The Aerial Elephant Dataset: A New Public Benchmark for Aerial Object Detection.

J.J. Naudé, D. Joubert
Innoventix
hannes@innoventix.co.za

Abstract

Aerial surveying is a key tool for effective wildlife management. However, the high costs associated with large scale surveys means that this tool is often underutilized. We believe that computer vision can be used to dramatically decrease the costs associated with surveying, while at the same time improving the consistency of results. We present the Aerial Elephant Dataset, a challenging dataset to enable research on game detection under real-world conditions. The dataset consists of 2101 images containing a total of 15 511 African bush elephants in their natural habitats, imaged with a consistent methodology over a range of background types, resolutions and times-of-day. A baseline algorithm for elephant detection is trained and tested to demonstrate the feasibility of the proposed task. The algorithm is used in a larger system, where false positive rejection and counting of densely spaced individuals is aided by a human-in-the-loop. We evaluate the performance of this system against traditional methods by performing surveys in tandem with professional human surveying crews and comparing results in terms of detections missed, man-hours spent and cost.

Figure 1. Example of a typical image from the dataset with 11 elephants present. Elephants are circled in blue and in cases of clusters containing multiple elephants in a single circle, the count is indicated. Zoomed views of two cow-calf groups are also provided.

1. Introduction

Wildlife population monitoring is crucial to general conservation, sustainable wildlife use and managing human wildlife interactions. Aerial surveying is an effective tool for monitoring populations within large areas.

However, such surveys are often expensive and arduous to conduct. Most aerial surveys still rely on the same methods devised five decades ago [26, 40]. Since then, digital cameras have become more affordable and computer vision software has advanced to such a state that those aerial survey methods can be augmented with these technologies to reduce costs and risk to human life while improving results.

Towards realizing this possibility, we have developed a low-cost image acquisition rig that can be used to gather geolocated aerial images of large areas with off-nadir angles of 35 degrees or less at rates of up to 240 km^2 per hour when capturing at 10 cm ground sample distance resolution. This rig is entirely self-contained and can easily be mounted in any aircraft with an observation aperture.

Having the ability to gather huge amounts of aerial data is meaningless if we do no also have the ability to process it. As a proof of concept we developed the Elephant Survey System (ESS) that allows for the semi-automated aerial surveying of African bush elephants (*Loxodonta Africana*). The same basic processing flow could be used to survey any species that is generally visible from the air during daytime.

In the process of developing and testing this system we have collected a large dataset of aerial images of various different ecosystems. We have also produced a large number

of annotations of elephants that are present in these images. We believe that this is a challenging dataset to test object detection algorithms against, as the target object is extremely scarce, fairly small in pixel terms, often partially occluded and generally appears against highly cluttered backgrounds with many natural distractors.

To show the viability of the problem we present a baseline network and evaluate its performance.

Lastly we present results from real-life comparison surveys where the ESS was evaluated head-to-head against two highly experienced aerial game surveying teams, using two different methodologies against two different background types. We show that ESS based surveying produces results that are comparable to or better than manual surveys at a fraction of the cost. In addition, machine vision based surveys will continue to improve as research progresses. It is in the hope of supporting such research that the Aerial Elephant dataset is released publicly at http://innoventix.co.za/AED.

2. Related work

For decades, the field of aerial animal surveys has seen very little change. Recently, this has started changing rapidly due to four drivers:

- widespread availability of high quality digital cameras,
- availability of affordable unmanned aerial vehicles,
- crowdsourcing and
- advances in computer vision.

High quality digital cameras make it possible to photographically record huge areas and analyze it at a later stage instead of having to carry an entire observation team on board every flight. This makes aerial wildlife surveys more viable as fewer personnel are required to be present on the actual flights in locations that are often remote and may place them in harm's way. The availability of unmanned aerial systems or drones extend this trend further and dramatically lowers cost for small scale surveys, although regulatory barriers and the limited endurance and payload capabilities of contemporary drones often mean that manned aircraft are still preferred for large scale surveys.

These two factors have caused the cost of data acquisition to plummet in recent years, often with the effect that organizations can gather more data than they can realistically process. Luckily recent trends in crowdsourcing and computer vision have made inroads toward providing a similar dramatic reduction in the cost of processing.

It has long been known that a group of non-expert individuals can be as accurate as an expert for certain tasks if their responses are aggregated in a suitable manner[13, 17]. The so-called 'wisdom of crowds' [49] means that as more individuals estimate some quantity of interest, random errors are suppressed and only systematic errors remain. Crowdsourcing has been used in a number of studies to divide a task over a large number of volunteers or workers and arrive at a reliable result.

However, while crowdsourcing can dramatically reduce the cost of manual processing, the scale of certain problems mean that automation is required. Wildfowl counts done by Gilmer *et al*. [18], Bajzak and Piatt [3] and Cunningham *et al*. [14] are examples of early attempts to use automated methods to count animals from images. Machine vision has since been successfully applied in a number of contexts where animals are highly aggregated and appear with reasonable contrast against a simple background [50, 20, 23, 46, 7, 8, 19, 6, 29, 1, 11] or where the machine vision system is used in concert with crowdsourced inputs [2]. However, these limitations prevent the use of computer vision in large scale minimally constrained surveys. These are still performed [9] using techniques which are essentially unchanged from those proposed forty years ago [26, 40].

Finally, advances in machine learning due to improved techniques[28, 21, 43, 32, 42], massively increased processing power and huge annotated databases [15, 31, 16] has brought us to the point where the use of computer vision no longer needs to be limited to relatively simple cases. Recent work has seen modern machine learning methods applied to great effect to count dugong [35]and wildebeest [51].

2.1. Related datasets

The vast majority of publicly available aerial and satellite imagery datasets are focused on urban areas. In such areas, typical tasks include building detection or segmentation[34, 10, 52, 36], road extraction [52, 47] and detection of people and common man-made landscape-features or objects [53, 44, 5, 39].

In satellite imagery specifically, there are also datasets focusing on cloud[38, 37] and ship[27] detection as well as some tackling the land-cover classification problem [48, 55, 54, 12, 22, 4, 25].

To the best of the authors' knowledge the only public dataset for animal detection in aerial imagery available today is the NOAA arctic seal dataset[41]. This data set contains about one million thermal/RGB image pairs, representing a 2016 aerial survey of sea ice habitat in U.S. waters of the Chukchi Sea, conducted by NOAA fisheries. Annotations indicate the locations of approximately 7000 seals in these images.

The NOAA dataset is similar to the aerial elephant dataset in also being an extremely imbalanced data set. Unlike the proposed set however, the backgrounds are extremely simple and thermal images are available to support detection. In our initial experiments long wave infrared

Figure 2. A photo of the acquisition rig mounted in a BushCat light sport aircraft.

thermal images were collected, but it soon became apparent that these would be all but useless for daytime animal detection on the African savanna.

3. Acquisition methodology

The images were gathered using Canon 6D consumer oriented digital single-lens reflex cameras. Consumer DSLRs were preferred over professional aerial survey equipment since the goal was to maximize affordability. Equipment failure, especially mechanical shutter failure, was an initial concern. However, after 5 years of operation and in excess of 500 000 total shutter activations, we have yet to experience a single failure. These cameras were mounted in a SkyReach BushCat light sport aircraft by means of a purpose built frame, and capture images through an aperture in the fuselage.

The frame accommodates three bodies, each equipped with an 85 mm lens. One camera is pointed straight down, while the other two are tilted to the left and right respectively by 20 degrees each. This arrangement maximizes the width of the imaged strip underneath the plane, while maintaining the viewing angle at 35 degrees from nadir or below.

The cameras are controlled via their USB interfaces by a Raspberry Pi single board computer running custom software that triggers synchronous captures at a programmable frequency. The capture process can be started or stopped using a mechanical switch mounted in the cockpit or via a web-interface.

The cameras have built in GPS receivers, so each image is automatically GPS-tagged in the EXIF metadata. These GPS tags are used to reconstruct the path of the survey aircraft during analysis and to remove any images which were not taken on transect legs (for example during turns) from the data to be processed.

The cameras as well as the controlling computer are powered from a Li-ion battery which can power the rig for more than 5 hours. The rig initially drew power from the aircraft supply, but after experiencing problems due to conducted EMI, was modified to be entirely self-contained. This means that only mechanical integration is required when mounting the rig in a new carrier aircraft.

We have found empirically that elephants can be reliably detected in imagery with a ground sample distance (GSD) of about 10 cm. Our setup allows us to acquire such images from a height of about 4000 feet above ground level (AGL). At this height, the effective search strip width is just in excess of 1500 m on the ground. This is dramatically larger than the strip width that can realistically be searched by human observers from any height.

In practice we have found that the neural net performs well on data at this resolution, but human observers (especially inexperienced ones) start to experience significant ambiguity when verifying the results. For this reason, we would often prefer to operate at 3000 feet, reducing both the GSD and the search strip width by a factor 0.75.

In some cases cloud cover has forced us to operate at even lower altitudes, so the dataset contains imagery at a variety of resolutions. Images are provided at their original resolution, with GSD specified in the metadata, so both multi-scale and single scale approaches can be experimented with.

4. Processing methodology

The ESS software is used to filter the set of images to be processed, removing images that were taken during initial positioning, returning to the landing strip or during turns between transect legs. Next the software calculates a set of detections for all of the selected images by using a deep neural net that will be described in more detail in Section 6.

As mentioned before, due to the massive discrepancy in base rates for the elephant and background classes, we typically end up with more false positives than true positives. These are filtered out by human operators. While this task may consume several hours for a large survey, it would be completely unfeasible without the help of the neural nets to reduce the data to a manageable volume. In practice we have found that the neural net reduced the amount of data that had to be manually inspected by more than two orders of magnitude.

Elephants are well known to be social animals that mostly occur in herds (lone bulls being the exception). We exploit this a-priori knowledge by inspecting the images with confirmed elephants closely, to verify that no additional elephants were missed in these images. This ensures that the resulting dataset is clean and boosts the accuracy of the final count for very little additional effort (since the vast majority of images do not contain any elephants).

Acquired images overlap slightly in the transverse direction and significantly in the flight direction. Due to this, it is necessary to register images relative to one another and use

the resulting transform to ensure that elephants appearing in multiple images are not counted multiple times. The ESS software suite contains functionality to support this registration task.

Once these tasks have been completed we can query the system for the total number of elephants seen and the locations of these observations.

5. Dataset description

The dataset consists of 2101 images containing a total of 15 511 elephants. It is split into training and test subsets with 1649 images containing 12455 elephants in the training set and 452 images containing 3056 elephants in the test set. The resolution of the images varies between 2.4 cm/pixel and 13 cm/pixel, but the nominal resolution for each image is specified in the accompanying metadata, so it is a simple matter to resample images to a consistent GSD. Because acquired images often overlap, the same individuals may sometimes be seen in 2 or 3 consecutive images. Care has been taken with the train/test split to ensure that such clusters of related images are not split, thus maintaining independence of the training and test sets.

These images were acquired over the course of 8 separate campaigns in different environments. These environments are

- Hluluwe-iMfolozi Park and Phinda Private Game Reserve in central KwaZulu-Natal, South Africa (Multiple flights spread over September 2014 to May 2015)
- The Northern Tuli Game Reserve in the Tuli block, Botswana (September 2015)
- NG26 concession in the Okavango Delta, Botswana (2 campaigns, September 2015 and July 2018.
- Bwabwata and Mudumu national parks in the Zambezi strip, Namibia (2 campaigns: August 2016 and February 2018)
- Madikwe game reserve in the North-West province, South Africa (2 campaigns: July 2017 and November 2018)

The dataset represents both dry-season and wet-season backgrounds in a variety of environments and captured over the full day from sunrise to sunset.

6. Baseline network

The class imbalance inherent in the game surveying problem means that false positives are a major problem. Even with extremely low false alarm rates, we typically end up with more false positives than true positives. This means that some manual verification will be needed before the data can be used.

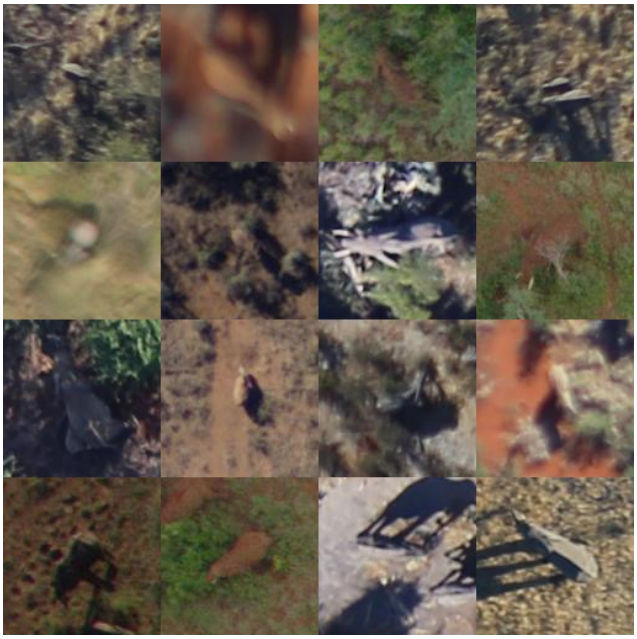

Figure 3. A selection of elephants from the dataset. The elephants in the top two rows were missed by the detector, while those in the bottom two rows were successfully detected. Note the large amount of variation in appearance, scale and background type. The missed elephants tend to be those where the image quality is very poor due to small scale, image blur or other sensor artifacts.

During manual verification we present a small crop (400x400 pixels) of the image, centered on the neural network's detection, and the operator is asked to mark the elephants present in the crop, or reject the detection if no elephants are present.

To evaluate the performance of a classifier we use the mean average precision (mAP). However, to have our figure of merit accurately reflect the real-world cost of verification we consider an elephant to have been successfully detected if it occurs within a Chebyshev distance of 200 pixels of the detection coordinate. This means that a single detection may include several elephants.

We generally expect some deterioration in performance near the edges of the image, as some of these elephants may only be partially included in the image and even if the subject elephant is completely inside the image, the network may be affected by the presence of padding in its nominal receptive field. For these reasons we exclude the outer edge (128 pixels) of each image from the analysis. Our evaluation code that computes our slightly customized mAP performance metric on the test set, given a set of detections is included with the dataset.

To perform detection we use a Mobilenet architecture [24] which has been modified to be fully convolutional as formalized in [33]. This net can do inference on an image

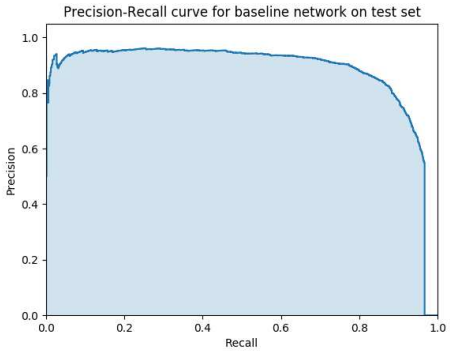

Figure 4. The precision-recall curve of the baseline network on the evaluation set.

of any size and produce a dense heatmap of classification results corresponding to strided locations within the input image. We subsequently perform connectivity on the regions with a threshold of 0.8 in the heatmap and report the center of each region as a single detection with a score equal to the maximum response within the region. We also check that regions have Chebyshev radii smaller than 200 pixels (*i.e*. the entire region falls inside the presented crop) and split them until this condition is satisfied.

Training is also performed in a fully convolutional manner, but all outputs within a guard region around known elephants are treated as 'don't care' outputs *i.e*. they are not back propagated. These outputs will contain the elephant image within their receptive fields, but it will not be centered. Since we do not want to waste resources trying to learn this arbitrary and noisy distinction, we only enforce our minimum requirements, which is that the closest output to the elephant center report a detection, while outputs with no elephant center within their receptive fields do not. The exact shape of the transition between these two states is of little importance to us. We use a focal loss [30] for the background class to focus the training effort on the hard examples.

We train on random crops of the source image chosen to be maximal squares. Data augmentation consists of random mirroring and random rotations in 90 degree increments, as well as some standard color perturbations.

The achieved precision-recall curve is shown in 4. We attain an average precision of 0.89 on the test set. Note that while this may appear quite high, the base rate discrepancy in the test-set is roughly a factor 50 smaller than in a typical survey in a high elephant density area where only 2% of images will contain elephants as opposed to 100% of the images in the test set. In a low density area the class imbalance will be even worse.

7. Comparison against human crews

The ESS system was benchmarked against the performance of human surveying teams on two separate occasions against two different experienced human survey teams using two different counting methodologies.

7.1. Fixed wing sample survey in Okavango Delta, Botswana

The first comparison was against a human crew from Elephants Without Borders performing a sample survey of elephants using a Cessna 206 in the NG26 concession in the Okavango delta in Botswana. This method requires 4 crew members (a pilot, a recorder and one observer on each side of the plane). Crew members count elephants within 200 m wide strips on the ground on either side of the plane, while ignoring any elephants outside this strip. The strip is visually delineated by two carefully placed rods attached to the wing struts. They are also equipped with digital cameras on fixed mounts which are used to take photos when herds are too large to be reliably counted in-flight.

Differences in aircraft performance meant that we could not closely synchronize the two surveys. Therefore we could only compare the estimated elephant densities yielded by the two methods rather than checking whether individual sightings correspond. In the two sorties that were evaluated, the ESS estimated density was respectively 1.98% and 5.32% higher than the density estimated by the human observers. It is a known result that manual aerial surveys miss an estimated 13% of elephants [45] so a natural interpretation is that the use of the ESS has reduced this miss rate somewhat. However, the sample size is not large enough to make this conclusion definitive.

7.2. Rotary wing full count in North-West province, South Africa

The second comparison was against another very experienced human crew, put together by Bassair aviation performing a full count of all large mammals in the Madikwe game reserve in South Africa using a Bell Jet Ranger III helicopter.

This count also involved flying a pre-planned set of transects, but in a full survey, unlike in a sample survey, the craft may deviate from transect legs and hover over or orbit around animals when necessary to obtain an accurate count. They would also fly low over dense thickets to flush out animals that are not clearly visible. This requires that they fly much lower (often below 100 feet) and slower and search a fairly narrow strip. Unlike the sample survey team, they do not have a clearly delineated strip, but since their transect legs are spaced 500 m apart, this requires that they visually search at least 250 m out on either side of the helicopter. They try to count an entire herd as an atomic unit at all

times and note location, numbers and herd composition. By tracking this information carefully and keeping all transect legs relatively short, they can avoid counting the same herd twice during a single sortie. However, the slow search rate implies that it takes multiple days to complete a count of any park of appreciable size.

Wherever possible, geographic features that impede animal movement are used as the boundaries for each day's search area, but it is impossible to reliably avoid some over- or under-counting due to inter-day movement of herds. To build confidence in the resulting numbers, the reserve is typically surveyed several times and the final counts compared. These factors conspire to make this a very expensive form of game counting.

In this comparison exercise, we took care to synchronize the two surveys to within a few minutes (by having the BushCat circle after each leg until the helicopter could complete the leg. This allowed us to do direct comparisons of sightings. We found that the ESS system had not missed a single herd detected by the human observers, and missed only 2 lone bulls, neither of which could be found in our imagery, so these misses are assumed to be due to visibility bias (an elephant standing directly underneath a large tree would be invisible from our point of view, but not necessarily from the point of view of the human observers in the low-flying helicopter). On the other hand, we detected some 8 lone bulls as well one herd of 9 elephants that were missed by the human crew.

In many cases where the same herds were spotted, our count of individuals would differ from that produced by the human team, but since they took no photos of these observations, we have no way of pinpointing whether the error is an over-count by the human team or visibility bias on the part of ESS.

Despite these differences, final counts on the two sorties compared came out to within 7.1% and 1.6% respectively.

While the verification and registration steps by human operators does still consume a significant amount of time, in all four sorties where comparisons were made we calculated that fewer man-hours were spent overall to obtain the ESS result than were spent to obtain the manual result, despite the fact that the ESS search strip is typically about 3 times wider. It is hoped that further research on this detection problem will reduce the time and expenditure in the verification phase even further.

8. Conclusion

This paper introduces the Aerial Elephant dataset, a challenging new benchmark for animal detection in aerial surveys. We also proposed a baseline algorithm and demonstrated that a system built around this algorithm is already capable of outperforming experienced human spotters. More work on this application has the potential to revolutionize aerial animal surveys by radically reducing costs and improving both precision and accuracy.

9. Acknowledgements

We gratefully acknowledge Paul Maritz for the vision and financial support that made this project possible. We also wish to thank Etelka Paxton, Ortwin Aschenborn, Pasquale Scaturro, John Bassi, Mike Chase and many others who made invaluable contributions.

References

[1] Amr Abd-Elrahman, Leonard Pearlstine, and Franklin Percival. Development of pattern recognition algorithm for automatic bird detection from unmanned aerial vehicle imagery. *Surveying and Land Information Science*, 65(1):37, 2005. 2

[2] C. Arteta, V. Lempitsky, and A. Zisserman. Counting in the wild. In *European Conference on Computer Vision*, 2016. 2

[3] D Bajzak and John Piatt. Computer-aided procedure for counting waterfowl on aerial photographs. *Wildlife Society Bulletin*, 18:125–129, 01 1990. 2

[4] Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, and Ramakrishna Nemani. Deepsat: a learning framework for satellite imagery. In *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*, page 37. ACM, 2015. 2

[5] Margherita Bonetto, Pavel Korshunov, Giovanni Ramponi, and Touradj Ebrahimi. Privacy in mini-drone based video surveillance. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 4, pages 1–6. IEEE, 2015. 2

[6] Marjolein Bruijning, Marco D Visser, Caspar A Hallmann, and Eelke Jongejans. trackdem: Automated particle tracking to obtain population counts and size distributions from videos in r. *Methods in Ecology and Evolution*, 9(4):965–973, 2018. 2

[7] Dominique Chabot, Christopher Dillon, and Charles Francis. An approach for using off-the-shelf object-based image analysis software to detect and count birds in large volumes of aerial imagery. *Avian Conservation and Ecology*, 13:15, 06 2018. 2

[8] Dominique Chabot and Charles M Francis. Computer-automated bird detection and counts in high-resolution aerial images: a review. *Journal of Field Ornithology*, 87(4):343–359, 2016. 2

[9] Michael J Chase, Scott Schlossberg, Curtice R Griffin, Philippe JC Bouché, Sintayehu W Djene, Paul W Elkan, Sam Ferreira, Falk Grossman, Edward Mtarima Kohi, Kelly Landen, et al. Continent-wide survey reveals massive decline in african savannah elephants. *PeerJ*, 4:e2354, 2016. 2

[10] Qi Chen, Lei Wang, Yifan Wu, Guangming Wu, Zhiling Guo, and Steven L Waslander. Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. *ISPRS Journal of Photogrammetry and Remote Sensing*, 147:42–55, 2019. 2

[11] Louis-Philippe Chrétien, Jérôme Théau, and Patrick Ménard. Visible and thermal infrared remote sensing for the detection of white-tailed deer using an unmanned aerial system. *Wildlife Society Bulletin*, 40(1):181–191, 2016. 2

[12] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018. 2

[13] Marquis de Condorcet. Essay on the application of mathematics to the theory of decision-making. *Reprinted in Condorcet: Selected Writings, Keith Michael Baker, ed*, 33, 1976. 2

[14] David J Cunningham, William H Anderson, and R Michael Anthony. An image-processing program for automated counting. *Wildlife Society Bulletin*, 24(2):345–346, 1996. 2

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 2

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

[17] Francis Galton. Vox populi (the wisdom of crowds). 2

[18] D.S. Gilmer, J.A. Brass, L.L. Strong, and D.H. Card. Goose counts from aerial photographs using an optical digitizer. 16:204–206, 01 1988. 2

[19] Luis Gonzalez, Glen Montes, Eduard Puig, Sandra Johnson, Kerrie Mengersen, and Kevin Gaston. Unmanned aerial vehicles (uavs) and artificial intelligence revolutionizing wildlife monitoring and conservation. *Sensors*, 16(1):97, 2016. 2

[20] GJ Grenzdörffer. Uas-based automatic bird count of a common gull colony. *International archives of the photogrammetry, Remote sensing and spatial information sciences*, 1:W2, 2013. 2

[21] 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. 2

[22] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *arXiv preprint arXiv:1709.00029*, 2017. 2

[23] 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. 2

[24] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 4

[25] Dino Ienco, Raffaele Gaetano, Claire Dupaquier, and Pierre Maurel. Land cover classification via multitemporal spatial data by deep recurrent neural networks. *IEEE Geoscience and Remote Sensing Letters*, 14(10):16851689, Oct 2017. 2

[26] G.M. Jolly. Sampling methods for aerial censuses of wildlife populations. *East African Agricultural and Forestry Journal*, pages 46–49, 1969. 1, 2

[27] Kaggle. Airbus ship detection challenge. https://www.kaggle.com/c/airbus-ship-detection/data. 2

[28] 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. 2

[29] Andrea S Laliberte and William J Ripple. Automated wildlife counts from remotely sensed imagery. *Wildlife Society Bulletin*, pages 362–371, 2003. 2

[30] 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 vision*, pages 2980–2988, 2017. 5

[31] 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. 2

[32] 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. 2

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 4

[34] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017. 2

[35] Frederic Maire, Luis Mejias Alvarez, and Amanda Hodgson. Automating marine mammal detection in aerial images captured during wildlife surveys: a deep learning approach. In *Australasian Joint Conference on Artificial Intelligence*, pages 379–385. Springer, 2015. 2

[36] Microsoft. Microsoft/canadianbuildingfootprints, Mar 2019. 2

[37] S. Mohajerani, T. A. Krammer, and P. Saeedi. A cloud detection algorithm for remote sensing images using fully convolutional neural networks. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5, Aug 2018. 2

[38] S. Mohajerani and P. Saeedi. Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery. *CoRR*, abs/1901.10077, 2019. 2

[39] T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *European Conference on Computer Vision*, pages 785–800. Springer, 2016. 2

[40] M Norton-Griffiths. *Counting Animals*. African Wildlife Leadership Foundation, 1978. 1, 2

[41] National Oceanic and Atmospheric Administration. "NOAA Arctic Seals". http://lila.science/datasets/arcticseals. 2

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[43] 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. 2

[44] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *European conference on computer vision*, pages 549–565. Springer, 2016. 2

[45] Scott Schlossberg, Michael Chase, and Curtice Griffin. Testing the accuracy of aerial surveys for large mammals: An experiment with african savanna elephants (loxodonta africana). *PLOS ONE*, 11:e0164904, 10 2016. 5

[46] AC Seymour, J Dale, M Hammill, PN Halpin, and DW Johnston. Automated detection and enumeration of marine wildlife using unmanned aircraft systems (uas) and thermal imagery. *Scientific Reports*, 7:45127, 2017. 2

[47] Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sebastien Benitez, and Uwe Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. 2012. 2

[48] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. *arXiv preprint arXiv:1902.06148*, 2019. 2

[49] James Surowiecki. *The wisdom of crowds*. Anchor, 2005. 2

[50] Colin J Torney, Andrew P Dobson, Felix Borner, David J Lloyd-Jones, David Moyer, Honori T Maliti, Machoke Mwita, Howard Fredrick, Markus Borner, and J Grant C Hopcraft. Assessing rotation-invariant feature classification for automated wildebeest population counts. *PloS one*, 11(5):e0156342, 2016. 2

[51] 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*, 2019. 2

[52] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 2

[53] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[54] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279. ACM, 2010. 2

[55] Xiaoxiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Hossein Bagheri, Jian Kang, Hao Li, Lichao Mou, Guicheng Zhang, Matthias Hberle, Shiyao Han, Yuansheng Hua, Rong Huang, Lloyd Hughes, Yao Sun, Michael Schmitt, and Yuanyuan Wang. So2sat lcz42, 2018. 2