Bird Object Detection Based on RT-DETR in Complex Scenarios

Wenxiao He,

I. Abstract

Bird conservation is essential for maintaining ecological balance and promoting economic development. In ornithological research, locating and recognizing individual birds are fundamental tasks. However, bird detection is challenging due to issues such as occlusion, small object size, blurriness, backlighting, and crowding. In this paper, we established TH-Birds, a dataset containing 708 images of birds, many of which are difficult to detect due to the aforementioned challenges. We annotated the dataset with bounding boxes and segmentation masks, and further categorized the images based on quality-related factors. Additionally, we propose a novel data augmentation technique to simulate branch occlusions and modify the loss function, Feature Pyramid Networks (FPN), and positive sample matching strategy of the RT-DETR model. These enhancements enable our model to achieve improved accuracy and recall in complex scenarios. Experimental results demonstrate a significant improvement in detection performance, particularly in occluded and cluttered environments.

II. Introduction

Avian biodiversity is an important factor in ecosystems and contributes to economic and culture development [1]. Currently, the rapid economic development leads to fragmentation and degradation of habitat and overhunting which eventually cause many bird species becoming endangered [1, 2].

Monitoring birds is the basis of further protection. Many relative organizations have lunched campaigns tracking quantities and behaviors. Xu Shi et al., investigated the migration of ***Clanga clanga*** and found the difference of migration patterns between adult and juvenile specimens [3]. Eric R. Gulson-Castillo et al., discussed the impacts of space weather on bird migration [4].

Manual monitoring is time-intensive and requires professional knowledge. Deep learning, however, can utilize computers' computing capability to learn the characteristics of birds, reaching a high accuracy. In recent years, methods of birds monitoring and identification based on deep learning are thriving. For example, Stefan Kahl and Stefan Kahl proposed a deep neural network BirdNET which can identify bird species by sound in an average precision of 0.791 for single-species recordings [5]. Within the domain of sound identification, methods such as Gaussian mixture model (GMM), support vector machine (SVM) were adopted [6, 7]. For birds images identification, many state-of-art deep learning designs were proposed. Potluri, H et al. applied ResNet (He et al., 2016) [8] on CUB-200-2011 dataset [9] and reached an accuracy of 96.5% [10]. Liu et al. presented a novel feature concentration Transformer (TransIFC) to extract the semantic information in birds images [11]. This design was tested on NABirds dataset [12] with an accuracy of 90.9%.

These researches focus on identification birds' species with high-quality sounds or images. However, in real-time monitoring, detection of locations of birds with complex backgrounds is the fundamental task. Currently, most of birds images datasets are designed for fine-grained classification such as CUB-200-2011 [9], NABirds [12] and DongNiao International Birds 10000 [13] rather than for object detection. Unfortunately, popular general-purposed object detection datasets MS COCO [14] and Pascal VOC [15] are lack of birds images photoed in field environment. Yuki Kondo et al. built a dataset of birds images for small object detection [16]. Using this dataset, Da Huo et al. developed a model combined with Swin Transformer and CenterNet ; Hao-Yu Hou et al. introduced ensemble fusion techniques to reach an average precision of 77.6% at an IoU threshold of 0.5 [17, 18, 19, 20]. Although [16] provides an alternative of MS COCO [14] and Pascal VOC [15], this dataset does not contain many occluded, crowded and backlighting samples and its backgrounds are restricted to a few certain scenes.

In most industrial cases, the object detection models are based on Convolution neural network or CNN which extracts the images' features with a succession of convolution layers and generates proposals of instances with a detection head. Fast-RCNN [21], for example, can utilize convolution network such as ResNet [8] and RoI(region of interests) Pooling to extract features and detect objects in certain regions of images.

In light of these issues, we proposed a dataset containing birds images in field environment TH-Birds. The dataset is publicly available at https://github.com/TamakoHe/TH-Birds. Sample images in the proposed dataset were all collected from various field environment including river, lake and mountainous region. Every bird in these images is annotated with its bounding box, instance segment along with notes regarding the characteristics of every instances. In addition, we proposed a new method of data augmentation simulating the trees' braches' occlusion and other complex scenarios. The FPN layer and the loss function of the RT-DETR [22] model were also redesigned for better performance in bird detection.

Experimental results demonstrated that this model increased the mAP by 1.3% on the TH-bird and MS-COCO dataset. Performance on the occluded and blurry images were also significantly increased. The results indicated that the proposed dataset and model have the potential to provide a new spotlight of wide bird detection and recognization. YOLOv2 [23] modified the detection head to avoid flattening and fully-connected layers that weakened the spatial features in RoI. In 2017, I. Guyon et al,

III. Method

A. Dataset

We used both public dataset and customized dataset to train and evaluate our model. Firstly, the MS-COCO [14] which contains 2.5 million annotated instances in 328k images categorized as 91 types. As our aim is to train a model that can not only locate birds in images but also avoid false detections that recognize non-bird objects as birds. All 80 types of training subset is used. However, as the MS-COCO dataset is lack of instances of birds that is difficult to be detected. We collected 708 birds image in field environment and manually labeled all these birds instances with their bounding boxes and segments. Many of these instances is hard to be detected because of factors such as occlusion. Besides, we also noted birds with the following criterias.

1) whether the instance is occluded.
2) whether the instance is not clear.
3) whether the instance is photographed in side direction.
4) whether the instance is photographed in backlighting condition.

A tabel describing our dataset based on these four criterias is shown in I.

	Instances amount
Occluded	262
Not clear	400
Not side direction	136
Backlighting	70
Total	1142

TABLE I: Summary of TH-Bird dataset

B. Data augmentation

In order to simulate a shooting scene in the field, we adapted and designed some data argument methods. In the existing method, we apply the random disorder in color zone, random horizontal flip, random crop and random extend.

In the field environment, birds may be occluded by branches or be unclear on photos because of movement. We designed the stochastic fuzzy algorithm and random occlusion algorithm to simulate.

1) Stochastic fuzzy algorithm: Firstly, extract the bounding boxes coordinates of birds instances and determine the area to be blurred. Then, apply a convolution kernel of a specific size. All weights of this kernel is one. Therefore, the convolution results can be derived as

$$I'(x,y) = \sum_{i=-k}^{k} \sum_{j=-k}^{k} K(i,j) \cdot I(x+i, y+j)$$

- $I(x,y)$: Original pixel value at (x,y).
- $I'(x,y)$: Blurred pixel value at (x,y).
- $K(i,j)$: Kernel value at (i,j).
- k: Half the kernel size (e.g., for a 3x3 kernel, $k=1$).

For example, to get average value from 3×3 regions to blur the instances. The weights are chosen as:

$$K = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

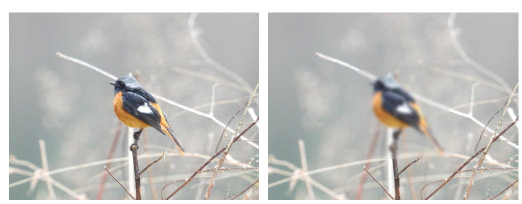

Fig. 1: An exaple of stochastic fuzzy algorithm

2) Random occlusion algorithm: Being blocked by branches is a very common circumstances in bird photography. We design an algorithm to generate simulated branches from data augmentation. To obtain the optimized color, the RGB color and standard deviation of each channel were counted in a branches dataset [24]. The statistic results are shown in II

Value \ Channels	Average value	Standard deviation
R	87.35	2.487
G	89.24	2.497
B	82.97	2.435

TABLE II: Average value and standard deviation from [24]

Fig. 2: Demonstrations of branches simulation algorithm

The for each sample image, the birds instances are found and determine whether the bird's width and height are greater 25 to enable branches augmentation. To draw a occlusion line, its "center" point is randomly selected within the bounding box. After that, the slope and x coordinates of the start and end points are determined. The y coordinates is derived with x coordinates, slop and bonding box limit. The thinness varies along the line with an average value determined by the size of the bird and a fixed standard deviation. The flow graph and the pseudo code are displayed in 3

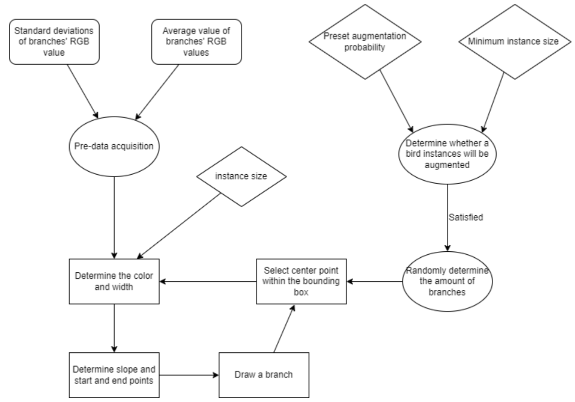

Fig. 3: The flow graph of random branches augmentation

C. *Model architecture*

Our model is based on DETRs Beat YOLOs on Real-time Object Detection or RT-DETR [22].

References

[1] A. C. Lees *et al.*, "State of the world & apos;s birds," *Annual Review of Environment and Resources*, vol. 47, no. Volume 47, 2022, pp. 231–260, 2022, ISSN: 1545-2050. DOI: https://doi.org/10.1146/annurev-environ-112420-014642. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-environ-112420-014642.

[2] I. 2024. "The iucn red list of threatened species." (2024), [Online]. Available: https://www.iucnredlist.org/.

[3] X. Shi, X. Wang, Q. Wei, Q. Lin, and L. Zhu, "Detour for the inexperienced? migration count data suggest mostly juvenile greater spotted eagles appear in coastal peninsulas in china," *Avian Research*, vol. 15, p. 100 183, 2024, ISSN: 2053-7166. DOI: https://doi.org/10.1016/j.avrs.2024.100183. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2053716624000264.

[4] E. R. Gulson-Castillo *et al.*, "Space weather disrupts nocturnal bird migration," *Proceedings of the National Academy of Sciences*, vol. 120, no. 42, e2306317120, 2023. DOI: 10.1073/pnas.2306317120. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2306317120. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2306317120.

[5] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "Birdnet: A deep learning solution for avian diversity monitoring," *Ecological Informatics*, vol. 61, p. 101 236, 2021, ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2021.101236. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954121000273.

[6] S.-h. Zhang, Z. Zhao, Z.-y. Xu, K. Bellisario, and B. C. Pijanowski, "Automatic bird vocalization identification based on fusion of spectral pattern and texture features," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 271–275. DOI: 10.1109/ICASSP.2018.8462156.

[7] T. Tuncer, E. Akbal, and S. Dogan, "Multileveled ternary pattern and iterative relieff based bird sound classification," *Applied Acoustics*, vol. 176, p. 107 866, 2021, ISSN: 0003-682X. DOI: https://doi.org/10.1016/j.apacoust.2020.107866. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X20309713.

[8] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1512.03385.

[9] X. He and Y. Peng, "Fine-grained visual-textual representation learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 520–531, Feb. 2020, ISSN: 1558-2205. DOI: 10.1109/tcsvt.2019.2892802. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2019.2892802.

[10] H. Potluri, A. Vinnakota, N. P. Prativada, and K. C. Yelavarti, "Bird species identification based on images using residual network," in *Advances in Information Communication Technology and Computing: Proceedings of AICTC 2022*, Springer, 2023, pp. 559–567.

[11] H. Liu, C. Zhang, Y. Deng, B. Xie, T. Liu, and Y.-F. Li, "Transifc: Invariant cues-aware feature concentration learning for efficient fine-grained bird image classification," *IEEE Transactions on Multimedia*, pp. 1–14, 2023. DOI: 10.1109/TMM.2023.3238548.

[12] G. Van Horn *et al.*, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015.

[13] J. Mei and H. Dong, *The dongniao international birds 10000 dataset*, 2020. arXiv: 2010.06454 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2010.06454.

[14] T.-Y. Lin *et al.*, *Microsoft coco: Common objects in context*, 2015. arXiv: 1405.0312 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1405.0312.

[15] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[16] Y. Kondo *et al.*, "MVA2023 Small Object Detection Challenge for Spotting Birds: Dataset, Methods, and Results," in *2023 18th International Conference on Machine Vision and Applications (MVA)*, https://www.mva-org.jp/mva2023/challenge, 2023.

[17] Z. Liu *et al.*, *Swin transformer: Hierarchical vision transformer using shifted windows*, 2021. arXiv: 2103.14030 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2103.14030.

[18] X. Zhou, D. Wang, and P. Krähenbühl, *Objects as points*, 2019. arXiv: 1904.07850 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1904.07850.

[19] D. Huo *et al.*, "Small object detection for birds with swin transformer," in *2023 18th International Conference on Machine Vision and Applications (MVA)*, 2023, pp. 1–5. DOI: 10.23919/MVA57639.2023.10216093.

[20] H.-Y. Hou *et al.*, "Ensemble fusion for small object detection," in *2023 18th International Conference on Machine Vision and Applications (MVA)*, 2023, pp. 1–6. DOI: 10.23919/MVA57639.2023.10215748.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "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, 2017. DOI: 10.1109/TPAMI.2016.2577031.

[22] Y. Zhao *et al.*, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 16 965–16 974.

[23] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[24] R. Universe, *Tree branch detection dataset*, https://universe.roboflow.com/fyp-lvrkf/tree-branch-detection, Accessed: 2024-11-19, 2024.