O1-1-3

17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

Distant Bird Detection for Safe Drone Flight and Its Dataset

Sanae Fujii Kazutoshi Akita[1] Norimichi Ukita[2]
Toyota Technological Institute, Japan
{sd21501[1], ukita[2]}@toyota-ti.ac.jp

Abstract

For the safe flight of drones, they must avoid the attacks of aggressive birds. These birds move very fast and must be detected far enough away. In recent years, deep learning has made it possible to detect small distant objects in RGB camera images. Since these methods are learning-based, they require a large amount of training images, but there are no publicly-available datasets for bird detection taken from drones. In this work, we propose a new dataset captured by a drone camera. Our dataset consists of 34,467 bird instances in 21,837 images that were captured in various locations and conditions. Our experimental results show that, even with the SOTA detection model, our dataset is sufficiently challenging. We also demonstrated that (1) several standard techniques for improving detection methods (e.g., data augmentation) are inappropriate for our challenging dataset, and (2) carefully-selected techniques can improve the detection performance.

1 Introduction

With the increased expectations to use drones in various fields such as delivery and agriculture, obstacle detection for drone-flight safety has been developed and put into practical use. For example, the Mavic 2 Pro [1] drone uses optical sensors that accurately detect objects at most 20 meters distance all around the drone. This sensor enables the drone to avoid collisions with most static obstacles, such as walls and trees.

Although such high-performance sensors, it is reported that aggressive birds, such as hawks, crash drones in flight [2, 3, 4]. Since these birds can fly at high speeds of 50 meters/second or more, a drone cannot avoid their attacks even if the drone can detect birds 20 meters away. Therefore, it is necessary to detect such birds distant enough to avoid their attacks.

To detect distant objects, sensors such as LiDAR [19] or millimeter-wave radar [20] are standard options. However, these sensors can only point or line scanning. It is difficult for these sensors to distinguish observed birds between aggressive ones and safe ones. Furthermore, these sensors are heavy and expensive. Thus these sensors are not suitable for drones. An RGB camera is another option. In recent years, many object detection methods allow us to find small distant objects even in RGB images. Since these methods are based on deep neural networks, they need a large annotated dataset for training. However, there is no publicly-available dataset for this purpose.

In this work, we propose a new dataset captured by a camera onboard a drone. Our dataset contains 34,467 manually annotated bird instances in 21,837 images with various backgrounds such as sky, forest, rice field, buildings, and so on. These bird instances were captured as close as 10 meters, while others at distances of 200 meters or more.

2 Related Work

2.1 Bird Dataset

Many of the publicly-available bird datasets (e.g., the Caltech-UCSD Birds 200 dataset [21], the 260 bird species dataset [22], and the NABirds dataset [23, 24]) are intended to identify bird species. Many of these datasets show large images of birds, which are not suitable for the purpose of detecting birds in the distance. On the other hand, Yoshihashi et al. [5] developed a dataset for detecting small distant birds with the aim of investigating the situation of bird strikes at wind farms. Sample images are shown in the top row of Fig. 1. In this dataset, a number of very small birds (about 100 × 100 pixels in a 4K image) were captured and annotated (e.g., “crow” and “hawk”). However, since images were taken only by fixedly-installed cameras, the background is not changed. With drones, a variety of backgrounds such as sky, forest, buildings, rice fields, and roads are observed. Birds are captured with those background scenes. Furthermore, motion blur is inevitable in images taken from moving drones. Therefore, this dataset [5] is not suitable for our purpose (i.e., distant bird detection from a drone that can be operated in various scenes).

2.2 Object Detection

Many modern CNN-based object detection algorithms (e.g., Faster R-CNN [7]) scatter anchor boxes in the CNN features, estimate the confidence score of each anchor box as an object, and then classify each box having a high confidence score to each object class. This two-stage detection algorithm has high performance but high computational cost. On the other hand, SSD [26], DSSD [6], and YOLO [27, 28] directly estimate the confidence score of each scattered anchor box’s class. These one-stage detection algorithms are

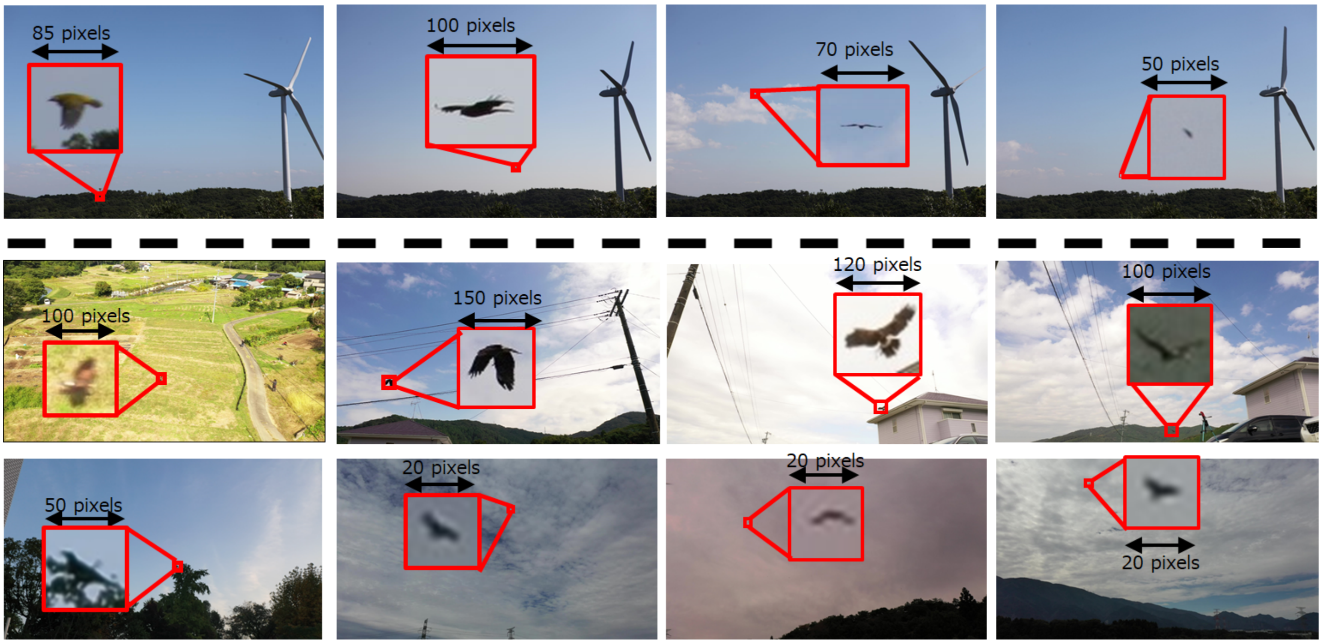


Figure 1: Comparison between a previous tiny-bird detection dataset [5] (above the dashed line) and ours (below the dashed line). Our proposed dataset has rich variations in the background and the posture of the birds.

computationally inexpensive and have potential applications in real-time detection on low-performance machines.

However, these object detection methods scatter anchor boxes densely in the image, resulting in multiple overlapping detections for a single object. Therefore, they use Non-Maximum Suppression (NMS) to remove overlapping results. Since NMS checks overlaps among all detection results, the computational cost increases significantly as the number of detections becomes larger. This is a critical problem for distant object detection because a huge number of tiny miss-detections are unavoidable in general for tiny object detection.

CenterNet [8] achieves object detection by estimating the object's center position and the box's width and height. The probabilities of object positions are represented as a heatmap in each object class. In its detection process, a lightweight 3 × 3 max-pooling is used to extract peaks from the heatmap for object detection. Since this peak extraction has sufficient performance as an alternative to NMS, it has the potential to be a real-time processing method even for detection in high-resolution images.

In this paper, we conduct experiments using CenterNet as one of the SoTAs for efficient object detection, while further improvement in tiny bird detection [9] is achieved with super resolution (SR) (e.g., single image SR [10, 11, 12] and video SR [13, 14]).

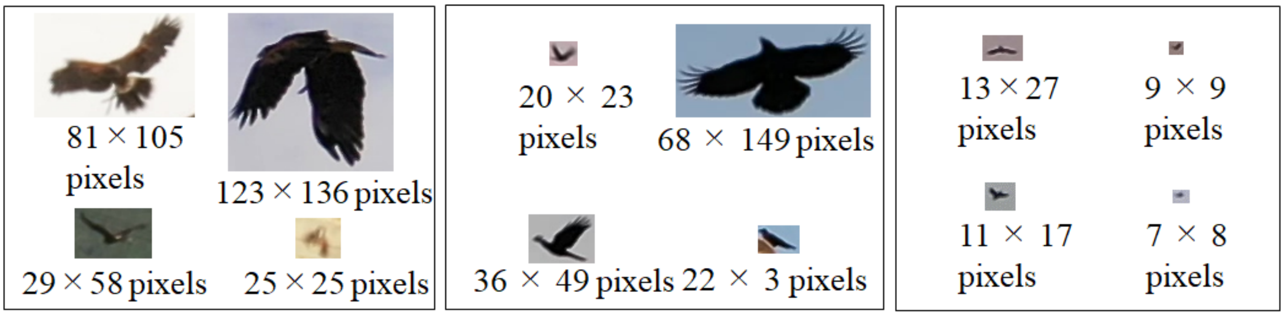
3 Dataset Development

3.1 Image Capturing

Since drones flight in various locations, images with various backgrounds such as in fields, mountains, and near houses are required for our dataset. Birds also need to be captured in various conditions, such as on the ground, in-flight, or downward from a high altitude. However, it is extremely unlikely that wild birds will appear in the images when the drone captures such locations. It is also difficult to capture a large number of bird images only in natural environments. In order to collect such a large number of bird images, we asked a hawker to fly hawks in various conditions. Examples of the collected images are shown in the middle row of Fig. 1. The resolution of these hawk images are relatively higher (between 50 × 50 pixels and 150 × 150 pixels).

Furthermore, several kinds of wild birds were also captured both in city areas and rural areas. Since such wild birds fly in a higher sky, the resolution of these bird images are relatively lower (between 10 × 10 pixels and 50 × 50 pixels). Examples of the images are shown in the bottom of Fig. 1.

In all images mentioned above, birds were captured from a short distance of at least 20 meters and more than 200 meters to ensure variation in size. The images were captured using a camera onboard the Mavic 2 Pro, with a resolution of 3,840 × 2,160 pixels with 30 fps.



Hawk Crow Wild bird

Figure 2: Examples of bird bounding-boxes with various scales in three classes.

3.2 Annotation

In our dataset, we needed to annotate distant birds. With a single frame, it is very difficult for annotators to find a small bird or distinguish a bird from other flying objects such as an insect or airplane. In a video, however, it is relatively easy to detect moving objects, even small ones, and classify them. Therefore, we used VATIC [29], an annotation tool for tracking, to annotate them. This tool allows us to annotate while checking a continuous sequence, making it easy to find small but moving objects and distinguish them between birds and other irregularities based on their movements. Either of the three object labels (i.e., Hawk, Crow, and Wild bird, as shown in Fig. 2) is provided to each bounding box. The interpolation function of VATIC between frames also makes it easy to obtain annotation information for the frames in between by annotating only two frames in a continuous sequence. This makes it possible to prepare a large dataset in a short time. Four people in total did the annotation, and the data was distributed so that the same annotator annotated scenes of similar genres to prevent negative effects caused by the bias of human annotation tendencies. Despite these measures, manual annotations frequently fail. Therefore, we double-checked all sequences to check for mistakes.

In total, our dataset has 21,837 images. These images are divided into 16,754 training images and 5,083 validation images. The number of instances (i.e., bird bounding-boxes) is 34,367. In the training and validation images, 24,919 and 9,548 instances are included.

4 Experiments

We conducted experiments with our dataset. As mentioned in Sec. 2.2, Centernet [8] is employed for the baseline. While the original Centernet is proposed with several variants of ResNet [18] (e.g., ResNet18 and ResNet101 [17]), the one employed in our experiments consists of the Hourglass network [16] for improving the performance of tiny object detection.

4.1 Training

In our dataset, many birds are in complex backgrounds that make detection challenging, such as forests and rice fields. Such hard-examples lead to training unstable. On the other hand, in the Yoshihashi dataset [5], most birds are in the sky background and easy to be trained. Therefore, Centernet was pretrained with the Yoshihashi dataset. For this pretraining, we used 10,081 images. After the pre-training, CenterNet was fine-tuned with our proposed dataset.

We used the Adam optimizer with $\beta = (0.9, 0.999)$ for both pre-training and fine-tuning. The learning rate is 2.5e-4. The mini-batch size is 32. Since the image is huge, it is divided to patch images each of which is 512 × 512 pixels due to memory constraints. Each patch image is fed into Centernet both in training and inference. Patch images are randomly cropped for data augmentation as described below.

Data augmentation: Random flip, random scaling, random cropping, and color jittering are standard data augmentation techniques. However, color jittering might lose the important appearance cues of tiny objects by disarranging their silhouettes. Conversely, smooth textures in negative regions (e.g., sky and clouds) might be changed to be similar to the silhouette of a tiny bird. In our dataset, many birds are captured tinily as shown in Fig. 2. Indeed, we empirically found that color jittering degrades the detection performance of Centernet.

Iterative hard-negative training: Since most regions in images are the sky in the Yoshihashi dataset and our proposed dataset, if we perform random cropping uniformly on these images, birds as positive samples and hard-negative samples that are likely to be misclassified to birds (e.g., leaves in forests and edges of turbines) are not trained enough. Therefore, we crop patch images randomly but with a high sampling rate in the area of these regions.

Since the bounding boxes of birds are known in the training dataset, patch images are sampled so that they contain the bird regions with a high sampling rate. The

Table 1: Evaluation results with mAP. (a) All data augmentation techniques are used. No hard-negative training. (b) All data augmentations except color jittering are used. No hard-negative training. (c) All data augmentations except color jittering are used. Iterative hard-negative training. DA, HN, and CJ mean data augmentation, hard-negative training, and color jittering, respectively.

	Fixed	Drone
(a) All DAs with no HN	66.72	61.70
(b) No CJ with no HN	71.33	41.51
(c) No CJ with HN	72.03	72.13

sampling rates of positives and negatives are 8 and 2, respectively. Negative samples are selected from textured regions for mainly sampling hard-negatives.

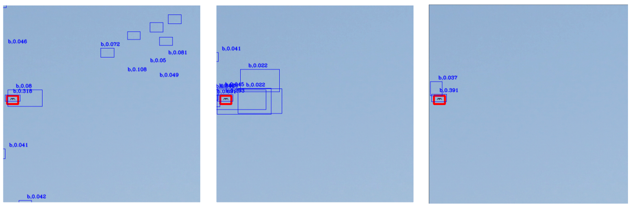
4.2 Inference

During inference, each image with 3,840 × 2,160 pixels is divided into 40 patch images without overlap at regular intervals. Each patch image is fed into Centernet for bird detection.

4.3 Results

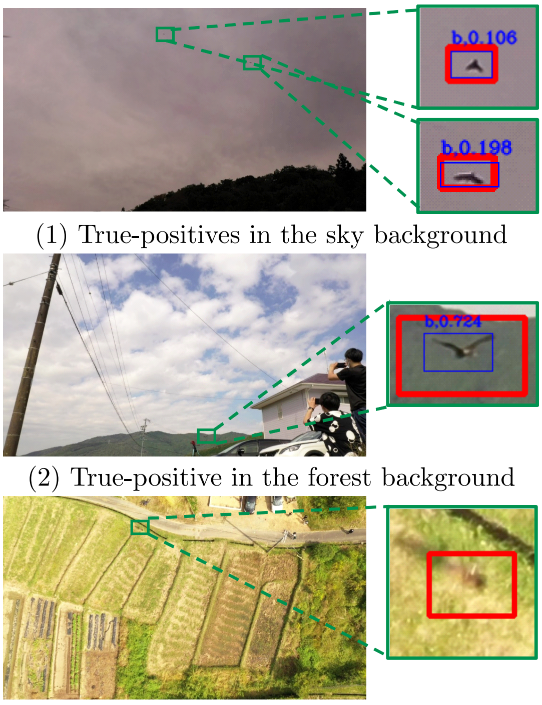
We conducted experiments with the Yoshihashi dataset [5] and our proposed dataset, which are called “Fixed” and “Drone” in Table 1, respectively. Among all detection results given by Centernet, if each detection satisfies the following two conditions, this detection is regarded as a positive detection: (1) the score of each detection is above a pre-defined threshold, and (2) its IoU with the ground-truth bounding-box of a bird is also above a threshold. These thresholds were determined with validation data selected from training data. These detection results are evaluated by mAP, as shown in Table 1. In the baseline method (a), all data augmentation techniques including random flip, random scaling, random cropping (between 0.6 and 1.4 times scaling in our experiments), and color jittering (CJ in Table 1) were employed, and no hard-negative training was achieved. In (b), different from (a), color jittering was not used. In addition to (b), iterative hard-negative training was performed in (c). In Table 1, we can see that (c) is the best. This result validates that the detection performance of Centernet can be improved by carefully-selecting data augmentation techniques and iteratively-learning hard-negatives. Examples shown in Fig. 3 shows the effect of hard-negative training and excepting color jittering.

Several examples of bird detections in our proposed dataset are shown in Fig. 4. In the top and middle examples, birds in textureless and textures regions are detected correctly, respectively. In the bottom example, it is difficult to detect such a blurred bird. These results validate that images with a variety of difficulty grades are included in our dataset.



(a) All DAs with no HN (b) No CJ with no HN (c) No CJ with HN

Figure 3: Effect of data augmentations (DAs) and hard-negative (HN) training.



(1) True-positives in the sky background

(2) True-positive in the forest background

(3) False-negative in the rice-field background

Figure 4: Detection results in our dataset. Red and blue bounding-boxes indicate the ground-truth and detection boxes, respectively.

5 Concluding Remarks

In this paper, we proposed a new dataset for detecting distant birds captured by a drone camera. Future work includes bird tracking in videos for improving detection robustness to intermittent false-positives in continuous frames. Moreover, even the high-speed CenterNet can only achieve about 6 fps on a GPU, which is not enough for real-time processing on a computer mounted on a drone. A more lightweight algorithm for real applications is needed. This work was supported by JSPS KAKENHI Grant Number 19K12129.

References

[1] Mavic 2-DJI: `https://www.dji.com/jp/mavic-2`

[2] Hawk vs. Drone! (Hawk Attacks Quadcopter) - YouTube: `https://www.youtube.com/watch?v=AhDG_WBIQgc`

[3] Territorial Bird Attacks Flying Drone - YouTube: `https://www.youtube.com/watch?v=1NY3Df_Wuc4`

[4] Phantom 3 get kidnapped by two eagles - YouTube `https://www.youtube.com/watch?v=FX3uOQiZsOA`

[5] R. Yoshihashi, et al.: "Bird detection and species classification with time-lapse images around a wind farm: Dataset construction and evaluation," *Wind Energy*, vol.20, no.12, pp.1983–1995, 2017.

[6] Cheng-Yang Fu, et al.: "Dssd: Deconvolutional single shot detector," *arXiv:1701.06659*, 2017.

[7] Ren, Shaoqing and He, Kaiming and Girshick, Ross and Sun, Jian.: "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol.39, no.6, pp.1137–1149, 2016.

[8] Zhou, Xingyi and Wang, Dequan and Krähenbühl, Philipp: "Objects as points," *arXiv:1904.07850*, 2019.

[9] Muhammad Haris and Gregory Shakhnarovich and Norimichi Ukita: "Task-Driven Super Resolution: Object Detection in Low-resolution Images," *arXiv*, 1803.11316, 2018.

[10] Muhammad Haris and Gregory Shakhnarovich and Norimichi Ukita: "Deep Back-Projection Networks for Super-Resolution," *CVPR*, 2018.

[11] Shuhang Gu, et al.: "AIM 2019 Challenge on Image Extreme Super-Resolution: Methods and Results," *ICCV Workshop*, 2019.

[12] Muhammad Haris and Gregory Shakhnarovich and Norimichi Ukita: "Deep Back-Projection Networks for Single Image Super-resolution," *arXiv*, 1904.05677,2018.

[13] Muhammad Haris and Gregory Shakhnarovich and Norimichi Ukita: "Recurrent Back-Projection Network for Video Super-Resolution," *CVPR*, 2019.

[14] Muhammad Haris and Gregory Shakhnarovich and Norimichi Ukita: "Space-Time-Aware Multi-Resolution Video Enhancement," *CVPR*, 2020.

[15] Law, Hei and Deng, Jia: "Cornernet: Detecting objects as paired keypoints," *ECCV*, 2018.

[16] Newell, Alejandro and Yang, Kaiyu and Deng, Jia: "Stacked hourglass networks for human pose estimation," *ECCV*, 2016.

[17] Xiao, Bin and Wu, Haiping and Wei, Yichen: "Simple baselines for human pose estimation and tracking," *ECCV*, 2018.

[18] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian: "Deep residual learning for image recognition," *CVPR*, 2016.

[19] Jansson, Samuel and Papayannis, Alexandros and Åkesson, Susanne and Tsaknakis, Georgios and Brydegaard, Mikkel: "Exploitation of Multi-Band Lidar for the Classification of Free-Flying Migratory Birds: A Pilot Study over Athens, Greece," *EPJ Web of Conferences*, 2016.

[20] Nilsson, Cecilia and Dokter, Adriaan M and Schmid, Baptiste and Scacco, Martina and Verlinden, Liesbeth and Bäckman, Johan and Haase, Günther and Dell ’ Omo, Giacomo and Chapman, Jason W and Leijnse, Hidde and others: "Field validation of radar systems for monitoring bird migration," *Journal of Applied Ecology*, vol.55, no.6, pp.2552–2564, 2018.

[21] P. Welinder and S. Branson and T. Mita and C. Wah and F. Schroff and S. Belongie and P. Perona: "Caltech-UCSD Birds 200," *California Institute of Technology*, CNS-TR-2010-001, 2010.

[22] "260 Bird Species," `https://www.kaggle.com/gpiosenka/100-bird-species`.

[23] "NABirds dataset," `https://dl.allaboutbirds.org/nabirds`.

[24] Grant Van Horn and Steve Branson and Ryan Farrell and Scott Haber and Jessie Barry and Panos Ipeirotis and Pietro Perona and Serge J. Belongie: "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection" *CVPR*, 2015.

[25] Wei Liu and Dragomir Anguelov and Dumitru Erhan and Christian Szegedy and Scott E. Reed and Cheng-Yang Fu and Alexander C. Berg: "Ssd: Single Shot MultiBox Detector" *ECCV*, 2016.

[26] Wei Liu and Dragomir Anguelov and Dumitru Erhan and Christian Szegedy and Scott E. Reed and Cheng-Yang Fu and Alexander C. Berg: "Ssd: Single Shot MultiBox Detector" *ECCV*, 2016.

[27] Joseph Redmon and Santosh Kumar Divvala and Ross B. Girshick and Ali Farhadi: "You Only Look Once: Unified, Real-Time Object Detection" *CVPR*, 2016.

[28] Redmon, Joseph and Farhadi, Ali.: "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[29] Vondrick, Carl and Patterson, Donald and Ramanan, Deva: "Efficiently Scaling up Crowdsourced Video Annotation" *International Journal of Computer Vision*, 2012.