

MVA2023 Small Object Detection Challenge for Spotting Birds: Dataset, Methods, and Results

Yuki Kondo¹Norimichi Ukita²Takayuki Yamaguchi³Hao-Yu Hou⁴Mu-Yi Shen⁴Chia-Chi Hsu⁴En-Ming Huang⁴Yu-Chen Huang⁴Yu-Cheng Xia⁴Chien-Yao Wang⁵Chun-Yi Lee⁴Da Huo⁶Marc A. Kastner⁷Tingwei Liu⁶Yasutomo Kawanishi^{8,6}Takatsugu Hirayama^{9,6}Takahiro Komamizu⁶Ichiro Ide⁶Yosuke Shinya¹⁰Xinyao Liu¹¹Guang Liang¹¹Syusuke Yasui¹²¹Toyota Motor Corporation, ²Toyota Technological Institute, ³Iwate Agricultural Research Center,⁴National Tsing Hua University, ⁵Institute of Information Science, Academia Sinica, ⁶Nagoya University,⁷Kyoto University, ⁸RIKEN, ⁹University of Human Environments, ¹⁰Independent Researcher,¹¹Xi'an Jiaotong University, ¹²Space shift inc.

Abstract

Small Object Detection (SOD) is an important machine vision topic because (i) a variety of real-world applications require object detection for distant objects and (ii) SOD is a challenging task due to the noisy, blurred, and less-informative image appearances of small objects. This paper proposes a new SOD dataset consisting of 39,070 images including 137,121 bird instances, which is called the Small Object Detection for Spotting Birds (SOD4SB) dataset. The detail of the challenge with the SOD4SB dataset ¹ is introduced in this paper. In total, 223 participants joined this challenge. This paper briefly introduces the award-winning methods. The dataset ², the baseline code ³, and the website for evaluation on the public testset ⁴ are publicly available.

1 Introduction

Object detection is one of the fundamental technologies in the field of machine vision. Its performance has been improved by Convolutional Neural Networks (CNN) [4, 5, 6, 7] and Vision Transformers (ViT) [8, 9, 10]. The performance of the object detection task is evaluated in several huge datasets such as COCO [11] and PASCAL VOC [12]. Compared with

¹Challenge site [1]: <https://www.mva-org.jp/mva2023/challenge>²Dataset: <https://drive.google.com/drive/u/2/folders/1vTHilelagbzP0795yh0dNUFh9u2XxZP->³Baseline code [2]: <https://github.com/IIM-TTIJ/MVA2023Small10bjectDetection4SpottingBirds>⁴Codalab [3] site: <https://codalab.lisn.upsaclay.fr/competitions/9594>

Yuki Kondo, Norimichi Ukita and Takayuki Yamaguchi are the MVA2023 Small Object Detection Challenge for Spotting Birds organizers. The other authors participated in the challenge.

Appendix.A contains the authors' team names and affiliations.

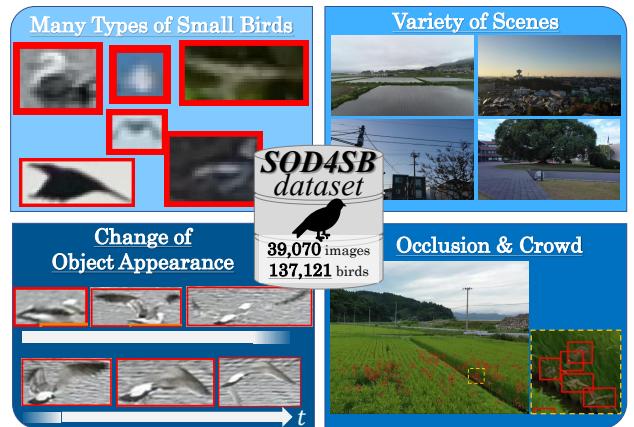


Figure 1: Overview of the Small Object Detection for Spotting Birds (SOD4SB) dataset. The SOD4SB dataset contains a wide variety of small bird types and a variety of scenes. In addition, the flight behavior of the birds and the movements of the UAVs as they film the scene can significantly change bird appearance, and flocking behavior can cause birds to occlude each other, which makes the SOD task even more challenging.

common object detection tasks, SOD [13, 14, 15] is still challenging due to the noisy, blurred, and less-informative image appearances of small objects. One of the reasons of the immaturity of SOD is a limited amount and variety of SOD datasets [16, 17, 18, 19, 20] and evaluation platforms [21, 22].

Considering the aforementioned issues in SOD, we organized the SOD challenge, *Small Object Detection Challenge for Spotting Birds*, with our new dataset of SOD for spotting birds. Compared with visually-simple objects that are targeted in the previous SOD

datasets (e.g., pedestrians [17, 23] and rigid objects such as vehicles and ships [19, 18, 16, 22] captured from bird-eye views), wild birds (i) change their moving paces and silhouettes, (ii) freely fly not on the ground plane but three-dimensionally, (iii) are often crowded, and (iv) are observed in front of a variety of background regions (e.g., sky, clouds, trees, mountains, and so on) in images. These properties make SOD for spotting birds difficult. Furthermore, various real-world applications use SOD for spotting birds, as mentioned in Sec. 2.

The contributions of this paper are as follows:

- The Distant Bird Detection dataset [24] is extended so that more amount and variety of wild birds are observed, as shown in Fig. 1. This extended dataset is called the Small Object Detection for Spotting Birds (SOD4SB) dataset. The baseline code is also provided with the dataset.
- The challenge-winning methods (five methods) are briefly introduced.

2 Why Wild Birds? Sample Applications of Small Bird Detection

Among all possible targets in SOD, this challenge focuses on wild birds. The application of the technology for recognizing birds in images is expected to be in the field of nature conservation and in bird damage prevention technology.

In the field of nature conservation, it is important to understand the status of bird habitats, but it has been necessary to conduct periodic surveys with the naked eye, which requires a great deal of labor. Ogawa et al. have developed a technology that automates the previously manual survey of bird populations through image recognition [25]. Such technology will significantly reduce labor and realize efficient nature conservation activities. We look forward to further technological development by applying and developing the recognition technology tested in this competition.

Next is the application to bird damage prevention technology. Damage caused by birds is not limited to primary industries such as agriculture [26] and fisheries [27] but also covers a wide variety of fields such as aircraft [28] and electric utility industry [29], and the amount of damage is enormous. As technology for avoiding damage, techniques that use sound and light to drive away birds are widely used, and in recent years, UAVs have been developed to drive away birds [30]. However, conventional birds control technologies are installed continuously over a certain period of time, regardless of whether birds are present or not. As a result, birds become accustomed to them, and their effectiveness is reduced or nonexistent [31]. Therefore, there is a need to develop a technology to control birds only when they are detected and to suppress the occurrence of habituation, but a technology that can recognize a wide range of field and minute birds has not been put to practical use. By combining technologies that detect the birds with birds control technologies, it is expected that technologies that can avoid the habituation of birds will be developed, thereby reducing bird damage in a wide range of fields.

One of the difficulties in developing SOD datasets for spotting birds is annotation cost. Even for human annotators, it is difficult to correctly annotate small birds flying against a background of highly-textured objects such as the leaves of trees. In the Drone vs. Bird Detection Challenge [22], although drones are annotated, wild birds are not, and A. Coluccia et al. believe that the annotation of such birds is an important future challenge. Furthermore, annotating all crowded birds is more erroneous and time-consuming.

While a few SOD datasets for spotting birds [32, 33, 34] are developed, these datasets have some limitations. In the Wind Farm dataset [32, 33], time-lapse images were captured from a limited number of fixed-view points. In the AirBird dataset [34], time-lapse images were captured from fixed-view cameras only around airports.

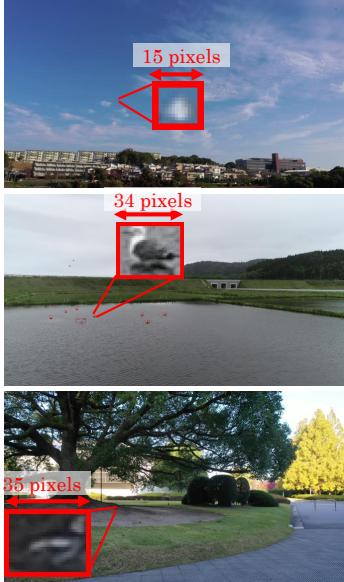
Our SOD4SB dataset, on the other hand, has a variety of images that are useful for various types of real-world applications, which are introduced in Sec. 2.

3.1 Collection

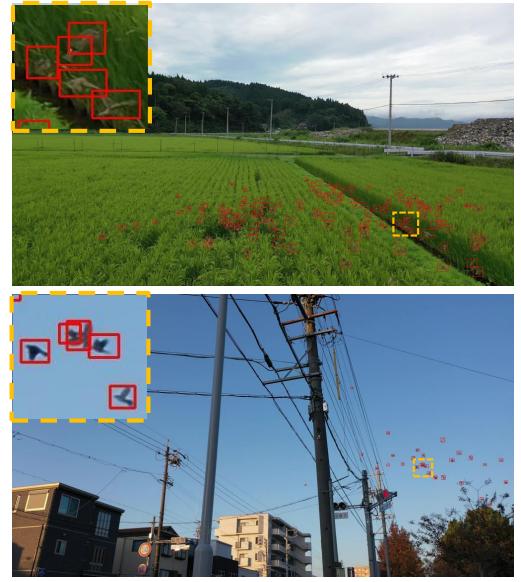
On-drone cameras were used for image collection. The drones used for filming were the DJI Mavic 2 Pro and the DJI Phantom 4 Pro V2.0. The camera captures videos at 30 fps, while temporal frames in the same video are regarded as independent frames in this year’s challenge. The image resolution is $3,840 \times 2,160$ pixels. The videos were captured in various locations such as urban areas, parks, forests, and fields under different weather conditions, as shown in Fig. 2 (a). Birds observed in the images are hawks, crows, waterfowls, sparrows, and so on. Several types of birds are crowded and mutually occluded, as shown in Fig. 2 (b). Due to the quick motion of the birds and drone, bird and background images are sometimes blurred, as shown in Fig. 1. Since most birds were located far from the drone, most bird instances in the images are considered to be small objects, as described in Sec. 3.4.

3.2 Annotation

We manually extracted temporal frames in which any birds are observed in each video. The extracted temporal frames were annotated so that trained annotators enclosed each bird instance by a bounding-box by the publicly-available video annotation tool,



(a) Various types of small birds and diverse scenes



(b) Crowded and mutually occluded birds

Figure 2: Samples of SOD4SB. As shown in (a) and (b), the birds in this dataset are not only small, but also require recognition in a variety of scenes and furthermore occlude each other, making it a challenging dataset for the SOD task.

VATIC [35]. The annotated bounding boxes were double-checked. While several types of wild birds are observed in the SOD4SB dataset, all types of birds are annotated as “bird” because it is difficult to correctly classify all small bird instances even by human annotators. In total, the SOD4SB dataset includes 39,070 images and 137,121 bird instances.

3.3 Splitting

The 39,070 annotated images and instances are split as follows:

- **Training subset** consists of 9,759 images with 29,037 annotated bird instances.
- **Public Test subset** consists of 9,699 images with 29,775 annotated bird instances.
- **Private Test subset** consists of 20,512 images with 78,309 annotated bird instances.

Temporal frames within the same video are considered independent for this challenge, so the images are shuffled.

The annotation data of the public test subset and both the images and annotation data of the private test subset are not publicly available.

3.4 Appropriateness for SOD

The quantitative validation is described in what follows. While there are several criteria of SOD,

our SOD4SB dataset is evaluated with two criteria. (i) In [36, 37], the pixel size of each instance is simply evaluated so that the instance is regarded as a small object if its size is less than 32^2 pixels. (ii) In [38], on the other hand, the criterion is defined with the relative sizes of objects compared to the image size so that the median of the relative sizes of all instances in each object category is less than 0.58% of the image size.

Adapting criterion (i) to the SOD4SB dataset, the number of instances of the corresponding small objects was 74,612, meaning 95.28% of the total. Adapting criterion (ii) to the SOD4SB dataset, the median object size relative to the image size was $0.002\% < 0.58\%$, meeting the requirement. Furthermore, the number of objects satisfying the relative size of 0.58% was 78,222, meaning 99.89%. Based on the above, we consider the SOD4SB dataset the specialized dataset for SOD, since objects that satisfy the definition of small objects are dominant in SOD4SB.

For intuitively validating the appropriateness of the SOD4SB dataset for SOD, the distributions of the object instance sizes in the SOD4SB dataset, the MS COCO validation dataset (val2017) [11] for generic object detection, and the AirBirds dataset [34] for SOD are shown in Fig. 3 along with the aforementioned small objects definition. Comparing the SOD4SB dataset with the COCO dataset, we can see that more objects clearly meet the definition of small objects. On the other hand, a comparison with the AirBirds dataset shows that the distribution seems to be more concentrated on smaller objects in the AirBirds dataset. How-

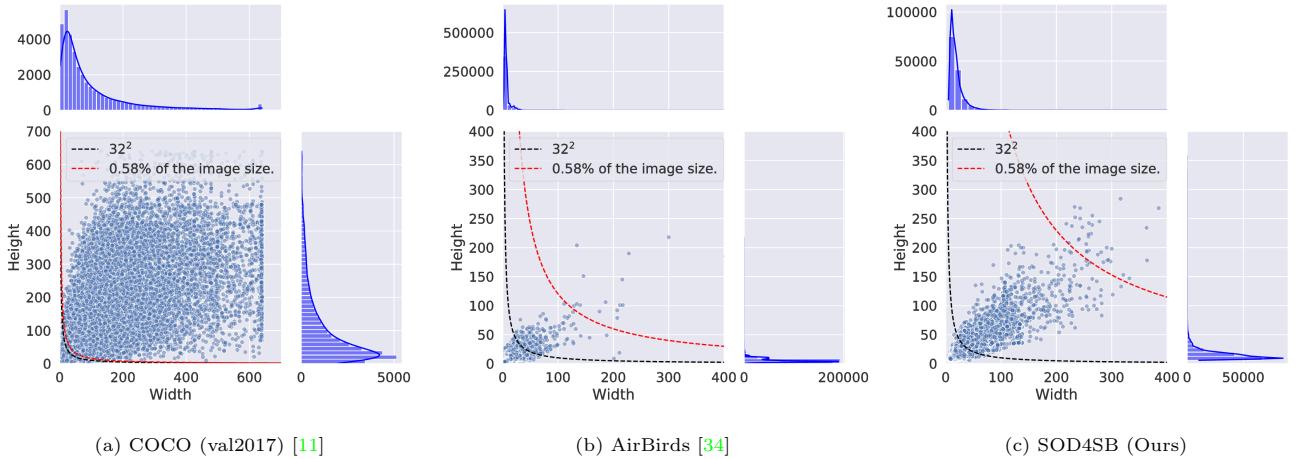


Figure 3: Comparison of object size distributions for (a) a generic object detection dataset, (b) an existing SOD dataset, and (c) the SOD4SB dataset. The dotted line is the definition of small object, and the image sizes used in the relative definitions are (a) 574×484 pixels, (b) $1,920 \times 1,080$ pixels, and (c) $3,840 \times 2,160$ pixels, respectively.

Table 1: Comparison of difficulty levels for datasets based on CenterNet [39] with the ResNet18 [40]-based backbone. COCO (val2017) scores taken from [39].

Dataset	AP@50	AP@25	AP@75
SOD4SB public test	46.4	59.5	5.4
SOD4SB private test	15.4	24.1	1.6
COCO (val2017) [11]	51.5	-	35.1

ever, when compared by criterion (ii) of small objects, the number of objects in the SOD does not change significantly. This is due to the different resolutions of the AirBirds and SOD4SB datasets. From this, the SOD4SB dataset is dominated by objects that satisfy the definition of small objects, and the distribution of object sizes is more diverse than AirBirds. When evaluating models on metrics such as AP, the SOD4SB dataset, which meets the definition of small objects but has a broad distribution, can appropriately evaluate models that can detect small objects of various sizes rather than overestimating models that can only detect extremely small objects.

4 MVA2023 Small Object Detection Challenge for Spotting Birds

We describe the details of the challenge using the SOD4SB dataset planned by the organizers.

4.1 Baseline Code

The organizers provide the baseline code [2] for the challenge. This code is developed based on CenterNet [39] with the ResNet18 [40]-based backbone pro-

vided in MMDetection [41]. The network is trained with hard negative mining to cope with an imbalance problem in which foreground pixels are significantly less than background pixels.

The results of bird detection using the baseline code are shown in Table 1. The detection performance on the SOD4SB public test and private test are much worse than Centernet’s AP@50 on the MS COCO validation (val2017) [11]: 46.4 and 15.4 vs. 51.5 [39]. The difference at AP@75 is even more pronounced. This comparison proves the difficulty in SOD on our SOD4SB dataset and this challenge.

4.2 Challenge Phases

The public and private test phases were given to participants. In the public test phase, the participants can evaluate their methods on the public test subset of the SOD4SB dataset by submitting the detection result to CodaLab [3]. In this phase, the participants can access to only images without annotations. In the private test phase, the organizer ran the code provided by each team for evaluation on the private test subset of the SOD4SB dataset. After the challenge ends also, CodaLab is publicly available for evaluation on the public test subset, as described in Note 4 in the footnote. Each team can evaluate their results at most two times per day for restricting HARKing [42] to the public test subset.

4.3 Challenge Categories and Ranking Criteria

Our challenge has two categories. In the development category, participants are requested to improve the AP@50 score on the private test subset. Only the score is evaluated. No technical novelty is appraised for

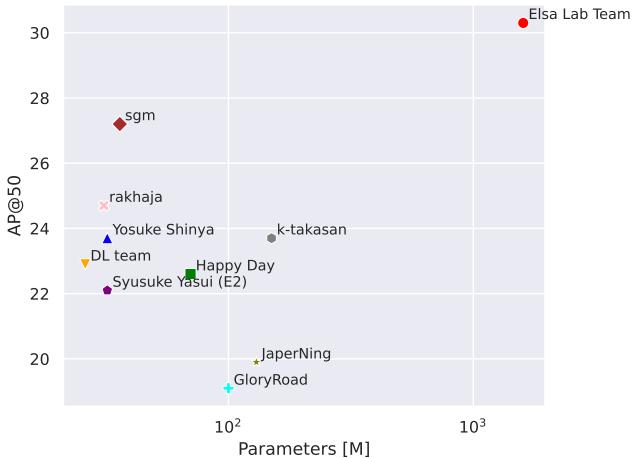


Figure 4: Parameters-AP@50 trade-off of winner’s solutions in SOD4SB private test subset. The horizontal axis is shown in log scale.

the ranking in the development category. In addition to the AP@50 score, technical novelties and methodological effectiveness are evaluated in the research category. This evaluation is done by three reviewers per paper in a double-blind manner and given an average score ranging from 0 to 5 points.

5 Challenge Results

223 participants joined this challenge. Based on the ranking criteria described in Sec. 4.3, five teams are selected as winners in both the development and research categories, as shown in Table 2. While the results of many teams are above those of the baseline code in AP@50, the results on the private test set are not as different as the public test compared to the baseline code scores and are significantly down from the public test scores for each method. This is presumably due to a gap between the data distribution of the public and private test subsets, but since both images were taken during the same time period and at the same location, this difference should be minute. Even under these conditions, since the AP@50 score of the Elsa Lab Team which won the first rank in the research category is above 30, their approach can be considered a method with high generalization capacity.

Figure 4 shows the relationship between the AP@50 score and model size of the challenge-winning 10 methods. The Elsa Lab Team achieved the highest AP with a huge number of parameters, while sgm combines an efficient model with a low number of parameters and a high AP.

6 Challenge Methods and Teams

This section briefly describes the methods proposed by the winning teams in Research Category.

6.1 Elsa Lab Team (Team1)

Elsa Lab Team (Team1) utilized an ensemble fusion method [43] that leverages the strengths of existing approaches to enhance the overall performance. To achieve this objective, their ensemble fusion method integrates variants from two model architectures: Cascade R-CNN [44] and CenterNet [39]. During the training phase, an assortment of backbones (e.g., InternImage [9] and ResNet [40]) and techniques (e.g., Normalized Wasserstein Distance (NWD) [45] and Copy-Paste (CP) [46]), are utilized to generate variants exhibiting diverse performance attributes. In the inference phase, additional variants are produced using techniques such as Slicing Aided Hyper Inference (SAHI) [47] and test time augmentation (TTA). By ensembling the variants and their predictions using the Weighted Box Fusion method (WBF) [48], a substantial improvement is attained as compared to each top-performing model.

Fig. 5 (a) illustrates an overview of their proposed framework, which consists of two distinct stages: the *data preparation stage* and the *model ensemble stage*. In the *data preparation stage*, they utilize the CP data augmentation technique to enrich the training data provided by SOD4SB. In this stage, the images from the SOD4SB dataset undergo cropping and augmentation with birds sourced from either the SOD4SB dataset or the Birds Flying dataset [49]. The augmented data are then forwarded to the *model ensemble stage*, where several model variants are developed and grouped together to form an ensemble using WBF. By combining the predictions from different variants, WBF enables the exploitation of these outputs to generate more precise final bounding boxes.

The performance of various ensembling methods and the top-performing single model are reported in Table 3 (b), while baseline results are shown in Table 3 (a). The WBF ensembling method surpasses all baselines in terms of AP scores, achieving an AP@50 score of 77.6. Fig. 5 (b) depicts the impact of WBF, where the bounding boxes from different predictions are ensembled, resulting in a more accurate prediction.

6.2 Happy Day

Happy Day team proposed a Swin Transformer [53] based network [51] with a hierarchical design for small object detection (Fig. 6), which improves the features learned by a neck network corresponding to the CenterNet [39] architecture to learn effective features for small objects. The key idea in Happy Day’s work is to change the size of the shifting windows of the neck to a smaller one, which contributes to capture the attentions of small objects inside the small windows, which reduces the parameters and leads to good performance for small object detection. In addition, to detect small objects precisely in location even through several up-and-down-samplings, Happy Day uses skip

Table 2: Quantitative evaluation results from public and private tests for this challenge. In the category column, “R” represents the research category and “D” represents the development category. Runtime indicates the inference time for one image when the mini-batch size is set to 1. The results of the public test are the final results after the challenge period.

Category	Rank	Team	Private Test			Public Test			Review	Params. [M]	Runtime [s/image]	GPU
			AP@50	AP@25	AP@75	AP@50	AP@25	AP@75				
R	1	Elsa Lab Team	30.3	42.9	7.5	77.6	84.0	22.5	4.33	1600	77.00	V100
	2	Happy Day	22.6	35.8	7.2	70.2	80.5	14.0	4.67	70	0.26	A100
	3	Yosuke Shinya	23.7	36.0	5.6	73.1	80.2	19.1	Secret	32	0.27	RTX3090
	4	DL team	22.9	31.8	5.3	73.1	80.2	19.1	2.67	26	0.09	RTX3090
	5	Syusuke Yasui (E2)	22.1	30.8	5.2	69.6	78.1	19.2	Secret	32	6.80	A100
D	1	sgm	27.2	35.8	7.2	73.7	80.3	20.4	-	36	1.00	A30
	2	rakhaja	24.7	32.1	5.6	69.3	74.3	18.3	-	31	4.39	RTX5000
	3	k-takasan	23.7	34.8	2.3	70.7	80.8	12.1	-	150	2.00	V100
	4	JaperNing	19.9	31.8	2.8	59.5	71.7	10.1	-	130	0.70	RTX2000
	5	GloryRoad	19.1	28.2	2.7	67.9	78.5	13.8	-	100	4.00	A100

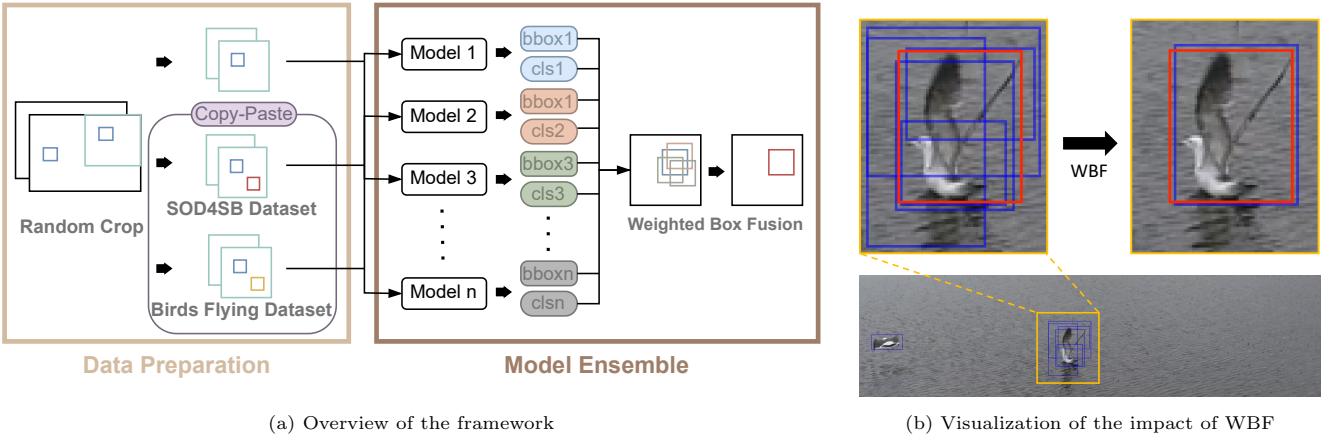


Figure 5: (a) An overview of the framework [43]; (b) Visualization of the impact of WBF: Comparison of the predictions before and after applying WBF (indicated by the blue boxes) against the ground truth (depicted by the red box).

Table 3: The AP(%) scores of: (a) baselines and (b) various ensemble methods evaluated on the SOD4SB testing set.

	Baseline Model	Backbone Network	AP@25	AP@50	AP@75
(a)	DetectoRS [50]	ResNet-50	48.3	34.6	3.8
	CenterNet [39] [2]	ResNet-18	61.6	49.1	7.1
	Cascade R-CNN [44]	ResNet-50	63.1	53.3	10.8
	Ensemble Method		AP@25	AP@50	AP@75
(b)	Top-Performing Single Model		80.3	73.7	18.3
	Pure NMS with no weight		67.3	61.6	19.5
	Weighted NMS		81.4	75.1	19.3
	Soft NMS		79.7	73.9	20.8
	WBF (Elsa Lab Team)		84.0	77.6	22.5

connections [54] for providing precise locations from backbone to the neck. The architecture of the proposed neck network is shown in Fig. 6.

The input consists of different scales of features from backbone, as shown in Fig. 6, from C5 (32× down-sampling) to C2 (4× down-sampling) with different sizes, where Ci ($2 \leq i \leq 5$) corresponds to i -th blocks of the backbone. The shifting windows are inside of the

Swin Transformer blocks in each stage, and the windows size is selected as a smaller one, with a default size of 2. Further, Happy Day adds skip-connection after Stages 1, 2, and 3 for merging features with the backbone outputs. The output of the neck is passed to the CenterNet head for predicting the center points, height, and width of objects in the image.

To evaluate the effectiveness and feasibility of the proposed network, Happy Day compared the proposed neck with the original CenterNet neck, by computing the AP metrics on the validation set of the SOD4SB dataset. Experiments indicate that the most metrics for object detection, including AP@50 and AP@75, of the proposed method surpassed the ones with the default CenterNet neck.

For the key idea in Happy Day’s work, the small windows size is expected to be an effective feature representation. Happy Day also performed ablation study by changing the windows sizes 2, 3, and 5 for experiments. They conducted experimental verification and found that a windows size of 2 significantly emphasized

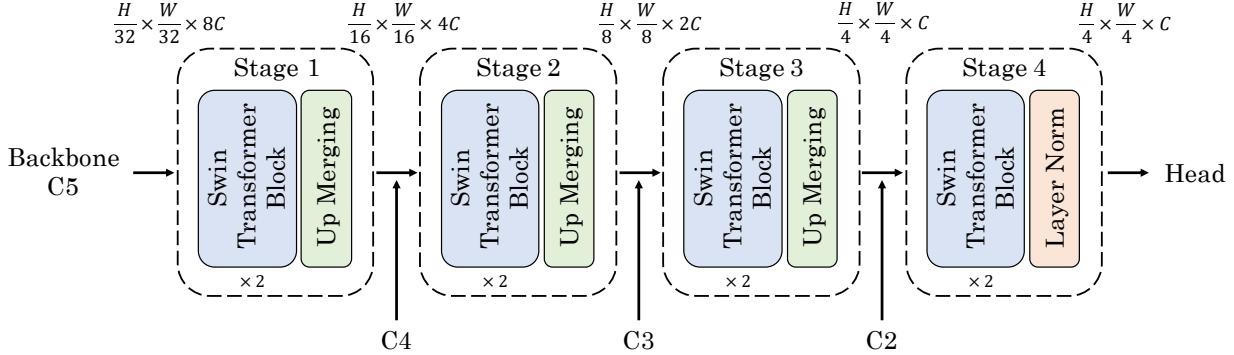


Figure 6: Happy Day team proposed neck network [51], the windows size 2 was used in each Swin Transformer Block, and Up Merging [52] Module upsamples features and merges with those extracted in the backbone to effectively recognize object features in SOD.

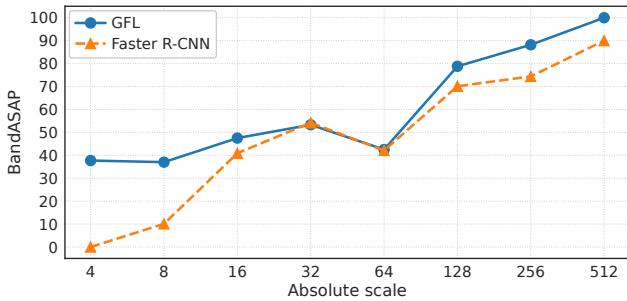


Figure 7: Results of BandASAP [55] metrics.

local information in comparison to other sizes, thereby leading to improved detection performance. This particular windows size is a crucial factor in efficiently computing local features, which play a vital role in small object detection.

6.3 Yosuke Shinya

Yosuke Shinya proposed BandASAP [55], which is a set of scale-wise metrics for object detection evaluation. For finer scale-wise evaluation than the COCO metrics [11, 56], it is based on ASAP [57]. For more reliable and intuitive evaluation than ASAP, the author proposed a filter bank consisting of triangular and trapezoidal band-pass filters.

The author trained GFL [58] and Faster R-CNN [4] with simple settings and selected GFL for the final submission because it is significantly better than Faster R-CNN. He analyzed the results using the proposed metrics (Fig. 7). BandASAP succeeds in highlighting differences between the methods. Although the author discussed a possible cause of low BandASAP₆₄, the cause of the remarkable differences between GFL and Faster R-CNN remains unknown. Further analysis in future research would lead to performance improvement in small object detection.

6.4 DL team

DL team proposed a DL Method (Fig. 8) to enhance the detection capability of small objects. By partitioning the images in the training set into smaller sub-images for training, the method enables training with a larger batch size within the same GPU memory. This enables the model to observe a broader range of features during a single training iteration, leading to a significant improvement in the model’s generalization ability and better capture of the details and features of small objects. Under comparable memory consumption conditions, the method achieves an improvement of over 7 percentage points. The partitioning method takes into account the dataset characteristics to adjust the overlap rate and annotation for the partitions. Furthermore, DL team discovers the importance of data augmentation in training small object detection networks. Consistent utilization of data augmentation in experiments enables the model to learn more features and details across diverse scenes, thereby enhancing the effectiveness of object detection. The partitioning approach generates a greater number of small object samples, while the data augmentation technique ensures sufficient diversification to simulate various variations in real-world scenarios, thus enhancing the model’s robustness. Finally, DL team integrate their method into the medium-scale YOLOv8 [5], evaluate it on the SOD4SB public test and achieve an AP@50 score of 73.3.

6.5 (E2) Syusuke Yasui

(E2) Syusuke Yasui proposed 5 simple but effective methods to do small object detection.

- NWD (Normalized Gaussian Wasserstein Distance) [45, 59] is an improved method for objects with small IoU loss. Introduced to more accurately

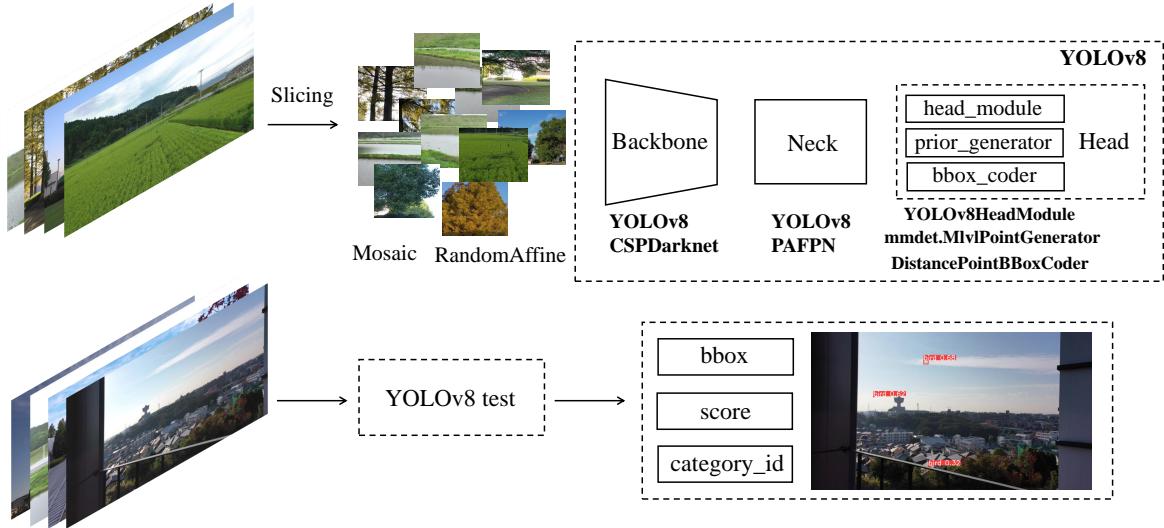


Figure 8: DL Method overview.

indicate bounding box distances for small objects. The optimal weight for loss is 3/4 of varifocal loss and 2 for refine.

- Probability Distribution Surface Models such as CenterNet [39] and VarifocalNet [60] that predict the pixels of the object you want to detect with a probability distribution are more suitable. This is also shown by the experimental results. If the detected object is small, the ratio of positive and negative is severely unbalanced, and it is generally difficult for the model to learn the detection points.
- Switch Hard Augmentation: By performing hard augmentation in the first half of learning and performing lightweight augmentation with only flip in the second half of learning, it is possible to create a stable, high-speed, and highly accurate model. The hard augmentation here is mosaic [61], mixup [62], affine transform. Because hard augmentation effectively increases positives.
- Multi scale train is a standard method of learning while changing the resolution with resize. Randomly selecting around 20% of the input resolution will increase the detection accuracy the most.
- Weight Moving Average is a method to limit the model weight by exponential moving average. The purpose of this is to avoid overfitting in a single backward step and not disproportionately overfitting, and it is effective even for small objects. Its value is 1e-4.

7 Conclusion

This paper proposes a new SOD dataset, the SOD4SB dataset, and reviews the MVA2023 Small Object Detection Challenge for Spotting Birds, which utilizes this dataset. The 223 participants were tasked with detecting small birds in a variety of scenes. The winners' methods performed remarkably well on the challenging SOD4SB dataset and provided several novel and progressive proposals to help solve the SOD task. We hope that the results of this challenge will help build a foundation for advanced UAVs applications.

This challenge is expected to be extended to Video SOD [63, 64] or Video Small Object Tracking [65, 66]. This extension is expected to promote research and development of more useful recognition processing for improving the accuracy of detecting small birds and for autonomous control of UAVs at a later stage. Furthermore, by setting constraints on inference time and the arithmetic unit, we expect to develop technologies that enable real-time inference on edge devices mounted on UAVs.

8 Acknowledgments

We would like to express our deepest gratitude to Zhao Kaikai, Riku Miyata, and Kazutoshi Akita at Toyota Technological Institute for their hard work in preparing for the dataset, base code, and website for this challenge. We also would like to thank Masatsugu Kidode at Nara Institute of Science and Technology for his helpful advice. We also appreciate code testers and annotators.

This challenge was supported by a donation from the MVA organization.

A Teams and Affiliations

MVA2023 Small Object Detection Challenge for Spotting Birds Organizers

Title:

MVA2023 Small Object Detection Challenge for Spotting Birds: Dataset, Methods, and Results

Members:

Yuki Kondo¹ (yuki_kondo_ab@mail.toyota.co.jp), Norimichi Ukita², Takayuki Yamaguchi³

Affiliations:

¹ Toyota Motor Corporation, Japan

² Toyota Technological Institute, Japan

³ Iwate Agricultural Research Center, Japan

Elsa Lab Team (Team1)

Title:

Ensemble Fusion for Small Object Detection

Members:

Hao-Yu Hou¹

(howard.hou.fan@elsa.cs.nthu.edu.tw), Mu-Yi Shen¹, Chia-Chi Hsu¹, En-Ming Huang¹, Yu-Chen Huang¹, Yu-Cheng Xia¹, Chien-Yao Wang², and Chun-Yi Lee¹.

Affiliations:

¹ Elsa Lab, Department of Computer Science, National Tsing Hua University, Taiwan.

² Institute of Information Science, Academia Sinica, Taiwan

Happy Day

Title:

Small Object Detection for Bird with Swin Transformer

Members:

Da Huo¹ (huod@cs.is.i.nagoya-u.ac.jp), Marc A. Kastner², Tingwei Liu¹, Yasutomo Kawanishi^{3,1}, Takatsugu Hirayama^{4,1}, Takahiro Komamizu¹, Ichiro Ide¹

Affiliations:

¹ Nagoya University, Japan

² Kyoto University, Japan

³ GRP, RIKEN, Japan

⁴ University of Human Environments, Japan

Yosuke Shinya

Title:

BandRe: Rethinking Band-Pass Filters for Scale-Wise Object Detection Evaluation

Members:

Yosuke Shinya¹ (<https://shinya7y.github.io/>)

Affiliations:

¹ Independent researcher, Japan

DL

Title:

Method to Achieve High Performance for Small Object Detection

Members:

Xinyao Liu¹ (2205124667@stu.xjtu.edu.cn), Guang Liang¹ (2204313319@stu.xjtu.edu.cn)

Affiliations:

¹ Xi'an Jiaotong University, China

(E2) Syusuke Yasui

Title:

More easy framework of small object detection

Members:

Syusuke Yasui¹ (syuchimu@gmail.com)

Affiliations:

¹ Space Shift Inc., Japan

References

- [1] Y. Kondo, N. Ukita, and T. Yamaguchi, “MVA2023 Small Object Detection Challenge for Spotting Birds.” <https://www.mva-org.jp/mva2023/challenge>, 2023.
- [2] K. Zhao, R. Miyata, Y. Kondo, and K. Akita, “Baseline code for SOD4SB by IIM-TTIJ.” <https://github.com/IIM-TTIJ/MVA2023SmallObjectDetection4SpottingBirds>, 2023.
- [3] A. Pavao, I. Guyon, A.-C. Letournel, X. Baró, H. Escalante, S. Escalera, T. Thomas, and Z. Xu, “Codalab competitions: An open source platform to organize scientific challenges,” *Technical report*, 2022.
- [4] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *NIPS*, 2015.
- [5] G. Jocher, A. Chaurasia, and J. Qiu, “YOLO by Ultralytics.” <https://github.com/ultralytics/ultralytics>, 2023.
- [6] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “CenterNet: Keypoint triplets for object detection,” in *ICCV*, 2019.
- [7] M. Tan, R. Pang, and Q. V. Le, “Efficientdet: Scalable and efficient object detection,” in *CVPR*, 2020.
- [8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *ECCV*, 2020.
- [9] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li, *et al.*, “Internimage: Exploring large-scale vision foundation models with deformable convolutions,” in *CVPR*, 2023.
- [10] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, “Eva: Exploring the limits of masked visual representation learning at scale,” in *CVPR*, 2023.
- [11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in *ECCV*, 2014.

- [12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *IJCV*, vol. 88, pp. 303–338, 2010.
- [13] N.-D. Nguyen, T. Do, T. D. Ngo, and D.-D. Le, “An evaluation of deep learning methods for small object detection,” *Journal of electrical and computer engineering*, vol. 2020, pp. 1–18, 2020.
- [14] Y. Liang, Y. Han, and F. Jiang, “Deep learning-based small object detection: A survey,” in *ICCAI*, 2022.
- [15] G. Chen, H. Wang, K. Chen, Z. Li, Z. Song, Y. Liu, W. Chen, and A. Knoll, “A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal,” *IEEE Transactions on systems, man, and cybernetics: systems*, vol. 52, no. 2, pp. 936–953, 2020.
- [16] J. Wang, W. Yang, H. Guo, R. Zhang, and G.-S. Xia, “Tiny object detection in aerial images,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3791–3798, IEEE, 2021.
- [17] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han, “Scale match for tiny person detection,” in *WACV*, 2020.
- [18] K. Behrendt, L. Novak, and R. Botros, “A deep learning approach to traffic lights: Detection, tracking, and classification,” in *ICRA*, 2017.
- [19] S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai, “isaid: A large-scale dataset for instance segmentation in aerial images,” in *CVPRW*, 2019.
- [20] B. Bosquet, M. Mucientes, and V. M. Brea, “Stdnet: A convnet for small target detection,” in *BMVC*, 2018.
- [21] X. Yu, Z. Han, Y. Gong, N. Jan, J. Zhao, Q. Ye, J. Chen, Y. Feng, B. Zhang, X. Wang, *et al.*, “The 1st tiny object detection challenge: Methods and results,” in *ECCVW*, Springer, 2020.
- [22] A. Coluccia, A. Fascista, A. Schumann, L. Sommer, A. Dimou, D. Zarpalas, F. C. Akyon, O. Eryuksel, K. A. Ozfuttu, S. O. Altinuc, *et al.*, “Drone-vs-bird detection challenge at ieee avss2021,” in *AVSS*, IEEE, 2021.
- [23] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, “Widerperson: A diverse dataset for dense pedestrian detection in the wild,” *IEEE Transactions on Multimedia*, 2019.
- [24] S. Fujii, K. Akita, and N. Ukita, “Distant bird detection for safe drone flight and its dataset,” in *MVA*, 2021.
- [25] K. Ogawa, Y. Lin, H. Takeda, K. Hashimoto, Y. Konno, and K. Mori, “Automated counting wild birds on UAV image using deep learning,” in *International Geoscience and Remote Sensing Symposium, IGARSS*, 2021.
- [26] R. DeHaven and R. Hothen, “Estimating bird damage from damage incidence in wine grape vineyards,” *American journal of enology and viticulture*, vol. 32, no. 1, pp. 1–4, 1981.
- [27] E. Spanier, “The use of distress calls to repel night herons (*nycticorax nycticorax*) from fish ponds,” *Journal of Applied Ecology*, pp. 287–294, 1980.
- [28] R. Hedayati and M. Sadighi, *Bird strike: an experimental, theoretical and numerical investigation*. Woodhead Publishing, 2015.
- [29] O. Hüppop, J. Dierschke, K.-M. EXO, E. Fredrich, and R. Hill, “Bird migration studies and potential collision risk with offshore wind turbines,” *Ibis*, vol. 148, pp. 90–109, 2006.
- [30] B. A. Grimm, B. A. Lahneman, P. B. Cathcart, R. C. Elgin, G. L. Meshnik, and J. P. Parmigiani, “Autonomous unmanned aerial vehicle system for controlling pest bird population in vineyards,” in *ASME International Mechanical Engineering Congress and Exposition*, 2012.
- [31] S. Mahesh, V. Vasudeva Rao, G. Surender, D. Kiran kumar, and K. Swamy, “Distress feeding of depredatory birds in sunflower and sorghum protected by bioacoustics,” *bioRxiv*, p. 200097, 2017.
- [32] R. Yoshihashi, R. Kawakami, M. Iida, and T. Naemura, “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.
- [33] R. Yoshihashi, R. Kawakami, M. Iida, and T. Naemura, “Construction of a bird image dataset for ecological investigations,” in *ICIP*, 2015.
- [34] H. Sun, Y. Wang, X. Cai, P. Wang, Z. Huang, D. Li, Y. Shao, and S. Wang, “Airbirds: A large-scale challenging dataset for bird strike prevention in real-world airports,” in *ACCV*, 2022.
- [35] C. Vondrick, D. Patterson, and D. Ramanan, “Efficiently scaling up crowdsourced video annotation: A set of best practices for high quality, economical video labeling,” *IJCV*, vol. 101, pp. 184–204, 2013.
- [36] A. Torralba, R. Fergus, and W. T. Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” *PAMI*, vol. 30, no. 11, pp. 1958–1970, 2008.
- [37] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, “Traffic-sign detection and classification in the wild,” in *CVPR*, 2016.
- [38] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao, “R-cnn for small object detection,” in *ACCV*, 2017.
- [39] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” *arXiv*, 2019.
- [40] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016.
- [41] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, “MMDetection: Open mmlab detection toolbox and benchmark,” *arXiv*, 2019.
- [42] N. L. Kerr, “Harking: Hypothesizing after the results are known,” *Personality and social psychology review*, vol. 2, no. 3, pp. 196–217, 1998.
- [43] H.-Y. Hou, M.-Y. Shen, C.-C. Hsu, E.-M. Huang, Y.-C. Huang, Y.-C. Xia, C.-Y. Wang, and C.-Y. Lee, “Ensemble fusion for small object detection,” in *MVA*, 2023.
- [44] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving

- into high quality object detection,” in *CVPR*, 2018.
- [45] J. Wang, C. Xu, W. Yang, and L. Yu, “A normalized gaussian wasserstein distance for tiny object detection,” *arXiv*, 2021.
- [46] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T.-Y. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, “Simple copy-paste is a strong data augmentation method for instance segmentation,” in *CVPR*, 2021.
- [47] F. C. Akyon, S. Onur Altinuc, and A. Temizel, “Slicing aided hyper inference and fine-tuning for small object detection,” in *ICIP*, 2022.
- [48] R. Solovyev, W. Wang, and T. Gabruseva, “Weighted boxes fusion: Ensembling boxes from different object detection models,” *Image and Vision Computing*, pp. 1–6, 2021.
- [49] Gareth, “Birds flying dataset.” www.kaggle.com/datasets/nelyg8002000/birds-flying, 2021.
- [50] S. Qiao, L.-C. Chen, and A. Yuille, “DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution,” in *CVPR*, 2021.
- [51] D. Huo, M. A. Kastner, T. Liu, Y. Kawanishi, T. Hirayama, T. Komamizu, and I. Ide, “Small object detection for bird with swin transformer,” in *MVA*, 2023.
- [52] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *CVPR*, 2016.
- [53] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *ICCV*, 2021.
- [54] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *MICCAI*, 2015.
- [55] Y. Shinya, “BandRe: Rethinking band-pass filters for scale-wise object detection evaluation,” in *MVA*, 2023.
- [56] T.-Y. Lin, P. Dollár, *et al.*, “COCO API.” <https://github.com/cocodataset/cocoapi>, Accessed on Apr. 14, 2023.
- [57] Y. Shinya, “USB: Universal-scale object detection benchmark,” in *BMVC*, 2022.
- [58] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, “Generalized Focal Loss: Learning qualified and distributed bounding boxes for dense object detection,” in *NeurIPS*, 2020.
- [59] C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, “Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 79–93, 2022.
- [60] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf, “Varifocalnet: An iou-aware dense object detector,” in *CVPR*, 2021.
- [61] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Scaled-yolov4: Scaling cross stage partial network,” in *CVPR*, 2021.
- [62] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *arXiv*, 2017.
- [63] A. M. Rekavandi, L. Xu, F. Boussaid, A.-K. Seghouane, S. Hoefs, and M. Bennamoun, “A guide to image and video based small object detection using deep learning: Case study of maritime surveillance,” *arXiv*, 2022.
- [64] B. Bosquet, M. Mucientes, and V. M. Brea, “Correlation-based convnet for small object detection in videos,” in *ICPR*, 2021.
- [65] Y. Zhu, C. Li, Y. Liu, X. Wang, J. Tang, B. Luo, and Z. Huang, “Tiny object tracking: A large-scale dataset and a baseline,” *Transactions on Neural Networks and Learning Systems*, 2023.
- [66] Z. Zhang, F. Wu, Y. Qiu, J. Liang, and S. Li, “Tracking small and fast moving objects: A benchmark,” in *ACCV*, 2022.