When a few clicks make all the difference: Improving weakly-supervised wildlife detection in UAV images

Benjamin Kellenberger, Diego Marcos, Devis Tuia
Wageningen University, the Netherlands
name.surname@wur.nl

Abstract

Automated object detectors on Unmanned Aerial Vehicles (UAVs) are increasingly employed for a wide range of tasks. However, to be accurate in their specific task they need expensive ground truth in the form of bounding boxes or positional information. Weakly-Supervised Object Detection (WSOD) overcomes this hindrance by localizing objects with only image-level labels that are faster and cheaper to obtain, but is not on par with fully-supervised models in terms of performance. In this study we propose to combine both approaches in a model that is principally apt for WSOD, but receives full position ground truth for a small number of images. Experiments show that with just 1% of densely annotated images, but simple image-level counts as remaining ground truth, we effectively match the performance of fully-supervised models on a challenging dataset with scarcely occurring wildlife on UAV images from the African savanna. As a result, with a very limited amount of precise annotations our model can be trained with ground truth that is orders of magnitude cheaper and faster to obtain while still providing the same detection performance.

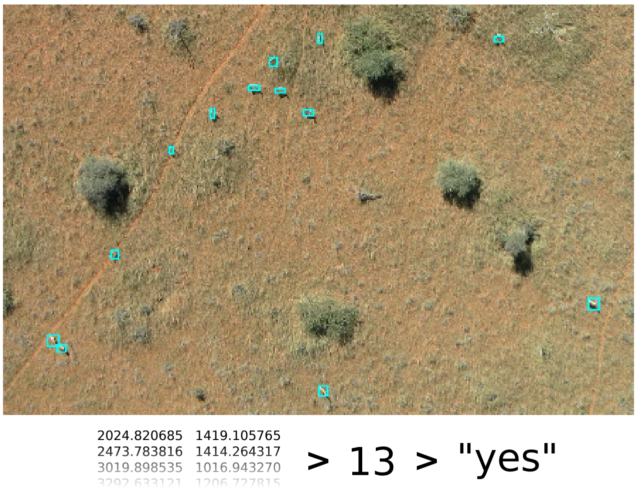

Figure 1: Large-scale UAV datasets pose substantial labeling efforts if bounding boxes or object coordinates (bottom left) are required. Our model reduces this tedium by resorting to inexpensive, image-level object counts (bottom middle) or simple object presence/absence (bottom right) for the most part, and requiring positions only for a handful of images.

1. Introduction

Object detection in images from Unmanned Aerial Vehicles (UAVs) recently experienced an uprising interest in the computer vision community [6, 7, 3, 18, 2]. Applications are manifold and range from security and safety [12] to animal conservation [1, 6]. Thanks to research advancements like Convolutional Neural Networks (CNNs) [16], automated detectors have shown significant increase in positional and classification accuracy of objects [24, 10, 22, 23].

Traditionally, object detectors are fully-supervised, which implies that the class, location in the image and dimensions of every object of interest is known during training. This high level of supervision leads to superior model performances, but comes at a substantial annotation cost: it requires precise bounding boxes to be drawn, and class labels to be given, to many objects in hundreds or thousands of images (as in Figure 1). This becomes especially tedious for datasets that contain large numbers of objects per image, such as crowd surveillance imagery [33], as well as cases where objects of interest are a rare sight. UAV images bear particularly high implications for ground truth cost and quality in this respect [18].

A promising direction to ease label complexity is Weakly-Supervised Object Detection (WSOD), where a model is only given image-level labels, but aims at localizing objects in the image nonetheless [5]. This concept is also related to density estimation [33] in that in both cases no bounding boxes or object positions are provided as a ground truth. Such models require much simpler and there-

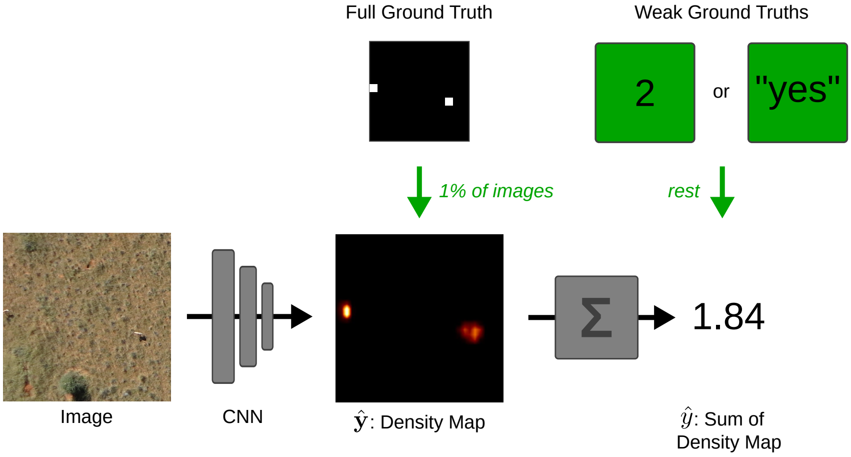

Figure 2: Overview of our proposed model. The detector predicts object locations via a heatmap and receives weak supervision, either through object counts or presence/absence of objects as a ground truth. In a small fraction of cases (*e.g.* 1%), we train on the full, positional ground truth to improve performances.

fore cheaper annotations, such as image-level labels in the WSOD case, and counts (number of class instance occurrences) in density estimation. Consider Figure 1: drawing bounding boxes for all objects in the UAV images would be tedious and hence expensive, especially when thousands of such images need to be labeled. In turn, providing object count estimates or even just the presence or absence of objects for the entire images can be done with just one interaction per image. The downside to WSOD models is typically a loss in accuracy: such models tend to miss small objects [19] or combine multiple close-by instances of the same class into one [8]. Count-based models need to be trained with a sufficient number and variation of objects per image, sometimes including images that contain no object at all, and they risk to focus on the wrong kind of object if two classes appear in similar numbers in the imagery, a scenario not unlikely in UAV images. For example, one may be interested in the number of cars in a top-down view image, but the detector could get confused due to similar numbers of motorcycles.

In this paper we propose to overcome these issues by combining the merits of both fully- and weakly-supervised detectors in a hybrid approach, as shown in Figure 2. In detail, we train a WSOD model with weak supervision (object count or binary presence/absence of objects) and then complement it with a small fraction of images where a full, positional ground truth is available. The intuition behind this is that few precise ground truth maps are comparably inexpensive to obtain, but sufficient for the model to focus on the object class of interest, so that predictions become more precise and ambiguities are reduced. We evaluate our models on a challenging UAV-derived dataset and show that, even with completely weak supervision, object localization is possible to a high accuracy, but that adding just 1% of full positional ground truth can match the performance of a model that has profited from 100% full supervision.

The rest of this paper is organized as follows: Section 2 addresses related work on the main topics; Section 3 explains the method in detail. Experiments are outlined in Section 4, results of which are discussed in Section 5. Finally, Section 6 concludes the work and outlines potential future research on the topic.

2. Related Work

2.1. Object counting

Object counting models recently gained a lot of attraction in the computer vision community [17, 13, 31, 34]. They generally fall into one of three categories: *detection-based* models [24, 22, 23, 9] first localize objects of a kind in an image and then simply return the number of detection. These models provide explicit localizations, but thus require large numbers of expensive bounding boxes. Also, they can be strongly affected by cases of occlusions and pose variations, since each object instance is handled as a binary contribution to the total count. *Regression-based* models [28, 27] directly predict an estimated number of objects from the image. They forgo expensive ground truth, but may struggle whenever they have to do extrapolation (*i.e.*, predicting more or less objects than seen in the training set) and cannot easily provide object positions. Finally, *density-based* models [13, 35] predict a spatial heatmap of localizations whose sum corresponds to the expected number of objects. These models basically combine the advantages of both detection- and regression-based approaches in that they only need count estimations but still yield a spatial prediction. If successful, these models can therefore be seen as a variant of WSOD, which is discussed below.

2.2. Weakly-Supervised Object Detection

The aim of WSOD is to localize objects in images with just image-level ground truth. The hope of such models is that they will learn which parts of the image to draw their attention to, and that these parts will then coincide with the location of the objects of interest. Several variants of WSOD models have been proposed: [5] retrieves reoccurring patterns across the dataset by clustering. [20] trains an object detector through active learning, querying an oracle for automatically predicted bounding boxes. [30, 29] both exploit prior knowledge (object sizes, texture compositions of objects) to facilitate identification of objects across images. [19] is closely related to heatmap-based models used in this study, but differs in that the model is only trained with weak supervision, and that the prediction is a grid of pseudo-probabilities that gets softmax-activated across all classes. The authors report good localization performance for vision benchmark datasets, but due to the softmax operator encouraging weight concentration to one sample their model is likely to be unable to localize multiple objects of the same category in one image, a crucial necessity for counting. [8] attempts to solve this drawback by disentangling multiple objects detected together, based on a count-based ground truth, but their method still relies on explicit detectors that may be hard to train and may fail under occlusions and on small targets.

2.3. Hybrid Approaches

Some works studied combining different levels of supervision, but typically focus on performance improvements, rather than on reducing annotation efforts. [12] employs both positions and counts for crowd density estimation, but needed a dataset with complete positional information for over 1.5 Million targets. [17] proposes a model that uses detection for low estimated object counts and switches to regression for crowded scenes. [21] detects objects with image-level labels and single point annotations; [4] extends this idea to semantic image segmentation. However, both approaches require all of the training images to be labeled this way. [34] is more similar to our model in that the authors also alternate between weak supervision (density) and object position ground truth for training, but also their scheme required all training set instances to be fully labeled.

3. Method

3.1. Density-based Object Detection

As a base model for all experiments we employ a deep fully-convolutional CNN that accepts an image and predicts a downsampled grid $\hat{\mathbf{y}}$ of size $N \times M$. This grid corresponds to the density map and is expected to contain high values in those locations where an instance of our object class of interest is to be found, and values close to zero elsewhere. A non-negative activation nonlinearity, such as a ReLU or sigmoid, ensures that predictions are bound to a range of zero or more. Let scalar $\hat{y} = \sum_{i \sim N} \sum_{j \sim N} (\mathbf{y}_{ij})$ be the sum over the output grid $\hat{\mathbf{y}}$ for each image. Our objective then is it to train the model to predict a grid whose sum $\hat{y}$ corresponds to the ground truth. In this paper we consider two levels of weak supervision: *(i.)* a scalar for the object count; *(ii.)* a binary variable denoting presence/absence of the object class of interest in the image. The latter is an extreme case that we use to see if a CNN can localize objects where the only knowledge we have is whether the class of interest is present in the image, eliminating the need for the annotator to count the objects to create the ground truth and thus greatly reducing the annotation effort.

To this end we employ different loss functions, depending on the level of weak supervision. In cases where the ground truth consists of an object count, we employ a Smooth $\ell 1$, also known as Huber loss [9]:

$$\ell(y, \hat{y}) = \ell_{Huber}(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{if } |y - \hat{y}| < \delta \\ |y - \hat{y}| - 0.5\delta & \text{otherwise} \end{cases} \tag{1}$$

where y is the object count in the image. Unlike an $\ell 2$ loss, the Huber loss employs an $\ell 1$ norm beyond a threshold δ (set to 1 in our experiments), which poses less of a restriction on outliers in the data. This is also useful for balancing loss strengths between samples with low and high counts, which is the case in our UAV images (wildlife tends to flock together, causing some images to have high counts and others to contain no animal at all).

In the binary case, we only have two possible labels: 0 (absence) and 1 (presence). We therefore use a conditional loss instead:

$$\ell(y, \hat{y}) = \begin{cases} 0 & \text{if } y = 1 \text{ and } \hat{y} \geq 1 \\ \ell_{Huber}(y, \hat{y}) & \text{otherwise} \end{cases} \tag{2}$$

In other words, the loss here pushes the sum over the predicted heatmap to zero whenever no object is present ($y = 0$), but is zero if the respective area is labeled as "has object" ($y = 1$) and the sum over the heatmap is at least one.

Note that Equations (1) and (2) are limited to a single object class, as is the case in our experiments, but an extension to the multi-class scenario would be straight-forward by predicting K heatmaps for K classes and individually applying the respective loss function.

3.2. Sparse Full Supervision

Our proposed method follows the same principle and employs the same loss functions for the most part. For the weakly labeled images the training loss therefore corresponds to either Equation (1) or (2), depending on whether a count- or presence/absence-based ground truth is available. However, for a small percentage of images we assume the availability of precise, positional ground truth for all objects they contain (left ground truth in Figure 2). For these images we replace the weakly-supervised loss with a spatially explicit binary cross-entropy loss:

$$\ell(\mathbf{y},\hat{\mathbf{y}}) = \sum_{i\sim N}\sum_{j\sim M} -\left(y_{ij}\log(\hat{y}_{ij}) + (1-y_{ij})\log(1-\hat{y}_{ij})\right) \tag{3}$$

where $\mathbf{y}$ and $\hat{\mathbf{y}} \in \mathcal{R}^{N\times M}$ are the ground truth and prediction (density) maps, respectively. $\mathbf{y}$ contains value one wherever an object is present, and zero elsewhere.

The primary intuition behind this is to spatially focus the model on the kind of object we would like to detect. While pure WSOD models tend to localize objects fairly well, they struggle estimating the objects' dimensions [19]. This is because the only limitation imposed on the model through the loss is the heatmap sum constraint. Providing a few fully-supervised cases, however, encourages the model to concentrate more mass on just one prediction grid cell, instead of spreading it out to the neighbors. This not only reduces spurious, shallow detections of undesired objects, but also helps disentangling two instances next to each other that otherwise would likely be combined.

Furthermore, in cases where multiple kinds of objects appear in similar quantities, it helps reducing chances of the model detecting the wrong object class. This may particularly happen with presence/absence ground truths that naturally allow for high ambiguities.

3.3. Evaluating Positional Accuracies

Most fully-supervised object detectors predict bounding boxes, which can easily be evaluated through the intersection-over-union with the nearest ground truth box. In our case this does not apply, as we only predict heatmaps. To assess the quality of the positional detection, we thus resort to a distance-based evaluation: we first threshold the heatmap at a given value and subject all remaining locations to a Non-Maximum Suppression (NMS) stage to filter multiple hits of the same object. The assessment then follows the principle of [14], which calculates the Euclidean distance of the respective prediction cell coordinate to the nearest ground truth location, and only accepts it as a true positive if it is below a maximum distance and if the ground truth object has not been detected by another grid cell (in which case it counts as a false positive).

This raises the question for the maximum distance threshold to be used. Especially in WSOD, we cannot expect the model to always perfectly predict the center position of an object. However, depending on the task, a pixel-precise localization may not be necessary after all, such as in our case of animal census. Consider Figure 3, where circles of different distances are shown. The smallest circle (blue) only encompasses just a part of the animal; this poses a severe restriction, as detectors are required to perfectly pinpoint the objects all the time. However, we would already be satisfied if the model detected *e.g.* the animal shadow, which lies outside the blue circle. Such a detection is thus only counted as a hit if we increase the distance threshold. Too large circles (brown, pink) in turn bear the risk of including false predictions outside of the animal. In essence, the positional tolerance is a function of the objective and the object size (in pixels). The effect of positional tolerance will be studied in the experiments.

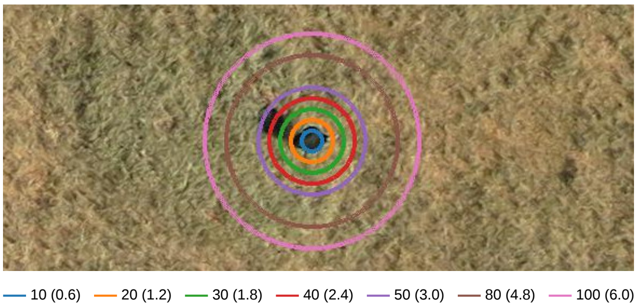

Figure 3: Concentric circles with varying radii over an animal in pixels. Equivalence in meters is reported in parentheses. Too small evaluation distances (blue) require models to precisely pinpoint the animal; too large (pink) risk including neighboring predictions. If we are also satisfied when the model predicts *e.g.* the animals' shadows, a radius in-between may be a viable choice.

4. Experiments

We now put the models with varying type and degree of supervision to the test. All in all, we compare models trained with the following levels of ground truth:

- Binary (baseline A): we regress the density map sum with the presence/absence of objects (Eq. (2)).
- Counts (baseline B): here, we provide the number of objects per image patch as a ground truth and use Eq. (1) as a loss.
- Binary + 1% (proposed A): this scenario adds full positional ground truth for 1% of the images (three images in our case) through Eq. (3), but presence/absence for the rest.
- Counts + 1% (proposed B): the same, but with counts as ground truth for the majority of images.
- Full dense (upper bound): we train the model with a complete set of positional ground truth (only Eq. (3) is used), and no counting or binary loss.

Set	# images with animals	 without	 total	# animals
train	159	291	450	830
val	35	41	76	118
test	45	83	128	235

Table 1: Properties of the Kuzikus UAV dataset.

4.1. The Kuzikus Dataset

For evaluation we resort to a set of UAV images acquired over the Kuzikus game reserve in Namibia[1]. Kuzikus is a private-owned wildlife park that stretches across 103km^2 and is home to around 3000 individuals of multiple large mammal species, such as black rhino, ostriches, zebras, and various ungulates [25, 14]. In 2014, the SAVMAP consortium[2] imaged parts of the park with a SenseFly eBee[3] UAV, equipped with a consumer-grade camera at a flying altitude of around 100m above ground. This resulted in 654 nadir RGB images with a resolution of 4-8cm. An initial localization of animals was obtained through a crowdsourcing campaign conducted by MicroMappers[4]; upon refinement of the labels, 1183 animals could be identified in the 654 images. This makes them a comparably rare sight and required manually examining every image for small targets. The dataset statistics are listed in Table 1; images are available at https://doi.org/10.5281/zenodo.609023.

4.2. Model Setup and Training

Our model for all experiments is based on a ResNet-18 [11], pre-trained on ImageNet [26], until the third-last layer: *i.e.*, we remove the original classification and average pooling layers and instead add two Multi-Layer Perceptrons (MLPs) that map from 512 to 1024 dimensions, and then to one (the heatmap), respectively. We replace the original BatchNorm layers with non-affine Instance Normalization variants [32], since we observed better stability for variable batch sizes during inference. Also, we reduce the first layer's stride to one to allow for a higher-resolution density map prediction. The density map output is scaled to $[0, 1]$ through a sigmoid activation.

To train the model we crop 16 random patches per epoch of size 512×512 from each UAV image in the training set and perform data augmentation by random flipping (both axes) and random 90° rotations. The model then predicts a density map of 32×32, which corresponds to a downsampling factor of 16. With our dataset's image resolution of around 4-8cm, one grid cell has a side length of roughly 0.6-1.3m, which still allows detecting animals below the average size of around 1.5m. This map gets summed and compared to the ground truth according to either loss function described above, depending on the degree of supervision (counts versus presence/absence of objects). We use mini-batches of four patches and employ the Adam optimizer [15] with a learning rate of 10^{-5} for the first 50, and 10^{-6} for the remaining 250 epochs. We also enable weight decay of 10^{-4}.

During inference we tile the images evenly into same-sized patches, evaluate them individually and stitch the predictions back together to one density map. For detection we filter the predicted heatmap and retain locations with value 0.01 or greater. These are subjected to NMS with a search radius of 2 prediction grid cells (32 pixels; about 1.92m). Evaluation is then performed on a distance basis as in [14], with thresholds of 30 pixels (ca. 1.8m) and 40 pixels (2.4m), respectively.

For the fully labeled images used by our proposed models we randomly select three images from the pool of images that contain at least one animal. The eventually selected images contain 1, 38, and 7 animals, respectively.

Model	MAE	MSE
Binary (baseline A)	1.96	22.29
Counts (baseline B)	**0.79**	1.63
Binary + 1% (proposed A)	1.22	4.99
Counts + 1% (proposed B)	**0.79**	**1.59**
Full dense	1.28	2.75

Table 2: MAE and MSE values for the models on the sum of the density maps.

5. Results and Discussion

Table 2 lists Mean Absolute Error (MAE) and Mean Squared Error (MSE) values on the heatmap sums over the test set. As can be seen, the count-based model with 1% full supervision yields the best score for both measures and is on a tie with the fully WSOD counts model in terms of MAE. Generally, it is not surprising that the count-based models outperform the others, even the model with full positional supervision, since they are the only ones explicitly trained on the objective of these measures, *i.e.* counting. The large MSE, but still reasonable MAE scores of the binary model imply that it predicts a few images wrong, but by a large difference (see Figure 4 top). Since animals are a minority compared to background, this indicates that the model generally manages to localize them, but overpredicts them by also assigning high confidences to their spatial surroundings. This is not surprising, given that the binary model did not have any constraints other than to predict at least a mass sum of one whenever an animal is present. Adding just 1% of full supervision significantly improves

[1] http://kuzikus-namibia.de/xe_index.html
[2] https://lasig.epfl.ch/savmap
[3] https://www.sensefly.com
[4] https://micromappers.wordpress.com

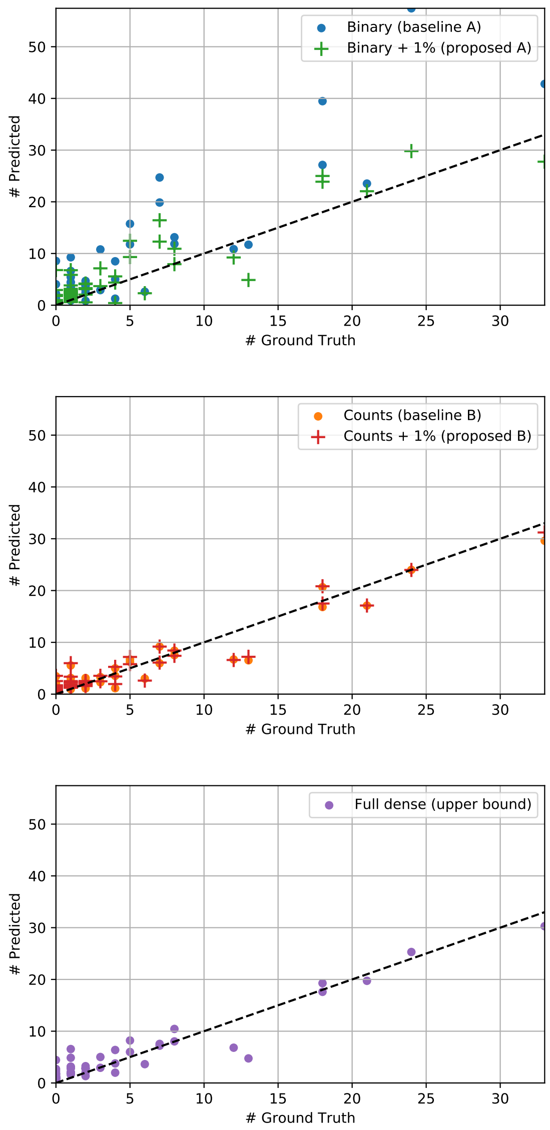

Figure 4: True animal counts versus predicted heatmap sums for the two binary models (top), the two count models (middle) and the fully-supervised model (bottom). In the binary case, 1% of full supervision (green) greatly reduces the overprediction of animal presence. For the count-based models the full supervision does not significantly improve the heatmap sum, but instead raises the precision (see Figure 6). The fully-supervised model's heatmap sum is similarly close to the optimum (dashed line).

the scores, although they do not reach those of the other models. In the case of the count-based models (Figure 4 middle), the WSOD-only model already performs similarly well as the fully-supervised one (Figure 4 bottom). However, here the 1% fully-supervised images have the effect of reducing false positives and hence improve localization performance, which can be seen in Figure 6.

Figure 5 shows example images from the test set (top row), together with heatmaps predicted by all five models. The images further contain the true animals' bounding boxes (cyan). As visible in the heatmaps, the CNNs do indeed manage to predict precise locations of animals, even if trained with weak supervision. The binary model (baseline A) assigns particularly large masses to the animals, which corresponds to the observations made in Table 2 and Figure 4. Also, it is more prone to falsely detecting background objects like tree trunks, such as those in the left part of the left image. However, adding just 1% of positions (proposed A) significantly reduces both effects and dampens predictions around the animals, as well as around background locations. This confirms that a bit of strong supervision is enough to solve both issues (overprediction of animals and false detections of the mentioned background objects).

In the case of the count-based models (baseline B, proposed B) the prediction strength is greatly reduced; activations over animals are smaller and with lower values. This shows that the counts pose more of a restriction to the model than the binary ground truth. However, despite this stronger learning signal also this baseline commits quite a few mistakes in the form of false positives, especially in the left image of Figure 5. Again, adding 1% of full supervision reduces the number of false detections. Visually, the upper bound (bottom row) nonetheless seems to provide the sharpest predictions (least spread in space around animals).

Figure 6 shows results for distance thresholds of 30 pixels (around 1.8m; top) and 40 pixels (ca. 2.4m; bottom), corresponding to the green and red circles in Figure 3, respectively. These distances are still fairly hard, as the models only get a small chance of predicting a true positive when *e.g.* concentrating mass on animal shadows. The curves show that, even under such conditions, all models manage to reach reasonably high recalls of ca. 70 to 90%, but with varying degrees of precision. In general, the binary model (baseline A) falls short the most, which goes in line with it not penalizing the wrong number of animals: despite NMS, it produces many false positives. The semi-supervised counterpart (proposed A) shows significant improvements in this respect, but does not come close to the levels of recall of the other models. Therefore, binary ground truth seems sufficient to get the models into a good direction, but not enough to really make them focus on the target objects. Higher percentages of full supervision might improve results in this respect, but we leave this to further studies.

For the count-based models, we already see a good performance with the WSOD-only model (baseline B), but only when using a more generous distance of 40 pixels. It seems that the counts-only model focuses more on shad-

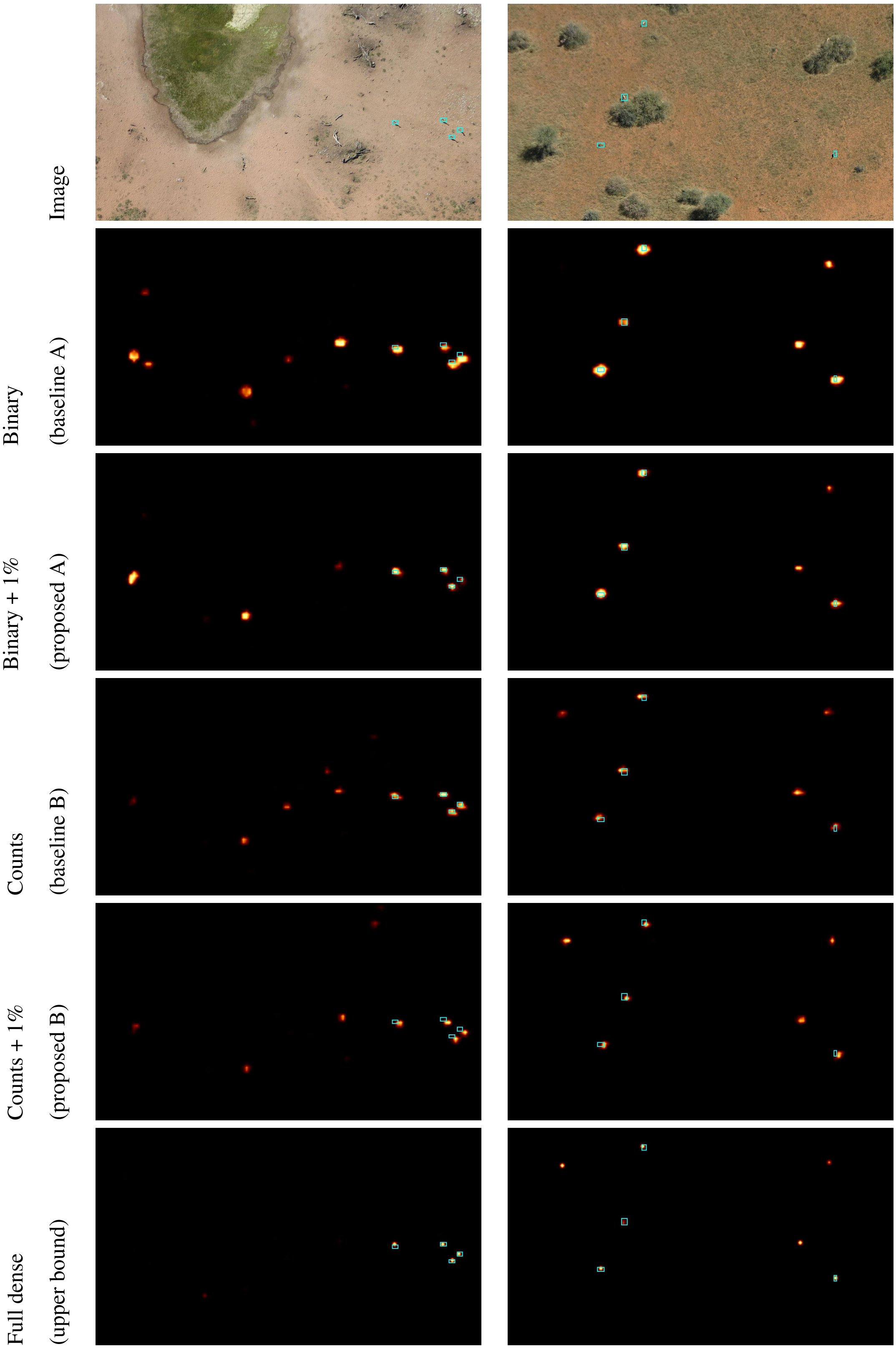

Figure 5: Crops of images from the Kuzikus dataset (top), along with predicted heatmaps for all the models.

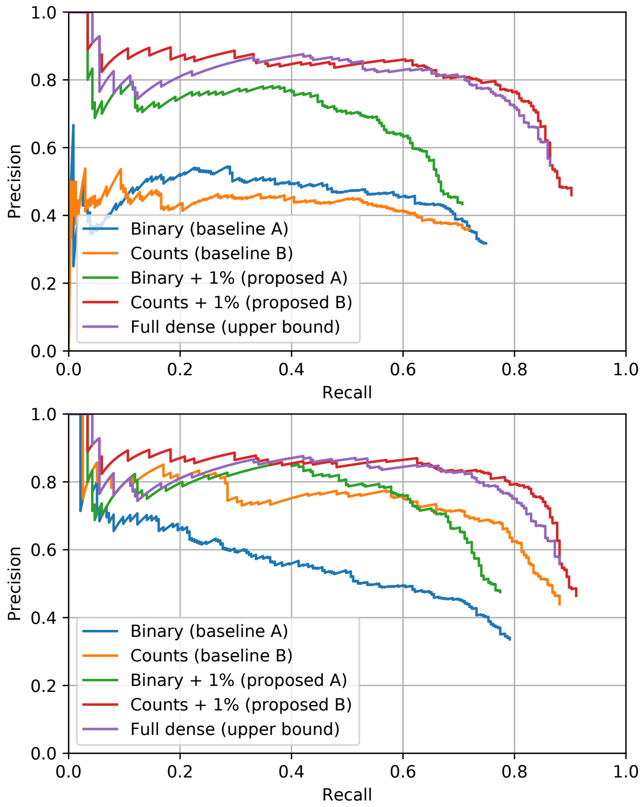

Figure 6: Precision-recall curves for the models on the Kuzikus test set, with a maximum ground truth distance of 30 pixels (1.8m; top) and 40 pixels (2.4m; bottom).

ows than on the animals, which corresponds to the slight spatial shifts of the heatmap peaks observable in Figure 5. A threshold of 30 pixels is insufficient to account for such misplacements. However, we get a substantial improvement in the model with counts and 1% of positional ground truth (proposed B); in this case the model surprisingly is on par with the one trained on 100% positions (upper bound). This is even the case with 30 pixels tolerance, indicating that a few labels of strong supervision are enough to steer the model towards detecting the actual animal itself, rather than its shadow. It may seem surprising that this model actually outperforms the upper bound for recalls of around 80-90%. Our intuition on this phenomenon is that a full set of positions may actually be counterproductive, possibly due to limitations from the prediction grid: the heatmap of the CNN is purposely downgraded in resolution, which may cause problems if the ground truth position lies *e.g.* on the border between two heatmap grid cells. In those cases assigning all mass to one of the two cells may be suboptimal for the model, especially when this happens frequently. Hence, the proposed model sees just enough positional ground truth to learn what to focus on, but not too much to cause problems due to spatial precision. All in all, the end-effect of this is that instead of having to densely label hundreds of images, weak image-based labels with a tiny fraction of full supervision gives equally good or even slightly better results, meaning that the labeling process becomes easier, faster, and thus less expensive, allowing for more or larger UAV datasets to be processed in a given time.

6. Conclusion

In this work we demonstrated how to reduce labeling efforts for training object detectors in Unmanned Aerial Vehicle (UAV) images. The proposed model resorts to Weakly-Supervised Object Detection (WSOD), which makes use of simple image-level labels like counts or presence/absence of objects. When augmented with just 1% of positional ground truth, the count-based variant effectively matches its counterpart trained on object positions in all images. Hence, we believe our strategy to be of major benefit to the labeling of large-scale UAV image datasets. To this end, further work is required on the role of the fully-supervised ground truth, in particular regarding the percentage of fully-supervised images, or else more sophisticated, perhaps model-driven strategies for selecting the images that require a positional ground truth. Also, model performance on objects of varying sizes could be worth investigating. Finally, improvements for the detection performance of the binary model could pave the way to lower the required labeling efforts even further.

References

[1] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *ECCV*, pages 483–498. Springer, 2016.

[2] Nazia Attari, Ferda Ofli, Mohammad Awad, Ji Lucas, and Sanjay Chawla. Nazr-CNN: Object Detection and Fine-Grained Classification in Crowdsourced UAV Images. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2016.

[3] Yakoub Bazi and Farid Melgani. Convolutional SVM Networks for Object Detection in UAV Imagery. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 56(6):3107–3118, 2018.

[4] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. Whats the point: Semantic segmentation with point supervision. In *ECCV*, pages 549–565. Springer, 2016.

[5] Hakan Bilen, Marco Pedersoli, and Tinne Tuytelaars. Weakly Supervised Object Detection with Convex Clustering. In *CVPR*, pages 1081–1089, 2015.

[6] Elizabeth Bondi, Debadeepta Dey, Ashish Kapoor, Jim Piavis, Shital Shah, Fei Fang, Bistra Dilkina, Robert Hannaford, Arvind Iyer, Lucas Joppa, and Milind Tambe. Airsim-w: A simulation environment for wildlife conservation with uavs. In *Proceedings of the 1st ACM Conference*

on Computing and Sustainable Societies (COMPASS). ACM, 2018.

[7] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking. In *ECCV*, pages 370–386, 2018.

[8] Mingfei Gao, Ang Li, Ruichi Yu, Vlad I. Morariu, and Larry S. Davis. C-WSL: Count-Guided Weakly Supervised Localization. In *ECCV*, pages 152–168, 2018.

[9] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[10] Kaiming He and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2961–2969, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778, 2016.

[12] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, and Dong Zhang. Composition loss for counting, density map estimation and localization in dense crowds. In *ECCV*, pages 532–546.

[13] Di Kang, Zheng Ma, and Antoni B. Chan. Beyond Counting: Comparisons of Density Maps for Crowd Analysis Tasks – Counting, Detection, and Tracking. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2018.

[14] Benjamin Kellenberger, Diego Marcos, and Devis Tuia. Detecting Mammals in UAV Images: Best Practices to Address a Substantially Imbalanced Dataset with Deep Learning. *Remote Sensing of Environment*, 216:139–153, 2018.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[16] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[17] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G. Hauptmann. DecideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation. In *CVPR*, pages 5197–5206, 2017.

[18] 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(1):47–59, 2016.

[19] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? – weakly-supervised learning with convolutional neural networks. In *CVPR*, pages 685–694, 2015.

[20] Dim P. Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. We don't need no bounding-boxes: Training object class detectors using only human verification. In *CVPR*, pages 854–863, 2016.

[21] Dim P. Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Training object class detectors with click supervision. In *CVPR*, pages 6374–6383, 2017.

[22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, pages 779–788, 2016.

[23] Joseph Redmon and Ali Farhadi. YOLO9000: Better, Faster, Stronger. In *CVPR*, pages 7263–7271, 2017.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, volume 28, pages 91–99, 2015.

[25] 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, 2017.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015.

[27] David Ryan, Simon Denman, Sridha Sridharan, and Clinton Fookes. An evaluation of crowd counting methods, features and regression models. *CVIU*, 130:1–17, 2015.

[28] Chong Shang, Haizhou Ai, and Bo Bai. End-to-end crowd counting via joint learning local and global count. In *ICIP*, pages 1215–1219, 2016.

[29] Miaojing Shi, Holger Caesar, and Vittorio Ferrari. Weakly supervised object localization using things and stuff transfer. In *ICCV*, pages 3381–3390, 2017.

[30] Miaojing Shi and Vittorio Ferrari. Weakly supervised object localization using size estimates. In *ECCV*, pages 105–121. Springer, 2016.

[31] Vishwanath A. Sindagi and Vishal M. Patel. A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation. *Pattern Recognition Letters*, 107:3–16, 2018.

[32] Dmitry Ulyanov and Andrea Vedaldi. Instance Normalization: The Missing Ingredient for Fast Stylization. 2016.

[33] Li Wang, Weiyuan Shao, Yao Lu, Hao Ye, Jian Pu, and Yingbin Zheng. Crowd Counting with Density Adaption Networks. pages 1–5, 2018.

[34] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene Crowd Counting via Deep Convolutional Neural Networks. In *CVPR*, pages 833–841, 2015.

[35] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016.