18th International Conference on Machine Vision Applications (MVA) Workshop
Small Object Detection Challenge for Spotting Birds 2023
Hamamatsu, Japan, July 23-25, 2023.

Ensemble Fusion for Small Object Detection

Hao-Yu Hou*[1], Mu-Yi Shen*[1], Chia-Chi Hsu*[1], En-Ming Huang*[1], Yu-Chen Huang*[1], Yu-Cheng Xia*[1], Chien-Yao Wang[2], and Chun-Yi Lee[1]

[1]Elsa Lab, Department of Computer Science, National Tsing Hua University, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan

Abstract

Detecting small objects is often impeded by blurriness and low resolution, which poses substantial challenges for accurately detecting and localizing such objects. In addition, conventional feature extraction methods usually face difficulties in capturing effective representations for these entities, as down-sampling and convolutional operations contribute to the blurring of small object details. To tackle these challenges, this study introduces an approach for detecting tiny objects through ensemble fusion, which leverages the advantages of multiple diverse model variants and combines their predictions. Experimental results reveal that the proposed method effectively harnesses the strengths of each model via ensemble fusion, leading to enhanced accuracy and robustness in small object detection. Our model achieves the highest score of 0.776 in terms of average precision (AP) at an IoU threshold of 0.5 in the MVA Challenge on Small Object Detection for Birds.

1 Introduction

Small object detection (SOD) has emerged as a pivotal task in computer vision (CV), as the ability to precisely detect small objects is essential for a wide range of applications, including surveillance [1], autonomous driving [2] and aerial image analysis [3]. Despite significant progress in CV, SOD remains a challenging task due to several factors. First of all, small objects tend to exhibit lower contrast and limited salient features, making it difficult to distinguish them from backgrounds. The presence of blurry and rapidly moving objects in datasets such as the Drone dataset [4] and the dataset from the Small Object Detection Challenge for Spotting Birds (SOD4SB) [5] further exacerbates this problem. As down-sampling and convolutional operations may blur the details of small objects, conventional feature extraction methods often face difficulties in capturing effective representations for such entities. Moreover, small objects often appear in complex backgrounds, which increases the difficulty of separating them from the surrounding clutter. Furthermore, even a slight deviation in the bounding box localization may result in a failure to enclose the target entirely. Due to the importance and difficultues of SOD, this research field has attracted attention in the past several years.

To address aforementioned challenges, researchers have explored two avenues: leveraging existing object detection models [6–23] or developing models specifically tailored for SOD [24–38]. Regarding the former, although most object detection models perform unsatisfactorily, some models such as CenterNet [7] and Cascade R-CNN [6] have achieved superior performance due to their distinctive architecture designs. Despite this, there is still room for improvement when compared to specifically tailored models for SOD. Over the past few years, researchers have attempted several directions for such models: (a) utilizing low-level features or applying image super-resolution methods to enhance the contrast and salient features of small objects [24, 25, 37, 38], (b) modifying downsampling or multi-scale feature fusion and prediction strategies (e.g., using dilated convolutions or adjusting model necks) to prevent the blurring of small object details during feature extraction [25–28], (c) employing attention mechanisms to select more relevant and important feature information [28–31], and (d) leveraging data augmentation as well as label design and assignment strategies (e.g., using Normalized Wasserstein Distance (NWD) or copy-paste based augmentation methods) to further enhance the performance [32–36]. Albeit effective, each of the above approaches has its own advantages and drawbacks. Further study and investigation is required to fully harness their potential in SOD tasks.

In light of the above issues, this paper aims to investigate an ensemble fusion method that leverages the strengths of existing approaches to enhance the overall performance. The rationale behind this is that by exploiting the diversity of these models, ensemble methods often allow for improved generalizability, leading to more robust and accurate predictions. To achieve this objective, our ensemble fusion method integrates variants from two model architectures: Cascade R-CNN [6] and CenterNet [7]. During the training phase, an assortment of backbones (e.g., InternImage [22] and ResNet [10]) and techniques (e.g., NWD [33] and Copy-Paste (CP) [32]), are used to generate variants exhibiting diverse performance attributes. In the inference phase, additional variants are produced using techniques such as Slicing Aided Hyper Inference (SAHI) [24] and test time augmentation (TTA). By ensembling the variants and their predictions using Weighted Box Fusion method (WBF) [39], a substantial improvement is attained compared to each top-performing model.

*These authors contributed equally to this work
Source code: `https://github.com/elsa-lab/MVATeam1`

To assess the efficacy of the proposed ensemble fusion approach, we perform comprehensive experiments and evaluations using the SOD4SB public dataset. The results suggest that the proposed method achieves a superior performance compared to any individual model incorporated into our ensemble, and even exceeds the top-performing baseline Cascade R-CNN on the SOD4SB public test dataset. Moreover, we offer a series of analysis to validate the effectiveness of each training, inference, and ensembling technique. The contributions of this work are summarized as follows:

- Implementation of different SOD training strategies, inference techniques, and ensemble methods.
- Evaluation of multiple ensemble fusion methods and their effectiveness on the SOD4SB dataset.
- A detailed analysis of the efficacy of the SOD strategies and techniques in ensemble methods.

2 Preliminary and Related Work

2.1 General-Purpose Object Detection Models

In the past decade, object detection has witnessed significant advancements driven by the use of deep neural network models. These models can be broadly categorized into two groups: one-stage and two-stage methods. The former [7, 8, 11–14, 18, 20, 21, 23] generally offers faster performance, while the latter [6, 9, 15–17, 19] tends to achieve greater accuracy. In SOD4SB, we employ two distinct architectures: Cascade R-CNN [6] and CenterNet [7], along with multiple backbones [10, 22]. Cascade R-CNN is utilized to yield more accurate predictions, while CenterNet is incorporated into our ensemble technique to further enhance the overall performance. They are described as follows.

Cascade R-CNN [6]. Cascade R-CNN is an extension of the Faster R-CNN [19] architecture, and is designed to address the issue arising from the use of a single Intersection over Union (IoU) threshold in object detection. Instead of employing a single R-CNN head, multiple R-CNN heads with incrementally increasing IoU thresholds are utilized in Cascade R-CNN. This enables the generation of progressively higher-quality predictions while effectively reducing false positives.

CenterNet [7]. CenterNet is a one-stage object detection architecture that relies on keypoint-based detection. Within CenterNet, a heatmap is initially generated using a fully convolutional network. The peaks detected within the heatmap serve as indicators for the centers of objects. In each corresponding location, CenterNet also predicts the center offsets and the sizes of the object to derive its accurate location and dimension. By utilizing keypoint estimation, CenterNet eliminates the need for additional post-processing techniques such as non-maximum suppression (NMS).

2.2 Small Object Detection Techniques

Various techniques have been developed to tackle the challenges inherent in SOD tasks. These techniques either adapt existing object detection methods or introduce new architectures to enhance performance. In this work, we adopt the techniques most suited for SOD4SB. These techniques are described as follows.

Normalized Wasserstein Distance (NWD) [33]. NWD is an approach developed for quantifying the degree of overlap between two bounding boxes, specifically targeting the issues encountered in anchor-based object detection techniques that rely on the IoU metric to evaluate the overlap between ground truth and predicted bounding boxes. While the IoU metric can be effective in standard cases, it can be overly sensitive to small object displacements within these methods. In extreme yet frequent cases where no overlap exists between the bounding boxes, IoU would result in a score of zero, which highlights the need for a more robust measure. To address this, NWD transforms bounding boxes into 2D Gaussian distributions and subsequently normalizes them. The Wasserstein distance employed in NWD effectively measures the distance between two probability distributions, irrespective of the presence or absence of overlap between them.

Copy-Paste (CP) [32]. CP is a method designed to enhance data efficiency. It involves selecting objects from one image, pasting them onto another one, while concurrently resizing them in a randomized manner.

Slicing Aided Hyper Inference (SAHI) [24]. SAHI is a technique developed to enhance the detection accuracy of high-resolution images containing small targets, as it increases the area ratio of objects relative to the image. During inference, SAHI divides the image into multiple regions using sliding windows, with each region and the entire image being processed separately for predictions. The predictions for each region are then combined and filtered using NMS to derive the final prediction outcome.

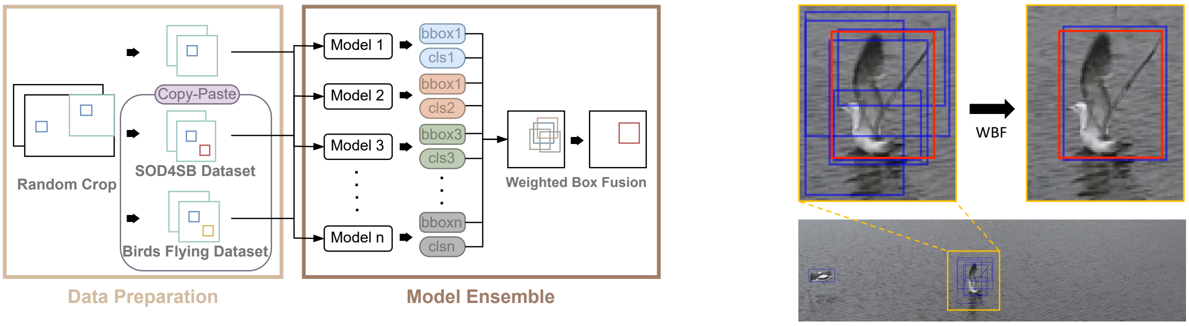
3 Methodology

3.1 Overview of the Framework

Fig. 1 (a) illustrates an overview of our proposed framework, which consists of two distinct stages: the *data preparation stage* and the *model ensemble stage*. In the *data preparation stage*, we utilize the CP data augmentation technique discussed in Section 2.2 to enrich the training data provided by SOD4SB. In this stage, images from the SOD4SB dataset undergo cropping and augmentation with birds sourced from either the SOD4SB dataset or the Birds Flying dataset [40]. The augmented data are subsequently forwarded to the *model ensemble stage*, where several model variants are developed and grouped together to form an ensemble. To leverage the strengths of various model variants, our framework incorporates the WBF as the ensembling method. By combining the predictions from different model variants, the WBF technique generates more precise final bounding box predictions.

3.2 Weighted Box Fusion (WBF) [39]

WBF is a technique for combining models with varying levels of bounding box prediction accuracy. By assigning a weight to each model variant within the



(a) Overview of the framework

(b) Visualization of WBF

Figure 1: (a) Overview of the framework; (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).

framework, WBF calculates the proposed bounding box coordinates using a weighted sum derived from the predictions of the model variants. Unlike NMS which selects only one box from a set of bounding boxes, WBF fuses all boxes based on an IoU threshold. Specifically, WBF ranks the predictions from different model variants according to their weights, and iteratively fuses the bounding boxes from higher-ranked models with lower-ranked ones in a continuous updating process. Predictions with insufficient IoU overlap with the current running prediction are discarded. This method is especially suitable for detecting small objects, where predicting precise bounding box positions is challenging, as it uses the information from each model variant. Fig. 1 (b) depicts the impact of WBF, where the bounding boxes from different predictions are ensembled, resulting in a more accurate prediction.

3.3 Ensemble Model Variant Preparation

To enhance the overall accuracy of the framework, a number of methods are implemented to generate diverse model variants, which serve as members of the ensemble. This process can be explained further for the training and inference phases. During training, the incorporation of various backbones and an assortment of augmentation methods strengthens the diversity of the model variants, which results in enhanced feature extraction capabilities. In the inference phase, the trained model variants are further subjected to diverse processing methods, such as SAHI and test time augmentation approaches. These methods enable the model variants to perceive objects at varying scales as well as augmenting the diversity of the input data. By employing these techniques, the overall prediction accuracy of the ensemble can be significantly enhanced.

3.3.1 Training Stage

Model Backbones. As a robust backbone is crucial for attaining high performance, two model backbones, ResNet [10] and InternImage [22], are selected. InternImage is a large-scale, convolutional neural network (CNN)-based backbone, which demonstrates superior adaptability across multiple datasets. Its DCNv3 operator achieves comparable results to the Vision Transformer [41], but with a reduced model size. In this study, InternImage-XL and InternImage-H are utilized to generate the ensemble model variants.

Table 1: The datasets used for model training in this study.

Dataset	# of Images	# of Birds	Img. Res.	Avg. BBox Res.
SOD4SB [5] Train	9,759	29,037	3840 × 2160	21 × 18 px
Drone [4]	47,260	60,971	3840 × 2160	37 × 31 px

Data Augmentation. To address the challenge of detecting small objects with sparse occurrences, this study adopts two data augmentation methods. The goal of these methods is to increase the number of objects, thereby enhancing the framework's generalizability and accuracy in detecting such object. The first method utilizes the Birds Flying dataset, which consists of 3,600 images of flying birds, to boost the object count. The second method involves copying objects from the SOD4SB dataset and superimposing them onto the training images, as illustrated in Fig. 1 (a).

3.3.2 Inference Stage

In the inference phase, the models trained in the preceding training phase are subjected to various methods, including SAHI and several TTA techniques such as image scaling and random flipping. These methods enable the alteration of input images, which in turn contributes to the enhancement of the prediction accuracy for the entire framework.

4 Experimental Results

4.1 Experimental Setup

Datasets. Table 1 outlines the SOD4SB [5] and Drone [4] datasets, which are provided by the Small Object Detection Challenge for Spotting Birds, and are utilized for pre-training, fine-tuning, and evaluating our models. The images from these datasets are augmented using the techniques described in Section 3.

Baselines. In this study, DetectoRS [17], CenterNet, and Cascade R-CNN are selected as the baseline models. DetectoRS is recognized for its notable performance on the Tiny Object Detection in Aerial Images Dataset [42], while CenterNet serves as the provided baseline for the SOD4SB challenge. On the other hand, Cascade R-CNN demonstrates commend-

Table 2: 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 [17]	ResNet-50	0.483	0.346	0.038
	CenterNet [7,43]	ResNet-18	0.616	0.491	0.071
	Cascade R-CNN [6]	ResNet-50	**0.631**	**0.533**	**0.108**
	Ensemble Method		AP@.25	**AP@.50**	AP@.75
(b)	Top-Performing Single Model		0.803	0.737	0.183
	Pure NMS with no weight		0.673	0.616	0.195
	Weighted NMS		0.814	0.751	0.193
	Soft NMS		0.797	0.739	0.208
	WBF (Ours)		**0.840**	**0.776**	**0.225**

Table 3: Comparison of the techniques in Section 3.3.1.

Model Architecture	CP	Backbone Network	NWD	AP@.25	**AP@.50**	AP@.75
Cascade R-CNN	✗	ResNet-50	✗	0.631	0.533	0.108
Cascade R-CNN	✗	ResNet-50	✓	0.774	0.681	0.158
Cascade R-CNN	✓	ResNet-50	✓	0.745	0.650	0.162
Cascade R-CNN	✗	InternImage-XL	✗	0.590	0.531	0.133
Cascade R-CNN	✗	InternImage-XL	✓	0.785	0.713	**0.190**
Cascade R-CNN	✗	InternImage-H	✓	**0.798**	**0.721**	0.182

able performance across various object detection tasks. The results for the baselines are listed in Table 2 (a).

Hyperparameter Setups. The evaluation of both the baselines and the proposed framework is conducted on the SOD4SB testing dataset. These models undergo pre-training on the Drone dataset for 140 epochs, followed by fine-tuning on the SOD4SB training dataset for an additional 40 epochs. The evaluation metric employed is the average precision score (AP) for the single bird class. Please note that the challenge utilizes an AP@0.5 score for evaluation purposes, and AP@(0.25, 0.75) scores are also provided for reference.

4.2 Quantitative Results

The "WBF" entry in Table 2 (b) reports the AP scores of our final model, which is based on Ensemble Fusion and incorporates various techniques during training and inference as described in Section 3. A notable observation is that our method surpasses all baselines in terms of AP scores. Our model even outperforms the best-performing baseline by approximately 45.59%, and achieves an AP@(0.50) score of 0.776.

4.3 Ablation Analysis

Training Strategies. In this section, we examine the effectiveness of various techniques outlined in Section 3.3.1 for generating model variants, with the results presented in Table 3. It is observed that incorporating NWD during training significantly enhances model accuracy. Moreover, the experimental results reveal that InternImage-H appears to be a more effective backbone compared to ResNet-50 and InternImage-XL. Furthermore, while the results show that applying the CP technique for data augmentation alone does not yield improvements, we discovered that ensembling models trained with the CP technique is able to boost accuracy. This advantage may be attributed to the increased data diversity introduced by the CP technique.

Table 4: Comparison of the techniques described in Section 3.3.2 in the inference stage. The "Final" column indicates whether the model is selected in the final ensemble.

Model Architecture	TTA	SAHI	AP@.25	**AP@.50**	AP@.75	Final
CenterNet + ResNet-18	✗	✗	0.616	0.491	0.071	✗
	✓	✗	0.636	0.514	0.076	✗
	✗	✓	0.606	0.487	0.066	✓
Cascade R-CNN + ResNet-50 + NWD	✗	✗	0.774	0.681	0.158	✓
	✓	✗	0.782	0.692	0.164	✗
	✗	✓	0.743	0.631	0.147	✓
Cascade R-CNN + ResNet-50 + NWD + CP	✗	✗	0.745	0.650	0.162	✓
	✓	✗	0.760	0.667	0.173	✗
	✗	✓	0.710	0.604	0.152	✓
Cascade R-CNN + InternImage-XL + NWD	✗	✗	0.785	0.713	0.190	✓
	✓	✗	0.790	0.725	**0.198**	✓
	✗	✓	0.780	0.676	0.158	✓
Cascade R-CNN + InternImage-H + NWD	✗	✗	0.798	0.721	0.182	✗
	✓	✗	**0.803**	**0.737**	0.183	✓
	✗	✓	0.790	0.713	0.184	✓

Inference Strategies. Table 4 presents the AP scores resulting from the application of various techniques described in Section 3.3.2 during the inference phase. It suggests that utilizing TTA enhances the AP score by approximately 2% ∼ 3%. In addition, the results reveal that employing SAHI may negatively impact performance. The "Final" column represents the models that are selected for inclusion in the ensemble.

Ensembling Methods. To determine the most effective ensembling method, we compare various approaches, with the outcomes reported in Table 2 (b). The AP scores of the top-performing single model (i.e., Cascade R-CNN + InternImage-H + NWD + TTA) are listed for reference. It is observed that ensembling methods significantly impact performance, but directly applying NMS to all models leads to a reduction in scores. This might result from interference caused by inferior models' bounding boxes with the more accurate ones derived from superior models. The results also reveal that employing weighted ensembling methods lead to improved performance, with WBF outperforming the other ensembling approaches.

5 Conclusion and Potential Applications

The ensemble fusion approach presented in this paper effectively addresses the challenges associated with detecting small objects by leveraging the strengths of multiple diverse model variants and fusing their predictions. The experimental results indicated that our approach is able to yield enhanced accuracy and robustness in small object detection, and achieve the highest score in the MVA2023 Challenge on Small Object Detection for Birds. Potential applications of the proposed method include wildlife monitoring, autonomous vehicle systems, robotics, and surveillance systems.

Acknowledgements

The authors acknowledge the National Science and Technology Council in Taiwan (grant number MOST 111-2223-E-007-004-MY3), NVIDIA Corporation and NVIDIA AI Technology Center for the donated GPUs, and the National Center for High-Performance Computing for the computational and storage resources.

References

[1] F. Pérez-Hernández, S. Tabik, A. Lamas, R. Olmos *et al.*, "Object detection binary classifiers methodology based on deep learning to identify small objects handled similarly: Application in video surveillance," *Knowledge-Based Systems*, vol. 194, p. 105590, 2020.

[2] M. Omachi and S. Omachi, "Traffic light detection with color and edge information," in *Proc. IEEE Int. Conf. on Computer Science and Information Technology*, Aug. 2009, pp. 284–287.

[3] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian, "Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1758–1770, 2020.

[4] S. Fujii, K. Akita, and N. Ukita, "Distant bird detection for safe drone flight and its dataset," in *Proc. Int. Conf. on Machine Vision Applications (MVA)*, 2021.

[5] Y. Kondo, N. Ukita, T. Yamaguchi, H.-Y. Hou *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)*, 2023, https://www.mva-org.jp/mva2023/challenge.

[6] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6154–6162.

[7] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv:1904.07850*, 2019.

[8] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," *arXiv:2107.08430*, 2021.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[11] C. Li, L. Li, H. Jiang, K. Weng *et al.*, "YOLOv6: A single-stage object detection framework for industrial applications," *arXiv:2209.02976*, 2022.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy *et al.*, "SSD: Single shot multibox detector," in *Proc. European Conf. on Computer Vision (ECCV)*. Springer, 2016, pp. 21–37.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[14] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Int. Journal of Computer Vision*, vol. 128, 2020, pp. 642–656.

[15] Z. Li, C. Peng, G. Yu, X. Zhang *et al.*, "Light-Head R-CNN: In defense of two-stage object detector," *arXiv: 1711.07264*, 2017.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He *et al.*, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944.

[17] S. Qiao, L.-C. Chen, and A. Yuille, "DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 208–10 219.

[18] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv:1804.02767*, 2018.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, 2015.

[20] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 9626–9635.

[21] C.-Y. Wang, A. Bochkovskiy, and M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv:2207.02696*, 2022.

[22] W. Wang, J. Dai, Z. Chen, Z. Huang *et al.*, "InternImage: Exploring large-scale vision foundation models with deformable convolutions," *arXiv:2211.05778*, 2022.

[23] S. Xu, X. Wang, W. Lv, Q. Chang *et al.*, "PP-YOLOE: An evolved version of YOLO," *arXiv:2203.16250*, 2022.

[24] F. C. Akyon, S. Onur Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," in *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, 2022, pp. 966–970.

[25] Z. Liu, G. Gao, L. Sun, and Z. Fang, "HRDNet: High-resolution detection network for small objects," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[26] J. Ding, N. Xue, G.-S. Xia, X. Bai *et al.*, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7778–7796, 2022.

[27] Y. Gong, X. Yu, Y. Ding, X. Peng *et al.*, "Effective fusion factor in fpn for tiny object detection," in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, 2021, pp. 1160–1168.

[28] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2021, pp. 2778–2788.

[29] D. Wang, Q. Zhang, Y. Xu, J. Zhang *et al.*, "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[30] Y. Zhao, L. Zhao, Z. Liu, D. Hu *et al.*, "Attentional feature refinement and alignment network for aircraft detection in sar imagery," *IEEE Trans. on Geoscience and Remote Sensing*, vol. PP, pp. 1–1, 12 2021.

[31] Z. Wei, D. Liang, D. Zhang, L. Zhang *et al.*, "Learning calibrated-guidance for object detection in aerial images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 2721–2733, 2022.

[32] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian *et al.*, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2917–2927.

[33] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized gaussian wasserstein distance for tiny object detection," *arXiv:2110.13389*, 2021.

[34] C. Xu, J. Wang, W. Yang, H. Yu *et al.*, "RFLA: Gaussian receptive field based label assignment for tiny object detection," in *Proc. European Conf. on Computer Vision (ECCV)*. Springer, 2022, pp. 526–543.

[35] C. Lee, S. Park, H. Song, J. Ryu *et al.*, "Interactive multi-class tiny-object detection," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 116–14 125.

[36] J. Yi, P. Wu, B. Liu, Q. Huang *et al.*, "Oriented object detection in aerial images with box boundary-aware vectors," in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, 2021, pp. 2150–2159.

[37] J. Noh, W. Bae, W. Lee, J. Seo, and G. Kim, "Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection," in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 9724–9733.

[38] W. Yang, X. Zhang, Y. Tian, W. Wang *et al.*, "Deep learning for single image super-resolution: A brief review," *IEEE Trans. on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.

[39] 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.

[40] "Birds flying dataset," www.kaggle.com/datasets/nelyg8002000/birds-flying, 2021.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. IEEE /CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3156–3164.

[42] J. Wang, W. Yang, H. Guo, R. Zhang, and G.-S. Xia, "Tiny object detection in aerial images," in *Proc. Int. Conf. on Pattern Recognition (ICPR)*, 2021, pp. 3791–3798.

[43] K. Zhao, R. Miyata, Y. Kondo, and K. Akita, "Baseline code for SOD4SB by IIM-TTIJ," 2023. [Online]. Available: https://github.com/IIM-TTIJ/MVA2023SmallObjectDetection4SpottingBirds