Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.078416

ARTICLE

A UAV Image Object Detection Algorithm Based on Deep Diverse Branch Block and Multi-Scale Auxiliary Feature

Wenfeng Wang[1,*], **Wenjie Fan**[1], **Fang Dong**[1], **Bin Zeng**[1], **Wenxin Yu**[1] and **Xiangping Deng**[2]

[1]School of Information Engineering, Jiangxi University of Water Resources and Electric Power, Nanchang, China
[2]Jiangxi Poyang Lake Water Control Project Construction Office, Nanchang, China

*Corresponding Author: Wenfeng Wang. Email: wangwf@nit.edu.cn

Received: 30 December 2025; Accepted: 08 April 2026; Published: 08 May 2026

ABSTRACT: Unmanned Aerial Vehicle (UAV) image object detection has been widely applied in many fields. However, compared with ordinary natural images, UAV images often exhibit complex backgrounds, a predominance of small objects, and significant variations in target scales, which cause traditional detection algorithms to easily suffer from missed or false detections with insufficient accuracy. To address these issues, this paper proposes a novel UAV image object detection algorithm named DMA-YOLO based on the YOLOv8s model, incorporating a deep diverse branch block and multi-scale auxiliary feature. First, a DF-C2f module integrating a deep diverse branch block and an adaptive fine-grained attention mechanism is designed to enhance small object detailed feature extraction. Second, a multi-scale auxiliary feature pyramid network (MAFPN) reconstructs the neck structure to strengthen multi-scale feature fusion and interaction, mitigating the impact of target scale variations. Finally, a dynamic detection head (DyHead) optimizes detection performance and model robustness, and the EIoU loss replaces the original CIoU to enhance bounding box regression accuracy and stability. Ablation experiments on the public VisDrone2019 dataset show that DMA-YOLO achieves a 3.7% increase in mAP50 and 2.8% in mAP50:95 compared with the baseline, with negligible changes in parameter counts and computational complexity. Comparative experiments with mainstream detection models and UAV-specific state-of-the-art algorithms confirm DMA-YOLO's superior detection accuracy. Further experiments on the RSOD dataset validate its generalization capability and stability across diverse data distributions, highlighting its applicability in complex UAV object detection scenarios.

KEYWORDS: Deep diverse branch block; fine-grained attention mechanism; MAFPN; DyHead; EIoU

1 Introduction

In recent years, UAV-based image object detection has been widely applied in such fields as military defense [1] and agricultural monitoring [2]. Traditional deep-learning based detection methods are divided into two-stage [3,4] and single-stage [5–8] approaches. Two-stage methods typically achieve higher accuracy but suffer from high computational cost, making them unsuitable for resource-constrained UAV platforms. In contrast, single-stage methods offer faster inference but often lag in detection accuracy. Moreover, UAV images are characterized by dense small objects, complex backgrounds, and large-scale variations, which significantly challenge core processes such as feature extraction, foreground-background separation, and multi-scale matching. This necessitates improving the accuracy of single-stage detection algorithms in UAV scenarios.

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

To address these issues, researchers have focused on optimizing key components of detection algorithms, including the backbone network, neck network, loss function, and detection head. The optimizations aim to enhance feature capture, multi-scale feature fusion, and overall detection performance. These optimizations have achieved some success in UAV image object detection, but small object detection remains challenging. In practical UAV scenarios, small objects have few pixels, leading to weak feature representation and high missed and false detection rates. Complex backgrounds also easily cause target noise confusion, which exacerbates detection errors. Most improved algorithms focus on feature fusion but do not sufficiently consider fine-grained feature extraction for small objects. This makes the low accuracy of small object detection a key bottleneck for the practical deployment of UAV-based detection technology.

To address these challenges, this paper proposes a UAV image object detection algorithm based on YOLOv8, namely DMA-YOLO, which integrates a deep diverse branch block and multi-scale auxiliary features to further enhance the model's perception capability and detection accuracy for small-scale objects. The main contributions of this study are summarized as follows:

(1) A DF-C2f module is designed by integrating a Deep Diverse Branch Block (DeepDBB) and an Adaptive Fine-Grained Attention Mechanism (AFGAM), which enhances the detailed feature capture capability of the backbone and elevates sensitivity toward small object features.
(2) A Multi-scale Auxiliary Feature Pyramid Network (MAFPN) is introduced into the neck to enable cross-scale interaction and efficient fusion of shallow and deep features, further strengthening multi-scale object feature representation.
(3) The Dynamic Head (DyHead) is adopted, which incorporates a multi-dimensional attention mechanism and a dynamic feature fusion strategy, thus improving detection accuracy and robustness under complex backgrounds.
(4) The CIoU loss is replaced with the EIoU loss, providing more reasonable gradient guidance and enhancing bounding box regression precision.

2 Related Work

Over the past few years, studies on UAV image object detection have mainly concentrated on tackling challenges including poor small-target detection performance, complex background disturbances, and insufficient multi-scale feature fusion capabilities. Current studies primarily refine mainstream algorithms like the YOLO series to enhance detection accuracy and mitigate practical issues such as false/missed detections and inadequate feature extraction.

To address the challenge of detecting small targets in UAV images, Liu et al. [9] integrated GhostNet into YOLOv5 to minimize detail loss during downsampling. Li et al. [10] reconstructed YOLOv8's C2f module with dilated residual convolutions to enhance contextual information fusion. Niu et al. [11] proposed VSTDet, utilizing orientation-selective coding and figure-ground segregation to improve feature representation. Ren et al. [12] proposed a scaled decoupled head that adaptively adjusts channel numbers to enhance small target feature representation, significantly boosting detection accuracy.

However, such methods still suffer from false and missed detections in complex backgrounds. Zhang et al. [13] improved RTMDet with a more robust feature extraction structure for better adaptability. Liang et al. [14] integrated a multi-path attention module into YOLOv7 to expand the receptive field and handle occlusions. Luo et al. [15] designed a reversible bidirectional feature pyramid with residual connections to preserve details and ensure stability. Drawing inspiration from visual pathways, Wang et al. [16] proposed RSVDet to suppress background interference via simulated ventral stream processing. Finally, Wang et al. [17] addressed target overlap and uneven distribution in YOLOv8n using space-to-depth layers and EMA attention.

Although effective in specific scenarios, these methods still struggle with small-object detection under multi-scale variations. Yi et al. [18] enhanced YOLOv8 with a dual-branch attention and an attention-guided bidirectional feature pyramid for better multi-scale fusion. Zhou et al. [19] designed a coordinate and global information aggregation module to integrate multi-scale contextual features, improving adaptability. Zhang et al. [20] utilized multimodal fusion and auxiliary super-resolution learning to distinguish features across different scales. Bakirci [21] validated the accuracy-efficiency balance of YOLOv8 in UAV traffic monitoring, offering a reference for practical optimization.

3 Baseline Algorithm

YOLOv8 [22] is a single-stage detector proposed by Ultralytics in 2023, consisting of Backbone, Neck, and Head. It has five scaled variants: n, s, m, l, x. Given the demand for both accuracy and real-time performance in UAV image detection, this study adopts YOLOv8s as the baseline. Despite newer versions such as YOLOv9 and YOLOv10, YOLOv8 is still widely used in academia and industry, especially in UAV vision tasks, due to its good reproducibility, stability, and mature community support. Its simple and flexible structure also facilitates targeted improvement, making YOLOv8s a representative and extensible baseline for this work.

The backbone consists of CBS, C2f, and SPPF. CBS first extracts initial features, then C2f fuses multi-scale feature maps to speed up extraction, and SPPF finally integrates the outputs of C2f. The neck uses PAN-FPN [23] to aggregate backbone features efficiently. The detection head adopts a task-decoupled design with separate losses for classification and box regression, enhancing convergence and detection performance. However, YOLOv8 has limitations: the C2f is weak in fine-grained feature extraction, failing to capture small object details. Besides, PAN-FPN insufficiently processes multi-scale features, easily neglecting shallow details and limiting cross-scale feature interaction.

4 Method

As previously mentioned, to address the shortcomings of YOLOv8, this paper proposes a novel UAV image object detection algorithm named DMA-YOLO. Fig. 1 illustrates its architecture.

The main improvements include: First, to remedy the deficiency in feature extraction of the original C2f module, a novel DF-C2f module is designed by integrating AFGAM and DeepDBB. This enhances the feature extraction capability without significantly increasing the parameter counts. Second, to alleviate the limited cross-scale feature interaction inherent in PAN-FPN, the neck is replaced with the MAFPN, thereby facilitating more comprehensive multi-level feature fusion and enhancing detection performance for small objects. Third, the DyHead is introduced to further enhance detection accuracy. Finally, the CIoU loss is replaced with the EIoU loss, providing a more effective bounding box regression strategy.

4.1 C2f Module Integrating Deep Diverse Branch Block and Adaptive Fine-Grained Attention

Though the C2f module handles standard-scale targets acceptably, it underperforms in UAV images with small, heavily occluded, or cluttered background targets, failing to preserve key discriminative clues for small objects and thus limiting detection accuracy.

To solve the above problems, this paper proposes the DF-C2f module, an improved feature extraction unit that integrates DeepDBB [24] and AFGAM [25] (see Fig. 2). It first processes input features via parallel DeepDBB units, which adopt a reparameterization strategy. The multi-branch structure captures rich multi-scale features in training, and each branch can be equivalently converted to a standard convolution. During inference, these branches are fused into a single convolution, maintaining the original structure and inference

speed of convolutional layers. The outputs of DeepDBB units are then concatenated and fed into AFGAM, an attention mechanism that uses global context to adaptively enhance fine-grained channel responses for small targets. This design enables DF-C2f to significantly boost small-object detection performance without sacrificing real-time inference.

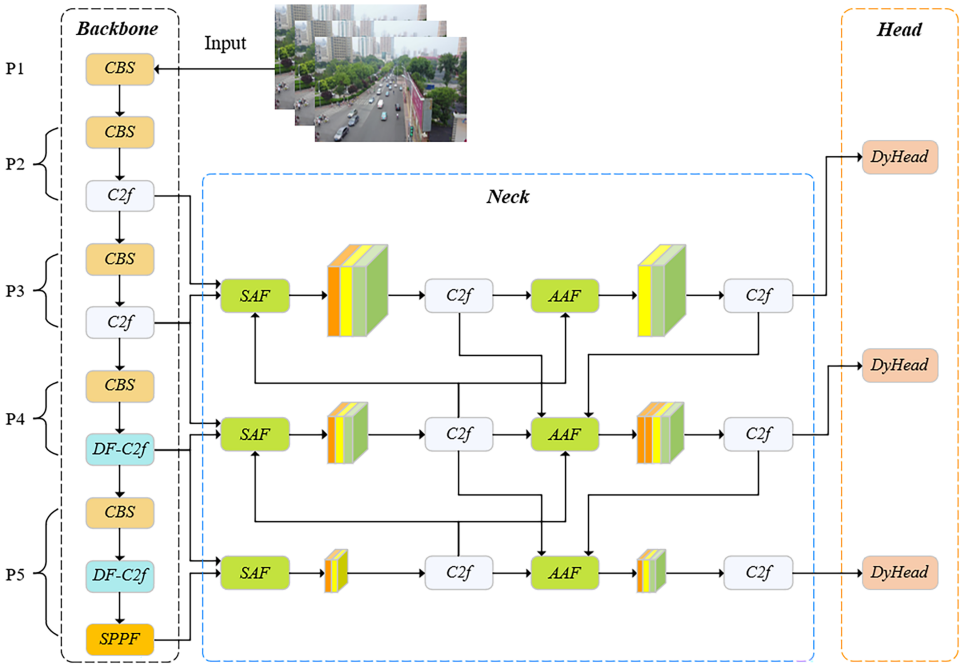


Figure 1: DMA-YOLO architecture.

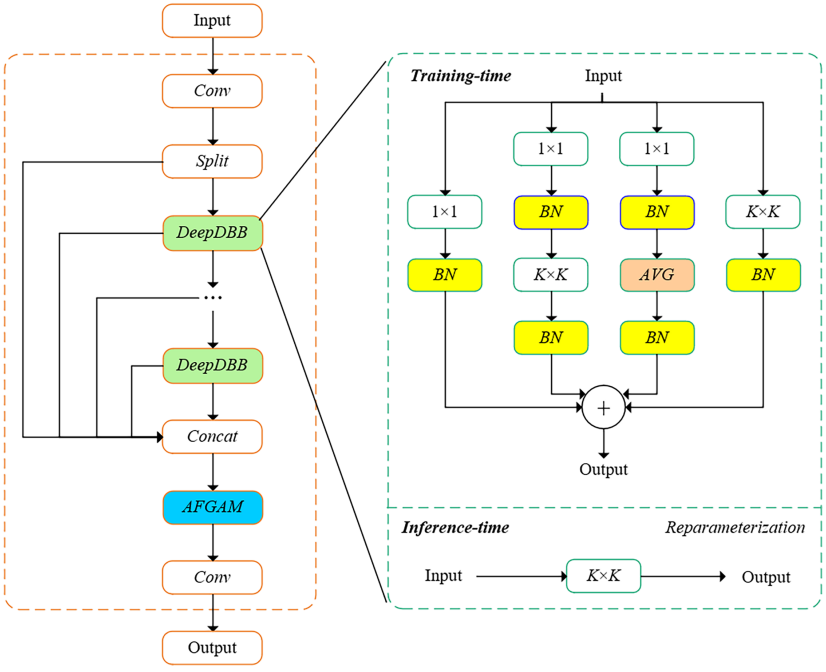


Figure 2: DF-C2f module.

DeepDBB adopts a multi-branch structure, integrating 1 × 1 convolution, $K \times K$ convolution, and average pooling operations. This structure expands the receptive field via parallel paths, effectively aggregates multi-scale spatial features, and yields richer representations to improve small-target detection. After training, all parallel branches are merged into a single convolution, fusing multi-receptive-field features. Batch normalization parameters are absorbed into adjacent convolutional layers. During inference, the model maintains a standard convolutional structure while preserving the powerful feature representation learned from multi-branch training.

Fig. 3 illustrates the pipeline of the AFGAM block, which improves performance by capturing global-local interactive features and optimizing feature weight allocation. First, the input feature map *FP* undergoes a global average pooling (*GAP*) operation to extract global contextual information, yielding channel-wise feature *U*. Second, a diagonal matrix and a band matrix are used to capture global dependency U_{gc} and local dependency U_{lc}, respectively, followed by cross-correlation to model their interaction. Finally, a learnable factor θ dynamically fuses global and local weights, which are applied to *FP* to obtain the refined feature map *FP**, enhancing sensitivity to small objects. Here, θ is initialized to −0.8. Following sigmoid mapping, asymmetric initial weights are assigned to the two cross-attention branches, guiding early-stage feature selection. These parameters are then adaptively updated via backpropagation for optimal cross-scale attention allocation, as formulated in Eqs. (1)–(5).

$$U = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} FP_n(i,j) \tag{1}$$

$$U_{lc} = \sum_{i=1}^{k} U \cdot b_i \tag{2}$$

$$U_{gc} = \sum_{i=1}^{c} U \cdot d_i \tag{3}$$

$$M = U_{gc} \cdot U_{lc}^{T} \tag{4}$$

$$FP^{*} = \left(\sigma\left(\sigma(\theta) \cdot \sigma\left(\sum_{j}^{c} M_{i,j} \right) + (1 - \sigma(\theta)) \cdot \sigma\left(\sum_{j}^{c} M_{i,j}^{T} \right) \right) \right) \otimes FP,\ i \in 1, 2, \ldots, c \tag{5}$$

here, b_i and d_i represent the weight matrices for local and global dependency ranges, respectively; c denotes the number of channels, M is the cross-correlation matrix, and σ represents the sigmoid activation function.

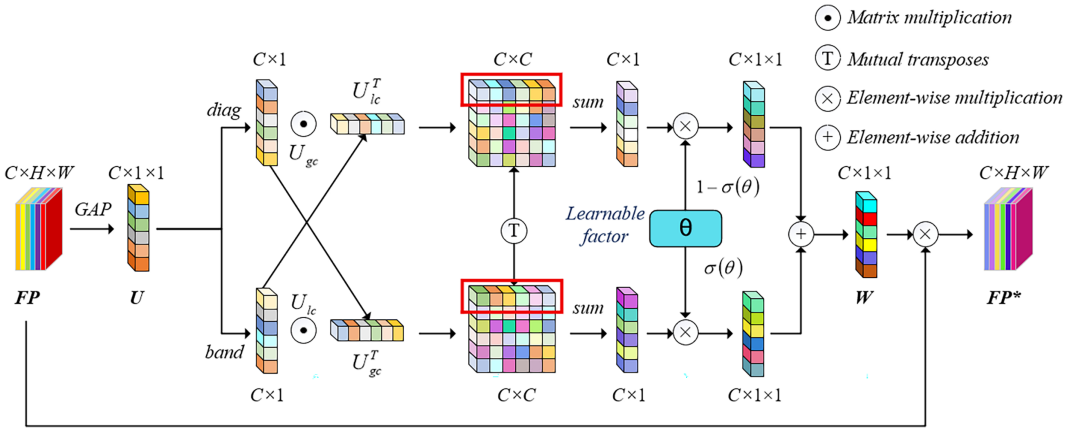


Figure 3: AFGAM block.

4.2 Multi-Scale Auxiliary Feature Pyramid Network

The YOLOv8 neck fuses multi-scale features from the backbone to boost semantic representation and detection performance. However, its PAN-FPN structure relies on deep feature-dominated fusion, leading to insufficient shallow feature retention and limited cross-scale interaction. Consequently, fine-grained details are seriously lost in UAV small object detection, degrading detection accuracy.

To address this issue, this paper introduces an improved MAFPN [26] to replace the original PAN-FPN. MAFPN uses two key components, Superficial Assisted Fusion (SAF) and Advanced Assisted Fusion (AAF), to achieve more efficient feature fusion. Through the collaboration of SAF and AAF, MAFPN integrates feature information from different layers more comprehensively and significantly improves the ability to detect small objects. The structures of SAF and AAF are shown in Fig. 4.

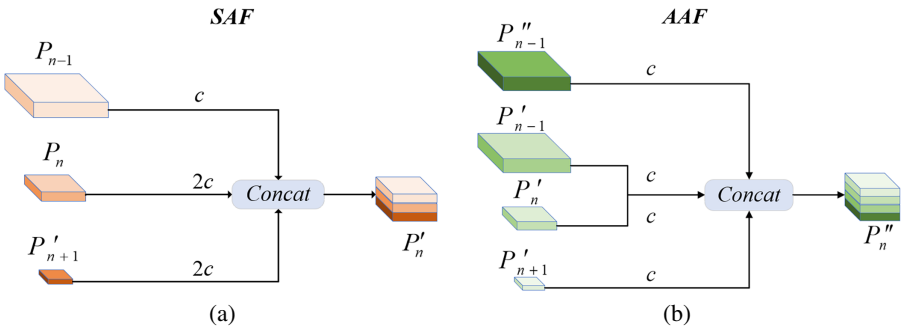


Figure 4: (**a**) SAF structure; (**b**) AAF structure.

To fully utilize the multi-scale features output by the backbone, this study classifies them into shallow and deep features based on their hierarchical depth: shallow features possess high spatial resolution and are mainly responsible for carrying target localization information, while deep features are rich in semantic information but have relatively low spatial resolution.

SAF enhances shallow feature representation via cross-hierarchical interaction. It fuses three inputs: early high-resolution features, current level features, and semantically enhanced features from deeper layers. All features are spatially aligned by upsampling or downsampling, unified in channel dimension with 1×1 convolution, and concatenated along channels. This strategy allows shallow features to retain high spatial precision while integrating rich semantic information, thus improving small target localization.

AAF further aggregates multi-scale enhanced features for consistent cross-hierarchical representation. It takes four inputs: current level features, high-resolution shallow features, low-resolution deep features, and enhanced features from shallower layers. After spatial alignment and channel normalization via 1×1 convolution, these features are concatenated to produce the final fused feature map. This mechanism promotes multi-scale feature collaboration and evidently improves detection for multi-scale targets.

4.3 Dynamic Detection Head

YOLOv8 employs an anchor-free decoupled head for multi-scale detection across resolutions, which performs well on ordinary objects. However, in UAV images with numerous small objects and complex backgrounds, the standard head struggles to capture key features, leading to frequent missed and false detections. Though it handles multi-resolution feature maps for objects of varying sizes, it fails to balance detection accuracy between small and large objects. For small targets, continuous pooling degrades critical information in low-resolution feature maps, severely reducing detection precision.

To address these issues, we adopt the dynamic detection head DyHead [27]. Using a dynamic attention, DyHead adaptively adjusts the feature tensor $F \in R^{L\times S\times C}$ across scale, space, and channel, then reweights it to strengthen small target feature representation. It consists of three core modules: scale-aware attention, spatial-aware attention, and task-aware attention.

By stacking attention mechanisms, DyHead unifies features across scales, spatial domains, and tasks for dynamic selection and optimization. It preserves fine-grained details while integrating global context, significantly enhancing small-target detection, as formalized in Eq. (6).

$$W(F) = \pi_C\left(\pi_S\left(\pi_L(F)\cdot F\right)\cdot F\right)\cdot F \tag{6}$$

here, $W(F)$ represents the final dynamic weight, $\pi_L(F)$ denotes the scale-aware attention module function, which achieves scale feature fusion by adaptively assigning weights to multi-scale feature maps. $\pi_S(F)$ denotes the spatial-aware attention module function, which optimizes the weights of the feature tensor in the spatial domain based on spatial position information. $\pi_C(F)$ denotes the task-aware attention module function, which implements task-oriented reweighting of channel features for the classification and regression branches of the detection task.

4.4 EIoU Loss

YOLOv8 uses CIoU for bounding box regression loss. Although CIoU outperforms IoU, it still has limitations. It cannot flexibly assign optimization weights between difficult targets and easy ones, leading to poor regression for small or occluded objects. Besides, the aspect ratio penalty in CIoU is not fully decoupled, causing unstable regression and reduced accuracy in some scenarios.

EIoU [28] further decouples the aspect ratio factors of predicted and ground truth boxes based on CIoU. It separately computes their width and height differences to increase overlap and make regression more efficient. It also introduces a center distance penalty to reduce alignment errors, allowing predicted boxes to converge more accurately to ground truth positions. This effectively solves the mismatch between boxes. Therefore, this paper adopts EIoU as the model loss function. Its calculation is shown in Eq. (7).

$$\text{EIoU} = 1 - \text{IoU} + \rho^2\left(b, b^{gt}\right)/c^2 + \left(w_p - w_t\right)^2/w_t^2 + \left(h_p - h_t\right)^2/h_t^2 \tag{7}$$

here, IoU refers to the Intersection over Union between the predicted bounding box and the ground truth bounding box. $\rho^2(b, b_{gt})/c^2$ is the center distance constraint term, representing the normalized Euclidean distance between the centers of the two boxes. $(w_p - w_t)^2/w_t^2$ is the width constraint term, indicating the width difference between the predicted and ground truth bounding boxes; $(h_p - h_t)^2/h_t^2$ is the height constraint term, denoting the height difference between them.

5 Experimental Results and Analysis

5.1 Experimental Environment

All experiments in this study are conducted on the same computing device. The device runs on the Windows 10 operating system, configured with an Intel Core i7-13700KF CPU, 32 GB of system memory, and an NVIDIA GeForce RTX 4080 graphics card with 16 GB of VRAM. Python 3.8 is adopted as the programming language, and the deep learning framework employed is PyTorch 2.3.0, integrated with CUDA 12.1 to enable parallel computing acceleration.

For model training, the total number of epochs is set to 300, 4 threads are allocated for data loading, input images are uniformly resized to 640 × 640 pixels, the batch size is configured as 8, and a fixed random seed of 0 is used throughout the process; SGD is selected as the optimizer with a momentum weight of

0.937, a linear learning rate scheduling strategy is implemented with an initial learning rate of 0.01 and a final learning rate decaying to 0.0005, and all models are trained from scratch without pre-trained weights; an early stopping mechanism is integrated with a patience threshold of 50, where training is automatically terminated if no significant improvement in detection accuracy is observed within 50 consecutive epochs, and a data augmentation strategy is adopted during training, including Mosaic augmentation, random flipping, random rotation (±20°), color jittering, and random cropping.

5.2 Datasets

Two public datasets, VisDrone2019 [29] and RSOD [30], are employed for experimental validation in this study. VisDrone2019, curated by the AISKYEYE team of Tianjin University, was acquired via diverse UAV platforms under varying scenarios, weather patterns, and lighting conditions. It covers ten object classes (Pedestrian, People, Bicycle, Car, Van, Truck, Tricycle, Awning-Tricycle, Bus, and Motorcycle) with a total of 8629 images, split into a training set (6741 images), a validation set (548 images), and a test set (1610 images). Owing to its scenario diversity and comprehensive object coverage, it serves as a critical benchmark for assessing model performance in UAV-based object detection. RSOD is a dedicated dataset for object detection tasks, encompassing 976 images in total. These images are categorized into four classes: 446 images containing 4993 aircraft, 165 images with 1586 oil tanks, 176 images featuring 180 overpasses, and 189 images including 191 playgrounds. For this study, the RSOD dataset is randomly divided into training, validation, and test subsets following a 7:1:2 ratio, resulting in 683, 98, and 195 images respectively.

5.3 Evaluation Metrics

In this study, we utilized precision (P), recall (R), mean average precision (mAP), parameters (Params), floating-point operations (FLOPs), and frames per second (FPS) as the performance metrics. Params reflects the model size in terms of weight count, FLOPs measure the computational complexity of the algorithm, and FPS represents the number of images processed by the model per second, which characterizes its real-time inference speed. To facilitate intuitive comparison and reading, the evaluation metric names and the optimal results in each experiment are highlighted in bold in Tables 1–6. The formulas are shown in Eqs. (8)–(10):

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{n} AP_i \tag{10}$$

5.4 Ablation Analysis on Improved C2f Modules

To clarify the individual contribution of each proposed component, we performed ablation experiments by separately incorporating DeepDBB and AFGAM into the original C2f module. As presented in Table 1, the DeepDBB module brings a 1.3% improvement in mAP50, and the AFGAM module yields a 0.8% performance gain. Moreover, the DF-C2f module, which integrates both structures simultaneously, achieves a 2.1% increase in mAP50. These results confirm that both designs effectively enhance feature representation and detection performance, and their complementary combination in DF-C2f further validates the effectiveness of our proposed improvement.

5.5 Ablation Experiments

To validate the effectiveness of each improvement, ablation experiments are conducted on the VisDrone2019 test set, with results shown in Table 2.

Table 1: Ablation study of different improved C2f modules on model performance.

Models	P%	R%	mAP50%	mAP50:95%
C2f	44.8	33.7	31.8	18.1
DeepDBB-C2f	45.7	34.6	33.1	19.2
AFGAM-C2f	45.2	34.8	32.6	19.0
DF-C2f	**46.2**	**35.6**	**33.9**	**19.7**

Table 2: Ablation experiments on the VisDrone2019 test set.

	Models	P%	R%	mAP50%	mAP50:95%	FLOPs	Params	FPS
A	YOLOv8s	44.8	33.7	31.8	18.1	28.5G	11.1M	216
B	YOLOv8s + DF-C2f	46.2	35.6	33.9	19.7	28.5G	11.4M	209
C	YOLOv8s + MAFPN	46.5	35.1	33.5	19.5	31.2G	11.2M	178
D	YOLOv8s + DyHead	45.4	34.1	32.6	18.8	**28.1G**	**10.8M**	215
E	YOLOv8s + EIoU	44.3	33.6	32.3	18.6	28.5G	11.1M	**224**
F	YOLOv8s + DF-C2f + MAFPN	46.6	35.9	34.6	20.2	31.2G	11.5M	188
G	YOLOv8s + DF-C2f + MAFPN + DyHead	47.1	36.2	35.1	20.6	31.2G	11.6M	195
O	YOLOv8s + DF-C2f + MAFPN + DyHead + EIoU	**47.7**	**36.8**	**35.5**	**20.9**	31.2G	11.6M	204

In Table 2, A denotes the baseline YOLOv8s; B, C, D, and E represent improved variants incorporating DF-C2f, MAFPN, DyHead, and EIoU, respectively, based on the baseline; F integrates both the DF-C2f and the MAFPN; G combines the three modules (DF-C2f, MAFPN, and DyHead); and O represents the final proposed model, i.e., the complete DMA-YOLO, obtained by further introducing the EIoU loss function upon G.

As shown in Table 2, B achieves improvements of 2.1% in mAP50 and 1.6% in mAP50:95 over the baseline. This demonstrates that the DF-C2f module, which integrates a diverse branch block and adaptive fine-grained attention, can more effectively extract small-target information. Benefiting from its reparameterized architecture, both computational cost and parameter count remain nearly unchanged. C achieves gains of 1.7% and 1.4%, confirming that MAFPN enhances detection accuracy through effective multi-scale feature fusion. Although it incurs slightly higher computation, its parameter count stays almost constant, still meeting real-time inference requirements on UAV platforms. D improves the two metrics by 0.8% and 0.7%, respectively, with slight reductions in both computational cost and parameter count compared to the baseline, suggesting that the introduced dynamic detection head (DyHead) enhances detection performance more efficiently. E incorporates the EIoU loss and achieves a 0.5% improvement in both metrics while maintaining the same parameter count and computational cost. F yields improvements of 2.8% and 2.1%, and G further increases these gains to 3.3% and 2.5%, fully validating the effectiveness of the proposed module integration strategy. Finally, the fully optimized Model O achieves the best performance with a 3.7% gain in mAP50 and a 2.8% gain in mAP50:95, comprehensively demonstrating the efficacy and synergistic optimization capability of the proposed improvements. Furthermore, to evaluate the performance of the

improved model across different object categories, this study conducted a class-wise comparative experiment between DMA-YOLO and the baseline YOLOv8s on the VisDrone2019 and RSOD datasets. The results are shown in Tables 3 and 4.

Table 3: The mAP50 and mAP50:95 values for each category on the VisDrone2019 dataset.

Categories	Validation				Test			
	mAP50%		mAP50:95%		mAP50%		mAP50:95%	
	YOLOv8s	DMA-YOLO	YOLOv8s	DMA-YOLO	YOLOv8s	DMA-YOLO	YOLOv8s	DMA-YOLO
Pedestrian	43.5	**46.3**	19.9	**21.7**	27.5	**28.8**	11.1	**11.7**
People	33.2	**36.0**	12.6	**14.1**	14.6	**16.4**	5.1	**5.77**
Bicycle	13.6	**13.7**	5.86	**6.15**	9.2	**10.5**	3.6	**4.35**
Car	80.0	**81.4**	57.3	**59.1**	71.6	**74.2**	45.0	**47.0**
Van	44.8	**48.8**	30.6	**34.6**	36.4	**42.9**	23.4	**28.2**
Truck	36.5	**40.0**	23.7	**27.8**	36.5	**43.6**	23.0	**28.4**
Tricycle	29.0	**31.1**	16.6	**18.3**	18.3	**21.1**	9.4	**12.0**
Awn-Tricy	15.6	**19.2**	10.1	**12.0**	17.8	**21.3**	9.8	**12.2**
Bus	57.0	**61.5**	41.7	**45.4**	56.2	**59.4**	38.5	**42.2**
Motor	44.8	**47.4**	20.0	**22.0**	29.5	**31.7**	12.0	**13.1**
All	39.8	**42.9**	23.9	**26.1**	31.8	**35.5**	18.1	**20.9**

Table 4: The mAP50 and mAP50:95 values for each category on the RSOD dataset.

Categories	Validation				Test			
	mAP50%		mAP50:95%		mAP50%		mAP50:95%	
	YOLOv8s	DMA-YOLO	YOLOv8s	DMA-YOLO	YOLOv8s	DMA-YOLO	YOLOv8s	DMA-YOLO
Aircraft	97.3	**97.4**	65.6	**67.0**	98.3	**98.4**	63.6	**64.3**
Oiltanks	97.1	**97.3**	78.9	**79.7**	98.5	**99.1**	79.1	**79.7**
Overpass	85.7	**92.2**	39.7	**44.1**	90.5	**94.2**	27.0	**34.1**
Playground	99.5	**99.5**	83.0	**88.8**	72.2	**79.6**	49.3	**54.6**
All	94.9	**97.2**	67.3	**69.8**	89.9	**93.1**	54.8	**57.9**

On the VisDrone2019 dataset, the mAP50 metrics for the Van, Truck, Awning-Tricycle, and Bus categories increased by 4.0%, 3.5%, 3.6%, and 4.5% on the validation set, and by 6.5%, 7.1%, 3.5%, and 3.2% on the test set. For the mAP50:95 metric, these three categories achieve steady improvements of 4.0%, 4.1%, and 3.7% on the validation set, and 4.8%, 5.4%, and 3.7% on the test set. Other categories also exhibit varying degrees of enhancement, strongly validating the effectiveness of the proposed strategy. In contrast, Bicycle and Tricycle show relatively minor gains, mainly due to their low sample proportion, small size, and similar appearance in the dataset, which increases detection difficulty.

On the RSOD dataset, the mAP50 for the Overpass category increased by 6.5% on the validation set, while Overpass and Playground increased by 3.7% and 7.4% on the test set, respectively. Regarding the mAP50:95 metric, these two categories increased by 4.4% and 5.8% on the validation set, and by 7.1% and 5.3% on the test set, showing substantial improvements. The Aircraft and Oiltank categories also exhibited improvements to varying degrees.

These results demonstrate that the improved model effectively enhances detection accuracy for small and scale-varying objects, significantly boosting overall performance and validating its effectiveness and generalization.

5.6 Comparative Experiments with Multiple SOTA Detection Algorithms

To further validate the superiority of the improved model, several current state-of-the-art object detection algorithms are selected for comparative experiments on the test set. All models are retrained and tested under identical experimental settings throughout the study. The experimental results are presented in Tables 5 and 6.

Table 5: Comparative experiments with state-of-the-art algorithms on the VisDrone2019 dataset.

Models	P%	R%	mAP50%	mAP50:95%	FLOPs	Params	FPS
SSD	/	/	19.3	13.6	85.4G	25.1M	54
Faster-RCNN	/	/	29.6	17.8	207.5G	41.4M	43
YOLOv5s	42.5	31.5	30.4	16.9	16.1G	7.2M	92
YOLOv6s	43.2	33.2	31.3	17.5	44.2G	16.5M	74
YOLOv7-tiny	44.6	31.9	27.8	14.2	**13.2G**	**6.2M**	193
YOLOv8s	44.8	33.7	31.8	18.1	28.5G	11.1M	216
YOLOv9s [31]	44.5	33.9	32.2	18.2	26.5G	7.3M	202
YOLOv10s [32]	44.6	34.6	32.4	18.4	21.4G	7.4M	**231**
YOLOv11s [33]	43.2	33.8	31.7	18.3	21.3G	9.4M	225
SDA-YOLO [34]	45.9	35.1	33.4	19.1	30.3G	10.9M	227
SD-YOLO [35]	46.9	36.2	33.6	19.5	37.7G	**4.3M**	192
YOLO-GE-s [36]	**49.2**	**38.4**	**36.8**	**22.3**	50.7G	13.0M	102
DMA-YOLO	47.7	36.8	35.5	20.9	31.2G	11.6M	204

Table 6: Comparative experiments with state-of-the-art algorithms on the RSOD dataset.

Models	P%	R%	mAP50%	mAP50:95%
SSD	/	/	78.8	41.9
Faster-RCNN	/	/	84.5	46.3
YOLOv5s	85.2	83.8	87.6	51.2
YOLOv6s	86.5	84.2	88.3	51.8
YOLOv7-tiny	85.1	81.2	85.7	49.5
YOLOv8s	87.1	85.4	89.6	54.8
YOLOv9s	89.3	84.7	90.0	55.3
YOLOv10s	90.2	85.1	90.5	55.5
YOLOv11s	89.4	85.3	89.8	55.4
SDA-YOLO	90.4	85.6	91.4	56.0
SD-YOLO	90.8	86.0	91.7	56.8
YOLO-GE-s	**95.1**	**88.7**	**94.5**	**59.2**
DMA-YOLO	93.2	87.5	93.1	57.9

Experimental results show that traditional detectors such as SSD and Faster R-CNN have low detection efficiency, high complexity, heavy computation, and large parameters, making them unsuitable for UAV platforms with real-time and hardware constraints. Among the YOLO series, YOLOv5 balances accuracy and model size but performs moderately. YOLOv6 improves accuracy at the cost of higher computation and parameters. YOLOv7 is lighter with less computation but loses accuracy. In comparison, newer

YOLO series models (YOLOv8, YOLOv9, YOLOv10, and YOLOv11) enhance detection performance while maintaining lightweight designs, though they still exhibit insufficient capability in detecting small objects in UAV scenarios.

To comprehensively evaluate the performance of DMA-YOLO, we compare it with recent advanced variants—SDA-YOLO, SD-YOLO, and YOLO-GE-s. SDA-YOLO balances accuracy and efficiency but offers no clear edge in either. SD-YOLO cuts Params to 4.3M yet raises FLOPs with minor accuracy gains. YOLO-GE-s boosts accuracy significantly but at the cost of 77.9% more FLOPs and 14.6% more Params. Its excessive overhead limits FPS to 102, failing to meet real-time needs for resource-constrained UAV platforms. In contrast, DMA-YOLO delivers competitive accuracy gains while maintaining 204 FPS, enabling efficient real-time inference for resource-constrained UAV small-object detection.

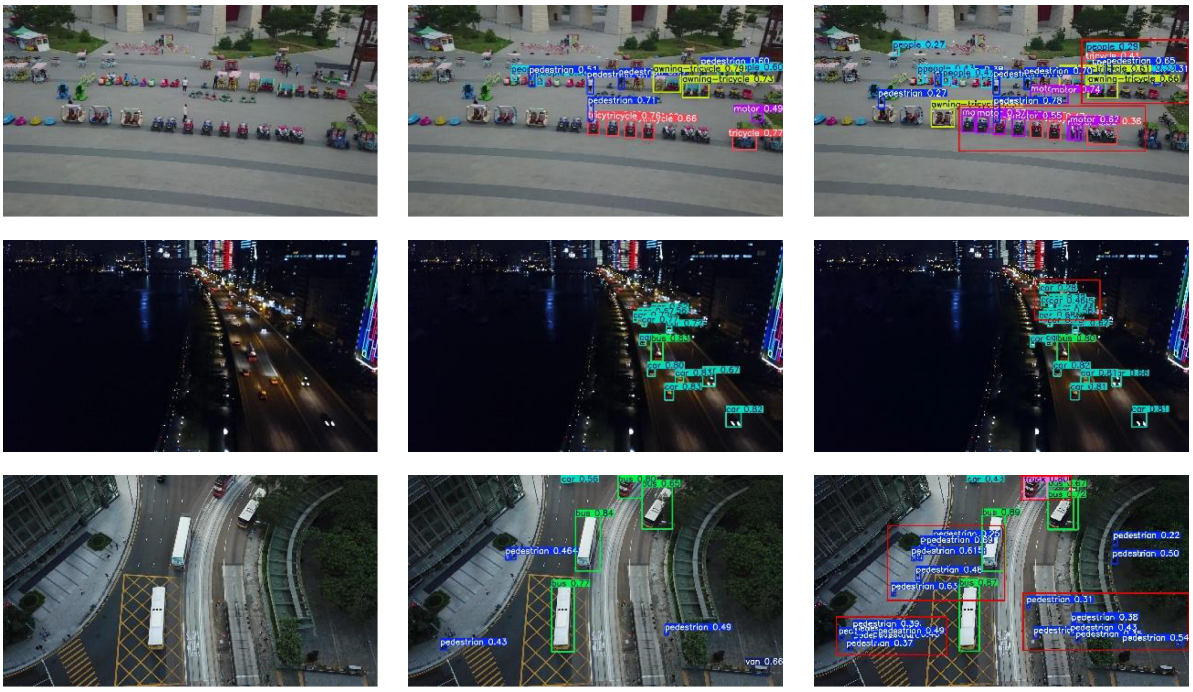
The proposed DMA-YOLO model demonstrates superior performance over mainstream detectors in comprehensive experiments. It attains excellent results in P, R, mAP50, and mAP50:95, with only 2.7G FLOPs and 0.5M parameters added over the baseline. Its 204 FPS inference speed satisfies the real-time demands of UAV platforms. With a favorable trade-off between accuracy and efficiency, DMA-YOLO is well-suited for edge deployment. Overall, it effectively improves small-object detection in UAV images while ensuring real-time inference, yielding high practical value and application potential.

5.7 Visualization of Object Detection Results

Several representative scenarios with predominant small objects, complex backgrounds, and significant scale variations of objects were selected from the VisDrone2019 dataset for visual comparison of detection results between DMA-YOLO and the baseline. To clearly demonstrate the detection advantages of DMA-YOLO, prominent red boxes are used to mark typical missed detections and false alarms of the baseline model in Fig. 5. The results are shown in Fig. 5.

As can be seen from Fig. 5, in small object detection scenarios, YOLOv8s exhibits obvious missed detections and false alarms, while the improved DMA-YOLO demonstrates superior detection performance in such scenarios, effectively alleviating the aforementioned issues. In complex nighttime low-light scenes, DMA-YOLO can identify more low-visibility targets at greater distances, indicating its stronger target perception capability. Finally, in dense multi-scale scenes, the baseline misidentified a truck as a bus and missed many small objects, whereas DMA-YOLO accurately identified corresponding targets and detected more objects. In summary, compared to the baseline, DMA-YOLO exhibits superior detection performance and adaptability in various complex UAV-based object detection tasks.

Fig. 6 visualizes the detection results of the baseline and DMA-YOLO algorithms on the RSOD dataset, where the proposed method again achieves more reliable and accurate object detection compared to the baseline.



(a) Original images (b) YOLOv8s (c) DMA-YOLO

Figure 5: Visualization of object detection results on the VisDrone2019 dataset under three challenging conditions. Red boxes denote typical missed detections and false alarms of the baseline: improved small-object detection with fewer misses and false positives (top); enhanced detection of distant, low-visibility targets in nighttime scenes (middle); and more accurate classification with better small-object recovery in dense multi-scale environments (bottom).

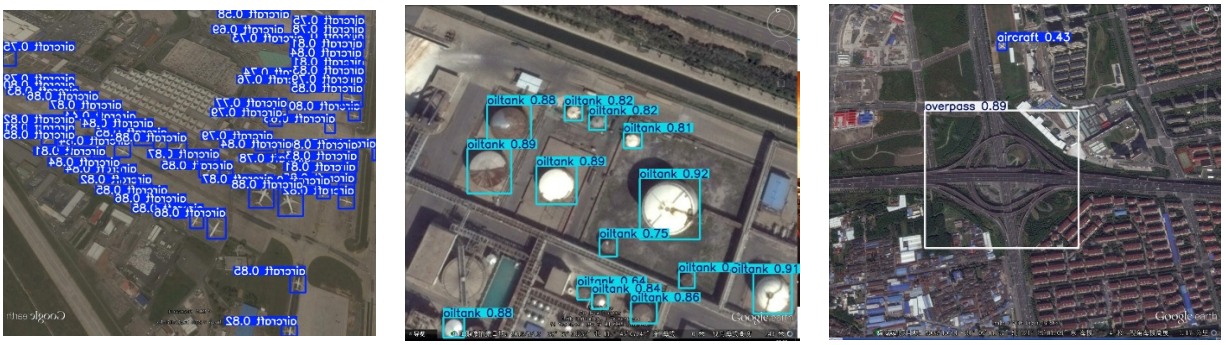


Figure 6: Visualization of object detection results on the RSOD dataset.

6 Conclusions

To address low small target detection precision, frequent missed detection, and high false alarm rates in UAV image analysis, this study proposes the lightweight DMA-YOLO architecture optimized based on YOLOv8s. This enhanced framework integrates four core improvement modules: the DF-C2f multi-branch structure for strengthened feature extraction; MAFPN for improved multi-scale feature fusion efficiency; a dynamic attention head for adaptive feature representation; and the EIoU loss function for optimized bounding box regression, enhancing target localization precision and overall model accuracy.

Despite these improvements, the current model still has room for further optimization on ultra-low-power edge devices, especially in real-world deployment with increasingly constrained computational resources. Future work will focus on developing more efficient convolutional modules to reduce computational cost and designing a lightweight neck network tailored for resource-constrained environments to lower model complexity, aiming to significantly improve the proposed method's engineering applicability in UAV systems with limited computing resources.

Acknowledgement: Not applicable.

Funding Statement: This work was supported in part by the Jiangxi Provincial Department of Water Resources Science & Technology Program Foundation (Grant Nos. 202325ZDKT17, 202426ZDKT13).

Author Contributions: Research conception and scheme design, Wenfeng Wang and Wenjie Fan; Algorithm effectiveness verification, Fang Dong and Bin Zeng; Formal analysis, Xiangping Deng and Wenxin Yu; Original draft writing, Wenjie Fan; Manuscript review and revision, Wenfeng Wang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: This study exclusively utilizes publicly available datasets. The source code can be obtained from the following repository: https://github.com/CCKaoYa/DMA-YOLO.git (accessed on 03 February 2026).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ma Q, Zhu B, Zhang H, Zhang Y, Jiang Y. Low-altitude UAV detection and recognition method based on optimized YOLOv3. Laser Optoelectron Prog. 2019;56(20):201006. doi:10.3788/LOP56.201006.
2. Castineira S, Delwar T, Duran R, Pala N. UAV-based agricultural monitoring and data acquisition system for precision farming. In: Proceedings of the Sensing for Agriculture and Food Quality and Safety XIII; 2021 Apr 12–17; Online. p. 55–65. doi:10.1117/12.2587914.
3. 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(6):1137–49. doi:10.1109/TPAMI.2016.2577031.
4. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):386–97. doi:10.1109/TPAMI.2018.2844175.
5. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot MultiBox detector. In: Computer vision—ECCV 2016. Cham, Switzerland: Springer International Publishing; 2016. p. 21–37. doi:10.1007/978-3-319-46448-0_2.
6. Tang S, Zhang S, Fang Y. HIC-YOLOv5: improved YOLOv5 for small object detection. In: Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan. p. 6614–9. doi:10.1109/ICRA57147.2024.10610273.
7. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, et al. YOLOv6: a single-stage object detection framework for industrial applications. arXiv:2209.02976. 2022.
8. Wang CY, Bochkovskiy A, Liao HM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696. 2022.
9. Liu S, Cao L, Li Y. Lightweight pedestrian detection network for UAV remote sensing images based on strideless pooling. Remote Sens. 2024;16(13):2331. doi:10.3390/rs16132331.
10. Li J, Chen Y, Niu M, Zhang X, Wang L. ADS-YOLO: a multi-scale feature extraction remote sensing image object detection algorithm. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:1234–45. doi:10.1109/ACCESS.2025.3538548.

11. Niu Y, Lin C, Jiang X, Qu Z. VSTDet: a lightweight small object detection network inspired by the ventral visual pathway. Appl Soft Comput. 2025;171(3):112775. doi:10.1016/j.asoc.2025.112775.
12. Ren Z, Yao K, Sheng S, Wang B, Lang X, Wan D, et al. YOLO-SDH: improved YOLOv5 using scaled decoupled head for object detection. Int J Mach Learn Cybern. 2025;16(3):1643–60. doi:10.1007/s13042-024-02357-3.
13. Zhang J, Zhang J, Zhou K, Zhang Y, Chen H, Yan X. An improved YOLOv5-based underwater object-detection framework. Sensors. 2023;23(7):3693. doi:10.3390/s23073693.
14. Liang K, Zhang W, Li F, Jiang Z, Zhang D. Improved YOLOv7 for small and overlapping objects detection. In: Proceedings of the 2023 IEEE 6th International Conference on Pattern Recognition and Artificial Intelligence (PRAI); 2023 Aug 18–20; Haikou, China. New York, NY, USA: IEEE; 2023. p. 153–7. doi:10.1109/PRAI59366.2023.10332096.
15. Luo F, Bian W, Jie B, Dong H, Fu X. ARBFPN-YOLOv8: auxiliary reversible bidirectional feature pyramid network for UAV small target detection. Signal Image Video Process. 2024;19(1):63. doi:10.1007/s11760-024-03661-9.
16. Wang X, Lin C, Pan Y, Wang R. RSVDet: remote sensing small object detection model inspired by neuronal mechanisms in visual pathways. IEEE Trans Circuits Syst Video Technol. 2026;36(4):4021–36. doi:10.1109/TCSVT.2025.3628010.
17. Wang Y, Zhang K, Wang L, Wu L. An improved YOLOv8 algorithm for rail surface defect detection. IEEE Access. 2024;12:44984–97. doi:10.1109/ACCESS.2024.3380009.
18. Yi H, Liu B, Zhao B, Liu E. Small object detection algorithm based on improved YOLOv8 for remote sensing. IEEE J Sel Top Appl Earth Observations Remote Sensing. 2024;17:1734–47. doi:10.1109/jstars.2023.3339235.
19. Zhou L, Liu Z, Zhao H, Hou YE, Liu Y, Zuo X, et al. A multi-scale object detector based on coordinate and global information aggregation for UAV aerial images. Remote Sens. 2023;15(14):3468. doi:10.3390/rs15143468.
20. Zhang J, Lei J, Xie W, Fang Z, Li Y, Du Q. SuperYOLO: super resolution assisted object detection in multimodal remote sensing imagery. IEEE Trans Geosci Remote Sens. 2023;61:5605415. doi:10.1109/TGRS.2023.3258666.
21. Bakirci M. Advanced aerial monitoring and vehicle classification for intelligent transportation systems with YOLOv8 variants. J Netw Comput Appl. 2025;237(B):104134. doi:10.1016/j.jnca.2025.104134.
22. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8 [Internet]. 2023 [cited 2026 Jan 1]. Available from: https://github.com/ultralytics/ultralytics.
23. Mei S, Shi Y, Gao H, Tang L. Research on fabric defect detection algorithm based on improved YOLOv8n algorithm. Electronics. 2024;13(11):2009. doi:10.3390/electronics13112009.
24. Wan D, Lu R, Hu B, Yin J, Shen S, Xu T, et al. YOLO-MIF: improved YOLOv8 with Multi-Information fusion for object detection in Gray-Scale images. Adv Eng Inform. 2024;62(3):102709. doi:10.1016/j.aei.2024.102709.
25. Sun H, Wen Y, Feng H, Zheng Y, Mei Q, Ren D, et al. Unsupervised bidirectional contrastive reconstruction and adaptive fine-grained channel attention networks for image dehazing. Neural Netw. 2024;176(3):106314. doi:10.1016/j.neunet.2024.106314.
26. Yang Z, Guan Q, Zhao K, Yang J, Xu X, Long H, et al. Multi-branch auxiliary fusion YOLO with re-parameterization heterogeneous convolutional for accurate object detection. In: Pattern recognition and computer vision. Singapore: Springer Nature; 2024. p. 492–505. doi:10.1007/978-981-97-8858-3_34.
27. Xie M, Wang W, Zhou B. A Four chaos target detection algorithm for rivers and lakes based on improved YOLOv5. J Nanchang Inst Technol. 2025;44(14):91–7. (In Chinese). doi:10.21203/rs.3.rs-5208143/v1.
28. Zhang YF, Ren W, Zhang Z, Jia Z, Wang L, Tan T. Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing. 2022;506(9):146–57. doi:10.1016/j.neucom.2022.07.042.
29. Zhu P, Wen L, Du D, Bian X, Fan H, Hu Q, et al. Detection and tracking meet drones challenge. IEEE Trans Pattern Anal Mach Intell. 2022;44(11):7380–99. doi:10.1109/tpami.2021.3119563.
30. Long Y, Gong Y, Xiao Z, Liu Q. Accurate object localization in remote sensing images based on convolutional neural networks. IEEE Trans Geosci Remote Sens. 2017;55(5):2486–98. doi:10.1109/TGRS.2016.2645610.
31. Wang CY, Yeh IH, Liao HM. YOLOv9: learning what you want to learn using programmable gradient information. arXiv:2402.13616. 2024. doi:10.48550/arXiv.2402.13616.
32. Sun H, Yao G, Zhu S, Zhang L, Xu H, Kong J. SOD-YOLOv10: small object detection in remote sensing images based on YOLOv10. IEEE Geosci Remote Sens Lett. 2025;22:8000705. doi:10.1109/LGRS.2025.3534786.

33. Luo C, Tang H, Li S, Wan G, Chen W, Guan J. YOLOv11s-CD: an improved YOLOv11s method for catenary dropper fault detection. IEEE Trans Instrum Meas. 2025;74:5043410. doi:10.1109/TIM.2025.3604118.
34. Yang Z, Xu W, Chen N, Chen Y, Wu K, Xie M, et al. SDA-YOLO: multi-scale dynamic branching and attention fusion for self-explosion defect detection in insulators. Electronics. 2025;14(15):3070. doi:10.3390/electronics14153070.
35. Qi S, Sun Y, Song X, Li J, Shang T, Yu L. SD-YOLO: a robust and efficient object detector for aerial image detection. IEEE J Sel Top Appl Earth Obs Remote Sens. 2025;18:20563–74. doi:10.1109/JSTARS.2025.3591493.
36. Yue M, Zhang L, Zhang Y, Zhang H. An improved YOLOv8 detector for multi-scale target detection in remote sensing images. IEEE Access. 2024;12:114123–36. doi:10.1109/ACCESS.2024.3444606.