THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 26, Number 2/2025, pp. 211–220
DOI: 10.59277/PRA-SER.A.26.2.12

UAV IMAGE SMALL TARGET DETECTION NETWORK VIA ATTENTION MECHANISM AND MULTI-SCALE FEATURE FUSION

Yanrong LI[1], Hua HUO[1,2], Liping WANG[1], Ge SAI[1], Liqun ZHAO[1]

[1] Henan University of Science and Technology, College of Information Engineering, Luoyang, Henan, China.
[2] Artificial Intelligence Laboratory, Henan Engineering Technology Research Center for Medical Big Data and Computational Intelligence, Luoyang, Henan, China.
Corresponding author: Hua HUO, E-mail: pacific_huo@126.com

Abstract. Aiming at the challenges of small-target feature extraction and low detection accuracy in object detection caused by dense small targets and large target scale variations in UAV aerial images, this paper proposes an Aerial Small-Target Detection Network with Attention Mechanism and Multi-Scale Feature Fusion (AMSFN). First, a Multi-Scale Dilated Fusion Module (MDFM) is designed to enhance the network's feature extraction capability through the combination of multi-scale feature fusion and attention mechanisms. Second, a Global Grouped Coordinate Attention Module (GGCA) is developed to capture multi-dimensional global information and strengthen the feature representation of small targets. Then, the Normalized Wasserstein Distance (NWD) loss function is combined with the Complete Intersection over Union (CIOU) loss function as the localization regression loss to accelerate network convergence. Finally, an additional target detection layer is introduced, and K-means++ is used to cluster anchor boxes. Experimental results on the VisDrone2019 dataset show that compared with YOLOv7, AMSFN improves mAP0.5 and mAP0.5:0.95 by 5.2% and 4.3%, respectively, significantly enhancing the detection accuracy of small targets.

Keywords*:* small object detection, image processing, loss function, attention mechanism.

1. INTRODUCTION

Target detection is a fundamental task in computer vision that aims to automatically identify and localize objects of interest in images or videos, typically by drawing bounding boxes and assigning category labels. It plays a vital role in numerous real-world applications. For instance, in security surveillance, it enables real-time identification of suspicious individuals or behaviors; in autonomous driving, it helps detect pedestrians, vehicles, and traffic signals to support safe decision-making; in industrial inspection, it detects defects or missing parts in products to improve quality control; and in precision agriculture, UAV based detection systems identify pest-affected areas for targeted pesticide application. Despite challenges such as complex backgrounds, illumination variations, and target occlusion, recent advances in detection algorithms have significantly improved their accuracy and robustness.

Stable flight of Unmanned Aerial Vehicles (UAVs) is not only the basis for safe UAV operations, but also a key prerequisite for ensuring high-quality image acquisition [1]. These UAV images are widely used in the fields of wildlife protection [2], environmental detection [3], geological exploration [4], and precision agriculture [5], which provide important data support for related work. However, small target detection in UAV images faces significant challenges: limited by imaging distance and sensor resolution, targets often appear at pixel scale, and suffer from complex backgrounds, dense targets, and drastic scale variations (spanning up to 10 times or more). Generic detectors perform well on natural scene datasets such as PASCAL VOC [6] and MS COCO [7], but poorly on UAV datasets such as VisDrone-2019 [8]. Traditional detection models such as YOLOv7 perform well in natural scenes, but in UAV scenarios due to insufficient feature extraction (fixed receptive fields are difficult to cover multi-scale targets), serious background noise interference (shallow feature redundancy leads to small target features being submerged), high sensitivity to localization loss (small targets are sensitive to coordinate bias), and a priori mismatch of anchor frames (preset

anchor frames are very different from the distribution of the UAV target), leading to significant degradation of detection accuracy. resulting in a significant degradation of detection accuracy. Aiming at the above problems, this paper proposes AMSFN, a UAV small target detection network based on the attention mechanism and multi-scale feature fusion, to improve the model's ability to sense and locate small targets through multilevel improvement. The main contributions of this paper are summarized:

1. The multi-scale dilated fusion module MDFM, by combining the dilated convolution module and the attention mechanism to expand the sensory field to capture a wider range of contexts, can extract and utilize feature information more effectively.

2. global grouped coordinate attention module GGCA, which can capture multi-dimensional global information, help the model to highlight small object feature information and enhance small target feature expression.

3. introducing the NWD loss function and combining it with CloU as a localization regression loss function to enhance the accuracy and robustness of small target detection while maintaining a high level of detection performance for medium and large targets. In addition, a small target detection layer is added to capture detailed information about small targets, and the anchor frames are clustered using the Kmeans++ clustering algorithm, which is more suitable for aerial photography scenarios from a UAV perspective.

Section 2 provides a comprehensive overview of related work in object detection. Section 3 presents a detailed description of the improved modules. Section 4 reports the experimental results, and Section 5 concludes the paper and discusses directions for future research.

2. RELATED WORKS

In the field of small object detection in UAV images, numerous studies have conducted in-depth explorations at the level of algorithm improvement. Ju et al. [9] integrated the Adaptive Feature Fusion with Attention Mechanism (AFFAM) into YOLOv3 to enhance feature perception by learning channel and spatial information. Li et al. [10] developed Cotton-YOLO based on YOLOv7, optimizing the backbone and head networks to improve the detection performance of foreign fibers. Huang et al. [11] proposed DC-SPP-YOLO, which improves the connection of the backbone network and the feature pooling method to enhance target feature learning. Zhang et al. [12] proposed MSFC-Net for the challenges of optical remote sensing images, innovating feature fusion and regression methods. Zhou [13] proposed YOLO-NL for real-time detection. Su et al. [14] proposed MPE-YOLO based on YOLOv8, which enhances small object detection capabilities through modules such as multi-level feature integration. These studies have significantly advanced the development of detection algorithms.

In addition, research in related fields has provided new insights for small object detection in UAV images. In the optimization of deep learning models, the semi-supervised LSTM with historical feature fusion attention model (HFFA-SSLSTM) proposed by Tang et al. [15] mines time-series features based on the characteristics of industrial process data, providing a reference for capturing the temporal variations of small objects in UAV images. Ning [16] discussed neural network pattern recognition methods under the edge computing framework, whose model compression and other strategies help balance the performance of detection algorithms and resource consumption on resource-constrained UAV platforms. Meni et al. [17] developed entropy-based loss terms to optimize neural networks from the perspective of entropy, providing new directions for optimizing loss functions and accelerating convergence in detection algorithms.

3. APPROACH

As a representative of single-stage object detection algorithms, YOLOv7 has achieved excellent performance in real-time detection, relying on the hierarchical feature aggregation strategy of ELAN (Efficient Layer Aggregation Network), the multi-scale feature fusion capability of the SPPCSPC module, and the dynamic label assignment strategy. However, in drone scenarios, its fixed receptive field struggles to cover small targets of varying scales, the insufficient fusion of shallow features and deep semantics leads to detail loss, and the anchor design mismatches the distribution of small targets in drone images, limiting the detection accuracy for small targets. To address this, the improved network proposed in this paper builds upon the

YOLOv7 architecture by introducing a Multi-Scale Dilated Fusion Module (MDFM) to enhance feature extraction, designing a Global Grouped Coordinate Attention (GGCA) mechanism to suppress feature redundancy, optimizing the loss function to improve localization sensitivity, and adding a small-target detection layer to adapt to the characteristics of drone images, forming a detection framework more suitable for drone scenarios, as shown in Fig. 1.

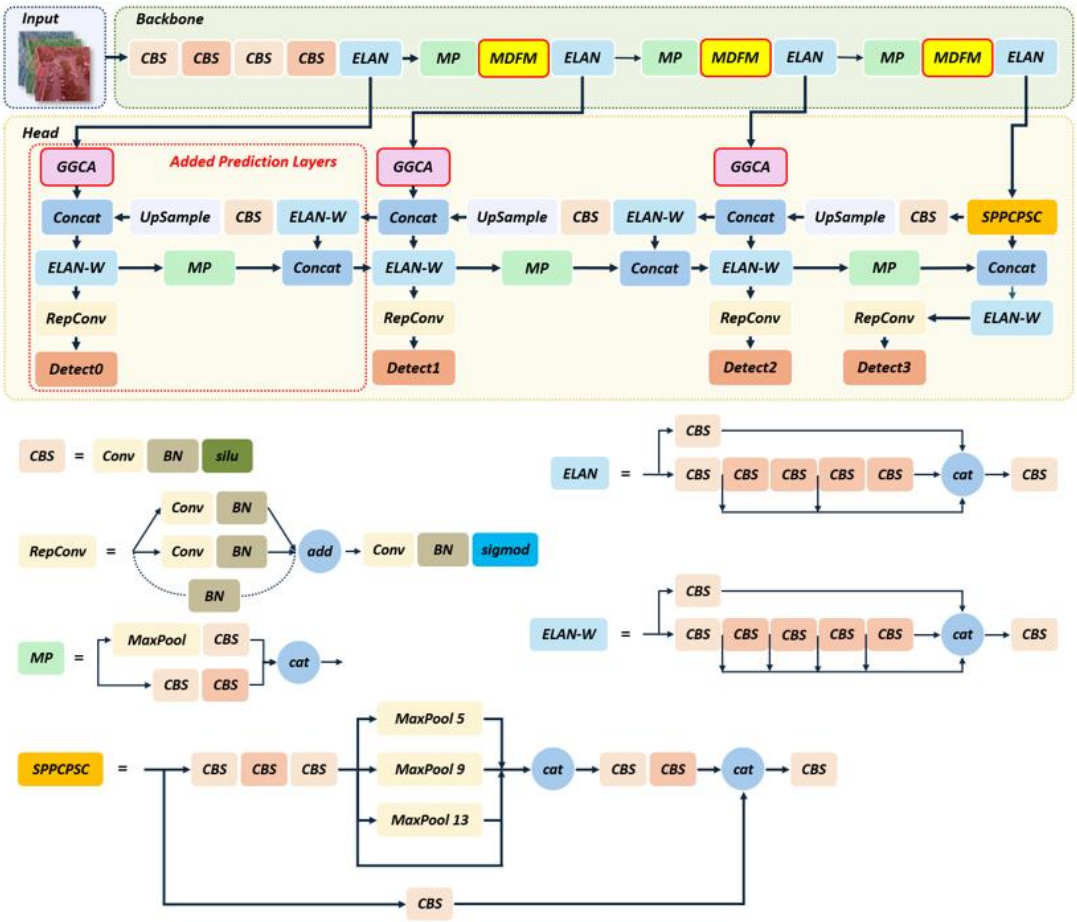


Fig. 1 – AMSFN structure.

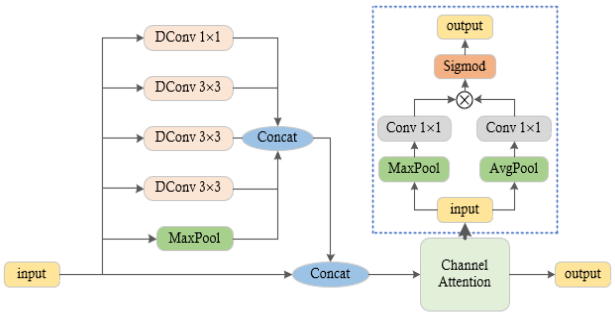


Fig. 2 – Multi-Scale Dilated Fusion Module.

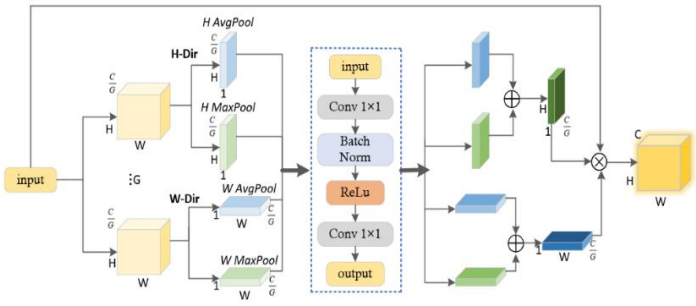


Fig. 3 –Global Grouped Coordinate Attention Module.

3.1. Multi-scale dilated fusion module

In the task of small-target detection from drone perspectives, optimizing feature extraction is crucial for improving detection performance. Due to the characteristics of small target scale and complex backgrounds in drone images, traditional single-scale feature extraction struggles to effectively capture target details and contextual relationships. To address this, our module employs multi-scale dilated convolutions to expand the

receptive field, covering target regions of different sizes without increasing computational complexity, thereby enabling fine-grained capture of multi-scale features for distant and close-range small targets. Meanwhile, a channel attention mechanism is introduced to adaptively learn the importance weights of each feature channel, suppressing background noise interference and enhancing key target features (such as textures and contours that are easily blurred from drone perspectives). Additionally, the module design incorporates cross-layer connections between pooling operations and original features to further preserve the complementarity between shallow detail information and deep semantic information, alleviating the "feature poverty" problem caused by low feature resolution in small targets.

Based on the above design, this paper proposes the Multi-Scale Dilated Fusion Module (MDFM) shown in Fig. 2. Through the collaborative mechanism of multi-branch structures and attention-guided feature interaction, it effectively enhances the discriminability and robustness of small-target features in drone images, providing richer high-resolution feature representations for subsequent detection heads and significantly improving the detection accuracy of small targets in complex backgrounds.

MDFM utilizes multi-scale dilated convolutions with varying dilation rates to extract spatial features. Unlike traditional methods, dilated convolutions expand the receptive field without sacrificing resolution. The first branch uses a 1×1 convolution for feature extraction at the same scale, followed by 3×3 convolutions with dilation rates of 6, 12, and 18 to broaden the receptive field. Global average pooling captures context, and channel attention is applied before fusing the feature maps, improving representation, as described in Equations (1) and (2)

$$C_1 = (\mathrm{DConv}_{all} \oplus \mathrm{Max}(\boldsymbol{X})) \oplus \boldsymbol{X} \tag{1}$$

where $\oplus$ denotes element-wise addition. $\boldsymbol{X}$ represents the input feature map, DConv_{all} denote the output of all dilated convolution operations performed on $\boldsymbol{X}$, and Max signify the result of applying the max-pooling operation to $\boldsymbol{X}$. Firstly, the max-pooling operation is applied to $\boldsymbol{X}$ to generate $\mathrm{Max}(\boldsymbol{X})$, which extracts the main features of $\boldsymbol{X}$ and reduces its dimensionality. Then, DConv_{all} is concatenated with $\mathrm{Max}(\boldsymbol{X})$ to fuse the dilated convolution features and the pooled features. Finally, the fused result is concatenated with the original input $\boldsymbol{X}$, enabling $\boldsymbol{C_1}$ to integrate both the processed features and the original information, thereby providing a more comprehensive feature representation for subsequent object detection tasks

$$C_2 = \sigma(\mathrm{Conv}(\mathrm{Max}(C_1)) \otimes \mathrm{Conv}(\mathrm{Avg}(C_1))) \otimes C_1 \tag{2}$$

where $\otimes$ denotes element-wise multiplication, σ is the Sigmoid function, "Conv" is a 1×1 convolution for feature extraction, Max and Avg are the maximum and average pooling operations. The maximum pooling selects the maximum value from a local area, while the average pooling calculates the mean value of the elements within a region. And $\boldsymbol{C_1}$ (an intermediate variable from prior multi-scale operations) is processed through these steps to generate $\boldsymbol{C_2}$, enhancing feature expressiveness for downstream tasks

3.2. Global grouped coordinate attention module

In small target detection scenarios from the UAV perspective, due to the tiny target scale and the complex and variable background, the model is highly susceptible to small target features being overwhelmed by background noise due to the lack of effective capture of global information. The Global Grouped Coordinate Attention (GGCA) module designed for this purpose constructs the attention graph along the spatial dimension (height and width) by deeply fusing the shared convolutional layer (shown as the blue dotted line in Fig. 3) with the coordinate attention mechanism. The mechanism can adaptively assign weights to the input feature maps to accurately enhance key features of small targets while suppressing the interference of irrelevant background information.

In addition, considering the computational bottleneck caused by the large amount of UAV image data, GGCA groups the input feature maps by channel dimension. This grouping strategy maintains the richness of feature expression while significantly reducing the computation amount of a single group, achieving a balance between computational efficiency and feature extraction capability. In this way, GGCA not only effectively strengthens the feature representation of small targets, but also captures multi-dimensional global information containing the spatial location relationship of targets, which significantly improves the accuracy and robustness of small target detection in UAV scenarios. The specific structure of GGCA is shown in Fig. 3.

Firstly, the input feature maps are partitioned into G groups, with each group containing C/G channels, based on the total number of channels

$$\boldsymbol{X} \in \mathbb{R}^{C\times H\times W} \tag{3}$$

where C denotes channel numbers; W represents the width of the feature map and H represents the height of the feature map, × denotes a weighting operation on the feature map. Based on this dimensional information, the grouped feature map can be represented as

$$\boldsymbol{X} \in \mathbb{R}^{G\times\frac{C}{G}\times H\times W} \tag{4}$$

Subsequently, global maximum pooling and global average pooling operations are performed on each group of feature maps along both the width and height dimensions

$$\boldsymbol{X}_{h,avg} = \mathrm{Avg}(\boldsymbol{X}) \in \mathbb{R}^{G\times\frac{C}{G}\times H\times 1},\ \ \boldsymbol{X}_{h,max} = \mathrm{Max}(\boldsymbol{X}) \in \mathbb{R}^{G\times\frac{C}{G}\times H\times 1} \tag{5}$$

$$\boldsymbol{X}_{w,avg} = \mathrm{Avg}(\boldsymbol{X}) \in \mathbb{R}^{G\times\frac{C}{G}\times 1\times W},\ \ \boldsymbol{X}_{w,max} = \mathrm{Max}(\boldsymbol{X}) \in \mathbb{R}^{G\times\frac{C}{G}\times 1\times W} \tag{6}$$

The feature description of each grouped feature map is achieved by means of a shared convolutional layer consisting of two 1×1 convolutional layers, a Relu activation function, and a batch normalization layer

$$\boldsymbol{Y}_{h,avg} = \mathrm{Conv}\big(\boldsymbol{X}_{h,avg}\big),\ \ \boldsymbol{Y}_{h,max} = \mathrm{Conv}\big(\boldsymbol{X}_{h,max}\big) \tag{7}$$

$$\boldsymbol{Y}_{w,avg} = \mathrm{Conv}\big(\boldsymbol{X}_{w,avg}\big),\ \ \boldsymbol{Y}_{w,max} = \mathrm{Conv}\big(\boldsymbol{X}_{w,max}\big) \tag{8}$$

The outputs from the convolutional layers are summed to produce height and width dimensions using a Sigmoid activation function

$$\boldsymbol{Z}_h = \sigma\big(\boldsymbol{Y}_{h,avg} + \boldsymbol{Y}_{h,max}\big) \in \mathbb{R}^{G\times\frac{C}{G}\times H\times 1},\ \ \boldsymbol{Z}_w = \sigma\big(\boldsymbol{Y}_{w,avg} + \boldsymbol{Y}_{w,max}\big) \in \mathbb{R}^{G\times\frac{C}{G}\times 1\times W} \tag{9}$$

The output feature map is obtained from the input feature map weighted by the attention weights

$$\boldsymbol{Z} = \boldsymbol{X} \times \boldsymbol{Z}_h \times \boldsymbol{Z}_w \in \mathbb{R}^{C\times H\times W} \tag{10}$$

The attention weights $\boldsymbol{Z}_w$ and $\boldsymbol{Z}_h$ are expanded in the width and height directions, respectively

3.3. Improvement of the loss function

YOLOv7 utilizes CIoU loss for bounding box regression, which is sensitive to minor positional deviations, particularly affecting small objects more than larger ones. To address this limitation, the NWD loss function, based on Wasserstein distance, is introduced. NWD models bounding boxes as 2D Gaussian distributions and measures similarity using normalized Wasserstein distance, reducing sensitivity to size and positional variations. This study combines NWD with CIoU to improve bounding box regression and enhance detection accuracy, with ablation experiments conducted to optimize the scale factors for CIoU and NWD coefficients

$$\mathrm{CIoU} = 1 - \left(IoU - \frac{\rho^2\big(b, b_{gt}\big)}{c^2} - \alpha v\right) \tag{11}$$

$$\mathrm{NWD}\big(N_a, N_b\big) = \exp\Big(-\frac{\sqrt{W_2^2\big(N_a, N_b\big)}}{C}\Big) \tag{12}$$

$$W_2^2\big(N_a, N_b\big) = \left\|\left(\begin{bmatrix} cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2}\end{bmatrix}^T, \begin{bmatrix} cx_b, cy_b, \frac{w_b}{2}, \frac{h_b}{2}\end{bmatrix}^T\right)\right\|_2^2 \tag{13}$$

$$\mathrm{Loss} = 1 - (a\cdot \mathrm{CIoU} + b\cdot \mathrm{NWD}) \tag{14}$$

where IoU measures the overlap between the predicted and real bounding boxes, $\frac{\rho^2(b,b_{gt})}{c^2}$ denotes the normalised centroid distance, α is a parameter, v is a measure the consistency of the aspect ratio, c represents the diagonal distance of the smallest enclosed area containing both the ground truth box and the prediction box, C is the amount of dataset categories, $W_2^2(N_a,N_b)$ represents the distance metric, N_a and N_b represent the Gaussian distributions modeled by $A = (cx_a, cy_a, w_a, h_a)$ and $B = (cx_b, cy_b, w_b, h_b)$, "·" represents scalar multiplication, "+" represents the addition operation

3.4 Optimization of the small target detection layer

YOLOv7 employs three detection heads of varying sizes (80×80×128, 40×40×256, 20×20×512) corresponding to different scale feature maps, where shallow layers with high spatial resolution are better suited for capturing small target details. To address the challenges of dense small target detection in UAV aerial imagery, this study introduces an additional small target detection layer with an output size of 160×160×64 (as shown by the red dashed box in Fig. 1).

To address the mismatch between YOLOv7's default anchor boxes and the target size distribution in UAV datasets, this paper employs the K-means++ algorithm for anchor optimization. This method clusters the width and height features of ground truth bounding boxes to identify a representative set of cluster centers, thereby minimizing the distance between each box and its assigned center. The clustering process can be regarded as an optimization problem, where the objective is to minimize the overall distance, such as IoU or Euclidean distance, under a fixed number of clusters through iterative partitioning. Compared to traditional K-means, K-means++ improves clustering stability and accuracy by carefully selecting initial cluster centers that are relatively far apart, providing prior information that better reflects the data distribution and enhancing the performance of the object detector.

4. TEST

4.1. Dataset & Test environment

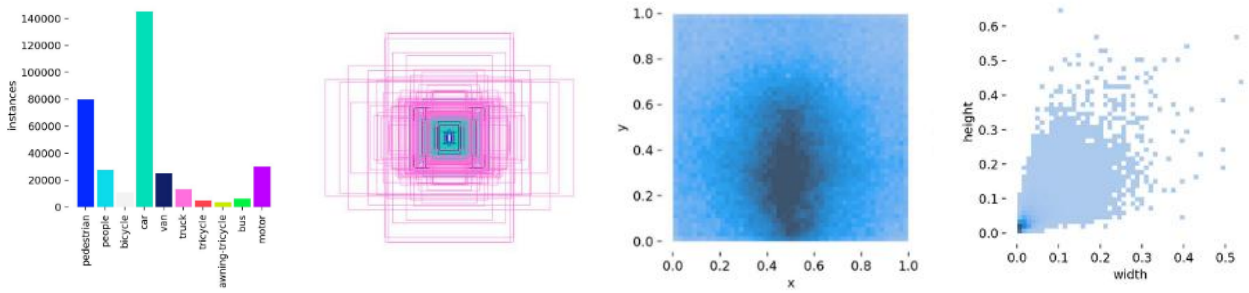
The VisDrone2019 dataset, collected by the AISKYEYE team at Tianjin University, comprises 10,209 UAV-captured images (6,471 for training, 548 for validation, 3,190 for testing) with 2.6 million annotations across 10 categories. It is designed for object detection and tracking in diverse aerial environments. The AI-TOD dataset [18], focused on small and medium-sized target detection, includes 28,036 images (11,214 for training, 2,804 for validation, 14,018 for testing) with 700,621 annotations across 8 categories. AI-TOD features smaller targets, averaging 12.8 pixels, making it particularly suitable for evaluating small object detection in aerial imagery. Both datasets are valuable for advancing research in UAV-based object detection, with VisDrone2019 offering broader environmental diversity and AI-TOD emphasizing small target challenges.

(a) (b)

Fig. 4 – Images from the VisDrone2019 dataset.

(a) (b)

Fig. 5 – Images from the AI-TOD dataset.



(a) (b) (c) (d)

Fig. 6 – Distribution status of labels and various indicators in the VisDrone2019 dataset: a) the number of each category in the training set; b) the dimensions and quantity of the boxes, demonstrating the distribution of the sizes and the corresponding amounts of the bounding boxes within the training set; c) position of the center point in relation to the whole image; d) the aspect ratio of the target relative to the whole image.

4.2. Ablation evaluation

To verify the practical effects of each improved module, ablation evaluation were conducted on the network model. Table 1 presents the experimental results of each module under different combinations. The first row contains the data of YOLOv7 without any added improved modules, serving as the baseline.

The experimental results indicate that when the multi-scale dilated fusion module (MDFM) is added alone, the mAP@50% improves to 0.424 and other metrics also show enhancements. After adding the global grouped coordinate attention module (GGCA), the detection accuracy experiences a slight change. When the improved CIoU loss combined with the NWD loss is incorporated, the detection accuracy drops to some degree. Upon adding the small object detection layer, the detection accuracy rebounds. As more modules are combined, the metrics exhibit different changes. When all modules are included, the mAP@50% reaches 0.461 and the detection accuracy is significantly improved.

Table 1

Ablation evaluation of each module

MDFM	GGCA	NWD	Added Layer	Kmeans++	mAp@50%	mAp@50:95%	P	R	Params/M
					0.409	0.240	0.537	0.435	37.2
√					0.424	0.251	0.555	0.45	37.91
	√				0.422	0.247	0.552	0.447	37.33
		√			0.412	0.245	0.549	0.445	37.21
			√		0.417	0.244	0.547	0.442	37.59
				√	0.413	0.241	0.544	0.438	37.21
√			√	√	0.439	0.267	0.570	0.461	38.39
	√		√	√	0.435	0.262	0.566	0.457	37.92
		√	√	√	0.433	0.261	0.563	0.454	37.59
√	√		√	√	0.456	0.280	0.586	0.463	38.61
√		√	√	√	0.452	0.277	0.582	0.463	38.39
	√	√	√	√	0.450	0.275	0.579	0.461	37.92
√	√	√	√	√	0.461	0.283	0.589	0.466	38.61

√ indicates the selected module

4.3. Find a suitable scaling factor in the loss function

We tested NWD (Normalized Wasserstein Distance) as the regression loss in YOLOv7 to improve small target detection. It worked well for small targets but reduced accuracy for medium and large ones. So, we kept the CIoU loss function. Combining NWD and CIoU with a 0.6/0.4 split improved overall performance, boosting small target detection while maintaining good results for larger ones. Table 2 shows the results.

4.4. Comparison evaluation

To evaluate the performance of the proposed model quantitatively, comparative evaluation were conducted using classical object detection models and popular YOLO series models. The VisDrone2019 dataset was employed under identical experimental conditions and parameter settings, with the results presented in Table 3. Compared to the YOLOv7 baseline model, the proposed model achieved significant

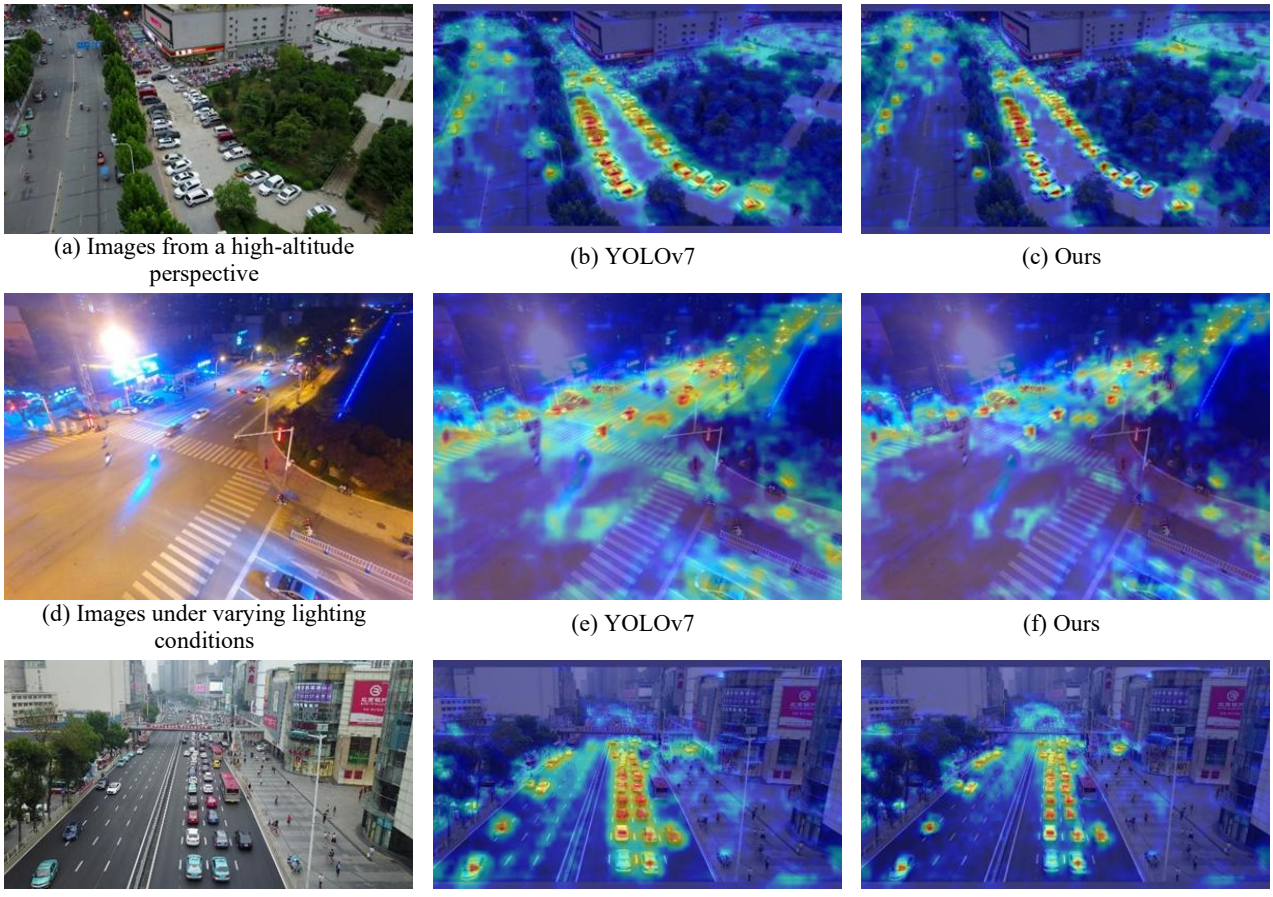
improvements: mAP@0.5 increased from 0.409 to 0.461, mAP@0.5:0.95 from 0.240 to 0.283, precision from 0.537 to 0.589, and recall from 0.435 to 0.466, accompanied by corresponding changes in parameter count and computational complexity. When compared to the two-stage algorithm Faster R-CNN, the mAP@0.5 improved from 0.383 to 0.461, with adjustments in parameter count and FLOPs. Against one-stage algorithms such as RetinaNet and SSD, as well as YOLO variants including YOLOX, YOLOv7-tiny, the proposed model demonstrated varying degrees of improvement in mAP@0.5.

Table 2

Results of combining CIoU and NWD with different ratio parameters on the model

CIoU	NWD	mAP@0.5	mAp@0.5:0.95	Precise	Recall
0	1	0.429	0.259	0.560	0.451
1	0	0.427	0.258	0.558	0.449
0.5	0.5	0.432	0.261	0.561	0.452
0.6	0.4	0.433	0.261	0.563	0.454
0.4	0.6	0.431	0.260	0.559	0.453
0.7	0.3	0.429	0.259	0.560	0.451
0.3	0.7	0.428	0.258	0.558	0.450
0.8	0.2	0.425	0.254	0.557	0.453
0.2	0.8	0.428	0.259	0.559	0.451

We used Grad-CAM to create heatmaps comparing YOLOv7 and AMSFN, shown in Fig. 7. The heatmaps highlight important regions, with darker colors indicating higher importance. AMSFN focuses more accurately on small object areas, reducing background noise. This improvement is due to AMSFN's attention mechanism, which enhances feature representation and improves detection accuracy.



(a) Images from a high-altitude perspective (b) YOLOv7 (c) Ours

(d) Images under varying lighting conditions (e) YOLOv7 (f) Ours

(g) Images with occluded objects (h) YOLOv7 (i) Ours

Fig. 7 – Comparison of AMSFN and YOLOv7 detection on VisDrone2019 dataset.

Table 3

Comparative evaluation results on the Visdrone2019 dataset

Model	mAP@0.5	mAp@0.5:0.95	P	R	Params(M)	GFLOPS(G)
YOLOv7	0.409	0.240	0.537	0.435	37.2	105.1
Faster R-CNN	0.383	0.239	0.516	0.388	137.1	137.1
RetinaNet	0.262	0.112	0.348	0.316	36.3	145.7
SSD	0.369	0.211	0.478	0.375	24.5	87.9
YOLOX [19]	0.385	0.197	0.511	0.466	6.3	12.9
YOLOv7-tiny	0.356	0.229	0.509	0.445	6.0	13.3
YOLOv5s	0.363	0.195	0.466	0.393	7.3	16.5
YOLOv8s	0.322	0.188	0.411	0.329	11.3	8.9
Gold-YOLO [20]	0.404	0.242	0.589	0.503	75.1	151.5
ours	0.461	0.283	0.589	0.466	38.6	127.2

4.5. Test on model generalization

To verify the versatility of the AMSFN network, model performance tests were conducted on the AI - TOD dataset. The detection results of AMSFN on this dataset are shown in Table 4. From the experimental data, AMSFN achieved an mAP@0.5 of 0.398, an mAP@0.5:0.95 of 0.269, a precision (P) of 0.473, a recall (R) of 0.412, a parameter quantity (Params) of 38.6M, and GFLOPS of 127.2G. Compared with other models, such as YOLOv7 with an mAP@0.5 of 0.363 and an mAP@0.5:0.95 of 0.240, and Faster R-CNN with an mAP@0.5 of 0.342 and an mAP@0.5:0.95 of 0.201, AMSFN performed better in key metrics such as mAP@0.5 and mAP@0.5:0.95.

This indicates that AMSFN outperforms other detection algorithms, mainly due to its lightweight feature extraction and attention mechanisms, which enhance the feature representation of small targets. The experimental results demonstrate the strong generalization ability of AMSFN across different datasets and scenarios, and cross - dataset evaluations further validate this, emphasizing the versatility of its design.

Table 4

Comparative evaluation results on the AI-TOD dataset

Model	mAP@0.5	mAp@0.5:0.95	P	R	Params(M)	GFLOPS(G)
YOLOv7	0.363	0.240	0.457	0.405	37.2	105.1
Faster R-CNN	0.342	0.201	0.404	0.373	137.1	137.1
RetinaNet	0.276	0.101	0.328	0.296	36.3	145.7
SSD	0.318	0.162	0.408	0.363	24.5	87.9
YOLOX	0.335	0.183	0.461	0.346	6.3	12.9
YOLOv7-tiny	0.371	0.213	0.488	0.415	6.0	13.3
YOLOv5s	0.331	0.170	0.423	0.353	7.3	16.5
YOLOv8s	0.309	0.167	0.385	0.287	11.3	8.9
Gold-YOLO	0.374	0.202	0.544	0.473	75.1	151.5
ours	0.398	0.269	0.473	0.412	38.6	127.2

5. CONCLUSION

Given the widespread application of unmanned aerial vehicles (UAVs), object detection in UAV images has become a key research area. This paper proposes an improved algorithm to address the challenges of difficult feature extraction and low detection accuracy for small targets in UAV aerial images. The Multi-Scale Dilated Fusion Module (MDFM) effectively integrates dilated convolution and attention mechanisms, expanding the receptive field to more proficiently capture and utilize small-target information. The Global Grouped Coordinate Attention Module (GGCA) excels at acquiring multi-dimensional global information, thereby highlighting small-target features. By introducing the Normalized Wasserstein Distance (NWD) loss function and combining it with CIoU as the localization regression loss, the algorithm enhances the accuracy and robustness of small-target detection while maintaining high-performance detection for medium and large targets. Additionally, a dedicated small-target detection layer is added, and anchor boxes are clustered using the Kmeans++ algorithm to make the model better adapted to UAV photography scenarios. Tests on the

VisDrone2019 dataset clearly demonstrate that the proposed algorithm significantly outperforms the baseline model YOLOv7. Key metrics such as mAP@0.5 and mAP@0.5:0.95 have been substantially improved, strongly verifying the algorithm's excellent effectiveness in enhancing small-target detection in UAV images.

Nevertheless, the algorithm still has limitations. In scenarios where multiple target categories with similar features coexist, misclassification may occur. Furthermore, missed detections remain an issue in images filled with extremely small targets. Future research will focus on further optimizing the network architecture, reducing the number of parameters, and improving both detection accuracy and speed to achieve more efficient object detection in UAV aerial images.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant No. 61672210, the Major Science and Technology Program of Henan Province under Grant No. 221100210500, and the Central Government Guiding Local Science and Technology Development Fund Program of Henan Province under Grant No. Z20221343032.

REFERENCES

[1] Chen YL, Li Y, Wen ZD, Bie XB. Robust neuroadaptive trajectory tracking for quadrotor UAV based on non-singular terminal sliding mode control. Proc Rom Acad Ser A. 2025; 26(1/2): 75–87. DOI: 10.59277/PRA-SER.A.26.1.10.

[2] Corcoran E, Winsen M, Sudholz A, et al. Automated detection of wildlife using drones: Synthesis, opportunities and constraints. Methods Ecol Evol. 2021; 12(6): 1103–1114.

[3] Capolupo A, Pindozzi S, Okello C, et al. Photogrammetry for environmental monitoring: The use of drones and hydrological models for detection of soil contaminated by copper. Sci Total Environ. 2015; 514: 298–306.

[4] Fugazza D, Scaioni M, Corti M, et al. Combination of UAV and terrestrial photogrammetry to assess rapid glacier evolution and map glacier hazards. Nat Hazards Earth Syst Sci. 2018; 18(4): 1055–1071.

[5] Zhang C, Kovacs JM. The application of small unmanned aerial systems for precision agriculture: A review. Precis Agric 2012; 13:693–712.

[6] Everingham M, Van Gool L, Williams CKI, et al. The PASCAL visual object classes (VOC) challenge. Int J Comput Vis. 2010; 88: 303–338.

[7] Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference. Zurich, Switzerland; September 6–12, 2014. Proceedings, Part V 13. Cham, Switzerland: Springer; 2014, pp. 740–755.

[8] Du D, Zhu P, Wen L, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019.

[9] Ju M, Luo J, Wang Z, Luo H. Adaptive feature fusion with attention mechanism for multi-scale target detection. Neural Comput Appl. 2021; 33(8): 2769–2781. DOI: 10.1007/s00521-020-05461-3.

[10] Li Q, Ma W, Li H, Zhang X, Zhang R, Zhou W. Cotton-YOLO: Improved YOLOv7 for rapid detection of foreign fibers in seed cotton. Comput Electron Agric. 2024; 219: 108752. DOI: 10.1016/j.compag.2024.108752.

[11] Huang Z, Wang J, Fu X, et al. DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. Inf Sci. 2020; 522: 241–258.

[12] Zhang T, Zhuang Y, Wang G, Dong S, Chen H, Li L. Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery. IEEE Trans Geosci Remote Sens. 2021; 60: 1–20. DOI: 10.1109/TGRS.2021.3124377.

[13] Zhou Y. A YOLO-NL object detector for real-time detection. Expert Syst Appl. 2024; 238: 122256. DOI: 10.1016/j.eswa.2023.122256.

[14] Su J, Qin Y, Jia Z, et al. MPE-YOLO: Enhanced small target detection in aerial imaging. Sci Rep. 2024; 14(1): 17799.

[15] Tang Y, Wang Y, Liu C, et al. Semi-supervised LSTM with historical feature fusion attention for temporal sequence dynamic modeling in industrial processes. Eng Appl Artif Intell. 2023; 117: 105547.

[16] Ning J. Neural network-based pattern recognition in the framework of edge computing. Rom J Inf Sci Technol. 2024; 27(1): 106–119.

[17] Meni MJ, White RT, Mayo ML, et al. Entropy-based guidance of deep neural networks for accelerated convergence and improved performance. Inf Sci. 2024; 680: 121239.

[18] Wang J, Yang W, Guo H, et al. Tiny object detection in aerial images. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE; 2021, pp. 3791–3798.

[19] Ge Z, Liu S, Wang F, et al. YOLOX: Exceeding YOLO series in 2021. arXiv preprint. arXiv:2107.08430, 2021.

[20] Wang C, He W, Nie Y, et al. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. Adv Neural Inf Process. Syst. 2023; 36: 51094–51112.

Received March 17, 2025