Small or Far Away? Exploiting Deep Super-Resolution and Altitude Data for Aerial Animal Surveillance

Mowen Xue dt20957@bristol.ac.uk
Theo Greenslade tg17437@bristol.ac.uk
Majid Mirmehdi majid@cs.bris.ac.uk
Tilo Burghardt tilo@cs.bris.ac.uk

Dept of Computer Science, University of Bristol, Bristol, BS8 1UB, UK

Abstract

Visuals captured by high-flying aerial drones are increasingly used to assess biodiversity and animal population dynamics around the globe. Yet, challenging acquisition scenarios and tiny animal depictions in airborne imagery, despite ultra-high resolution cameras, have so far been limiting factors for applying computer vision detectors successfully with high confidence. In this paper, we address the problem for the first time by combining deep object detectors with super-resolution techniques and altitude data. In particular, we show that the integration of a holistic attention network based super-resolution approach and a custom-built altitude data exploitation network into standard recognition pipelines can considerably increase the detection efficacy in real-world settings. We evaluate the system on two public, large aerial-capture animal datasets, SAVMAP and AED. We find that the proposed approach can consistently improve over ablated baselines and the state-of-the-art performance for both datasets. In addition, we provide a systematic analysis of the relationship between animal resolution and detection performance. We conclude that super-resolution and altitude knowledge exploitation techniques can significantly increase benchmarks across settings and, thus, should be used routinely when detecting minutely resolved animals in aerial imagery.

1. Introduction

Motivation and Aerial Surveys. Collecting regular wildlife census information through timely and accurate population surveillance [61, 19, 51] is crucial in understanding how animal populations move and change [35], and how conservation efforts can be conducted to counter-act environmental degradation [60] and species decline [7, 9]. Whilst surveillance on human populations opens a multitude of ethical concerns, surveillance for the purpose of animal protection is an ethical imperative. Manual surveys [57, 53] are often expensive though, have limited site access [66], and may even expose staff to poacher threats [43] or transport risks [35]. Recently, the use of unmanned aerial vehicles (UAVs) with ultra-high resolution cameras has emerged as a cost-effective alternative for

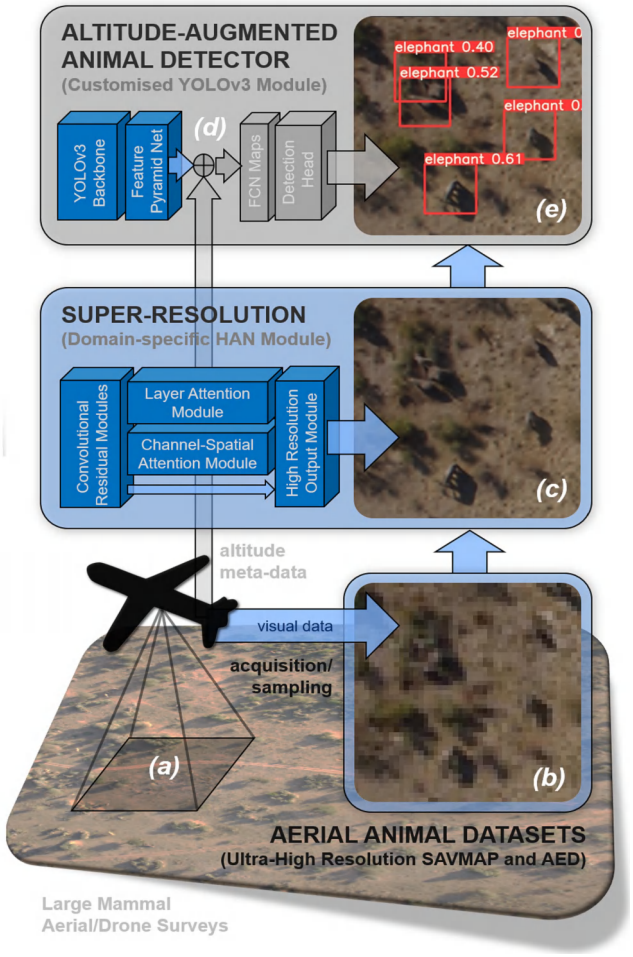


Figure 1. **Conceptual Overview.** Our approach integrates super-resolution and altitude data exploitation directly into deep animal detection pipelines for aerial survey applications. ***(a)*** During aerial image capture both ultra-high resolution RGB stills (blue) and associated altitude data (grey) is recorded. ***(b)*** We experiment with resulting original and systematically downsampled visuals. ***(c)*** A domain-trained holistic attention network is used to super-resolve the imagery, enhancing minutely resolved animal depictions. ***(d)*** Altitude data is then used as additional network input to effectively constrain the valid animal scale and appearance. ***(e)*** A custom-trained baseline network (YOLOv3) finally performs animal detection on the altitude-aware super-resolved inputs. For the public SAVMAP and AED datasets the setup proves highly effective, improving benchmarks beyond baselines and prior works.

various survey types [66, 19, 65]. However, the detection of scarce and often minutely resolved animals in vast amounts of highly variable environmental image content (see Fig. 1*(a)*) still poses a significant challenge.

Animal Detection in Aerial Imagery. Recent deep learning approaches that perform object detection for animal recognition in aerial imagery have been applied with some success [50, 24, 14], but in contrast to other fields of vision, like image classification [4], autonomous driving [45], benchmarks lack well behind human performance. In fact, even tiny object detection in less cluttered and variable environments poses an ongoing challenge for current object detection methods [71]. Whilst topics like dataset imbalance and animal scarcity [24], domain variability and transfer [66, 22], and semi-supervision and active learning [50, 25, 23] have well been studied in the domain of aerial animal detection, the problem of minute animal resolution [66] has not been tackled explicitly to date. As stated, animal appearance information often resides in only a few dozen pixels per animal for the majority of aerial datasets. Similar to recent works in other domains [12, 52, 2], we observed that domain-trained super-resolution (SR) techniques can recover valid animal-specific appearance information in many cases. In addition, we noted that virtually all aerial datasets are tagged with altitude information which could be used as an inference-basis to relate and constrain expected animal sizes in the images.

Paper Concept. Bringing these two ideas together, we propose directly combining deep object detectors with super-resolution techniques and altitude data in a single recognition pipeline as shown in Fig. 1. In particular, we will show that the integration of a holistic attention network (HAN) super-resolution approach and a altitude data exploitation network integrated into a baseline detector pipeline can significantly increase the detection efficacy of tiny animal detection in aerial imagery.

Main Contributions. Our key contributions can be summarised as follows: 1) We introduce a new animal detection approach for aerial surveys that integrates HAN-based super-resolution and altitude data into a baseline detector pipeline. 2) We evaluate our method on the two main, large aerial-capture animal datasets SAVMAP and AED, outperforming baselines and the state-of-the-art. 3) We perform detailed ablations and provide a systematic analysis of the relationship between animal resolution and detection performance in the system.

2. Background

2.1. Deep Learning for Animal Detection

Deep Object Detectors. Deep learning for object detection forms an active and extensive research field in computer vision. Many detector designs have been proposed in the recent past [63, 5, 69] with most common approaches following either a one-stage [47, 34, 36] or two-stage network design [16, 18, 15, 49]. Applications to animal biometrics [29] most often encompass species detection in camera trap imagery [54, 41, 67, 70] or manual photography [8, 11], as well as in captive settings [72] and agricultural inspection [55].

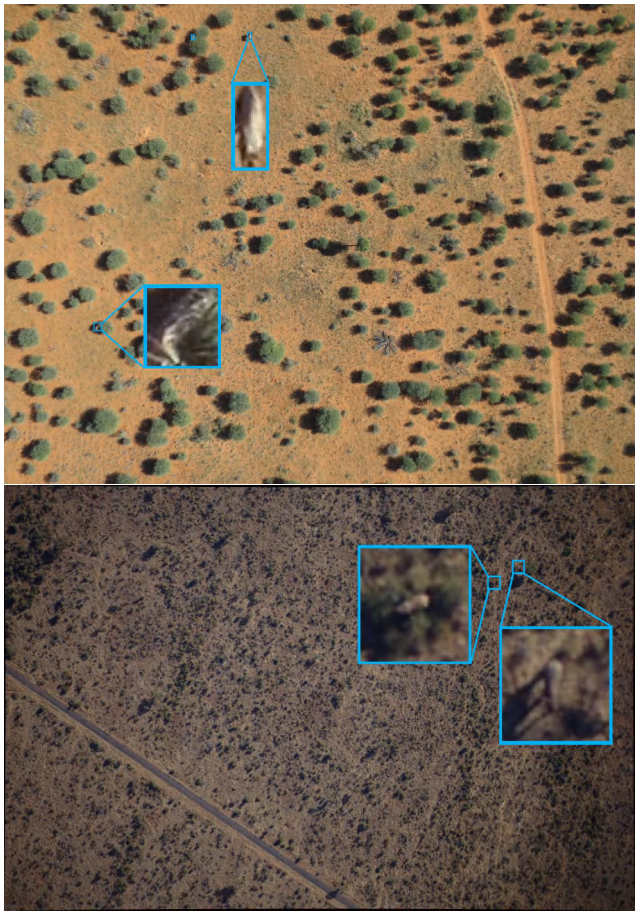


Figure 2. **Challenging Aerial Animal Imagery.** Representative images in original resolution selected ***(top)*** from the 654 ultra-high resolution SAVMAP images produced by high-flying drones, and ***(bottom)*** from the 2,074 images of the AED elephant dataset. Ground truth animal annotations are highlighted and some zoom-in is provided for better appreciation of the visuals. Note the vastness of the depicted environment and the low and challenging animal resolutions associated to the data.

Aerial Animal Detection. Research into applying object detection to aerial imagery of animals is still rare, but recent works have picked up pace with a focus on species such as whales [17], cattle [3, 1], and other large mammals [50, 24]. Utilising the SAVMAP aerial dataset [48] containing large African mammals in vast savannah environments (see Fig. 2*(top)*), Kellenberger et al. [24] showed recently that deep learning approaches outperform traditional vision techniques [50] in this domain, too. Their sys-

tem used a ResNet-18 backbone with two multi-layer perceptrons (MLPs) added with nonlinear activations (ReLU), dropout regularization, and Softmax activation. To further improve the system, curriculum learning, hard negative mining, a border class, and a new Census-oriented evaluation protocol were introduced by the group. They showed that 1) this model could still be effective when trained using weakly-supervised learning with only a small number of fully annotated labels [25], and 2) how this model could be transferred to new datasets using active learning [23].

Apart from SAVMAP, the Aerial Elephant Dataset (AED) [39] has recently been released as another public large-scale aerial drone dataset for animal detection, albeit providing slightly higher animal resolutions (see Fig. 2*(bottom)*). To benchmark this dataset, MobileNet [20] was adapted to a fully convolutional architecture by the authors [39] to perform image segmentation and ultimately elephant detection. Most recently, Duporge et al. [14] applied deep object detectors to find elephants in non-public, high-resolution satellite data (copyrighted by Maxar Technologies and the European Space Agency) to compare manual detection benchmarks against a standard two-stage Faster Region Convolutional Neural Network (F-RCNN) model. Due to the limitations of accessing satellite datasets, this paper will focus evaluation efforts on the publicly available drone datasets SAVMAP and AED to aid transparent scientific comparability and universal reproducibility. To the best of our knowledge, none of the published animal detection approaches for aerial data has so far utilised altitude data or addressed very low animal resolution explicitly.

Tiny Object Detection in Aerial Images. Current research into tiny object detection [71] has resulted in many frameworks to address this 'few pixel detection challenge'. Most aerial datasets used to evaluate approaches contain vehicles, such as cars or planes, captured via satellites [44, 56]. Approaches that address the resolution challenge include the use of feature pyramid networks [33], hard mining methods [58] and more recently, attention base mechanisms [32]. Benchmarking network architectures for tiny object detection, a very recent review [37] concluded that F-RCNN and YOLOv3 ('You-Only-Look-Once') [46, 47] currently perform strongest for the task across most performance statistics. Torney et al. [64] even showed that YOLOv3 could achieve comparable results with human annotators when detecting wildebeest in images from the Zooniverse-driven [59] Serengeti Wildebeest Count using aerial data from low-flying drones [38]. Although similar in content to usual survey data like SAVMAP, these images are taken from an up to five times lower altitude and show animals resolved significantly larger than for most common surveys. Nevertheless, following these leads [37, 64] we base our core architecture around a YOLOv3 backbone to leverage its noted applicability to wildlife detection and its proven performance on detecting tiny object content.

2.2. Deep Learning for Super-Resolution

Only recently has tiny object detection in aerial images been addressed via the application of super-resolution [10]. This approach builds on a long-standing research thread in performing super-resolution via deep learning initiated by works such as SRCNN [13]. Many architectures followed this ground-breaking work using more and more modern deep learning techniques to improve these results. The recursive convolutional networks DRCN [26] and DRRN [62] were introduced, a pyramidal framework was used in the LapSR network [30], and a generative adversarial network (GAN) was used in work on SRGAN [31]. Generally, residual learning and particularly deep architectures with large receptive fields produce particularly strong super-resolution performance [40].

Attention mechanisms have further improved benchmarks in recent years [73, 68, 27, 21]. Thus, in this research we utilise the current HAN super-resolution approach [40], which indeed incorporates two attention components – a layer attention module (LAM) and a channel-spatial attention module (CSAM). It performs state-of-the-art single image super resolution (SISR). Again, to the best of our knowledge, super-resolution has never been investigated for detecting animals in aerial imagery.

3. Datasets

This work is evaluated on two of the biggest public aerial drone datasets containing animals, that is the SAVMAP dataset [48] and the AED dataset [39]. Both represent real-world conditions that realistically reflect the challenges of performing aerial surveys. They cover African landscapes with manually annotated labels for the location of wildlife.

SAVMAP: The SAVMAP dataset was taken in the Kuzikus Wildlife Reserve between May 12 and May 15 2014. Kuzikus is a private wildlife park covering an area of $103km^2$ located in eastern Namibia. There are more than 20 species and 3,000 large mammals in this park, including Common Elands (Taurotragus oryx), Greater Kudus (Tragelaphus strepsiceros), Gemsboks (Oryx gazella), Hartebeests (Alcelaphus buselaphus), Gnus (Connochaetes gnou and Connochaetes taurinus) and others [42]. The dataset covers five flights using an ultra-high resolution Canon Powershot camera fixed on the aircraft. It contains 654 images, resolved at 3000×4000 pixels. Following Kellenberger et al [24], we used their training-validation-test split of $70\%-10\%-20\%$. Original MicroMapper crowd sourced labels [42] were error-corrected and improved by Kellenberger et al. [24]. The detection ground truth is formed by 1,183 tight animal bounding boxes with an average size of 25×23 pixels (see Fig. 2*(top)*).

AED: This dataset contains aerial images of African elephants (Loxodonta Africana) captured between 2014 and 2018. Resolving animals at slightly higher resolutions compared to SAVMAP, the dataset has 2,074 images containing 15,581 elephants [39] with a train-test split of $80\% - 20\%$. Canon 6D digital single-lens reflex cameras were attached to a SkyReach BushCat light-sport aircraft to produce the dataset. To maximise the image size, imagery was acquired using three cameras: one pointing straight down, and two pointing out left and right by 20 degrees each, all controlled by a Raspberry Pi to synchronise the capture. The images were acquired in 5 different wildlife reserves in Africa in 8 separate campaigns: Hluhluwe-iMfolozi Park and Phinda Private Game Reserve in central KwaZulu-Natal, South Africa, the Northern Tuli Game Reserve in the Tuli Block, Botswana, NG26 concession in the Okavango Delta, Botswana, Bwabwata and Mudumu national parks in the Zambezi strip, Namibia and the Madikwe game reserve in the North-West province, South Africa. The dataset contains images captured at various times of day from sunrise to sunset and in both the wet season and dry season. The labels provided in the dataset are coordinates of the centre of the elephants in the images. For comparability with a bounding box paradigm, we also defined approximate bounding boxes of 100×100 pixels around the centre coordinates even though some elephants (e.g. juveniles) in the dataset take up a smaller area (see Fig. 2*(bottom)*).

4. Method

Fig. 1 outlines the proposed recognition pipeline: acquired ultra-high resolution RGB input is first super-resolved via the HAN Resolution Enhancement Module and then fed forward into an Altitude-augmented Module which performs animal detection.

4.1. HAN Resolution Enhancement Module

Inspired by RCAN [73], we use HAN resolution enhancement (see Fig. 1*(c)*) exactly as described in [40] with its four fundamental parts: 1) a feature extraction backbone, followed by 2) LAM and 3) CSAM holistic feature weighting and, finally, 4) the image reconstruction block generating super-resolved content. To perform feature extraction, first a convolutional layer extracts shallow features from the low-resolution input, which are subsequently passed through a backbone made up of several residual groups to form content appearance features. Then, the LAM examines correlations between layers to emphasise hierarchical features in the image adaptively. These feature correlations are often ignored by other current SISR methods that use CNNs which often results in texture details in output images being smoothed, which is avoided here. To incorporate channel-spatial attention dependencies, we also use CSAM to selectively capture more informative features by learning

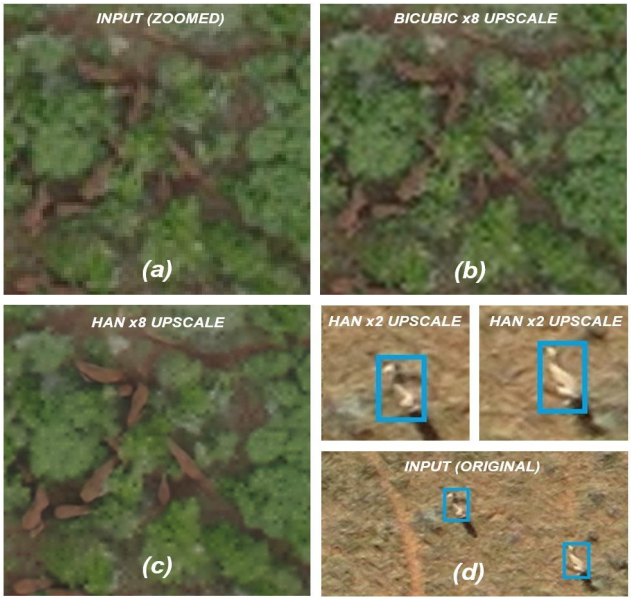


Figure 3. **HAN Super-Resolution of Aerial Animal Content.** Qualitative example depictions of super-resolution results applied to aerial animal imagery. ***(a)*** First, we show an AED data patch sampled at low animal resolution upscaled via ***(b)*** an 8-fold bicubic interpolation baseline, which is visually outperformed by ***(c)*** our domain-trained HAN super-resolution module producing a significantly clearer image at 8-fold upscale. ***(d)*** Secondly, we show a 2-fold upscaling application of our super-resolution component to a SAVMAP image patch with ground truth annotations superimposed. Note the enhancement of animal recognisability across examples. Our quantitative results in Tables 1 and 2 show that these qualitative observations regarding super-resolution are indeed aligned with improved detection performance.

across all channels. The super-resolved output finally produced by the reconstruction block (see Fig. 3 for examples) serves as input to a subsequent detection module.

4.2. Altitude-augmented Module

Our proposed altitude-augmented module explicitly incorporates altitude information by feature concatenation leading into the detection head. Incoming super-resolved content is processed via the DarkNet [47] backbone and a subsequent feature pyramid network. The latter provides content analysis at different scale levels - our key ingredient to enhancing tiny object detection. The resulting feature map is flattened into a long vector. We concatenate the altitude data, for example altitude $\mathcal{A} = 1496.68$ (in metres) represented as a 32bit float scalar, at the end of the feature vector to form the final feature representation, which is mapped via two fully connected layers before being fed into the detection head. The main components of the altitude-augmented module is outlined in Fig. 1*(d/e)*). The code for this work is available at `https://github.com/Mowen111/SALT`.

5. Implementation and Experimental Setup

Domain-specific Super-Resolution Training. We first downsampled the training image portion of the AED dataset – which as discussed is resolved higher than SAVMAP – by factors 2, 4 or 8, and then use these truly low-resolution images and corresponding higher-resolution images to train the super-resolution module. The initial learning rate was set to 1e-4. We used step decay with a learning rate decay factor of 0.5. The optimiser used for training was Adam [28] where beta1 was set to 0.9, beta2 was 0.999, and epsilon was 1e-8. We use weight decay with a value of 0.0001. We train the system for 50 epochs in total which resulted in a PSNR score on the validation dataset of 37.31, 31.68 and 27.30 for networks that scale by a factor of 2, 4 and 8 respectively.

Altitude-augmented Training. With a domain-trained super-resolution module in hand, we apply the learned resolution upscaling to both the SAVMAP and AED datasets. Subsequently, per dataset we train the altitude-augmented module on the resolution-enhanced visual input (sub-patched at a resolution of 512×512 pixels) and associated altitude meta-data using an initial learning rate of 1e-4, weight decay of 0.001, and momentum of 0.9. Standard Stochastic Gradient Descent (SGD) [6] optimisation is used to train the network for 1,000 epochs (see Fig. 4).

Inference Pipeline. The testing portion of each of the datasets are used and divided into patches (at a resolution of 512×512 pixels). For each patch, super-resolution is applied and before feeding the enhanced visual together with altitude meta-data into the altitude-augmented module to yield animal detections. Following [47], for visualisations and tests we use a detection confidence threshold of 0.1 across all compared setups.

Evaluation Metrics. Available state-of-the-art baselines for the SAVMAP and AED differ with respect to the exact evaluation metrics used. In order to address this, we provide two mean average precision (mAP) measures for evaluation, derived slightly differently and allowing for comparability with different previous works.

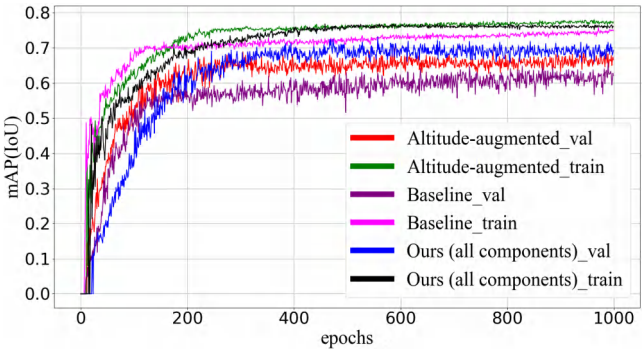


Figure 4. **Training Evolution.** Depicted is the development of SAVMAP training and validation mAP(IoU) results over 1,000 epochs of the SGD optimisation process for our baseline (YOLOv3 [47]) ***(purple and magenta)***, the altitude-augmented model ***(red and green)***, and our proposed model ***(blue and black)***.

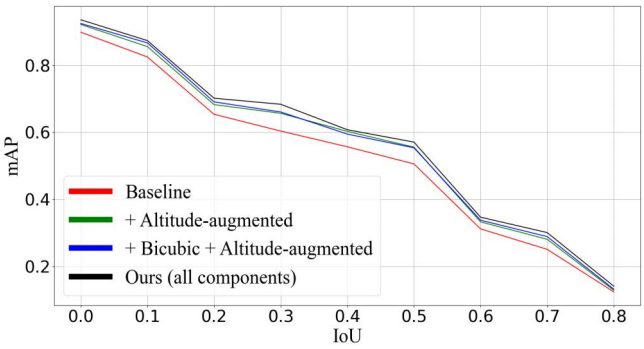


Figure 5. **Result Consistency.** Expectedly, mAP on SAVMAP test data declines as the IoU threshold for detection increases. However, our proposed method consistently performs better than other methods across the entire spectrum.

First, in line with most object detection evaluation paradigms [47], we provide mAP detection success judged based on the intersection-over-union (IoU) [47] between ground truth and candidate detection marked as mAP(IoU). In essence, a detection is a true positive only if the IoU to a ground truth annotation is above a threshold, which was set to 0.3 in this paper. If the predicted bounding box does not have a high enough IoU with any ground truth bounding box, it is classified as a false positive. Additionally, if there is no detection with a high enough IoU for a ground truth bounding box, that is a false negative.

On the other hand, AED [39] has so far been benchmarked based on Chebyshev distances rather than IoU calculations. The Chebyshev distance can be computed as $d = max(|x1 - x2|, |y1 - y2|)$, where $(x1, y1)$ and $(x2, y2)$ are the coordinates of the center points of the detection and ground truth, respectively. Following [39], the detection threshold is set to 200 pixels maximally accepted distance and the metric is marked as mAP(Che) in this paper. Note that the scale of detections is ignored in this metric.

6. Results

Applying our trained framework to a single SAVMAP frame takes 3.17 seconds for the full super-resolution and detection inference process on a system with a GTX1660 Titan GPU, 16GB RAM and Intel i7-10750H CPU.

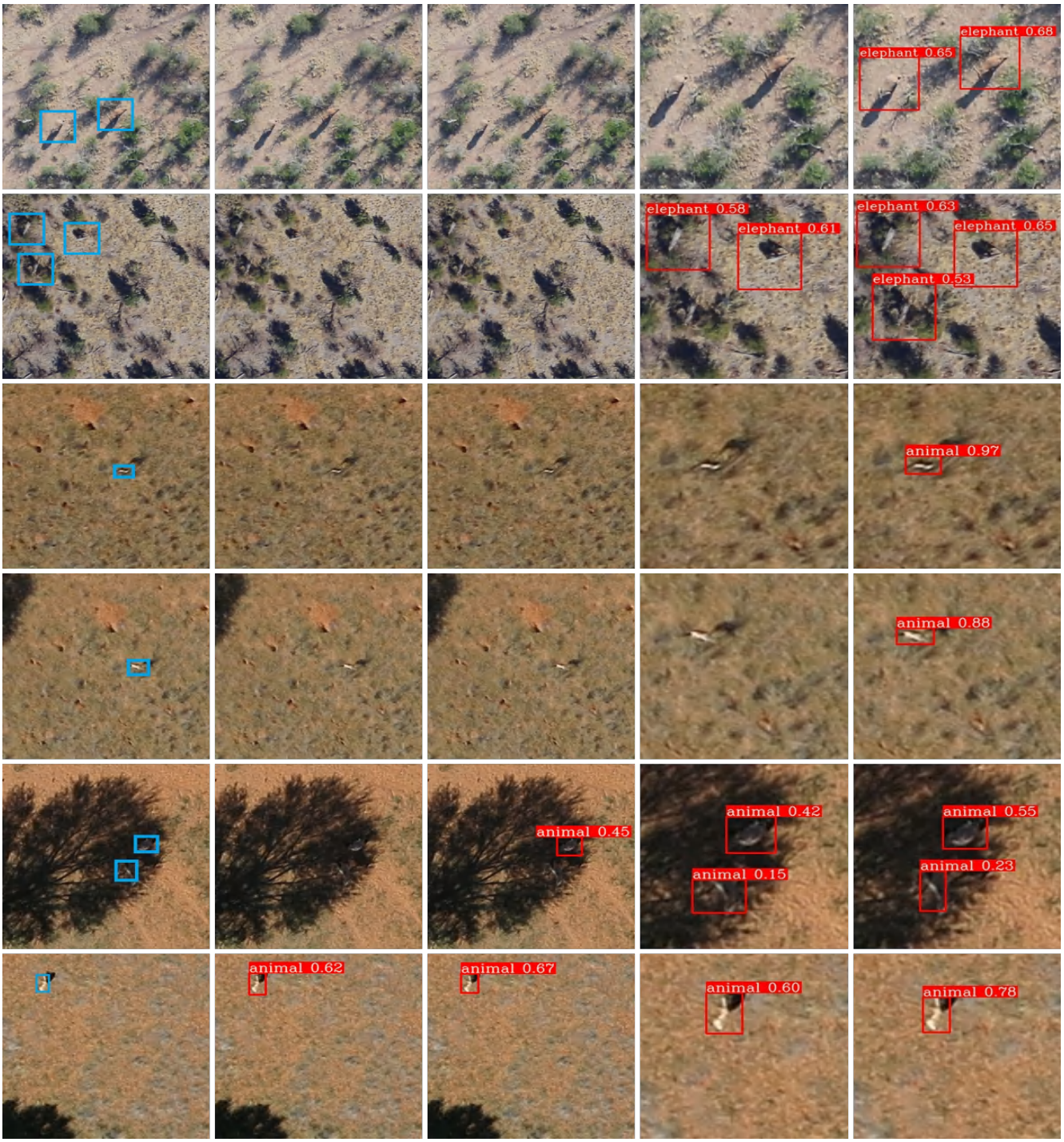
Comparative results for both AED and SAVMAP datasets are shown in Table 1 whilst result independence from the choice of IoU threshold is exemplified in Fig. 5. We report two different mAP outcomes underpinned by IoU and Chebychev calculations, respectively. Results show that the performance of benchmarks published so far (rows 1 and 6) can be improved upon by utilising up-to-date

Dataset	Row	Method	Operational Resolution	**mAP**(IoU)	**mAP**(Che)
SAVMAP [48]	1	Kellenberger et at.[25]	512×512	0.588	—
	2	Baseline		0.654	0.855
	3	+ Altitude-augmented		0.683	0.875
	4	+ Bicubic + Altitude-augmented	$512 \times 512 \rightarrow 1024 \times 1024$	0.691	0.886
	5	**Ours (all components)**		**0.702**	**0.892**
AED [39]	6	Naude et at.[39]	512×512	—	0.890
	7	Baseline		0.721	0.915
	8	+ Altitude-augmented		0.755	0.934
	9	+ Bicubic + Altitude-augmented	$512 \times 512 \rightarrow 1024 \times 1024$	0.763	0.946
	10	**Ours (all components)**		**0.778**	**0.955**

Table 1. **Result Overview.** We compare mAP results (showing both IoU and Chebychev calculations as discussed in Sec. 5) on the testing portion of each of the two datasets SAVMAP and AED. Previously published state-of-the-art benchmarks are given in rows 1 and 6. We use the state-of-the-art standard YOLOv3 [47] as our baseline (see rows 2 and 7). Augmenting baseline with altitude meta-data (see Sec. 4.2) can further improve animal detections (rows 3 and 8). Scaling image resolutions up using bicubic interpolation before detection consistently improves benchmarks again (see rows 4 and 9). Finally, rows 5 and 10 quantify results for the use of all proposed components, i.e. domain-specific super-resolution described in Sec. 4.1 feeding into altitude-augmented detection. This approach can demonstrably and consistently outperform the other techniques across both datasets.

Dataset	Method	Operational Resolution	Scale	**mAP**(IoU)	**mAP**(Che)
SAVMAP [48]	Baseline	256×256	1/2 to 1	0.612	0.795
	+ Altitude-augmented			0.653	0.824
	+ Bicubic + Altitude-augmented	$256 \times 256 \rightarrow 512 \times 512$		0.661	0.834
	Ours (all components)			**0.670**	**0.839**
	Baseline	128×128	1/4 to 1	0.568	0.732
	+ Altitude-augmented			0.603	0.755
	+ Bicubic + Altitude-augmented	$128 \times 128 \rightarrow 512 \times 512$		0.625	0.759
	Ours (all components)			**0.642**	**0.768**
	Baseline	64×64	1/8 to 1	0.552	0.695
	+ Altitude-augmented			0.594	0.734
	+ Bicubic + Altitude-augmented	$64 \times 64 \rightarrow 512 \times 512$		0.599	0.755
	Ours (all components)			**0.615**	**0.762**
AED [39]	Baseline	256×256	1/2 to 1	0.652	0.872
	+ Altitude-augmented			0.688	0.901
	+ Bicubic + Altitude-augmented	$256 \times 256 \rightarrow 512 \times 512$		0.698	0.911
	Ours (all components)			**0.703**	**0.915**
	Baseline	128×128	1/4 to 1	0.435	0.662
	+ Altitude-augmented			0.485	0.695
	+ Bicubic + Altitude-augmented	$128 \times 128 \rightarrow 512 \times 512$		0.495	0.701
	Ours (all components)			**0.532**	**0.712**
	Baseline	64×64	1/8 to 1	0.312	0.534
	+ Altitude-augmented			0.384	0.585
	+ Bicubic + Altitude-augmented	$64 \times 64 \rightarrow 512 \times 512$		0.452	0.612
	Ours (all components)			**0.475**	**0.633**

Table 2. **Resolution Analysis.** Artificially downsampling 512×512 pixel test images via bicubic interpolation systematically simulates acquisition at lower and lower animal resolution. We show here that reconstruction back to 512×512 pixel resolution via our suggested approach can maintain detection performance in these scenarios best. Superior results are consistent across datasets and mAP benchmarks for both IoU and Chebychev calculations.



Ground Truth | *Baseline* | *Altitude-augmented* | *Bicubic + Altitude-augmented* | *Ours (all components)*

Figure 6. **Detection Examples across Methods.** Ground truth annotations of animals (blue) in test patches from the AED (rows 1-2) and SAVMAP (rows 3-6) datasets are shown in the leftmost column. Detections (red) and associated confidence values produced by the various methods are given in subsequent columns. The 2 rightmost columns are shown 2-fold super-resolved by associated methods in accordance with their effective operational resolution. The first 4 rows show examples where only our full approach of combined HAN super-resolution and altitude utilisation allows for correct detection. The positive effect of trivial scale-up is exemplified in row 2. Row 5 depicts a case where altitude information is critical to focus the detector on expected animal sizes; the addition of super-resolution then solves the detection problem fully. Finally, row 6 shows a common case where off-the-shelf YOLOv3 baseline is adequate. However, note that even this case shows an improvement in detector confidence for our full approach found consistently across all examples depicted.

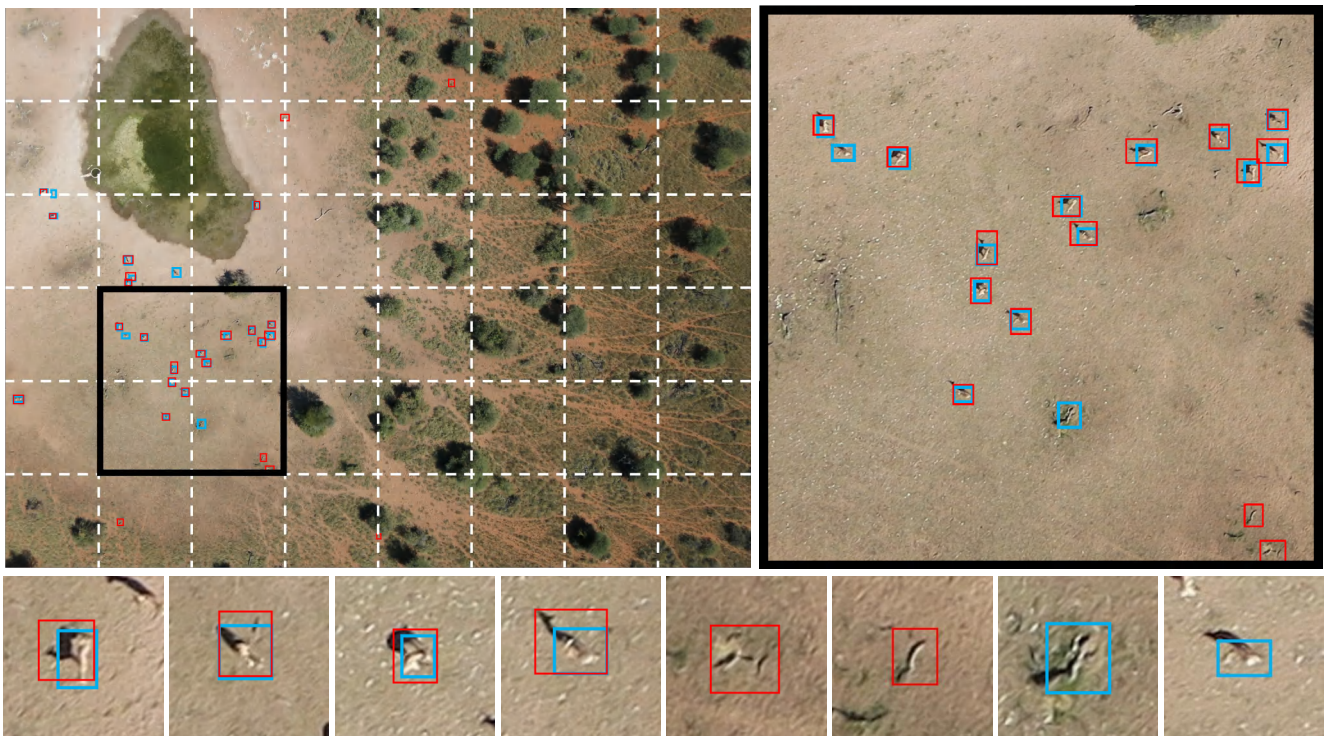


Figure 7. **Application and Limitations - Complete SAVMAP Frame Example.** ***(top left)*** Visualisation of ground truth annotations (blue) and detections by our proposed approach using all components (red) on one full SAVMAP test data frame shown here in original resolution. ***(top right)*** HAN super-resolved and zoomed-in area detail with ground truth and detections covering four selected SAVMAP frame subcells. ***(bottom)*** Further zoom into some of the true positives, false positives and false negatives for best visual appreciation. Note the visual similarity of animals and other structures. Super-resolution and altitude data exploitation can only address these visual ambiguities to some extent given that even manual animal identification is extremely difficult. Therefore, additional sensor information and methodological advances are required to resolve these ambiguities further and improve survey data processing beyond the results shown in this paper.

YOLOv3 detectors off-the-shelf (see rows 2 and 7), confirming efficacy arguments in [64, 37] on our datasets.

Next, we demonstrate that the proposed use of altitude meta-data information (see Sec. 4.2) can consistently benefit detection performance as shown in rows 3 and 8. The addition of domain-specific super-resolution (see Sec. 4.1), as shown in rows 5 and 10 of Table 1, outperforms these approaches and naive bicubic interpolation shown in rows 4 and 9, respectively. Increasing the size of the images fed into the detector from 512×512 to 1024×1024 increases every metric regardless of the applied method. We experimented with further upscaling, but found no significant effect. Essentially though, utilising both deep super-resolution and altitude information demonstrably and consistently outperforms other techniques across both datasets. Fig. 6 provides qualitative examples of this superior detection performance highlighting scenarios in which the proposed concepts succeed in improving detection. Fig. 7 depicts animal detection using our proposed approach on an entire SAVMAP data frame reflecting on its performance and pointing out remaining challenges. In order to investigate the efficacy of super-resolution on input image sizes further and quantify the limits of the approach, we downsampled the test portions of the SAVMAP and AED datasets across factors $\times 2$, $\times 4$ and $\times 8$ via bicubic interpolation to systematically simulate acquisition at even lower and lower animal resolution. We then reconstructed the original size via multi-scale domain-specific super-resolution. Results are presented in Table 2. The benchmarks demonstrate that reconstruction via our suggested approach can maintain detection performance in low resolution scenarios best across all settings tested.

7. Conclusion

Aerial animal surveillance is an essential tool to study biodiversity and protect animal populations – which constitutes an ethical imperative. Here, we addressed the problem of tiny animal resolutions for the first time explicitly: we combined HAN super-resolution with altitude data exploitation and showed that the integration of these components into standard recognition pipelines can systematically increase the detection efficacy on real-world animal datasets. We conclude that the techniques investigated are useful tools for aerial census and conservation automation.

References

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

[2] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[3] Jayme Barbedo, Luciano Koenigkan, Thiago Santos, and Patrícia Santos. A study on the detection of cattle in uav images using deep learning. 12 2019.

[4] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Pietro Lió, Marco Gori, and Stefano Melacci. Entropy-based logic explanations of neural networks. *CoRR*, abs/2106.06804, 2021.

[5] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context r-cnn: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In Yves Lechevallier and Gilbert Saporta, editors, *International Conference on Computational Statistics (COMPSTAT)*, pages 177–186, Heidelberg, 2010. Physica-Verlag HD.

[7] Philippe Bouché, Iain Douglas Hamilton, George Wittemyer, Aimé J Nianogo, Jean Louis Doucet, Philippe Lejeune, and Cédric Vermeulen. Will elephants soon disappear from west african savannahs? *PloS one*, 6:e20619, 06 2011.

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

[9] Gerardo Ceballos, Paul R. Ehrlich, Anthony D.Barnosky, Andrés García, Robert M.Pringle, and Todd M.Palmer. Accelerated modern human-induced species losses: Entering the sixth mass extinction. *Science Advances*, 1:e1400253, 06 2015.

[10] Luc Courtrai, Minh-Tan Pham, and Sébastien Lefèvre. Small object detection in remote sensing images based on super-resolution with auxiliary generative adversarial networks. *Remote Sensing*, 12(19), 2020.

[11] Anne-Sophie Crunchant, Monika Egerer, Alexander Loos, Tilo Burghardt, Klaus Zuberbühler, Katherine Corogenes, Vera Leinert, Lars Kulik, and Hjalmar S. Kühl. Automated face detection for occurrence and occupancy estimation in chimpanzees. *American Journal of Primatology*, 79(3):e22627, 2017.

[12] Dengxin Dai, Yujian Wang, Yuhua Chen, and Luc Van Gool. Is image super-resolution helpful for other vision tasks? In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. 01 2014.

[14] Isla Duporge, Olga Isupova, Steven Reece, David W. Macdonald, and Tiejun Wang. Using very-high-resolution satellite imagery and deep learning to detect and count african elephants in heterogeneous landscapes. *Remote Sensing in Ecology and Conservation*, 12 2020.

[15] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

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

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 06 2014.

[19] 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, 02 2018.

[20] 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. *CoRR*, abs/1704.04861, 2017.

[21] Yanting Hu, Jie Li, Yuanfei Huang, and Xinbo Gao. Channel-wise and spatial feature modulation network for single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, PP:1–1, 05 2019.

[22] Benjamin. Kellenberger. *Interactive machine vision for wildlife conservation*. PhD thesis, Wageningen University, 2020.

[23] Benjamin Kellenberger, Diego Marcos, Sylvain Lobry, and Devis Tuia. Half a percent of labels is enough: Efficient animal detection in uav imagery using deep cnns and active learning. *IEEE Transactions on Geoscience and Remote Sensing*, PP:1–10, 08 2019.

[24] Benjamin Kellenberger, Diego Marcos, and Devis Tuia. Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning. *CoRR*, abs/1806.11368, 2018.

[25] Benjamin Kellenberger, Diego Marcos, and Devis Tuia. When a few clicks make all the difference: Improving weakly-supervised wildlife detection in uav images. pages 1414–1422, 06 2019.

[26] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[27] Jun-Hyuk Kim, Jun-Ho Choi, Manri Cheon, and Jong-Seok Lee. Ram: Residual attention module for single image super-resolution. *ArXiv*, abs/1811.12043, 2018.

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[29] Hjalmar S Kühl and Tilo Burghardt. Animal biometrics: quantifying and detecting phenotypic appearance. *Trends in ecology and evolution*, 28 7:432–41, 2013.

[30] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[31] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[32] Jeong-Seon Lim, Marcella Astrid, Hyun-Jin Yoon, and Seung-Ik Lee. Small object detection using context and attention. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 181–186, 2021.

[33] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[34] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 07 2018.

[35] Julie Linchant, Jonathan Lisein, Jean Semeki, Philippe Lejeune, and Cédric Vermeulen. Are unmanned aircraft systems (uas) the future of wildlife monitoring? a review of accomplishments and challenges. *Mammal Review*, 45, 09 2015.

[36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2016.

[37] Yang Liu, Peng Sun, Nickolas Wergeles, and Yi Shang. A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications*, 172:114602, 2021.

[38] Payal Mittal, Raman Singh, and Akashdeep Sharma. Deep learning-based object detection in low-altitude uav datasets: A survey. *Image and Vision Computing*, 104:104046, 2020.

[39] Johannes Naude and Deon Joubert. The aerial elephant dataset: A new public benchmark for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[40] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 191–207, 2020.

[41] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Ali Swanson, Craig Packer, and Jeff Clune. Automatically identifying wild animals in camera trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115, 03 2017.

[42] Ferda Ofli, Patrick Meier, Muhammad Imran, Carlos Castillo, Devis Tuia, Nicolas Rey, Julien Briant, Pauline Millet, Friedrich Reinhard, Matthew Parkan, and Stéphane Joost. Combining human computing and machine learning to make sense of big (aerial) data for disaster response. *Big data*, 4, 03 2016.

[43] Cathleen O'Grady. The price of protecting rhinos - conservation has become a war, and park rangers and poachers are the soldiers, 2020.

[44] Jiangmiao Pang, Cong Li, Jianping Shi, Zhihai Xu, and Huajun Feng. R2-CNN: Fast tiny object detection in large-scale remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57:5512–5524, 08 2019.

[45] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multimodal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7077–7087, June 2021.

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

[47] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[48] Friedrich Reinhard, Matthew Parkan, Timothée Produit, Sonja Betschart, Beatrice Bacchilega, Morgan L. Hauptfleisch, Patrick Meier, Consortium SAVMAP, and Stéphane Joost. Near real-time ultrahigh-resolution imaging from unmanned aerial vehicles for sustainable land use management and biodiversity conservation in semi-arid savanna under regional and global change (savmap). 3 2015.

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[50] Nicolas Rey, Michele Volpi, Stéphane Joost, and Devis Tuia. Detecting animals in african savanna with uavs and the crowds. *Remote Sensing of Environment*, 200:341–351, 10 2017.

[51] Andrew Rylance, Susan Snyman, and Anna Spenceley. The contribution of tourism revenue to financing protected area management in southern africa. *Tourism Review International*, 21:139–149, 07 2017.

[52] Mehdi S. M. Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through

automated texture synthesis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

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

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

[55] Wen Shao, Rei Kawakami, Ryota Yoshihashi, Shaodi You, Hidemichi Kawase, and Takeshi Naemura. Cattle detection and counting in uav images based on convolutional neural networks. *International Journal of Remote Sensing*, 41(1):31–52, 2020.

[56] Jacob Shermeyer and Adam Van Etten. The effects of super-resolution on object detection performance in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[57] Yu Shiu, Peter H. Wrege, Sara Keen, and Elizabeth D Rowland. Large-scale automatic acoustic monitoring of african forest elephants' calls in the terrestrial acoustic recordings. *The Journal of the Acoustical Society of America*, 135:2334, 04 2014.

[58] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[59] Robert Simpson, Kevin R. Page, and David De Roure. Zooniverse: Observing the world's largest citizen science platform. In *Proceedings of the 23rd International Conference on World Wide Web*, page 1049–1054, 2014.

[60] Ketil Skogen, Håvard Helland, and Bjørn Kaltenborn. Concern about climate change, biodiversity loss, habitat degradation and landscape change: Embedded in different packages of environmental concern? *Journal for Nature Conservation*, 44, 06 2018.

[61] Timothy J. Smyser, Richard J. Guenzel, Christopher N. Jacques, and Edward O. Garton. Double-observer evaluation of pronghorn aerial line-transect surveys. *Wildlife Research*, 43:474–481, 10 2016.

[62] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[63] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[64] Colin J. Torney, David J. Lloyd-Jones, Mark Chevallier, David C. Moyer, Honori T. Maliti, Machoke Mwita, Edward M. Kohi, and Grant C. Hopcraft. A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods in Ecology and Evolution*, 10, 03 2019.

[65] Cédric Vermeulen, Philippe Lejeune, Jonathan Lisein, Prosper Sawadogo, and Philippe Bouché. Unmanned aerial survey of elephants. *PLOS ONE*, 8:1–7, 02 2013.

[66] Dongliang Wang, Quanqin Shao, and Huanyin Yue. Surveying wild animals from satellites, manned aircraft and unmanned aerial systems (uass): A review. *Remote Sensing*, 11:1308, 06 2019.

[67] Marco Willi, Ross Pitman, Anabelle Cardoso, Christina Locke, Alexandra Swanson, Amy Boyer, Marten Veldthuis, and Lucy Fortson. Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, 10, 10 2018.

[68] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[69] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *CoRR*, abs/2106.09018, 2021.

[70] Xinyu Yang, Majid Mirmehdi, and Tilo Burghardt. Great ape detection in challenging jungle camera trap footage via attention-based spatial and temporal feature blending. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 255–262, 10 2019.

[71] Xuehui Yu, Zhenjun Han, Yuqi Gong, Nan Jan, Jian Zhao, Qixiang Ye, Jie Chen, Yuan Feng, Bin Zhang, Xiaodi Wang, Ying Xin, Jingwei Liu, Mingyuan Mao, Sheng Xu, Baochang Zhang, Shumin Han, Cheng Gao, Wei Tang, Lizuo Jin, Mingbo Hong, Yuchao Yang, Huan Luo, Qijun Zhao, and Humphrey Shi. The 1st tiny object detection challenge: Methods and results. *CoRR*, abs/2009.07506, 2020.

[72] Lei Zhang, Helen Gray, Xujiong Ye, Lisa Collins, and Nigel Allinson. Automatic individual pig detection and tracking in pig farms. *Sensors*, 19(5), 2019.

[73] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.