71% over 42 distinct categories. The accuracy improves from this baseline when we take into account how viewpoint variance impacts identification (i.e. a $\pm 45\%$ degree shift in yaw is acceptable for giraffes and plains), which achieves a "fuzzy" accuracy of 87.11%.

4.5. Annotation Background Segmentation

Since the annotation background network was trained on noisy, patch-based data – and with the lack of fully-segmented species ground-truth in WILD – we cannot provide a true qualitative segmentation metric for the model's performance. However, looking at Figure 8, the background segmentation network performs well on various annotations of a known species to classify regions of the image as background and foreground. In this figure, the binary output masks of the background classification network are combined with their associated input annotations. Something to note is the lack of distinction between class instances (i.e. same-class animals in the annotation will not be masked out). Some of our previous work in [6] shows that overall identification matching accuracy improves when a background mask is used for feature weighting.

4.6. Annotation of Interest (AoI) Classification

Finally, the AoI classifier achieves an overall accuracy of 72.75% on the held-out test data (521 true positives, 1,268 true negatives, 506 false positives, 164 false negatives) when using a confidence threshold of 84%. Figure 7 shows ROC curves for each species. Ironically, the performance of plains zebras shines under this classification task; this further supports the claim that the background annotations for plains zebras in WILD are not good identification exemplars. The AoI classification is objectively the worst-performing component of the detection pipeline as it struggles with the ambiguity of the overall concept. The primary goal, however, of AoI selection is to reduce the overall number of poor annotations that are passed along to an identification pipeline. From this point-of-view, the AoI classifier correctly eliminates over half of the data required for processing at the cost of only 164 missed positive AoIs.

5. Conclusion

In this paper we evaluated five detection components against WILD, a new dataset of real-world animal sightings that focuses on challenging detection scenarios. Future work can be focused on improvements to WILD, improving the accuracy of annotation of interest (AoI) classification, and performing a comprehensive identification performance study. We also intend to formally evaluate the automated cleaning procedure in a future publication.

Figure 8. Annotation background segmentation results.

References

[1] T. Berger-Wolf, J. Crall, J. Holberg, J. Parham, C. Stewart, B. L. Mackey, P. Kahumbu, and D. Rubenstein. The great grevys rally: The need, methods, findings, implications and next steps. Technical report, Grevy's Zebra Trust, 2016. 1

[2] T. Y. Berger-Wolf, D. I. Rubenstein, C. V. Stewart, J. Holmberg, J. Parham, and J. Crall. Ibeis: Image-based ecological information system: From pixels to science and conservation. In *Bloomberg Data for Good Exchange Conference, New York, NY, USA*, 2015. 2

[3] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference*, volume 4 of *SciPy '10*, pages 3–10, 2010. 00548 Oral Presentation. 2

[4] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015. 2

[5] M. J. Chase, S. Schlossberg, C. R. Griffin, P. J. Bouch, S. W. Djene, P. W. Elkan, S. Ferreira, F. Grossman, E. M. Kohi, K. Landen, P. Omondi, A. Peltier, S. J. Selier, and R. Sutcliffe. Continent-wide survey reveals massive decline in african savannah elephants. *PeerJ*, 4(1):e2354, 2016. 1

[6] J. P. Crall. *Identifying Individual Animals using Ranking, Verification, and Connectivity*. phdthesis, Department of Computer Science, Rensselaer Polytechnic Institute, 2017. 7

[7] S. Dieleman, J. Schlter, C. Raffel, E. Olson, S. K. Snderby, D. Nouri, and others. Lasagne: First release., 2015. DOI: 10.5281/zenodo.27878. 2

[8] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 02424. 5

[9] T. Forrester, W. J. McShea, R. W. Keys, R. Costello, M. Baker, and A. Parsons. eMammalcitizen science camera trapping as a solution for broad-scale, long-term monitoring of wildlife populations. In *North America Congress for Conservation Biology*, 2014-07. 2

[10] K. Hara, M.-Y. Liu, O. Tuzel, and A.-m. Farahmand. Attentional network for visual object detection. *arXiv:1702.01478 [cs]*, 2017. 2

[11] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 3

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167 [cs]*, 2015. 3

[13] A. Irwin. *Citizen Science: A Study of People, Expertise and Sustainable Development*. Environment and Society. Routledge, 1995. 5

[14] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2083–2090, 2013. 2

[15] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 4

[16] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr. Microsoft COCO: Common objects in context. *arXiv:1405.0312 [cs]*, 2014. 5

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *Proceedings of the 2016 European Conference on Computer Vision*, pages 21–37. Springer, 2016. 2, 3

[18] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 2, 4

[19] J. Parham and C. Stewart. Detecting plains and grevy's zebras in th real world. In *2016 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 1–9, 2016. 3, 6

[20] J. R. Parham. Photographic censusing of zebra and giraffe in the nairobi national park. msthesis, Department of Computer Science, Rensselaer Polytechnic Institute, 2015. 2

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640:1–10, 2015. 2, 3

[22] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv:1505.04597 [cs]*, 2015-05-18. 4

[23] D. I. Rubenstein, C. V. Stewart, T. Y. Berger-Wolf, J. Parham, J. Crall, C. Machogu, P. Kahumbu, and N. Maingi. The great zebra and giraffe count: The power and rewards of citizen science. Technical report, Kenya Wildlife Service, 2015-07. 1

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *arXiv:1409.0575 [cs]*, 2014. 5

[25] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013. 2

[26] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009. 1

[27] D. Yoo, S. Park, J.-Y. Lee, A. S. Paek, and I. So Kweon. AttentionNet: Aggregating weak directions for accurate object detection. In *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pages 2659–2667, 2015. 00012. 2