WS-YOLO: A single-stage detector designed for aircraft detection by enhancing fine-grained segmentation

Yuzhu Lei[1], Jun Li[2], Lei Zhang[1], Guoming Song[2]
leiyuzhu@stu.cdut.edu.cn, suiyuanlj2006@gmail.com
edu_1313299@163.com, songwell@cdtu.edu.cn
[1]The Fifth Research Institute of telecommunications technology Co.Ltd., Chengdu, China
[2]School of Computer Engineering College, Chengdu Technological University, Chengdu, China

Abstract. With the advancement of aerospace technology, small object detection in satellite remote sensing has become a major research focus globally, with aircraft detection regarded as a fundamental task. Advances in sensor technology enable the capture of more detailed aircraft features, yet simultaneously make fine-grained detection increasingly challenging. To address this, this paper proposes WS-YOLO, a novel architecture incorporating the Enhanced Fine-grained Module. Experimental results on the CORS-ADD and MAR20 datasets show that WS-YOLO achieves mPA improvements of 0.44%–4.95% and 0.44%–1.78%, respectively, compared to six existing YOLO variants, demonstrating its effectiveness in detecting small aircraft targets under high fine-grained conditions.

Keywords: Aircraft Detection , Remote Sensing Images , YOLO

1 Introduction

Remote sensing images (RSIs) provide a reliable source of data for aircraft detection, owing to their extensive spatial coverage and ability to capture detailed target-level features [1]. Compared with conventional imaging modalities, satellite-based RSIs can effectively capture aircraft in typical scenarios such as airports and aprons, thereby improving detection accuracy and robustness. Consequently, aircraft detection using satellite remote sensing imagery has attracted significant research attention and found widespread applications in resource monitoring, urban planning, environmental conservation and natural disaster early warning [2].

Aircraft are strategically important yet physically small targets, posing significant challenges for precise identification in object detection tasks. The ability to accurately localize and classify such small objects is a key performance metric for evaluating aircraft detection systems [3]. Although advances in sensor technology and the availability of high-resolution RSIs have provided richer visual details, they have also introduced greater complexity in fine-grained aircraft detection due to increased scene clutter and background interference.

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365272

Traditional aircraft detection methods are largely based on handcrafted feature engineering and heuristic parameter tuning. In recent years, deep learning–based object detectors—such as one-stage models and two-stage frameworks —have become dominant approaches [4]. Nevertheless, achieving accurate fine-grained aircraft detection in remote sensing imagery remains challenging. Unlike generic object detection, this task demands enhanced discrimination capability to distinguish small aircraft from complex backgrounds under varying imaging conditions.

The principal contributions of this paper are as follows:

- We propose WS-YOLO, a YOLO-based architecture that effectively balances global contextual modeling and local aircraft feature extraction, thereby improving detection performance in cluttered scenes.
- We introduce an Enhanced Fine-grained Module that suppresses irrelevant background features during feature extraction, thus enhancing the model's sensitivity to subtle aircraft characteristics.
- Extensive experiments on the CORS-ADD and MAR20 datasets demonstrate that WS-YOLO outperforms six state-of-the-art YOLO variants in terms of both mAP and F1 score.

2 Related Work

Xu et al. [5] addressed the challenges of small target sizes in remote sensing imagery, where existing methods commonly suffer from low recall and high false-negative rates. They proposed a deep learning-based aircraft detection algorithm based on YOLOv8. Ouyang et al. [6] noted that object detection is challenging due to minimal inter-class feature differences. To address this, they proposed a Prototype Contrastive Learning Detector for fine-grained object detection in remote sensing imagery. Jiang et al. [7] designed the ASCPA-EMF network to address feature sparsity, incomplete feature representation, high intra-class variation, and small inter-class differences in fine-grained aircraft recognition within SAR imagery. They integrated this network with YOLOv8n to construct the EM-YOLO framework for fine-grained recognition. Vasavi S et al. [8] proposed an enhanced CNN model to address the typically low inference speed and low accuracy in aircraft detection from optical images.

3 Methodology

3.1 WS-YOLO

This section introduces WS-YOLO, the proposed aircraft detection framework. Section III-A-a describes its overall architecture, and Section III-A-b details the Enhanced Fine-grained Module.

3.1.1 Structure

As shown in Fig. 1, WS-YOLO is a one-stage, convolutional neural network–based object detector tailored for aircraft detection in RGB remote sensing imagery. The model operates in an

end-to-end manner and comprises three standard components: a backbone, a neck, and a detection head.

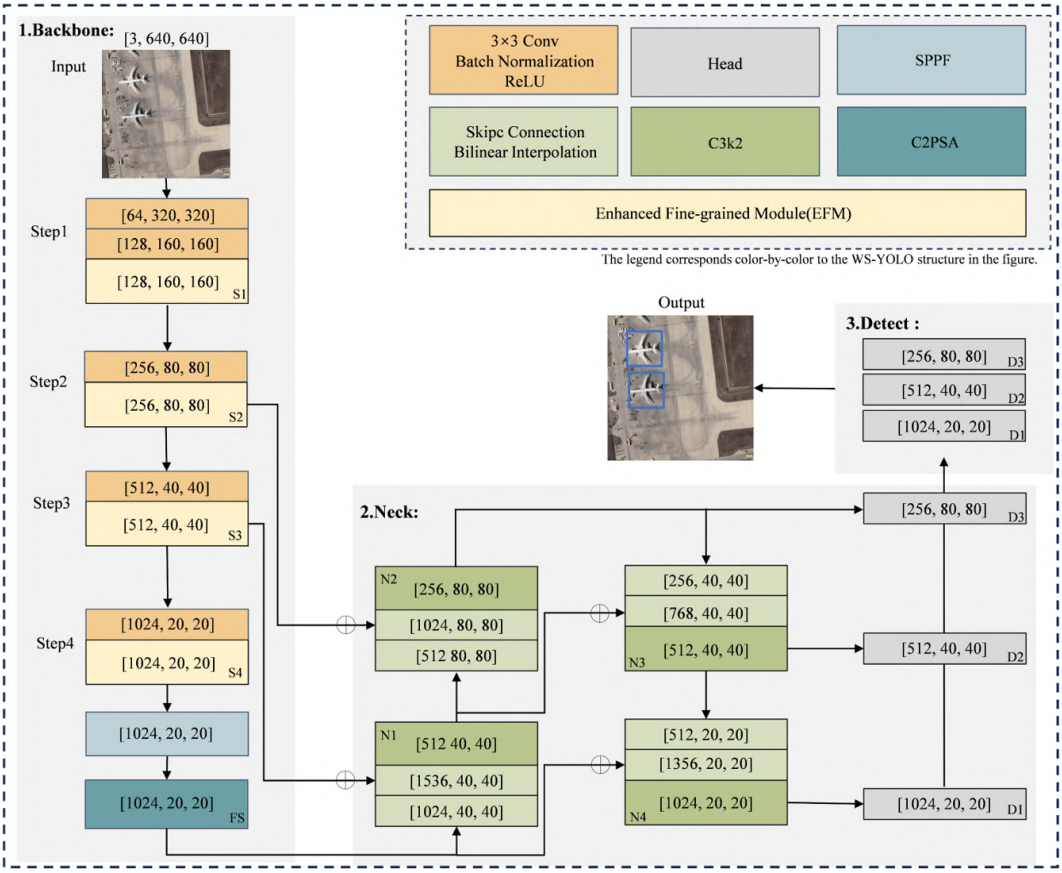

Fig. 1. WS-YOLO Architecture

WS-YOLO learns discriminative aircraft features $X \in R[C,H,W]$, where $H \times W$ denotes the spatial dimensions of the feature map and C represents the number of channels. The backbone network performs iterative down sampling through four stages, producing intermediate feature maps S_1, S_2, S_3, S_4, for $i = 1,2,3,4$. Subsequently, the SPPF and C2PSA modules are applied to S_4 to extract multi-scale contextual features, which serve as the high-level representation, denoted as FS.

The feature maps FS, S_2, S_3 are fed into the Neck for multi-scale fusion. The high-level feature FS undergoes successive up sampling $N_i \in [1,2,3,4]$ to restore it to the same dimensions as the specified input. Notably, skip connections are applied between S_2 and N_2, S_3 and N_1, FS and N_4, and N_1 and N_3 to better fuse aircraft features across multiple scales.

Finally, the feature maps from N_2, N_3, and N_4 are fed into three detection heads to predict air-

craft at multiple scales: $H_1 \in R[4C, H/8, W/8]$ for small-scale objects, $H_2 \in R[8C, H/16, W/16]$ for medium-scale objects, $H_3 \in R[16C, H/32, W/32]$ for large-scale objects. These multi-scale features are processed by task-specific heads to generate the final detection outputs.

3.1.2 Enhanced Fine-grained Module

Aircraft in remote sensing imagery exhibit significant scale variation, and their fine details are often obscured by complex backgrounds. Recent efforts [5]–[8] have demonstrated that enhancing fine-grained feature representation is crucial for detecting small aircraft in cluttered scenes. To address these challenges, we propose the Enhanced Fine-grained Module (EFM). Specifically, EFM incorporates dense skip connections that concatenate shallow, high-resolution features with deeper semantic representations, thereby preserving fine spatial details during multi-scale fusion. Additionally, the Enhanced Block preserves channel-wise contextual information to enrich feature representation. As shown in Fig. 2, the EFM comprises three core components: (1) a Fine-grained Block that captures local structural cues through dense feature propagation, (2) an Enhanced Block that models channel-wise dependencies via contextual recalibration, and (3) a 1×1 convolutional layer that adjusts the channel dimensionality for computational efficiency without compromising representational power.

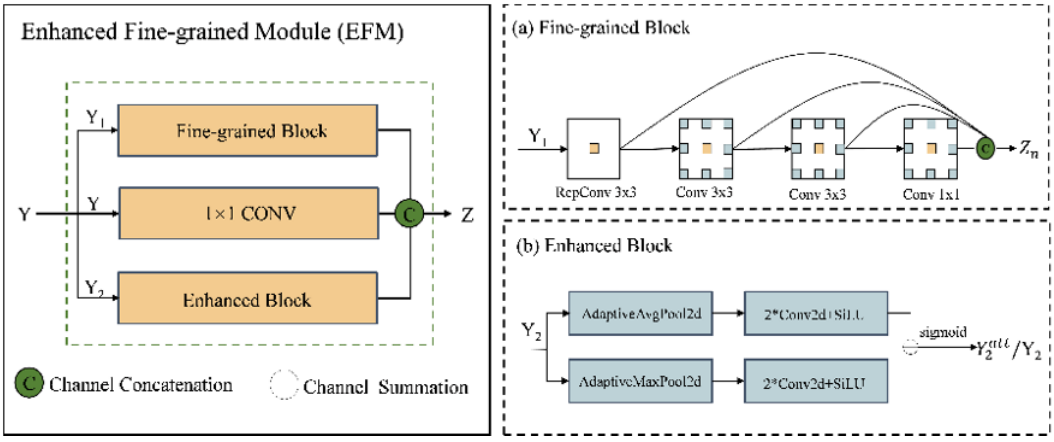

Fig. 2. Enhanced Fine-grained Module Structural Diagram

Perform a splitting operation on the input feature $Y \in R^{B\times 2C\times H\times W}$, decomposing it into two branches Y_1 and Y_2, where $Y_1 \in R^{B\times C\times H\times W}$ and $Y_2 \in R^{B\times C\times H\times W}$. This approach reduces the computational load of parameters without compromising feature accuracy, as illustrated in (1).

$$Y = [Y_1 : Y_2] \tag{1}$$

Progressive feature refinement is applied to the first branch Y_1. Specifically, Y_1 is first processed by a RepConv layer to obtain the initial fine-grained feature Z_0. The RepConv layer typically

integrates a 3×3 convolution, 1×1 convolution, and an identity connection within a single convolutional block. Subsequently, deep feature extraction is performed through a densely connected block comprising $n-1$ convolutional layers.

Let Z_k denote the output of the k-th intermediate convolutional layer with a total of $n-1$ such layers. The output Z_k is obtained by applying a 3×3 convolution to the output Z_{k-1} of the previous layer, progressively extracting deeper features while mitigating the loss of small-object features. Finally, a 1×1 convolutional layer is employed to fuse the channels of the resulting feature map, yielding the final output of this branch Z_n, as expressed in (2) and (3).

$$Z_k = F_{Conv}^{k} Z_{k-1}, k = 1, \ldots, n \tag{2}$$

$$Z_n = F_{Conv_{1\times1}}(Z_{n-1}) \tag{3}$$

Z_n represents the final feature map of this branch, which enhances the representational capability of the features by incorporating cross-channel information mixing.

Channel-wise attention is applied to the second branch Y_2 to adaptively emphasize informative channel features while suppressing irrelevant or noisy channels. Specifically, adaptive average pooling and adaptive max pooling are performed on Y_2 to aggregate spatial information. The resulting pooled features are then fed into a multilayer perceptron composed of two convolutional layers interleaved with SiLU functions. Subsequently, a channel attention weight vector is generated by applying the Sigmoid activation function. Finally, this weight vector is multiplied element-wise with each channel of the original feature Y_2 to produce the enhanced feature representation. The expression formulae are presented in (4).

$$Y_2^{att} = Y_2 \odot \sigma(MLP(Avg(Y_2) + Max(Y_2))) \tag{4}$$

Here, $\odot$ denotes element-wise multiplication, σ represents the Sigmoid activation function. Finally, the aforementioned six feature maps are concatenated along the channel dimension to produce the output of the Enhanced Fine-grained Module, as shown in (5).

$$Z = ||_{k=0}^{n} Z_k \tag{5}$$

Here, $||$ denotes feature layer fusion and concatenation. Z_k represents the output of the k-th intermediate convolutional layer. Z denotes the feature map extracted by the Enhanced Fine-grained Module.

3.2 Dataset

To investigate research methods for aircraft detection in remote sensing images, this paper selects two aircraft datasets: the CORS-ADD dataset and the MAR20 dataset.

CORS-ADD Dataset: The CORS-ADD dataset [9] collects 5,486 high-resolution remote sensing images from multiple satellite sensors, including WorldView series, Jilin-1, IKONOS, and Google Earth, containing a total of 32,285 aircraft instances. Most images have a resolution of

640×640 pixels. We partition the dataset into training, validation, and test sets in a 6:2:2 ratio, resulting in 3,762, 862, and 862 image patches, respectively. Due to its multi-source sensor origins, the dataset exhibits diverse and complex scene characteristics. Sample images are shown in Fig. 3.

Fig. 3. Partial samples of the CORS-ADD Dataset.

MAR20 Dataset: The MAR20 dataset [10] comprises 3,842 high-resolution remote sensing images collected via Google Earth from 60 airports across countries such as the United States and Russia. The majority of these images are sized at 800×800 pixels, with a small portion being 900×900 pixels. We partition this dataset into training, validation, and test sets in a ratio of 6:2:2, resulting in 2,306, 768, and 768 image patches respectively. This dataset is characterized by its high level of fine granularity. Sample images are shown in Fig. 4.

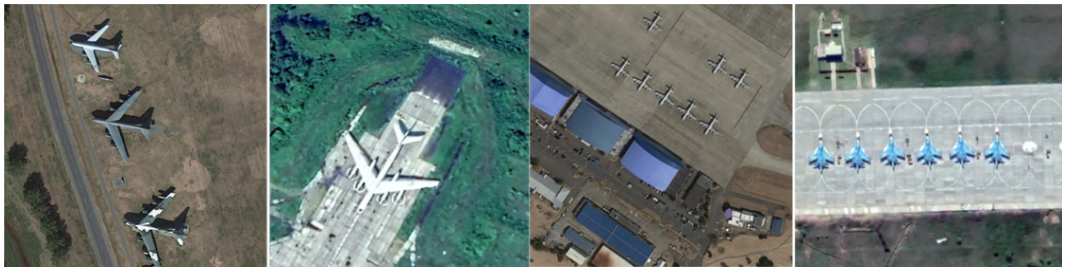

Fig. 4. Partial samples of the MAR20 Dataset.

4 Experimental and Analysis

4.1 Experimental setup

The experiments were conducted on an NVIDIA GeForce RTX 2060 (12GB) GPU and an Intel Core i5-12400F CPU. Both the baseline model and WS-YOLO were trained under identical experimental configurations. The batch size was set to 8, and the number of epochs was set to 100. Input size is specified in Section III-B. The learning rate was set to 0.01 with a weight decay of 0.0005. Random seed was set to 2025. Stochastic Gradient Descent (SGD) served as the optimizer for the

entire network, with the SGD momentum set to 0.9. Furthermore, Bounding Box Loss [11], Classification Loss [12], and Distribution Focal Loss [13] were employed as loss functions, collectively forming the aircraft detection loss function. Following each training iteration, testing was conducted on the validation set, with the most effective weights selected to evaluate their performance on the test set. Notably, all models were trained from scratch without pre-trained weights or data augmentation to ensure a fair and controlled comparison.

4.2 Evaluation Metrics

To more intuitively present the experimental results of the proposed method, this paper selects four commonly used evaluation metrics in the field of object detection: Precision, Recall, mean Average Precision, and F1-score, as defined in (6) to (9). These four evaluation metrics are also widely adopted in aircraft detection tasks. Therefore, this paper employs these four metrics to comprehensively analyze the experimental results.

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

$$mAP = \frac{AP_1 + AP_2 + AP_3 + ... + AP_n}{n} \tag{8}$$

$$F1score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

4.3 Ablation Study

On CORS-ADD, recall improves by 1.74% and F1-score by 0.59%; on MAR20, recall increases by 1.18% and F1-score by 0.96%, as summarized in Table 1. These gains indicate a notable reduction in missed detections while maintaining precision, which is further supported by a slight improvement in mAP. The performance boost is attributed to our newly introduced EFM Module, which suppresses distracting background features during feature extraction to better capture subtle details critical for fine-grained aircraft detection. Importantly, this enhancement incurs only a modest increase in computational cost and maintains a parameter count nearly identical to that of the baseline, ensuring that the proposed model retains real-time inference capability.

4.4 Detection results for the CORS-ADD Dataset

The numerical results of the CORS-ADD Dataset are shown in Table 2. The comprehensive values of the WS-YOLO model are superior to those of six baseline networks. Specifically, the overall performance in recall improved by 1.74% to 7.14%, while the overall performance in mAP increased by 0.44% to 4.95%. Regarding the F1 score, the overall performance improved by 0.59%

Table 1: ABLATION STUDY on the MAR20 Dataset AND the ON CORS-ADD Dataset

Dataset	Methods	Precision(%)	Recall(%)	mAP(%)	F1(%)	GFLOPs	FPS	Parameters
CORS-ADD	Baseline	92.94	77.74	84.98	84.66	6.3	267.07	2,582,347
	Baseline +EFM	91.91	79.48	85.42	85.25	6.8	223.20	2,490,371
MAR20	Baseline	97.48	95.04	98.40	96.25	6.3	242.76	2,582,347
	Baseline +EFM	98.22	96.22	98.92	97.21	6.8	180.87	2,490,371

to 5.00%. It is evident that the proposed model enhances aircraft detection accuracy without increasing the number of parameters, while maintaining low computational cost of 6.8 GFLOPs and high inference speed of 267 FPS.

Fig. 5 displays representative aircraft detection results. Compared with six baseline networks, the WS-YOLO model yields more accurate and comprehensive detections. As shown in the first and fourth rows of Fig. 5, the proposed model successfully detects small-scale aircraft targets that other models miss. In the second and third rows of Fig. 5, it demonstrates strong discriminative capability against background clutter and ground objects that visually resemble aircraft. This advantage stems from the Enhanced Fine-grained Module (EFM) integrated into the WS-YOLO architecture. Overall, WS-YOLO exhibits superior visual detection performance relative to all six baselines, maintaining high accuracy and robustness even in scenes with abundant fine-grained aircraft details.

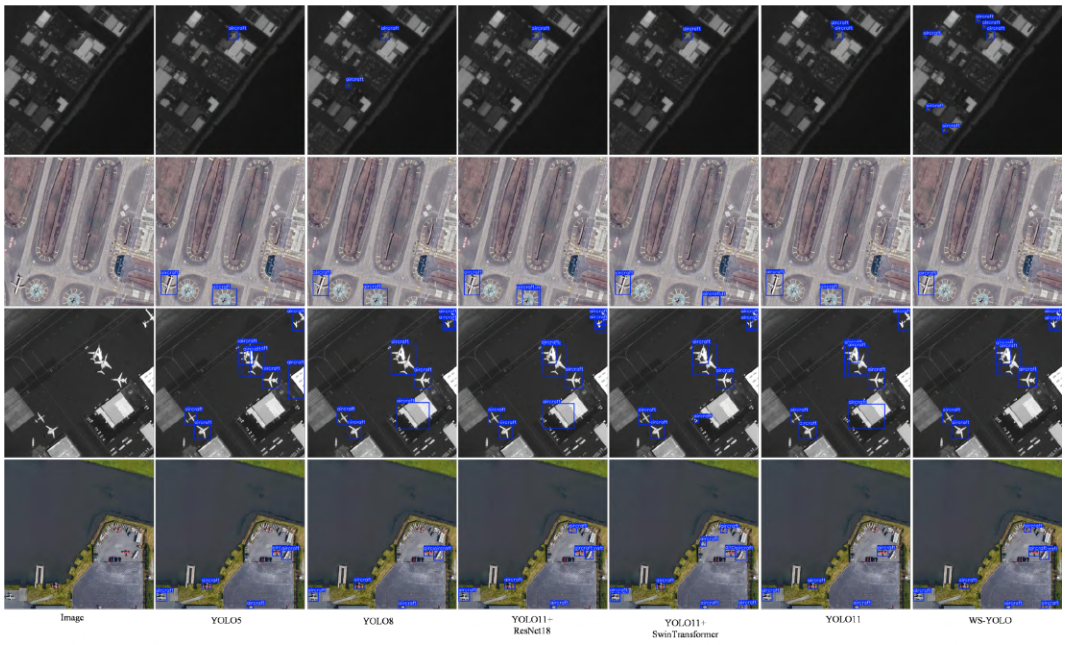

Fig. 5. Partial aircraft detection results on the CORS-ADD Dataset.

Table 2: Quantitative Comparison of Different Models on the CORS-ADD Dataset

Methods	Precision(%)	Recall(%)	mAP(%)	F1(%)	GFLOPs	FPS	Parameters
YOLO5	90.63	73.26	80.47	81.03	6.1*	140.09	**2,448,090**
YOLO8	90.82	74.96	81.92	82.13	**5.8**	165.47	2,508,539
YOLO11+ResNet18	91.73	75.35	82.56	82.74	33.6	119.19	13,056,003
YOLO11+SwinTransformer	90.11	72.34	79.67	80.25	77.6	35.42	29,715,709
YOLO11	**92.94**	77.74*	84.98*	84.66*	6.3	**267.07**	2,582,347
WS-YOLO (ours)	91.91*	**79.48**	**85.42**	**85.25**	6.8	223.20*	2,490,371*

Note: Models in bold denote the optimal numerical values, while those marked with an asterisk indicate the second-ranked models numerically.

Table 3: Quantitative Comparison of Different Models on the MAR20 Dataset

Methods	Precision(%)	Recall(%)	mAP(%)	F1(%)	GFLOPs	FPS	Parameters
YOLO5	96.43	91.57	97.14	93.94	6.1*	100.93	**2,448,090**
YOLO8	96.55	94.44	98.18	95.48	**5.8**	128.86	2,508,539
YOLO11+ResNet18	96.50	95.27*	98.48*	95.88	33.6	115.16	13,056,003
YOLO11+SwinTransformer	96.09	93.70	97.87	94.88	77.6	24.76	29,715,709
YOLO11	97.48*	95.04	98.40	96.25*	6.3	**242.76**	2,582,347
WS-YOLO (ours)	**98.22**	**96.22**	**98.92**	**97.21**	6.8	180.87*	2,490,371*

Note: Models in bold denote the optimal numerical values, while those marked with an asterisk indicate the second-ranked models numerically.

4.5 Detection results for the MAR20 Dataset

The numerical results on the MAR20 Dataset are presented in Table 3. The comprehensive performance of the WS-YOLO model surpasses that of the six baseline networks. Specifically, the overall recall improved by 0.95% to 4.65%, and the overall mAP increased by 0.44% to 1.78%. In terms of F1 score, the overall performance improved by 0.96% to 3.27%. It is evident that the proposed model enhances aircraft detection accuracy without increasing the number of parameters, while maintaining low computational cost of 6.8 GFLOPs and high inference speed of 242 FPS.

Fig. 6 displays representative aircraft detection results. Compared to the six baseline networks, WS-YOLO detects more aircraft instances with fewer missed targets. As clearly shown in the second, third, and fourth rows of Fig. 6, the model effectively suppresses interference from complex background clutter and accurately localizes aircraft. In the first row of Fig. 6, WS-YOLO is the only method that successfully identifies both aircraft, whereas the other models either miss one target or produce false detections. Overall, the proposed WS-YOLO demonstrates qualitatively superior performance, exhibiting higher detection completeness and robustness in challenging scenes.

5 Conclusions and Future Work

The proposed WS-YOLO, equipped with the Enhanced Fine-grained Module, demonstrates robust and competitive performance on both the CORS-ADD and MAR20 datasets. As remote sensing technology continues to advance, aircraft detection has become an increasingly important research

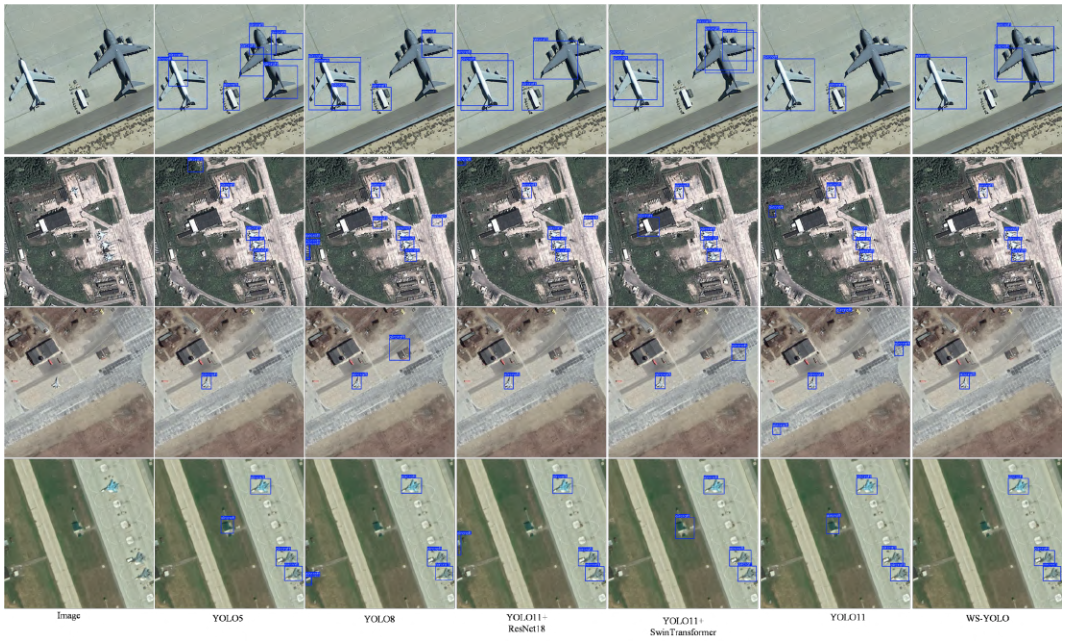

Fig. 6. Partial aircraft detection results on the MAR20 Dataset.

topic because it supports a wide range of practical applications, including airport monitoring, traffic analysis, emergency response, and large-scale scene understanding. However, the fine-grained nature of aircraft targets, their relatively small size, and the presence of complex background interference still make accurate detection a challenging task. In this study, WS-YOLO improves the representation of subtle aircraft characteristics while preserving efficient inference, showing that a lightweight detector can still achieve strong results in demanding remote sensing scenarios. Overall, the experimental results confirm the effectiveness of integrating fine-grained feature enhancement into a one-stage detection framework. From a broader perspective, this work also highlights the continuing value of lightweight and deployable detection models for real-world remote sensing applications, where computational resources and response time are often constrained. These characteristics are particularly important for practical monitoring systems that require stable and timely detection performance. In future work, we aim to further address remaining challenges in aircraft detection, extend the framework to multi-class remote sensing scenarios, and evaluate its generalizability to other object categories while identifying potential limitations and practical adaptation requirements.

Acknowledgments

The authors thank the colleagues and advisors who provided helpful suggestions during the writing of this paper. We also acknowledge our institution for its support in providing research

facilities and resources.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] Shi T, et al. Progressive class-aware instance enhancement for aircraft detection in remote sensing imagery. Pattern Recognition. 2025;164:111503.

[2] Panneerselvam RK, Bandi S, D SD. A Deep Learning Approach to Aircraft Detection and Classification in Satellite Imagery Using YOLOv8. In: 2025 6th International Conference for Emerging Technology (INCET). IEEE; 2025. p. 1-6.

[3] Zhou W, Cai C, Srigrarom S, Xu H, Liu R, Li C. SAD-YOLO: A Small Object Detector for Airport Optical Sensors Based on Improved YOLOv8. IEEE Sensors Journal. 2025;25(11):20513-22.

[4] Wu J, Zhao F, Yao G, Jin Z. FGA-YOLO: A one-stage and high-precision detector designed for fine-grained aircraft recognition. Neurocomputing. 2025;618:129067.

[5] Xu SL, Chen Z, Zhang H, Su H. Improved Aircraft Target Detection Algorithm for Remote Sensing Images with YOLOv8. In: 2023 3rd International Conference on Electronic Information Engineering and Computer Science (EIECS). IEEE; 2023. p. 317-21.

[6] Ouyang L, Guo G, Fang L, Ghamisi P, Yue J. PCLDet: Prototypical Contrastive Learning for Fine-Grained Object Detection in Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing. 2023;61:1-11.

[7] Jiang L, Zang J, Zhang Y. EM-YOLO: A Fine-grained Recognition Model for Aircraft Targets in SAR Images. In: 2025 International Conference on Communication Networks and Smart Systems Engineering (ICCNSE). IEEE; 2025. p. 255-9.

[8] S V, Kalahasti T, Kotturu MS. Military Aircraft Detection from Optical Satellite Images Using Convolutional Neural Networks. In: 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS). IEEE; 2024. p. 1-6.

[9] Shi T, et al. Complex Optical Remote-Sensing Aircraft Detection Dataset and Benchmark. IEEE Transactions on Geoscience and Remote Sensing. 2023;61:1-9.

[10] Wenqi Y, et al. MAR20: A Dataset for Military Aircraft Target Recognition in Remote Sensing Images. Journal of Remote Sensing. 2024;27(12):2688-96.

[11] Tong Z, Chen Y, Xu Z, Yu R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. arXiv preprint arXiv:230110051. 2023.

[12] Li Q, Jia X, Zhou J, Shen L, Duan J. Rediscovering BCE Loss for Uniform Classification. arXiv preprint arXiv:240307289. 2024.

[13] Li X, et al. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. arXiv preprint arXiv:200604388. 2020.