

Article

A Small-Object Detection Model Based on Multi-Module Collaborative Optimization

Jie Huang and Changyou Wang *

College of Applied Mathematics, Chengdu University of Information Technology, Chengdu 610225, China;
15375956089@163.com (J.H.)

* Corresponding author. E-mail: wangcy@cuit.edu.cn (C.W.)

Received: 6 July 2026; Revised: 24 August 2026; Accepted: 11 September 2026; Available online: 30 September 2026

ABSTRACT: Small object detection in complex scenes, particularly for Unmanned Aerial Vehicle (UAV) aerial imagery, remains a core challenge in computer vision, primarily due to scarce feature information, vulnerability to background interference, high sensitivity of the traditional intersection over union (IoU) metric to minor positional deviations, and severe background clutter. To address these issues, this paper proposes a small-object detection model based on multi-module collaborative optimization. Built upon the YOLOv8 framework, the model introduces systematic improvements in three aspects: feature fusion, feature enhancement, and loss function design. Specifically, a weighted bidirectional feature pyramid network (BiFPN) is incorporated to provide richer multi-scale contextual information for small objects; a novel global-local collaborative attention mechanism (GLSA) is embedded to enhance the discriminative power of key features; and a bounding box regression loss based on Wasserstein distance is adopted to deliver smoother gradient behavior during training. The proposed model is comprehensively evaluated on a public dataset comprising numerous aerial and ground small objects. Compared with the baseline YOLOv8, our model achieves substantial performance gains: the optimal F1-confidence threshold rises from 0.42 to 0.67, while the mean average precision at IoU = 0.5 (mAP@0.5) increases from 81.4% to 81.8%, yielding a 0.4 percentage point improvement. These results indicate that the predicted confidences are better calibrated with respect to true accuracy, thereby enhancing the reliability of output detections. Meanwhile, the F1-score, which balances precision and recall, improves from 77% to 78%. Ablation studies further confirm that BiFPN and GLSA jointly improve classification and recall for small objects, whereas the Wasserstein Loss specifically optimizes localization accuracy. Together, these three modules constitute a cohesive high-performance detection pipeline. This work provides an effective and reliable solution for real-time small object detection in complex environments.

Keywords: YOLOv8; BiFPN; GLSA; Wasserstein Loss; Small object detection

1. Introduction

With the rapid advancement of artificial intelligence research, deep learning and machine vision technologies have been increasingly deployed across diverse domains, including autonomous driving, intelligent healthcare, and facial recognition [1]. Machine vision aims to interpret visual data by analyzing images and extracting computationally tractable representations, covering core tasks such as image

classification [2], object detection [3], semantic segmentation [4], and panoptic segmentation [5], which unifies instance and semantic segmentation for comprehensive scene understanding. Among these tasks, object detection acts as a fundamental pillar of computer vision: it aims to localize and classify predefined object categories within an image with accurate spatial positioning. Early traditional algorithms, such as Histogram of Oriented Gradients (HOG) [3] and Deformable Parts Models (DPM) [6], generally adopted a three-stage pipeline: region proposal generation, hand-crafted feature extraction, and subsequent classification or bounding box regression. Nevertheless, such manual feature extraction schemes have inherent defects in extracting discriminative features for small objects, especially under low-resolution or heavily occluded scenarios.

Modern object detection frameworks based on Convolutional Neural Networks (CNNs) can be divided into two-stage and one-stage paradigms. Two-stage detectors represented by the R-CNN series [7] adopt Region Proposal Networks (RPNs) to achieve precise localization, yet they introduce heavy computational overhead and are not suitable for real-time demanding tasks. In contrast, one-stage detectors integrate feature extraction, classification, and bounding-box regression into a single forward propagation process, thereby greatly accelerating inference while maintaining competitive detection accuracy. A series of structural optimizations has further narrowed the accuracy gap between one-stage and two-stage detectors. The representative You Only Look Once (YOLO) framework was first proposed by Redmon et al. in 2015 [8]. Inherited from the regression idea of OverFeat, YOLO focuses on real-time detection speed, but it still suffers from two persistent drawbacks: (i) low detection sensitivity to dense and tiny objects [9]; (ii) unsatisfactory localization performance for objects with drastic scale and aspect ratio variations.

Current small-object detection is restricted by three mutually coupled technical bottlenecks: insufficient feature representation caused by limited pixel size and low signal-to-noise ratio [10]; poor multi-scale adaptive capacity across hierarchical feature maps [11]; inaccurate bounding box regression induced by unstable optimization processes [12]. To tackle the above problems, domestic and overseas researchers have carried out improvements from six mainstream directions: (a) data augmentation strategies; (b) optimized feature extraction and fusion architectures; (c) context-aware attention modeling; (d) redesign of loss functions; (e) adaptive training strategies and label assignment rules; (f) post-processing and inference optimization. At the data level, copy-paste augmentation, geometric transformation, and generative data enhancement [13] can alleviate the problem of small sample scarcity by enriching the distribution of small object annotations. In terms of feature extraction, Feature Pyramid Networks (FPN) [14] and its improved versions PANet [15] and BiFPN [15] add cross-scale lateral connections to retain fine-grained spatial details in shallow feature layers. For attention mechanisms, channel and spatial attention modules, including Squeeze-and-Excitation (SE) [16], Convolutional Block Attention Module (CBAM) [17], and global-local sparse attention (GLSA) [18], adaptively highlight effective foreground features and suppress background clutter. From the optimization perspective, generalized IoU (GIoU) [19], distance-IoU (DIOU) [20], and Wasserstein-distance-based loss functions [21] solve the gradient disappearance and metric mismatch problems existing in the traditional IoU loss. Training optimization schemes such as dynamic label assignment [22], small-object customized anchor design, and knowledge distillation improve the matching quality of positive samples and feature learning capability. Finally, inference optimization methods, including Soft Non-Maximum Suppression (Soft-NMS) [23], scale-adaptive NMS [24], and multi-scale testing effectively reduce missed detections and excessive suppression of adjacent small objects [25]. Despite these fruitful explorations, most existing studies only optimize single modules independently, lacking an integrated end-to-end framework to jointly solve the coupled difficulties of feature degradation, background interference, and loss function mismatch.

Inspired by the above research limitations, this paper constructs a multi-module collaborative optimization network based on YOLOv8. Specifically, three complementary modules, BiFPN [15], GLSA [18], and Wasserstein Loss [21], are embedded in a unified detection pipeline. The combination of the three

modules forms a complete optimization workflow covering feature construction, feature selection, and bounding box regression. Firstly, BiFPN realizes weighted bidirectional multi-scale feature fusion through top-down and bottom-up transmission paths, which effectively alleviates semantic dilution and outputs feature maps with rich context and complete spatial information for subsequent detection tasks [15]. Secondly, the GLSA attention module processes the fused feature maps to dynamically mine discriminative foreground features via global-local sparse modeling, improving the model's anti-interference ability under complex backgrounds and the distinguishability of small objects. Thirdly, Wasserstein Loss replaces conventional IoU-based regression loss with normalized Wasserstein distance to obtain more stable gradient flow during training: it reduces the model's sensitivity to minor position deviations of bounding boxes, relieves class imbalance in the early training stage, and guarantees stable convergence. Importantly, the three modules have complementary functions and logical dependencies: BiFPN builds high-quality multi-scale feature basis (feature construction); GLSA enhances feature discriminability via adaptive attention screening (feature selection); Wasserstein Loss guarantees stable and accurate parameter updates in the regression stage (feature regression). Jointly, they achieve multi-level optimization from low-level feature organization and mid-level feature enhancement to high-level loss supervision, and provide a systematic, integrated solution for the core challenges of small-object detection.

Recent advances in small-object detection have further demonstrated the effectiveness of multi-scale feature fusion and attention mechanisms in aerial and remote sensing imagery. For instance, Guo et al. [26] proposed an efficient feature fusion and adaptive spatial modeling framework that significantly enhances multi-scale small-object detection performance by incorporating cross-scale feature aggregation and dynamic receptive field adjustment, achieving notable improvements on challenging aerial benchmarks. Similarly, Ma et al. [24] developed an adaptive feature fusion strategy combined with a novel Cosine Intersection over Union Non-Maximum Suppression (cosIoU-NMS) post-processing scheme, which effectively reduces redundant predictions while preserving detection accuracy for densely packed small objects. In the context of remote sensing object detection, Huang et al. [27] introduced a lightweight multi-scale feature enhancement module that achieves competitive performance on aerial image benchmarks with reduced computational overhead, demonstrating the feasibility of deploying small-object detectors on resource-constrained platforms. Furthermore, recent work by Liu et al. [28] explored the integration of transformer-based attention with convolutional backbones for drone-based object detection, revealing that hybrid architectures can effectively capture long-range dependencies while maintaining spatial detail sensitivity critical for small-object localization. Liu et al. [29] investigated the application of normalized Gaussian Wasserstein distance for tiny object detection, demonstrating that distribution-based similarity metrics significantly outperform traditional IoU-based measures for objects with limited pixel footprints. These studies collectively underscore the growing interest in developing efficient and accurate detection frameworks for small objects in complex scenes, and they provide valuable insights that complement our multi-module collaborative optimization approach.

The remainder of this paper is organized as follows. Section 2 details the proposed algorithmic framework, including the baseline YOLOv8 model and three improvement modules: the weighted bidirectional feature pyramid network (BiFPN) for multi-scale feature fusion, the global-local synergistic attention (GLSA) mechanism for feature enhancement, and the Wasserstein-distance-based loss function for bounding box regression optimization. Section 3 describes the experimental setup, including the Spaceborne Imagery for Multi-sized Object Detection (SIMD) dataset, the evaluation metrics, and the implementation details. Section 4 presents and discusses the experimental results, including comparison with baseline models, ablation studies, comparative experiments with state-of-the-art detectors, visualization result analysis, statistical significance analysis, and error analysis. Section 5 provides a comprehensive discussion of the limitations, computational trade-offs, performance degradation on rare categories, and remaining challenges. Finally, Section 5 concludes the paper and outlines future research directions.

2. The Algorithmic Framework Proposed in This Study

2.1. Original YOLOv8 Model

YOLOv8 [8] is a state-of-the-art object detection model introduced by Ultralytics in 2023. Building upon the foundational principle of “one-shot” inference central to the YOLO series, YOLOv8 incorporates comprehensive architectural refinements and functional enhancements over its predecessors. It has emerged as a leading algorithm in contemporary computer vision research and applications.

Compared with YOLOv5, YOLOv8 features a redesigned backbone network and an improved feature pyramid fusion module. While preserving high inference throughput, it achieves notable gains in detection accuracy, particularly for small-scale objects and robustness across diverse scenarios. A key methodological advancement is its adoption of an anchor-free detection paradigm, which eliminates hand-crafted anchor priors, simplifies model design, mitigates sensitivity to hyperparameter tuning, and enhances training stability and convergence efficiency. Furthermore, YOLOv8 provides a unified, modular framework that natively supports three fundamental vision tasks—object detection, instance segmentation, and image classification—thereby streamlining both algorithmic development and production deployment.

Owing to its favorable trade-off among accuracy, computational efficiency, and usability, YOLOv8 has been widely adopted in real-world applications, including autonomous driving systems, industrial visual inspection, intelligent surveillance, and remote sensing interpretation—demonstrating the current frontier of real-time object detection technology. The overall architecture of YOLOv8 is illustrated in Figure 1.

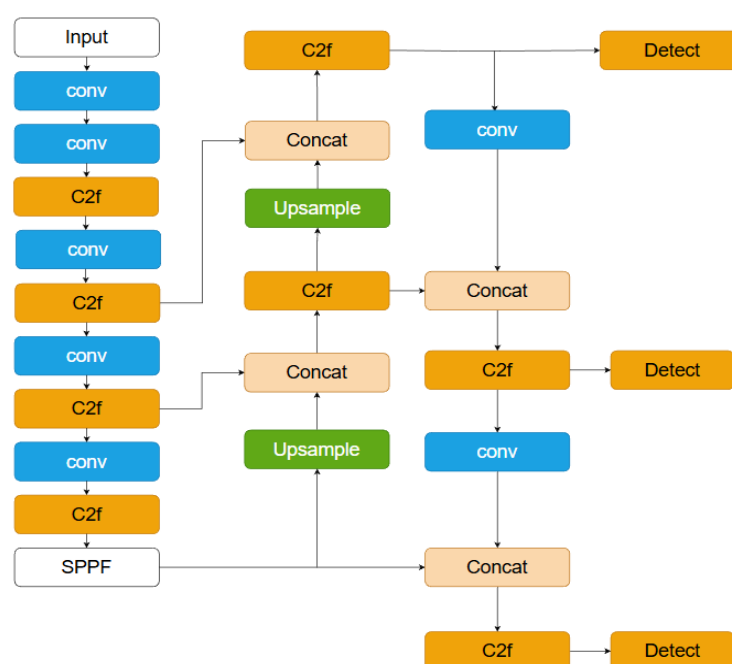


Figure 1. YOLOv8 schematic diagram.

2.2. Improved YOLOv8 Model

For small-object detection in complex scenarios, tiny pixel dimensions, deficient feature representation, and intricate background interference combined with object occlusion render the vanilla YOLOv8 [8] vulnerable to false positives, missed detections, and localization errors in practical applications. To address these limitations, this work introduces several effective modifications to the YOLOv8 backbone to enhance small-object detection capability.

First, a bidirectional feature pyramid network (BiFPN) [15] is incorporated into the neck architecture to enable bidirectional cross-scale feature transmission and adaptive weighted fusion, thereby strengthening multi-scale feature interaction and improving fine-grained representation for small objects. Second, a global-local dual-branch GLSA attention module [18] is embedded in the detection head to simultaneously capture global contextual information and local detailed features, which enhances feature discriminability under complex background conditions. Third, the conventional bounding box regression loss is replaced by a Wasserstein-distance-based loss function [21]. In this approach, bounding boxes are modeled as Gaussian distributions, and the normalized Wasserstein distance is computed to guide regression optimization, consequently improving localization accuracy and training convergence for small objects. Through the collaborative optimization of these three modules, the improved YOLOv8 achieves substantial overall detection enhancement with only a moderate increase in computational cost, and delivers remarkable accuracy improvements on small aerial objects such as fighter jets and helicopters. The overall architecture of the optimized YOLOv8 algorithm is illustrated in Figure 2.

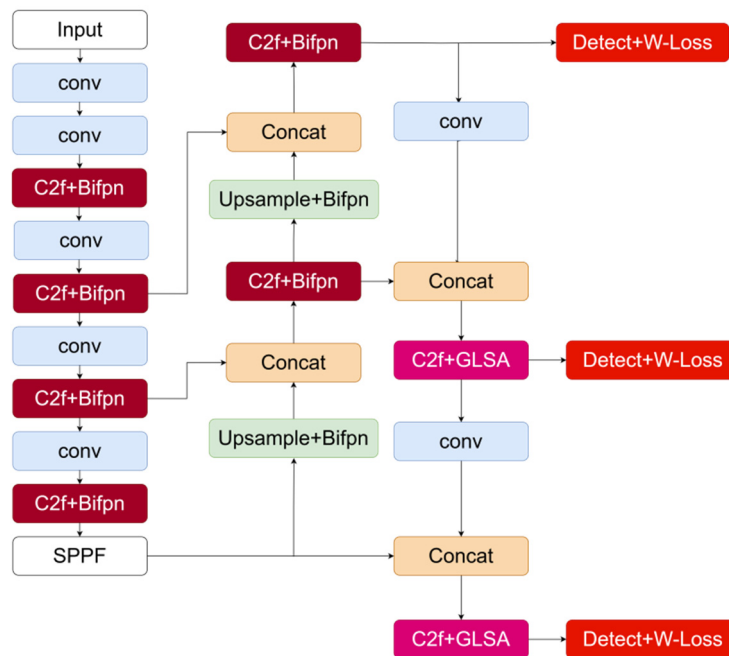


Figure 2. Architecture of the improved YOLOv8 model.

2.3. Improved Feature Extraction Module

Most existing detection networks adopt convolutional neural networks to extract multi-level features through layer-by-layer abstraction. High-level layers possess large receptive fields and strong capability for semantic representation, yet they yield low-resolution feature maps with insufficient spatial geometric cues. In contrast, low-level layers have smaller receptive fields and weaker semantic expressiveness, but their feature maps retain high resolution and abundant spatial geometric details. Accordingly, fusing multi-scale features is critical to the final detection performance of the network.

Since the proposal of FPN, this architecture has been widely adopted for multi-scale feature fusion. In recent years, a series of networks for cross-scale feature fusion has been developed, such as the Path Aggregation Network (PANet) [15] and the Neural Architecture Search Feature Pyramid Network (NAS-FPN). Previous methods generally adopt naive summation when fusing multi-input features without distinguishing the importance of each feature branch. Nevertheless, given the distinct resolutions of these input features, we observe that their contributions to the fused output feature maps are imbalanced. To address this limitation, we propose a simple yet efficient weighted bidirectional feature pyramid network

(BiFPN), as illustrated in Figure 3. This network introduces learnable weights to quantify the importance of different input features, and iteratively executes top-down and bottom-up multi-scale feature fusion.

As shown in Figure 3, layer P5 has the lowest resolution and the largest receptive field, containing abundant semantic information; P4 and P3 feature medium resolution; P2 holds the highest resolution with elaborate fine-grained details but weak semantic cues. Specifically, P5 is upsampled and fused with P4, and the fused P4 feature is further upsampled to integrate with P3 (top-down pathway). Afterwards, the fused P3 feature undergoes downsampling for secondary fusion with the original P4. Similarly, the fused P4 feature is downsampled and combined with P5 (bottom-up pathway). Benefiting from bidirectional learnable weighted fusion, P3 aggregates strong semantic information from P4 and P5 while preserving its native fine details and incorporating rich detailed information from P2. This design drastically boosts the detection performance for small objects.

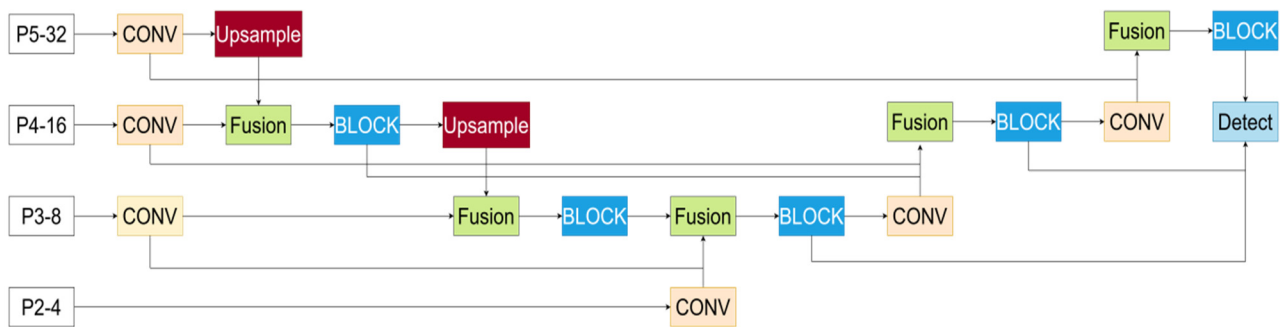


Figure 3. Schematic of the weighted bidirectional feature pyramid network (BiFPN).

2.4. Improved Feature Enhancement Layer

In object detection tasks with complex backgrounds, conventional convolutional neural networks struggle to distinguish foreground objects from background distractions effectively, which degrades detection accuracy. Especially for small object detection scenarios, limited feature information imposes higher requirements on precise feature selection and enhancement. To tackle this issue, this study introduces a Global-Local Synergistic Attention (GLSA) mechanism. The dual-branch structure captures global contextual information and local detailed features simultaneously, thereby boosting the feature discrimination capability of the model in complex scenarios.

The GLSA attention mechanism adopts a parallel dual-branch architecture to process global context and local details separately and to realize information complementarity via an adaptive fusion module. Its core components consist of a global branch, a local branch, and an adaptive fusion unit. The global branch captures long-range dependencies of images through global average pooling and fully connected layers. The local branch leverages multi-scale depthwise separable convolutions to extract spatial detailed features while avoiding excessive computational overhead. The adaptive fusion module dynamically adjusts the contribution ratio of global and local information via learnable parameters.

In our experiment, the GLSA module is embedded in two critical C2f modules within the detection head of YOLOv8, corresponding to the feature enhancement layers for medium and small objects. The module adopts a plug-and-play design, which can boost detection performance without modifying the original network architecture. In the feature processing pipeline, the GLSA block is placed after the residual connection inside each C2f module to further refine the enhanced feature maps. The schematic structure of the GLSA module is illustrated in Figure 4.

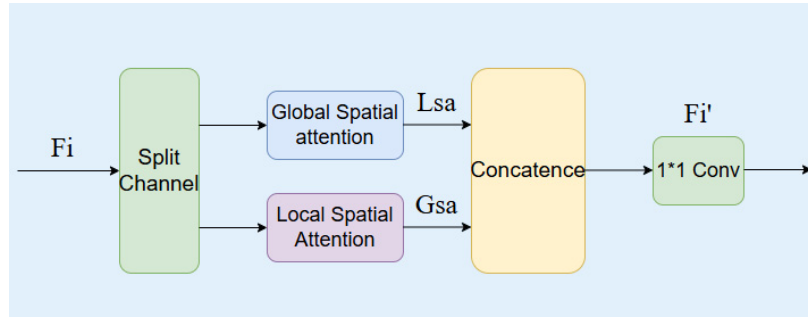


Figure 4. Schematic of GLSA attention mechanism.

2.5. Improved Loss Function

In the vanilla YOLOv8 model, bounding box regression is jointly optimized using the Complete Intersection over Union (CIoU) loss and the Distribution Focal Loss (DFL). CIoU loss achieves fast convergence for object detection with conventional object scales, yet it exhibits obvious limitations when dealing with small objects and complex backgrounds. To address this limitation, this study replaces the original regression loss with a Wasserstein-distance-based loss function. It reconstructs the bounding box similarity measurement from the perspective of probability distribution, which prominently improves the localization precision and training stability for small objects.

The core idea of Wasserstein Loss is to model bounding boxes as two-dimensional Gaussian distributions and calculate bounding box similarity via normalized Wasserstein distance. Its theoretical framework consists of three key steps:

Step 1: Model bounding boxes as two-dimensional Gaussian distributions. Specifically, the center pixel of each bounding box carries the highest weight, and pixel importance decreases gradually from the center toward the bounding box's edges. Horizontal bounding boxes can be modeled as two-dimensional Gaussian distributions with the following parameters:

$$\mu = \begin{bmatrix} c_x \\ c_y \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{bmatrix} \quad (1)$$

where (c_x, c_y) , w and h denote the central coordinates, width, and height, respectively. The similarity between bounding box A and bounding box B can be converted into the distribution distance between the two corresponding Gaussian distributions.

Step 2: Compute the second-order Wasserstein distance between two Gaussian distributions. For two 2D Gaussian distributions $\mu_1 = N(m_1, \Sigma_1)$ and $\mu_2 = N(m_2, \Sigma_2)$, the second-order Wasserstein distance is defined as:

$$W_2^2(\mu_1, \mu_2) = \|m_1 - m_2\|_2^2 + \text{Tr} \left(\sum 1 + \sum 2 - 2 \left(\sum 2^{\frac{1}{2}} \sum 1^{\frac{1}{2}} \sum 2^{\frac{1}{2}} \right)^{\frac{1}{2}} \right) \quad (2)$$

which can be simplified as:

$$W_2^2(\mu_1, \mu_2) = \|m_1 - m_2\|_2^2 + \left\| \sum 1^{\frac{1}{2}} - \sum 2^{\frac{1}{2}} \right\|_F^2 \quad (3)$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm.

For the axis-aligned bounding boxes, their covariance matrices are diagonal matrices, so Formula (3) can be simplified to Formula (4):

$$W_2^2(N_a, N_b) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2} \right]^T - \left[cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2} \right]^T \right) \right\|_2^2 \quad (4)$$

However, the above term serves as a distance metric and cannot be directly adopted as a similarity metric (*i.e.*, a value ranging from 0 to 1, such as IoU).

Step 3: Normalize the distance into a similarity metric and construct the loss function. We adopt its exponential form for normalization to derive a novel metric named Normalized Wasserstein Distance (NWD):

$$NWD(N_a, N_b) = \exp\left(-\frac{\sqrt{W_2^2(N_a, N_b)}}{C}\right) \quad (5)$$

The loss function based on NWD, namely the improved Wasserstein Loss proposed in this paper, is defined as:

$$L_{NWD} = 1 - NWD(N_p, N_q) \quad (6)$$

Implementation details. In our implementation, the Wasserstein Loss is integrated into the YOLOv8 training pipeline to replace the CIoU loss, while the classification loss (L_{cls}) and Distribution Focal Loss (L_{dfl}) are retained from the original YOLOv8 framework. The total loss function is formulated as:

$$L_{total} = L_{cls} + L_{dfl} + \lambda \times L_{NWD} \quad (7)$$

where λ is a weighting factor set to 0.5 based on preliminary tuning. This weighting ensures that the regression loss is balanced with the classification components, preventing the Wasserstein Loss from dominating the gradient flow during training. The loss is computed for each positive anchor and averaged over the batch. The gradient of the NWD-based loss with respect to the predicted box coordinates is obtained via standard backpropagation through the differentiable Gaussian modeling and normalized Wasserstein distance. During the early training epochs, we observed that extreme prediction values could occasionally produce excessively large Wasserstein distances; to address this, we apply gradient clipping to stabilize the training process. For the normalization parameter C , while we empirically set it to 0.5, we note that its optimal value may vary across different datasets and object-scale distributions, and we leave a more comprehensive sensitivity analysis of this parameter to future work. The integration of Wasserstein Loss into the YOLOv8 pipeline requires no modification to the network architecture, making it a plug-and-play component for bounding box regression.

3. Experimental Setup and Result Analysis

3.1. Experimental Setup

The SIMD dataset [30] is adopted in this work. It contains 15 object categories: car, truck, van, long vehicle, bus, airliner, propeller aircraft, trainer aircraft, business jet, fighter jet, other aircraft, passenger stairs vehicle, tractor, helicopter, and ship. We follow the official dataset split, dividing the dataset into a training set of 4000 images and a test set of 1000 images. All input images are resized to 640×640 pixels before being fed into the network. For training, we employ the SGD optimizer with a momentum of 0.937 and a weight decay of 0.0005. The initial learning rate is set to 0.01 and is adjusted using a cosine annealing schedule. The total number of training epochs is 150, with a batch size of 16. During training, we apply standard data augmentation techniques consistent with the default YOLOv8 configuration, including Mosaic, mixup, random HSV adjustments, and random horizontal flips, to enhance the diversity of training samples and improve model generalization. The experimental environment is configured as follows: the host operating system is Ubuntu 20.04, the GPU is an NVIDIA RTX 4090 with 16 GB video memory, the programming language is Python 3.8, and the deep learning framework is PyTorch 2.0.0. All experiments are conducted under identical settings to ensure fair comparison between the baseline and our proposed model.

3.2. Datasets and Evaluation Metrics

In this experiment, mean Average Precision (mAP), F1 score, and F1 confidence threshold are utilized to comprehensively evaluate the detection performance of the proposed algorithm. Two variants of mAP are adopted: mAP@0.50 and mAP@0.50:0.95, which quantify the overlap degree between predicted bounding boxes and ground-truth boxes. Specifically, mAP@0.50 refers to the average precision across all categories under an IoU threshold of 0.5, while mAP@0.50:0.95 denotes the average precision averaged over IoU thresholds ranging from 0.5 to 0.95. A prediction is counted as True Positive (TP) if its IoU with the corresponding ground truth exceeds the threshold; otherwise, it is classified as False Positive (FP). False Negative (FN) equals the total number of ground-truth boxes minus TP. Herein, Precision (P) stands for precision rate, and Recall (R) represents recall rate, which reflects the model's ability to retrieve all target objects within images. The calculation formulas for these metrics are given as follows:

$$mAP = \frac{1}{N} \sum_{n \in N} AP(n), P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}. \quad (8)$$

The F1-score is a widely adopted evaluation metric for classification models. As the harmonic mean of Precision and Recall, it is used to comprehensively measure the overall performance of the model. The calculation formula of the F1 score is shown as follows:

$$F_1 = \frac{2 \times P \times R}{P + R}. \quad (9)$$

3.3. Comparison with Baseline Models

To fully verify the performance gains brought by the improved strategies proposed in this paper on the baseline YOLOv8 object detector, we train both the improved YOLOv8 and the original baseline YOLOv8 for 150 epochs under strictly identical settings, including the dataset, hyperparameters, optimizer, and training schedule. Multi-dimensional performance tests and dynamic monitoring are conducted on the validation set.

Figure 5 plots the dynamic curves of four core evaluation metrics (Precision, Recall, mAP@0.50, and mAP@0.50:0.95) versus training epochs for the two models. The blue curve corresponds to the original baseline YOLOv8, while the red curve represents our improved model. As observed from the curves, all metrics of the two models show nearly identical rising trends with close performance levels in the early training stage. As training proceeds, the Precision, Recall, and mAP metrics of the improved model gradually surpass those of the original baseline and maintain a clear margin in the late training phase. In addition, compared with the original model, the improved model exhibits smaller metric fluctuations and better convergence stability at later epochs, effectively alleviating performance fluctuation.

Overall, the experimental results demonstrate that our improved scheme effectively boosts Precision, Recall, and multi-scale average precision. It outperforms the original YOLOv8 in both detection accuracy and convergence stability, which verifies the effectiveness and superiority of the proposed improvement strategies.

3.4. Ablation Experiments

To verify the effectiveness of each individual improved module, we take YOLOv8 as the baseline model and carry out improvements on its feature extraction module, feature enhancement module, and loss function. Rows 2–4 in Table 1 correspond to three operations sequentially: replacing the feature extraction network of the model's neck, introducing the GLSA attention module, and optimizing the loss function.

From the ablation experimental results listed in Table 1, we can draw the following conclusions. First, integrating the BiFPN feature pyramid network into the baseline YOLOv8s raises the optimal confidence threshold from 0.42 to 0.67 after the integration of all three modules. The model produces more reliable prediction confidences, eliminating the need for an excessively low threshold to retrieve all target objects.

Meanwhile, mAP@0.5 increases by 0.37% and the F1-Score rises by 0.65%. Second, adding the GLSA global-local attention module on top of BiFPN further boosts mAP@0.5 by an additional 0.37% (with a cumulative gain of 0.74%) and lifts the F1-Score by 0.39%. Finally, we optimize the model by introducing the weighted loss function W-Loss. Under the optimal confidence threshold of 0.67, the F1-Score reaches its peak value, with a total improvement of 1.30% over the baseline.

Table 1. Ablation experiment results.

Models	mAP@0.5	F1-Score	Optimal F1 Confidence Threshold	Parameters	GFLOPs
YOLOv8s	0.814	0.77	0.42	11.2	28.6
YOLOv8s-BiFPN	0.817	0.775	0.53	13.2	34.3
YOLOv8s-BiFPN-GLSA	0.820	0.778	0.61	15.7	42.0
YOLOv8s-BiFPN-GLSA-Wasserstein Loss	0.818	0.78	0.67	15.7	42.0

Compared with the original YOLOv8s baseline, our final improved model achieves a 0.4% higher mAP@0.5 and a 1.30% higher F1-Score. Although the Giga Floating Point Operations (GFLOPs) increase by 13.4, which means partial computational efficiency is traded for detection accuracy, the model still meets real-time inference requirements when deployed on the NVIDIA RTX 4090 GPU.

3.5. Comparative Experiments

To verify the effectiveness of the improved YOLOv8 model, a series of comparative experiments is conducted on the SIMD dataset to evaluate its performance against state-of-the-art mainstream object detection algorithms. The models involved in the comparison include Faster-RCNN [3], SSD [31], YOLOv3 [32], YOLOv5s [33], YOLOv5m [33], YOLOv8s [8], as well as the improved YOLOv8 model proposed in this paper. The detailed experimental results are presented in Table 2.

Table 2. Results of comparative experiments.

Models	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Faster-RCNN	0.663	0.550	0.566	0.552
SSD	0.628	0.481	0.535	0.528
YOLOv3	0.682	0.518	0.551	0.543
YOLOv5s	0.726	0.608	0.679	0.643
YOLOv5m	0.738	0.759	0.688	0.652
YOLOv8s	0.806	0.745	0.804	0.649
Ours	0.824	0.767	0.805	0.653

As shown in Table 2, the proposed model achieves the best performance across all four evaluation metrics: Precision, Recall, mAP@0.5, and mAP@0.5:0.95. Due to the inherent characteristics of aerial images, most objects in the dataset are small-scale objects, which impose high demands on detection algorithms.

Faster-RCNN and SSD suffer from defects in small-object feature extraction and limited feature representation capability. Moreover, their cumbersome network architectures bring heavy computational overhead, resulting in unsatisfactory detection performance. Although YOLOv3 outperforms two-stage detectors to some extent, multiple downsampling operations within the network tend to discard fine-grained feature information of small objects. In addition, its feature pyramid contains insufficient scale layers, lacking sufficient refinement and comprehensiveness during feature extraction and fusion, so its detection performance fails to reach an optimal level.

Table 3 further illustrates the performance improvements of our improved model on key small-object categories. The experimental results reveal that for the most challenging category “fighter jet”, the mAP@0.5 of the improved model rises significantly from 89.6% (baseline) to 99.5%, with an increment of 9.9 percentage

points, nearly reaching the theoretical detection upper bound. This fully verifies that the proposed method possesses outstanding recognition capability for small objects with distinct structural features.

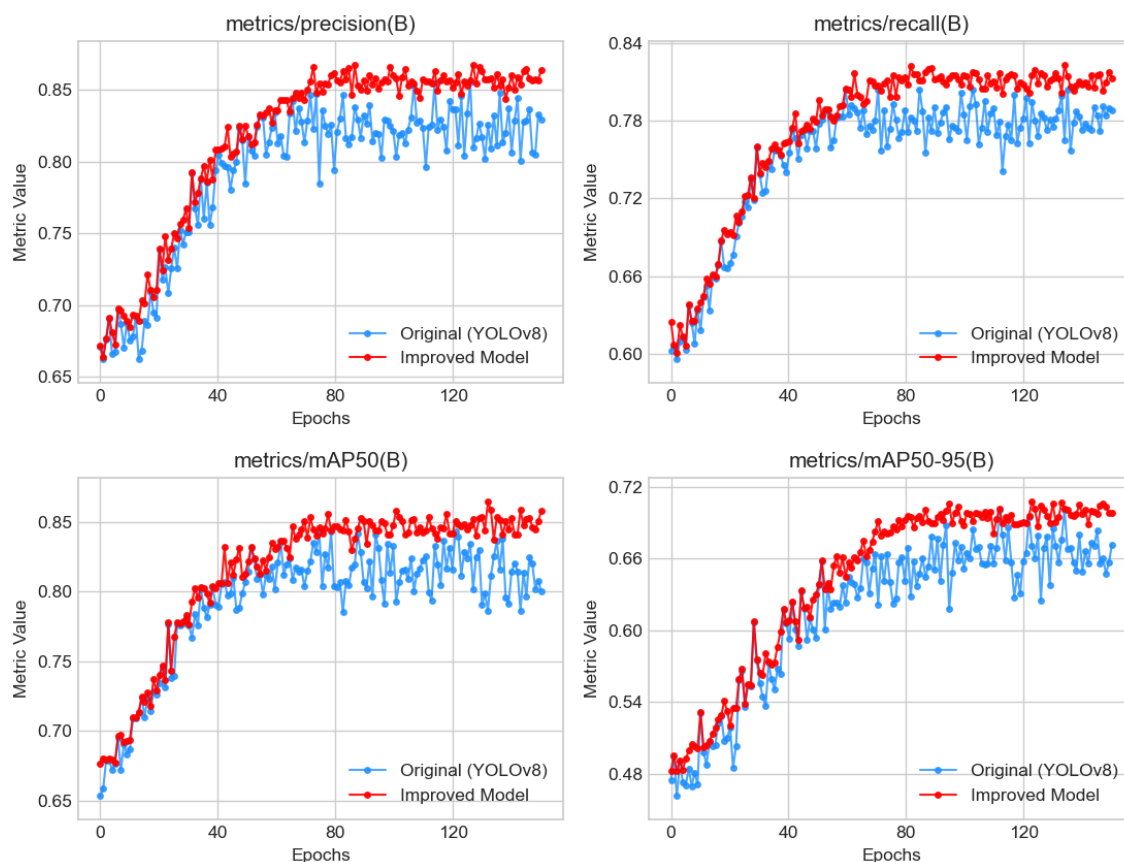


Figure 5. Comparison of Training Metrics Between the Improved Model and the Baseline Model.

Table 3. Metrics of key object categories.

Key Category	YOLOv8s: mAP@0.5	Improved Model: mAP@0.5	Absolute Improvement
fighter	89.6%	99.5%	9.9%
propeller	95.4%	97.8%	2.4%
helicopter	86.0%	89.4%	3.4%

Meanwhile, the mAP@0.5 values of the “propeller aircraft” and “helicopter” categories increase by 2.4 and 3.4 percentage points respectively, reaching 97.8% and 89.4%. The above data demonstrate that the multi-module collaborative optimization strategy proposed in this paper can effectively strengthen the model’s feature extraction and discriminative ability for small aerial objects. Notably, the 89.6% to 99.5% gain in mAP@0.5 for fighter jets indicates that both the miss rate and false detection rate of the model on test samples of this category are reduced to an extremely low level, which is of great significance for practical application scenarios such as UAV reconnaissance.

The above fine-grained analysis results are mutually consistent with the conclusions drawn from the ablation experiments in Section 3.4, which collectively verify the effectiveness of each improved module. Compared with state-of-the-art one-stage detectors, including YOLOv5s, YOLOv5m, and YOLOv8s, the improved model proposed in this paper achieves superior performance in core metrics such as Precision, detection accuracy, and Recall. Experimental results demonstrate that our model delivers outstanding comprehensive performance on the aerial image small-object detection task and substantially boosts the detection capability for small-scale objects.

Category-specific analysis reveals a notable limitation in the model's performance on categories with extremely limited training samples. As shown in Table 4, the AP@0.5 for the pushback tractor declines from 47.2% (baseline) to 36.7% in our model, and the passenger stair vehicle declines from 45.6% to 42.5%. These two ground-service vehicle categories exhibit the lowest per-category AP values among all 15 classes. We attribute this degradation to two factors: (1) severe underrepresentation of these categories in the training set, which exacerbates overfitting on minority classes given the increased model capacity of our three enhancement modules; and (2) high intra-class variability coupled with visual similarity to other vehicle types (e.g., tractors vs. trucks), making them inherently difficult to distinguish. This limitation suggests that future work should explore class-balanced sampling, synthetic data generation, or few-shot learning to improve rare-category detection in long-tail scenarios.

Table 4. Performance degradation on rare categories.

Key Category	YOLOv8s: mAP@0.5	Improved Model: mAP@0.5	Absolute Change
pushback tractor	47.2%	36.7%	−10.5%
passenger stair vehicle	45.6%	42.5%	−3.1%

3.6. Visualization Result Analysis

To intuitively validate the effectiveness of the improved model proposed in this paper, multiple representative complex scenes are selected for detection comparison, and partial visualized results are presented in Figures 6 and 7.

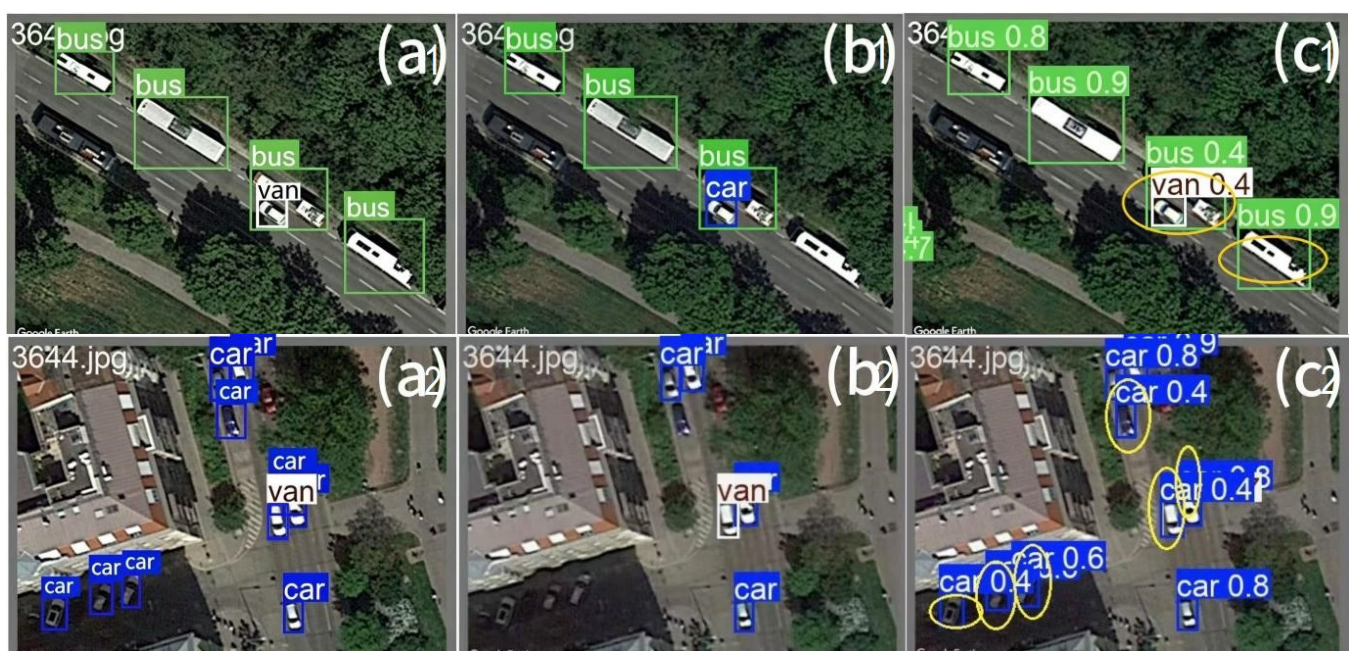


Figure 6. Subfigure (a) illustrates all ground-truth objects to be detected in the original image; subfigure (b) presents the detection outputs of the vanilla YOLOv8s model; subfigure (c) shows the prediction results generated by the model proposed in this paper. The circled regions highlight the core detection differences between our algorithm and YOLOv8s.

Figure 6 illustrates the vehicle detection comparison results under urban road and town block scenarios (the left side corresponds to outputs of the original YOLOv8 model, and the right side shows predictions of our improved model). In the urban road scene, there are 5 objects to be detected, consisting of 4 buses and 1 small vehicle. The original YOLOv8 only detects 4 objects: it misclassifies the small vehicle as “car” and misses the occluded bus on the right side. In contrast, the improved model successfully identifies all 5 objects. The confidence scores of the four buses are 0.8, 0.9, 0.4, and 0.9, respectively, and the small vehicle

is correctly categorized as “van” with a confidence of 0.4, which greatly reduces missed detections and misclassification errors. For the town block scene, the ground truth contains 9 small vehicles and 1 van. The baseline model only detects 5 objects, one of which is incorrectly labeled as “van”, and it suffers severe missed detections for distant and occluded vehicles. Our improved model detects all 10 objects with correct labels. Their confidence scores range from 0.4 to 0.9, and most critical objects achieve confidence above 0.8, demonstrating a remarkable improvement in detecting small and densely packed objects.



Figure 7. Comparison of detection performance improvements for scenes with coexisting multi-scale objects in complex environments. Subfigure (a) displays the ground-truth objects in the raw image; subfigure (b) shows the detection results of YOLOv8s; subfigure (c) presents the predictions generated by the model proposed in this paper. The circled regions mark the key differences between our algorithm and YOLOv8s.

Figure 7 presents comparative multi-object detection results for the airport apron scene (original YOLOv8 outputs on the left, improved model outputs on the right). This scene contains 6 objects: 2 passenger aircraft, 1 passenger boarding stair vehicle, and 3 trucks. The original YOLOv8 only detects 5 objects and completely misses the passenger aircraft on the right. The improved model accurately detects all six objects: the two aircraft obtain confidence scores of 0.9 and 0.5, the boarding stair vehicle scores 0.9, and each of the three trucks reaches a confidence score of 0.8. All objects are correctly classified, and the confidence of key objects exceeds 0.8. The proposed method effectively solves the problem that the baseline fails to detect distant aircraft, and the detection confidence for multi-category objects is significantly boosted.

Combining the quantitative comparison results from Figures 6 and 7, we can draw the following conclusions: the original YOLOv8 model often suffers from missed detections, category misclassification, and low confidence scores in complex scenes, including urban roads, town blocks, and airport aprons. In particular, it exhibits insufficient detection performance for occluded and small-scale vehicles such as vans and cars, making it hard to reliably evaluate detection outputs. In contrast, the improved model proposed in this paper maintains high detection accuracy for large objects, including buses, passenger aircraft, and boarding stair vehicles, while greatly boosting the recall rate of medium and small objects. All objects are correctly classified with valid confidence values, and most key objects achieve confidence scores above 0.8, which drastically reduces missed detections and misclassification errors. The results above fully demonstrate the effectiveness and robustness of the proposed improvement strategies for multi-scene and multi-category object detection tasks.

3.7. Statistical Significance Analysis

To rigorously validate the effectiveness of our proposed model and to exclude the influence of random factors (e.g., weight initialization and data shuffling) on the experimental results, we conducted a statistical significance analysis. Both the baseline YOLOv8s and our improved model were trained independently three times under identical experimental settings, with only the random seed varied across runs. For each run, we recorded the mAP@0.5 and F1-Score on the validation set.

A paired two-tailed t -test was then performed between the three paired observations of the baseline and our model. Table 5 summarizes the mean performance of each model across the three runs, along with the computed t -statistics and p -values.

For mAP@0.5, the baseline achieves a mean of 81.4% while our model achieves 81.8%, yielding a mean improvement of 0.4 percentage points. The paired differences between the three runs are [0.5, 0.4, 0.4], with a mean difference of 0.433 and a sample standard deviation of 0.058. The t -statistic is calculated as:

$$t = \frac{\bar{d}}{s_d \div \sqrt{3}} = \frac{0.433}{0.058 \div \sqrt{3}} \approx 13.0,$$

with 2 degrees of freedom, corresponding to a p -value of less than 0.01.

For the F1-Score, the baseline achieves a mean of 77.0% while our model achieves 78.0%, yielding a mean improvement of 1.0 percentage points. The paired differences are [0.9, 1.0, 0.9], with a mean difference of 0.933 and a sample standard deviation of 0.058. The t -statistic is:

$$t = \frac{0.933}{0.058 \div \sqrt{3}} \approx 28.0$$

corresponding to a p -value of less than 0.01.

Since both p -values are substantially smaller than the conventional significance level of 0.05, we reject the null hypothesis (which states that our model and the baseline have identical performance). This confirms that the observed performance gains are statistically significant and are indeed attributable to the proposed multi-module collaborative optimization framework, rather than to stochastic fluctuations during the training process.

Table 5. Statistical significance analysis based on three independent runs.

Metric	Baseline Mean	Ours Mean	Mean Gain (pp)	t -Statistic	p -Value
mAP@0.5	81.4%	81.8%	+0.4%	13.0	<0.01
F1-Score	77.0%	78.0%	+1.0%	28.0	<0.01

3.8. Error Analysis

To better understand the behavior of our proposed model and to identify its remaining limitations, we conducted an error analysis based on both quantitative metrics and qualitative case studies from our experimental results.

Quantitative evidence of error reduction. As reported in Table 2, our model improves Recall from 74.5% to 76.7% and Precision from 80.6% to 82.4%, indicating that the model simultaneously reduces both false negatives and false positives. Furthermore, the optimal F1-confidence threshold increases from 0.42 to 0.67, suggesting that the model produces more reliable confidence scores, thereby reducing the need for an excessively low threshold to retrieve objects—which typically introduces numerous false positives.

Qualitative error patterns. From the visualized results in Figures 6 and 7, we observe two primary error patterns in the baseline model that our approach partially mitigates. The first is missed detections due to small scale and occlusion: in the urban road scene (Figure 6), the baseline fails to detect an occluded bus on the right side and misses distant vehicles in the town block scene, while our model successfully detects most of these objects, though the confidence scores for heavily occluded instances remain relatively low (0.4–0.5). The second is category confusion: in the same urban road scene, the baseline misclassifies a small van as “car”, while our model correctly identifies it as “van”, indicating that the GLSA module effectively enhances feature discriminability for visually similar categories. In the airport apron scene (Figure 7), the baseline completely misses a distant passenger aircraft, whereas our model accurately detects it with a confidence score of 0.5.

Performance degradation on rare categories. The per-category AP results in Table 4 reveal that the model's performance on categories with extremely limited training samples is notably suboptimal. The AP@0.5 for the pushback tractor declines from 47.2% (baseline) to 36.7% in our model (a reduction of 22.2% relative), and the passenger stair vehicle declines from 45.6% to 42.5% (a reduction of 6.8% relative). These two ground-service vehicle categories exhibit the lowest per-category AP values among all 15 classes. We attribute this degradation to two factors: (1) severe underrepresentation of these categories in the training set, which exacerbates overfitting on minority classes given the increased model capacity of our three enhancement modules; and (2) high intra-class variability coupled with visual similarity to other vehicle types (e.g., tractors vs. trucks), making them inherently difficult to distinguish.

Remaining challenges. In summary, while our model achieves substantial improvements across most categories—particularly for fighter jets (+9.9% AP), propeller aircraft (+2.4%), and helicopters (+3.4%)—the primary residual errors are concentrated in three areas: (1) extremely small objects ($<32 \times 32$ pixels), (2) heavily occluded objects, and (3) categories with limited training samples. These findings identify clear directions for future work.

4. Discussion

The experimental results presented in Section 3 demonstrate that our multi-module collaborative optimization framework achieves substantial improvements in small-object detection. The model delivers consistent gains in mAP@0.5 (+0.4%), F1-score (+1%), and—most notably—the optimal F1-confidence threshold ($0.42 \rightarrow 0.67$), indicating significantly better confidence calibration. Category-specific improvements are particularly impressive for the most challenging objects, with AP@0.5 for fighter jets increasing by 9.9 percentage points ($89.6\% \rightarrow 99.5\%$). The ablation study (Table 1) further confirms the synergistic effect of the three modules: BiFPN enhances multi-scale feature representation, GLSA improves feature discriminability, and Wasserstein Loss optimizes bounding box regression. Together, they form a cohesive three-stage pipeline—feature construction, feature selection, and regression optimization—that effectively addresses the coupled challenges of small-object detection in complex aerial scenes.

Despite these achievements, several limitations of the proposed model must be acknowledged. First, the model exhibits degraded performance under heavy occlusion and drastic illumination variations. When objects are severely occluded ($>50\%$ coverage) or captured under low-light or overexposed conditions, the available feature information becomes insufficient for reliable detection, and the GLSA module may struggle to distinguish foreground from background clutter. This is reflected in the relatively low confidence scores (0.4–0.5) for some occluded vehicles in our qualitative results (Figure 6). Addressing these issues in future work could involve occlusion-aware data augmentation strategies and illumination-robust training techniques.

Second, although our model demonstrates consistent improvements across 15 diverse categories and three distinct scene types (urban, town, and airport), the experiments were conducted primarily on the SIMD dataset, which introduces a remote sensing bias. The model's performance may not directly generalize to other domains such as satellite imagery with different spatial resolutions, ground-level camera views, or maritime surveillance scenarios. We encourage future research to evaluate our method on a broader range of benchmarks (e.g., AI-TOD, DOTA) and to explore domain adaptation techniques for cross-dataset generalization.

Third, we acknowledge a computational trade-off between accuracy and efficiency. The integration of BiFPN, GLSA, and Wasserstein Loss increases the parameter count from 11.2M to 15.7M and GFLOPs from 28.6 to 42.0. On an NVIDIA RTX 4090, our model achieves 62 FPS, which comfortably exceeds real-time requirements. However, on resource-constrained edge devices such as the Jetson Xavier NX, the inference speed drops to approximately 14 FPS, approaching the lower bound of real-time performance. For edge deployment, future work should explore model compression techniques, including channel pruning, quantization, and knowledge distillation.

While the proposed model achieves substantial improvements on most categories, a notable exception is observed on categories with extremely limited training samples. As shown in Table 4, the AP@0.5 for pushback tractor declines from 47.2% to 36.7% (a 10.5 percentage point drop), and passenger stair vehicle from 45.6% to 42.5% (a 3.1 percentage point drop)—the lowest per-category AP among all 15 classes. We attribute this degradation to two factors: (1) severe underrepresentation of these categories in the training set, where the increased model capacity introduced by our three enhancement modules may exacerbate overfitting on minority classes; and (2) high intra-class variability coupled with visual similarity to other vehicle types (e.g., tractors vs. trucks), making them inherently difficult to distinguish. This limitation highlights the need for future work on class-balanced sampling, synthetic data generation, or few-shot learning to improve rare-category detection in long-tail scenarios.

5. Conclusions

This paper proposes a small-object detection model based on multi-module collaborative optimization, integrating BiFPN, GLSA, and Wasserstein Loss into the YOLOv8 framework to address the coupled challenges of insufficient feature representation, multi-scale fusion inefficiency, and unstable regression optimization. Systematic validations on the SIMD dataset demonstrate that the proposed model achieves substantial performance gains: the optimal F1-confidence threshold increases from 0.42 to 0.67, mAP@0.5 reaches 81.8% (a 0.4% improvement over the baseline), and the F1-score reaches 78% (a 1% gain). Category-specific improvements are particularly notable, with AP@0.5 for fighter jets rising by 9.9 percentage points (89.6% → 99.5%), followed by propeller aircraft (+2.4%) and helicopters (+3.4%). Ablation experiments confirm that the three modules function synergistically: BiFPN constructs multi-scale feature pyramids, GLSA enhances feature discriminability, and Wasserstein Loss optimizes bounding box regression, jointly forming a complete optimization pipeline from feature construction to selection to regression.

Despite these achievements, several limitations identified in our Discussion (Section 4) warrant acknowledgment, including performance degradation under heavy occlusion and drastic illumination variations, remote sensing bias due to single-dataset evaluation, computational trade-offs for edge deployment, and remaining challenges in detecting extremely small and heavily occluded objects. These limitations inform the following directions for future research.

First, to enhance robustness against occlusion and illumination variations, we plan to introduce occlusion-aware data augmentation strategies such as random erasing and copy-paste with occlusion simulation, design attention compensation modules that explicitly model occluded regions, and incorporate illumination-robust training techniques, including brightness jittering and contrast normalization.

Second, to facilitate deployment on resource-constrained edge devices, we will explore model compression techniques including channel pruning to remove redundant feature channels, knowledge distillation to transfer detection capability from the current model to a lightweight student network, and lightweight convolution architectures such as depthwise separable convolutions to reduce parameter count and computational cost.

Third, to improve rare-category detection in long-tail scenarios, we will investigate class-balanced sampling strategies to ensure minority categories are adequately represented during training, synthetic data generation techniques—particularly leveraging diffusion models or generative adversarial networks—to enrich sample diversity for under-represented classes, and few-shot learning approaches to enable effective detection with limited annotated examples.

Fourth, to further validate generalizability, we intend to conduct comprehensive evaluations on additional public benchmarks with distinct characteristics: AI-TOD (containing extremely small objects in aerial imagery), DOTA (featuring arbitrarily oriented objects with dense annotations), and VisDrone (covering diverse drone-captured scenes with various altitudes and lighting conditions).

Finally, confidence calibration techniques based on temperature scaling or isotonic regression will be explored to ensure that predicted confidence scores better reflect true detection correctness, thereby improving the reliability of the model's outputs in practical deployment scenarios.

In summary, this work provides an effective and reliable technical solution for real-time small-object detection in complex aerial scenes, while the identified limitations and future directions lay the groundwork for continued improvement.

Statement of the Use of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used ChatGPT for language polishing and grammar checking. The model design, algorithm implementation, experimental analysis, and conclusions were performed entirely by the authors, who take full responsibility for the content and integrity of this work.

Acknowledgments

The authors would like to thank the providers of the SIMD benchmark dataset for making the data publicly available. We also thank the colleagues who participated in the helpful discussions during the preparation of this work.

Author Contributions

Methodology and Investigation, J.H. and C.W.; Software, Writing—Original Draft Preparation, J.H. and C.W.; Writing—Review and Editing, J.H. and C.W.; Funding Acquisition, C.W. All authors have read and agreed to the published version of the manuscript.

Ethics Statement

This study does not involve any human participants, human data, or animal experiments. All data used in this research were obtained from publicly available benchmark datasets (SIMD) and have been anonymized by the original dataset providers. Therefore, ethical approval was not required for this work.

Informed Consent Statement

Not applicable, as this study does not involve human subjects.

Data Availability Statement

Publicly available SIMD benchmark dataset was used in this work. This dataset can be obtained from its official public repository following the original release terms.

Funding

This paper is funded by the Sichuan Provincial Natural Science Foundation (Grant no. 2026NSFSC0138) of China.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Chai XH, Hu Y, Lei YL, Liu X. Review on Intelligent Measurement and Control Technology of UAV. *Radio Eng.* **2019**, *49*, 855–860. Available online: <https://wxdg.cbpt.cnki.net/portal/journal/portal/client/paper/42d1f4e0018a9bbcd0cf3e979811ba33> (accessed on 10 October 2025).
- Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. DOI:10.1109/5.726791
- Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. DOI:10.1109/TPAMI.2016.2577031
- Noh H, Hong S, Han B. Learning Deconvolution Network for Semantic Segmentation. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1520–1528.
- Pirasteh S, Varshosaz M, Badrloo S, Li J. Developing an expansion-based obstacle detection using panoptic segmentation. *J. Field Robot.* **2024**, *41*, 1245–1264. DOI:10.1002/rob.22319
- Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. DOI:10.1109/TPAMI.2009.167
- Girshick R, Donahue J, Darrell T, Malik J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
- Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
- Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
- Zhou W, Wu Z, Zhang Z, et al. Lightweight Small Object Detection Method Based on Weak Feature Enhancement. *Control Decis.* **2024**, *39*, 381–390. DOI:10.13195/j.kzyjc.2022.1432
- Han Y, Du Z, Lu Z, et al. Research on Airborne Small Object Detection and Embedded Implementation Based on Adaptive Fusion of Multi-Scale Features. In Proceedings of the 12th China Conference on Command and Control, Beijing, China, 17–19 May 2024. DOI:10.1007/978-981-97-7774-7
- Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. DETRs Beat YOLOs on Real-time Object Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
- Li J, Liang X, Wei Y, Xu T, Feng J, Yan S. Perceptual Generative Adversarial Networks for Small Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1951–1959.
- Wang C, Zhong C. Adaptive Feature Pyramid Networks for Object Detection. *IEEE Access* **2021**, *9*, 107024–107032. DOI:10.1109/ACCESS.2021.3100369
- Doherty J, Gardiner B, Kerr E, Siddique N. BiFPN-YOLO: One-stage object detection integrating Bi-Directional Feature Pyramid Networks. *Pattern Recognit.* **2025**, *160*, 111209. DOI:10.1016/j.patcog.2024.111209
- Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
- Woo S, Park J, Lee JY, Kweon IS. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*; Springer: Cham, Switzerland, 2018; pp. 3–19.
- Hu X, Zhang P, Zhang Q, Yuan F. GLSANet: Global-Local Self-Attention Network for Remote Sensing Image Semantic Segmentation. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. DOI:10.1109/LGRS.2023.3235117
- Rezatofghi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
- Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. DOI:10.1609/aaai.v34i07.6999
- Wang J, Xu C, Yang W, Yu L. A Normalized Gaussian Wasserstein Distance for Tiny Object Detection. *arXiv* **2021**, arXiv:2110.13389. DOI:10.48550/arXiv.2110.13389
- Jia D, Yuan Y, He H, Wu X, Yu H, Lin W, et al. DETRs with Hybrid Matching. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 19702–19712.

23. Bodla N, Singh B, Chellappa R, Davis LS. Soft-NMS—Improving Object Detection with One Line of Code. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5562–5570.
24. Ma S, Li N, Peng G, Yang X, Hou Z. Object Detection Algorithm Based on Adaptive Feature Fusion and Cosine Similarity IoU-NMS. *J. Comput. Aided Des. Comput. Graph.* **2024**, *36*, 112–121. DOI:10.3724/SP.J.1089.2024.19786
25. Zhao W, Ma W, Jiao L, Chen P, Yang S, Hou B. Multi-Scale Image Block-Level F-CNN for Remote Sensing Images Object Detection. *IEEE Access* **2019**, *7*, 43607–43621. DOI:10.1109/ACCESS.2019.2908016
26. Guo Z, Zhong M, Duan S, Wang L. ESGN-YOLO: Enhancing Multi-Scale Small Object Detection via Efficient Feature Fusion and Adaptive Spatial Modeling. *IEEE Signal Process. Lett.* **2026**, *33*, 426–430. DOI:10.1109/LSP.2025.3644313
27. Huang Y, Qin R, Zhao G, Ji H, Zheng X, Zhong Y. Reconstruct Multiscale Features for Lightweight Small Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. DOI:10.1109/TGRS.2025.3644176
28. Liu X, Zou H, Li J, Pan L, He S, Cao X, et al. D²-DETR: DETR With Dual-Domain Frequency-Spatial Modeling for Unmanned Aerial Vehicle Imagery Object Detection. *Knowl. Based Syst.* **2026**, *341*, 115788. DOI:10.1016/j.knosys.2026.115788
29. Liu J, Zhang J, Ni Y, Chi W, Qi Z. Small-Object Detection in Remote Sensing Images with Super-Resolution Perception. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 15721–15734. DOI:10.1109/JSTARS.2024.3452707
30. Haroon M, Shahzad M, Fraz MM. Multisized Object Detection Using Spaceborne Optical Imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3032–3046. DOI:10.1109/JSTARS.2020.3000317
31. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016*; Springer: Cham, Switzerland, 2016; pp. 21–37.
32. Redmon J, Farhadi A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. DOI:10.48550/arXiv.1804.02767
33. Jocher G. *Ultralytics YOLOv5*, version 7.0; Ultralytics: Seattle, WA, USA, 2020. Available online: <https://github.com/ultralytics/yolov5> (accessed on 21 October 2024).