A Walk through a Digital Savanna: Aerial Wildlife Detection with Synthetic Data

Benjamin Kellenberger, Derek van de Ven, Devis Tuia*

Abstract

We investigate the applicability of computer-rendered training data for aerial wildlife detection using deep learning. To this end, we generate rendered images and ground truth from the AirSim-W environment and mix it with real drone images and labels to train a RetinaNet detector for detecting mammals in the African savanna. Despite the visual dissimilarity between both data sources, our model is able to detect the animals with high recall and good precision, which results in significantly less investments required into creating ground truth labels.

I. Introduction

The loss rate of terrestrial vertebrate mammal species has been increasing for the last few decades[1], due to land degradation (1), poaching (2), and other causes. A crucial requirement for wildlife conservation in this context is the ability to monitor endangered species populations over large areas, such as wildlife reserves. Unmanned Aerial Vehicles (UAVs) that can be programmed to remotely and safely acquire aerial images over comparably vast areas are increasingly used for this purpose (3). UAVs bear the promise of non-laborious wildlife censuses, especially if combined with image analysis through Computer Vision (CV) methodologies, such as the popular Convolutional Neural Networks (CNNs; (4)). However, despite recent efforts and unprecedented accuracies of CNNs for this task (5; 6) said models still require a wealth of training images. First tools to support the annotation process are appearing (7), but still a large part of the images has to be annotated tediously by hand. Unless the need for manually created training data declines, the usefulness of CNNs for such large-scale tasks thus remains limited.

In this work, we attempt to address this problem by resorting to computer-rendered images over a simulated environment to substitute as much of the hand-labelled training data as possible. Using synthetic data for CV model training has been proposed for scenarios where real training data is hard or impossible to obtain, *e.g.* for accident scenarios for training self-driving cars (8), but also to reduce annotation efforts required (9), as in our case. In this initial study, we employ AirSim-W (10), which contains a simulated environment of the African savanna, and use it to create training imagery to train a deep CNN for the task of wildlife detection from a UAV perspective. We assess the amount of UAV-derived training data that can be substituted with rendered images so that we are still able to train a CNN-based wildlife detector to satisfying accuracy.

Benjamin Kellenberger and Devis Tuia are with the Environmental Computational Science and Earth Observation Laboratory, EPFL, Switzerland. Derek van de Ven is with Laboratory of Geo-Information Science and Remote Sensing, Wageningen University, the Netherlands.

*corresponding author: devis.tuia@epfl.ch

[1] https://www.bbc.com/news/science-environment-54091048, accessed September 15, 2020.

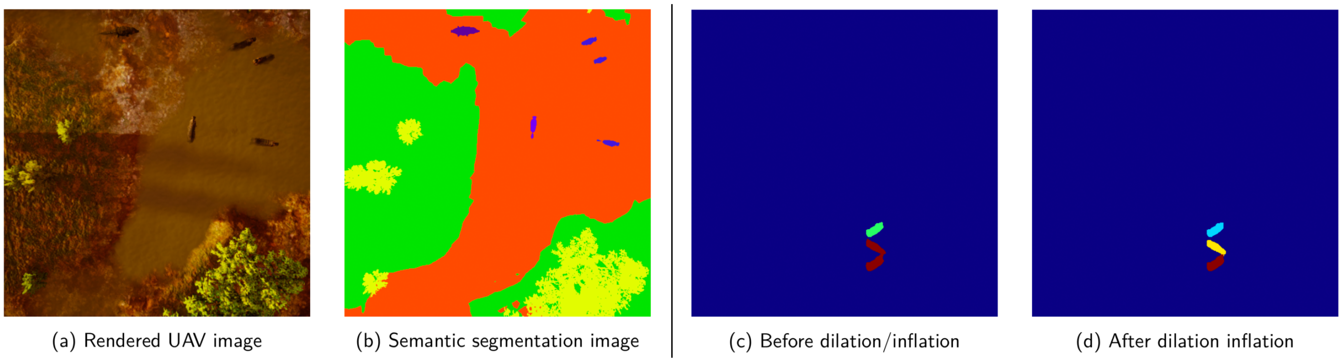
(a) Rendered UAV image (b) Semantic segmentation image (c) Before dilation/inflation (d) After dilation inflation

Fig. 1. Rendered image from the AirSim-W environment (a), with segmentation ground truth (b). Since we are interested in individual animals, we discarded non-animal ground truth and used mathematical morphology to separate connected segmentation masks (c and d).

II. Data generation and preparation

a) Rendered data: in a first step, we acquired rendered images from AirSim-W, simulating a UAV flight path that starts from the centre of the virtual African savanna environment, then proceeds to a random point within a 1000×1000 metres square around the centre via a diagonal path, with picture taken every 3-5 metres, at altitudes of 20m to 60m above ground in 10m steps. This resulted in 5000 images with accompanying segmentation maps of wildlife (Figure 1).

In order to obtain bounding boxes required for object detection, we processed the segmentation ground truth as follows: we first discarded all non-animal label classes, leaving us with animal and background pixels. This occasionally resulted in two animals' masks being connected to each other, when the individuals were standing close-by (Figure 1 (c)). To resolve this issue, we eroded the masks with a 3×3 square kernel, assigned instance codes to the now separated masks, and dilated them again. This allowed us to trivially infer bounding boxes based on minimum bounding rectangles. The total data set contained 4562 animals.

b) Real data: we resorted to the Kuzikus dataset (11)[2], which consists of 654 UAV images of size 4000×3000 pixels and a ground resolution of 4 to $8cm$. We split the images into 8000 selected patches of size 800×600, containing a total of 1518 animals in 735 of them. An example image (with predictions) can be seen in Figure 3.

c) Model training: we trained a RetinaNet (12) with ResNet-18 as feature extractor (13) on three data set combinations: *(i)* exclusively rendered data (5000 images); *(ii)* a mixture of rendered (5000) and real (1000) images; and *(iii)* exclusively real images (8000). For the mixture, we examined two training modes, one with a combined data set (*i.e.*, mixing both rendered and real images from the start), and one where we train the model on the rendered data for 20 epochs, but then add the real images and fine-tune for another seven. We trained the model on batches of four images for 20 epochs, using the Adam optimizer (14) with a learning rate of 10^{-5} and no weight decay. We performed data augmentation through random horizontal flipping and Gaussian blurring in 50% of the cases each. During testing, we applied non-maximum suppression and retained all predictions with confidence 0.01 or greater, to ensure a high recall. Predictions with an IoU ≥ 0.2 with the ground truth were treated as positives; n double detections as one true and $n-1$ false positives.

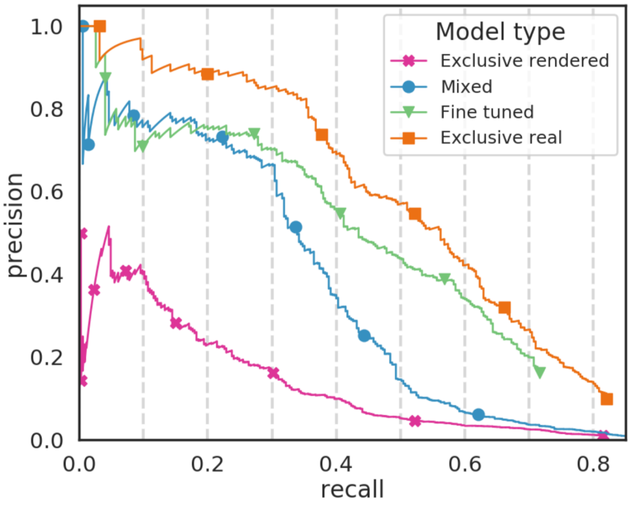

Fig. 2. Precision-recall curves on the test set for the four models.

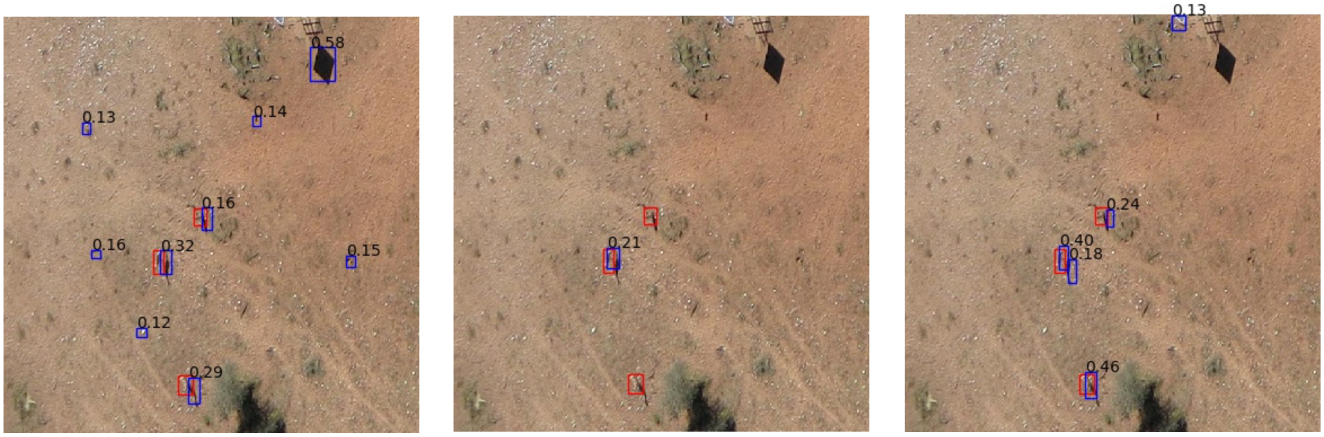

Fig. 3. Prediction results (blue) and IoU values with ground truth (red) for the “exclusive rendered” (left), mixed (middle), and exclusive real (right) model.

III. Results and Discussion

Figure 2 shows the precision-recall curves on the held-out test set (1500 real images, 345 animals) for all four models; Figure 3 shows a visual example for three models. Quite surprisingly, the “exclusive rendered” model (purple) found up to 85% of the animals, albeit with a low precision, and even managed to predict correctly sized bounding boxes. The equally low precision of the mixed model (blue) at high recall values indicates that there still is a strong discrepancy between the synthetic AirSim-W and real Kuzikus data, but this can be dramatically improved for free by performing a two-step fine-tuning approach (green). In this case, the performance is not far behind the model trained on exclusively real data (orange), but requires only a fraction of real images and annotations.

[2] Images can be downloaded at: https://doi.org/10.5281/zenodo.1204408.

IV. Conclusion

We presented a first case study on using synthetic images from a virtual, computer-generated environment to train detectors for wildlife in UAV imagery. This task is highly challenging, primarily due to the tedium involved in creating image annotations. In our experiments we were able to show that this workload can be significantly reduced by replacing a great part of the real images with synthetic ones, which are available for free. Results show only a marginal drop in precision and recall, compared to the upper bound. Future works may improve over these figures by including domain adaptation strategies like image-to-image translations as in (15).

References

[1] G. Beca, M. H. Vancine, C. S. Carvalho, F. Pedrosa, R. S. C. Alves, D. Buscariol, C. A. Peres, M. C. Ribeiro, M. Galetti, High mammal species turnover in forest patches immersed in biofuel plantations, Biological Conservation 210 (2017) 352–359.

[2] A. A. Rija, R. Critchlow, C. D. Thomas, C. M. Beale, Global extent and drivers of mammal population declines in protected areas under illegal hunting pressure, PloS one 15 (8) (2020) e0227163.

[3] J. C. Hodgson, S. M. Baylis, R. Mott, A. Herrod, R. H. Clarke, Precision wildlife monitoring using unmanned aerial vehicles, Scientific reports 6 (1) (2016) 1–7.

[4] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: NeurIPS, 2012, pp. 1097–1105.

[5] B. Kellenberger, D. Marcos, D. Tuia, Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning, Remote sensing of environment 216 (2018) 139–153.

[6] J. A. Eikelboom, J. Wind, E. van de Ven, L. M. Kenana, B. Schroder, H. J. de Knegt, F. van Langevelde, H. H. Prins, Improving the precision and accuracy of animal population estimates with aerial image object detection, Methods in Ecology and Evolution 10 (11) (2019) 1875–1887.

[7] B. Kellenberger, D. Tuia, D. Morris, AIDE: Accelerating image-based ecological surveys with artificial intelligence, Methods Ecol. Evol. (in press).

[8] A. Gambi, M. Mueller, G. Fraser, Automatically testing self-driving cars with search-based procedural content generation, in: SIGSOFT International Symposium on Software Testing and Analysis, 2019, pp. 318–328.

[9] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, A. M. Lopez, The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes, in: CVPR, 2016, pp. 3234–3243.

[10] E. Bondi, D. Dey, A. Kapoor, J. Piavis, S. Shah, F. Fang, B. Dilkina, R. Hannaford, A. Iyer, L. Joppa, et al., Airsim-w: A simulation environment for wildlife conservation with uavs, in: ACM SIGCAS Conference on Computing and Sustainable Societies, 2018, pp. 1–12.

[11] N. Rey, M. Volpi, S. Joost, D. Tuia, Detecting animals in african savanna with UAVs and the crowds, Remote Sensing of Environment 200 (2017) 341–351.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: ICCV, 2017, pp. 2980–2988.

[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[14] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: ICLR, 2015.

[15] O. Tasar, Y. Tarabalka, A. Giros, P. Alliez, S. Clerc, Standardgan: Multi-source domain adaptation for semantic segmentation of very high resolution satellite images by data standardization, in: CVPRw, 2020, pp. 192–193.