MS-YOLO: Multi-Scale Feature Fusion Based Small Object Vehicle Detection Approach for Aerial Images

Bao Liu, Jiangfan Wang, Jinlei Huang

***Abstract*—A multi-scale feature fusion based small object vehicle detection approach (MS-YOLO) is proposed to address the problem of poor feature extraction ability caused by the complex backgrounds and dense objects. Due to the vehicle small object is easy to be interfered by the background, it is difficult to locate the detection model. this paper designs a parallel self-attention module (PAM) to suppress redundant non-singular feature expressions and focus on the most relevant vehicle information. The PAM module is embedded into the feature layers of different scales in the backbone network, optimizing the feature extraction ability of the network by adaptively allocating the weights of channels and space. Moreover, the multi-branch feature pyramid network (MB-FPN) is proposed to integrate the feature information of different resolutions, which effectively solves the problem that the small vehicle object are prone to generate false information in the upsampling process. Finally, the Focaler-CIOU loss function is introduced to address the problem of sample imbalance. Experimental verification on the Aerial dataset confirms that the proposed method achieves the best detection performance compared to classic detection algorithms such as YOLOv5, YOLOv7, YOLOX, YOLOv8, YOLOv11, Fast R-CNN, and SSD.**

***Index Terms*—small object vehicle detection, multi-scale feature fusion, parallel self-attention, MB-FPN.**

I. Introduction

The utilisation of high-elevation aerial detection systems has gained significant traction in both military and civilian contexts, encompassing a wide range of applications such as battlefield reconnaissance, urban traffic management, road safety monitoring, field rescue, and numerous others [1]. Unfortunately, the traditional detection methods such as the HOG [2], SIFT [3], and SURF [4] cannot adapt the practical application requirements since the complex algorithm processes, large parameter quantities, and the performance in real time is unsatisfactory. In recent years, the mainstream approach to object detection has gradually become that based on deep learning [5-7], such as Faster R-CNN [8], FPN [9], YOLO [10-13], SSD [14], and RetinaNet [15], which get the good detection performance for large-scale objects. However, the feature information in the small-scale objects gradually disappears with the increase of downsampling times, and even only a few pixels are left in the final layer, which is not conducive to the subsequent detection.

The scholars have made a significant number of contributions to the improvement of the performance of small object detection. In terms of feature fusion, Tsung Yi Lin [9] proposed the feature pyramid network (FPN) to improve the multi-scale fusion ability. The PANet [16] further integrate deep and shallow information of feature maps. The ASFF [17] and AugFPN [18] are used the FPN module from different perspectives to enhance the feature fusion capability. Song [19] used the Bi-FPN module instead of the PANet module to improve the fusion degree feature. Jiang [20] designed multi-scale feature extraction block (MSFEM) and bidirectional dense feature pyramid network (BDFPN) to achieve efficient multi-scale information fusion. Moreover, Zhang [21] improved the YOLOX algorithm by combining the convolutional block attention module [22] (CBAM) module for small target detection of aerial vehicles. In proposing the BCC-Yolov8n model for infrared small targets, Xiang [23] suggested an enhancement to the neck network, building upon the reference network Yolov8. This model incorporates an attention mechanism, with the objective of addressing the challenges posed by low detection accuracy and missed detections in complex traffic scenarios. In summary, these algorithms have achieved significant improvements in the low-altitude aerial vehicle images object detection. However, there is currently limited research on aerial images captured by high-altitude drones (above 120 meters). The high-altitude aerial vehicle images are highly susceptible to background interference with low pixel rates and difficulty in extracting feature information, posing significant challenges to the aerial vehicle detection.

In response to the above problems, this paper proposes a multi-scale feature fusion based small object vehicle detection approach (MS-YOLO), the primary contributions of which are as follows.

(1) This paper proposes a parallel self-attention module (PAM) to address the problem of background interference by

Manuscript received December 4, 2024; revised April 11, 2025.

This work is supported in part by grant for Beilin District Science and Technology Plan Project (GX2231), the Key Research and Development Program of Shaanxi (2021GY-131), and Yulin Science and Technology Plan Project (CXY-2020-037).

Bao Liu is an associate professor of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (corresponding author, +86-18149067968, e-mail: xiaobei0077@163.com).

Jiangfan Wang is a postgraduate student of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail:893598776 @qq.com).

Jinlei Huang is a postgraduate student of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710054, China (e-mail: 1169852151@qq.com).

focusing on the object information through the parallel spatial and channel attention mechanisms. The PAM module is embeded into different scale feature layers of the backbone network to improve the feature extraction capability of the model.

(2) A multi-branch feature pyramid network (MB-FPN) is presented to address the issue of false information generated during the traditional upsampling process. The proposed MB-FPN module greatly improves the small object detection capability by fully integrating the deep and shallow information.

(3) In response to the problems of imbalanced samples and differences between the foreground and background in the aerial images, the Focaler-CIOU loss function is introduced to distinguish different regression samples through linear interval mapping, which improve the accuracy of bounding box regression.

The remainder of this paper is structured as follows. In Section II, we introduce the YOLOv5 network structure and related work. Section III proposes the MS-YOLO model and the components of each module in detail. In Section IV, the experimental environment and parameter configuration are introduced. Section V conducts the experimental comparison and result analysis to demonstrate the applicability and effectiveness of the method. Finally, Section VI is our conclusion.

II. Related Works

A. System process YOLOv8 network structure

The YOLOv8 [24] proposed by Glenn Jocher is one of the popular one-stage object detection models in the YOLO family, as shown in Fig. 1. It employs a novel SOTA model, facilitating applications such as object detection, image classification, instance segmentation, and object tracking within the domain of computer vision. The standard YOLOv8 network is comprised of three components: backbone, neck, and head. The backbone network has changed the preprocessing of the Cross Stage Part (CSP) structure [25] from 3 convolutions to 2 convolutions. The C3 structure of YOLOv5 has been replaced with a C2f structure with richer
gradient flow by drawing on the design concept of YOLOv7. ELAN module stacking. The neck network adopts path aggregation network (PANet) structure, which is the FPN+PAN structure. The FPN and PAN respectively convey strong semantic information and localization features to enhance network feature fusion capabilities. The prediction layer introduces three decoupling heads of different sizes to predict the positions and categories of large, medium, and small-scale objects.

B. Advantage of CBAM

The CBAM is the lightweight module that simultaneously adds attention mechanism in both channel and spatial dimension, which can effectively enhance information transmission between networks. Specifically, We sequentially multiply M_c and M_S with the given feature map X and perform adaptive feature refinement to obtain the feature map X_{out} with the added attention, as shown in Fig. 2. " $\otimes$ " represents point multiplication operation. Therefore, this paper parallelizes the spatial and channel attention mechanisms after absorbing the ideas of the CBAM and introduces them into the backbone network, which can effectively improve the ability of the backbone network to suppress the background interference.

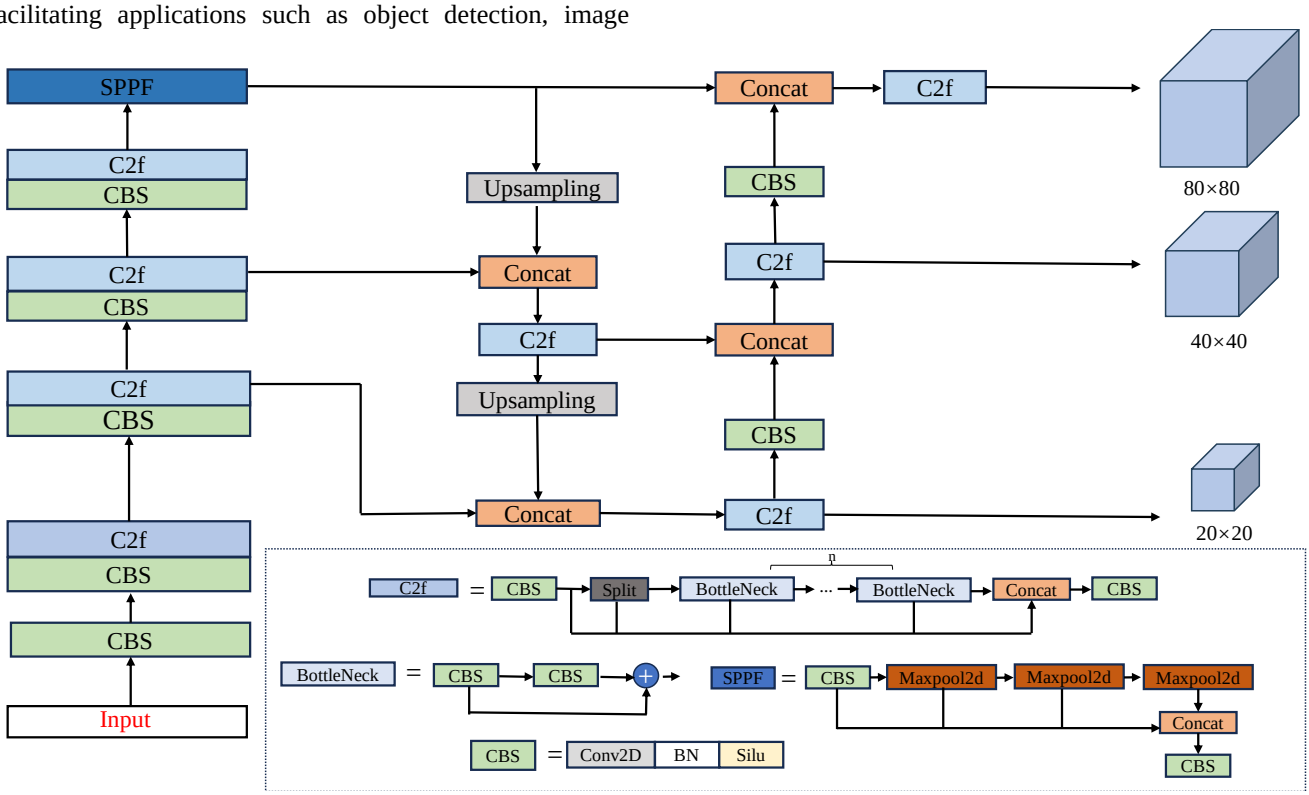


Fig. 1 Network structure of YOLOv8

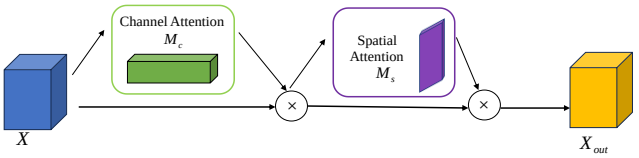


Fig. 2 The CBAM module schematic diagram

III. Proposed Methed

Due to the susceptibility of aerial vehicle objects to background interference and the detection algorithm is difficult to accurately locate the vehicle position， a MS-YOLO network has proposed, as shown in Fig. 3. The PAM attention mechanism module is proposed in the backbone with a view to reducing the adverse effects of background noise on object localization. Moreover, in order to deal with the false information generated by upsampling on the original FPN module, this paper proposes the MB-FPN module to avoid upsampling operations while fusing features of different resolutions. Finally, the Focaler-CIoU loss function is introduced to alleviate the problem of imbalanced aerial vehicle samples.

A. PAM module for feature extraction

The background interference in aerial images of vehicles can introduce a large amount of noise in the feature extraction process, resulting in the detection system being unable to correctly recognize and locate aerial images of vehicle [26]. This paper designs the PAM module based on the CBAM module，aiming to improve the feature extraction capability of the model, as shown in Fig. 4. On the one hand, the PAM inherits the idea of the CBAM and consists of spatial attention mechanism and channel attention mechanism. The spatial attention mechanism is concerned with the capture of dependency relationships at differing positions within the image, whilst the channel attention mechanism is concerned with the capture of correlation between differing channels in the feature map. On the other hand, the PAM models the spatial and channel dimensions separately by parallelizing the channel and spatial attention mechanisms to obtain richer and more accurate feature representations. We embed the PAM module into the feature layers with downsampling multiples of {4, 8, 16, 32} in the YOLOv8 backbone network, effectively reduce the noise interference in the feature extraction process and the highlight important features of the object vehicle.

Firstly, the PAM module adaptively calculates weights to obtain the refined feature M_c and M_s through the H-channel and spatial attention mechanisms. The two features are concatenated in terms of channels, and the output features are obtained by 3×3 convolution. The output feature map X_{out} can be written as

$$X_{out} = M_c + M_s \tag{1}$$

The H-channel attention mechanism obtains the global information from the feature map X by the global average pooling (GAP) module, and proportionally reduce and expand the feature map dimensions through the Hardswitch activation function [27] and 1×1 convolution, respectively. The sigmoid function is used to obtain the weight A_c and multiply them with feature X to obtain the weighted feature M_c .

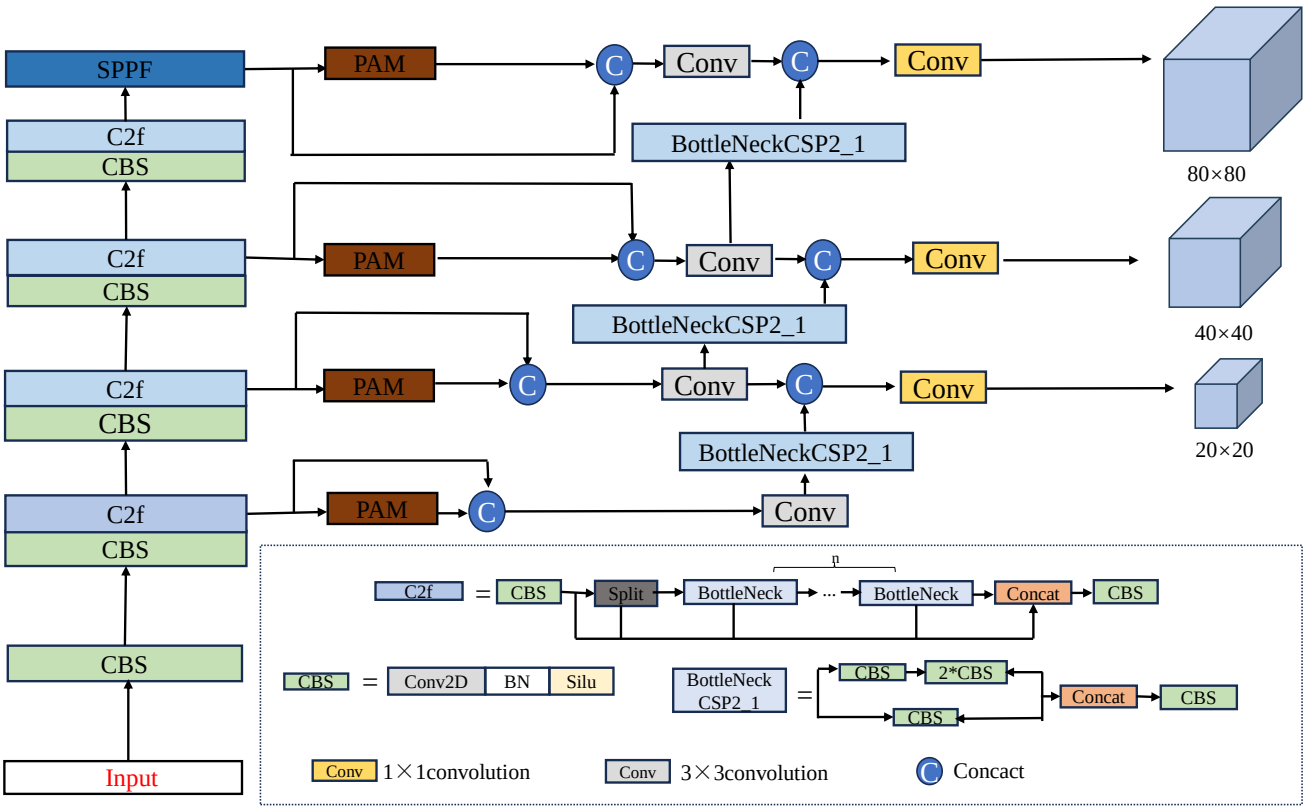


Fig. 3 Network structure of MS-YOLO

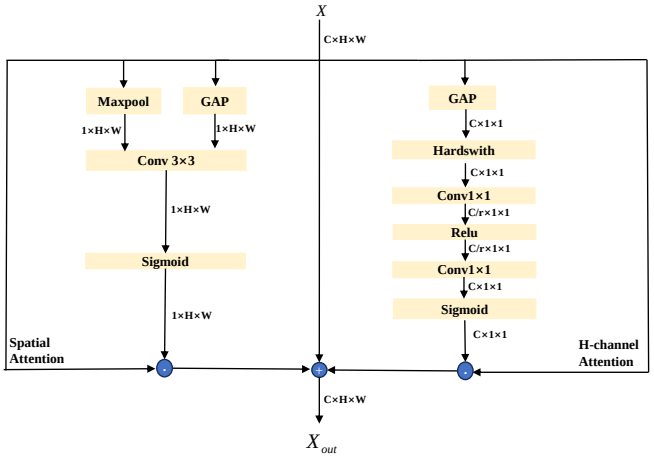


Fig. 4 Structure of PAM module

$$M_c = A_c \odot X \tag{2}$$

$$A_c = \delta(f_{1\times1}(R(f_{1\times1}(Hs(g(X))))) \tag{3}$$

$$g(X) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{[i,j]} \tag{4}$$

where $g(X) \in R_c$ is the GAP layer; $Hs(\bullet)$ is the Hardswitch activation function; $R(\bullet)$ is the Relu activation function; $X_{[i,j]}$ is the i-th row and j-th column of X; H and W represent the length and width of X, respectively; δ is the sigmoid activation function; "$\odot$" is point multiplication operation; $f_{1\times1}(\bullet)$ is 1×1 convolution.

The spatial attention mechanism can obtain the average and maximum values of each channel in the feature map and compress the number of channels in the image to 1. Then we use 3 × 3 convolution and the sigmoid activation function to fuse spatial information and activate spatial weights, which are multiplied with feature X points to obtain the weighted feature map M_s. The spatial weight A_s and the weighted feature map M_s are denoted by

$$M_s = A_s \odot X \tag{5}$$

$$A_s = \delta(f_{3\times3}(AP(X), MP(X)) \tag{6}$$

where $f_{3\times3}(\bullet)$ is 3×3 convolution, $AP(\bullet)$ is the average pooling operation, and $MP(\bullet)$ is the maximum pooling operation.

In summary, the PAM module can parallelize channel and spatial attention mechanisms while adaptively allocating target weights, thereby suppressing background interference and redundant information and improving the detection performance of the model.

B. MB-FPN module for feature fusion

The traditional FPN networks integrate the deep feature maps with the strong semantic features and the shallow feature maps with the strong texture information through the top-down and bottom-up paths, which can improve the network performance without affecting inference speed or increasing memory consumption. However, they are prone to generating false information when low resolution feature maps are upsampled. The false information makes the already small features of the object more chaotic, which is not conducive to subsequent recognition and detection. Therefore, the MB-FPN module is proposed, as shown in Fig.5. The feature maps $\{C_1, C_2, C_3, C_4\}$ are the feature layers downsampled from the feature extraction network $\{4, 8, 16, 32\}$, and the feature maps $\{P_1, P_2, P_3, P_4\}$ are the feature layers that $\{C_1, C_2, C_3, C_4\}$ have passed by the PAM module. Compared with the traditional FPN, the MB-FPN can integrate multi-scale information, avoid upsampling operations, and enrich the network's multi-scale expression ability.

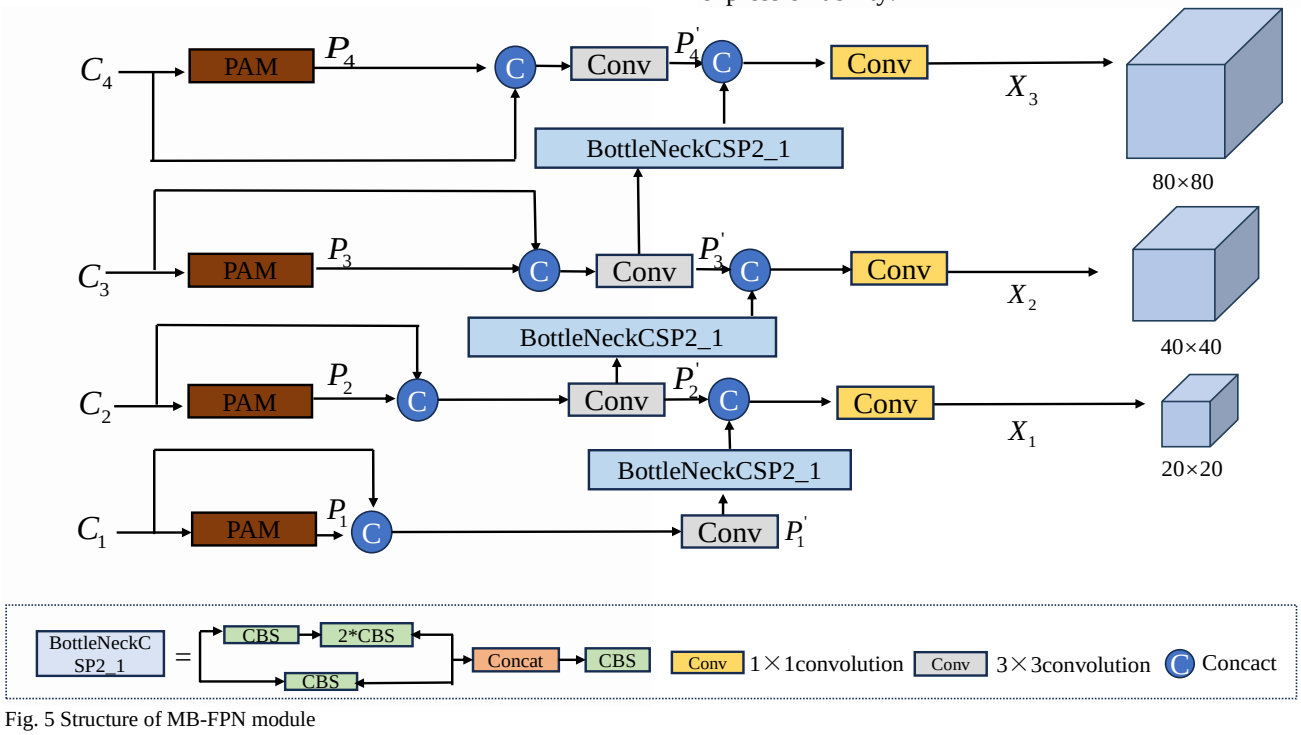


Fig. 5 Structure of MB-FPN module

Firstly, we concatenate features P_1 and C_1 to obtain richer shallow information. The fused feature $P_1^{'}$ is denoted by

$$P_1^{'} = f_{3\times3}(P_1 \oplus C_1) \tag{7}$$

Afterwards, we fuse the fused feature maps of P_2 and C_2 with the downsampled feature map of $P_1^{'}$ to obtain the feature map $P_2^{'}$. The feature map $P_2^{'}$ is subjected to 1×1 convolution to obtain the final output feature map X_1. The output feature X_1 is represented as

$$X_1 = f_{1\times1}((f_{3\times3}(P_2 \oplus C_2)) \oplus (Ds(P_1^{'})) \tag{8}$$

where $Ds(\bullet)$ represents downsampling the feature map.

Next, we fuse features C_3, P_3, and the downsampled features from $P_2^{'}$ to obtain the feature map $P_3^{'}$. The output feature X_2 is obtained by 1×1 convolution and denoted by

$$X_2 = f_{1\times1}((f_{3\times3}(P_3 \oplus C_3)) \oplus (Ds(P_2^{'})) \tag{9}$$

Similarly, the calculation process for output feature X_3 is represented as

$$X_3 = f_{1\times1}((f_{3\times3}(P_4 \oplus C_4)) \oplus (Ds(P_3^{'})) \tag{10}$$

C. Focaler-CIOU loss function for object location

The bounding box regression loss function is of pivotal significance for the domain of object detection. The positioning accuracy of object detection is contingent on the efficacy of the bounding box regression loss function. The YOLOv8 model employs the $CIoU$ [28] as the regression loss function for the purpose of determining the distance between the true and predicted boxes. The $CIoU$ loss function comprehensively considers the aspect ratio, center point distance, overlap area between the predicted box and the true box, improving the accuracy of object localization. However, the problem of significant background differences in aerial images leads to a highly imbalanced state between the object and background. The Focaler-IoU method is an algorithm that utilises linear interval mapping to address the challenges posed by imbalanced datasets, thereby enhancing the efficacy of regression models. Therefore, this paper applies the Focaler-IoU to the IoU loss function.

According to the difficulty of object detection, the object is divided into difficult samples and simple samples. We define general samples as simple samples, while small objects or objects that are difficult to accurately locate are considered difficult samples. The Focaler-IoU method uses linear interval mapping to reconstruct IoU loss. For detection tasks that mainly focus on simple samples, it is necessary to focus on simple samples during the regression process. On the contrary, it is necessary to focus on difficult samples when difficult samples dominate the regression process. The Focaler-IoU method can be formulated as

$$L_{Focaler-IoU} = 1 - IoU^{focaler} \tag{11}$$

$$\text{If } IoU < u, \text{ then } IoU^{focaler} = 0 \tag{12}$$

$$\text{If } d \ll IoU < u, \text{ then } IoU^{focaler} = \frac{IoU - d}{u - d} \tag{13}$$

$$\text{If } IoU > u, \text{ then } IoU^{focaler} = 1 \tag{14}$$

where IoU is the ratio of intersection to union, and $d, u \in [0,1]$. It can be seen that the Focaler-IoU method can focus on different detection tasks in different regression samples by adjusting the values of d and u. Therefore, this paper introduces the Focaler-CIoU loss function to alleviate the imbalance of positive and negative samples. the Focaler-CIoU loss function is defined as

$$L_{Focaler-CIoU} = L_{CIoU} + IoU - IoU^{Focaler} \tag{15}$$

where L_{CIoU} is the loss of $CIoU$, and $IoU^{Focaler}$ denotes the reconstructed IoU loss.

C. Dataset introduction

The Aerial dataset is sourced from aerial images of Spanish roundabouts, and mainly includes four types of objects: cars, buses, cycles, and trucks. This paper selects captured images from multiple scenes and uses data augmentation methods such as geometric transformation and brightness adjustment to improve the richness of the data, as shown in Fig. 6. We used labelimg software for manual annotation to construct the standard vehicle dataset, and randomly divided the dataset into training, validation, and testing sets in 8:1:1 ratio to obtain the best model. The reasonable partitioning of the dataset can prevent overfitting of the network and improve the accuracy of model training. This paper uses the K-means++ algorithm [29] to recluster the object categories. The anchor box parameters are shown in Table I.

TABLE I
ANCHOR BOX PARAMETERS

Feature map size	Receptive field size	Anchor
20×20	Big	(114, 91), (155, 199), (375, 327)
40×40	Middle	(31, 62), (63, 46), (58, 118)
80×80	Small	(11, 14), (15, 31), (32, 23)

D. Dataset experiment environment and parameters

To ensure the rigor of the experiment, all experiments are conducted in the same environment. The specific environmental configuration is shown in Table II.

Fig. 6 Aerial vehicle dataset for different scenarios

TABLE II
EXPERIMENTAL ENVIRONMENT CONFIGURATION

Parameter	Configuration	Parameter	Configuration
GPU	NVIDIA GeForce RTX 3060	CPU	Intel(R) Core(TM) i7-9750H CPU @2.6GHz
Image size	640×640	Learning rate	0.01
Operating System	Windows10	Epochs	200
CUDA	12.1	Momentum	0.937
Python	3.9.12	Weight decay	0.0005
Torch	2.2.1	Batch size	16

E. Evaluation index

In order to comprehensively and intuitively evaluate the performance of the MS-YOLO model, this paper selects common indicators such as precision, recall, mAP, F1, and FPS to evaluate the model. Precision is defined as the probability of detecting correctly among all detected objects. The recall refers to the probability of correctly identifying all positive samples. The average precision (AP) value is indicative of the precision of all recall rates. The mean average of each AP category is denoted by mAP. The following formulae are employed to calculate this mean average.

$$\text{precision} = \frac{TP}{FP+TP} \tag{16}$$

$$\text{recall} = \frac{TP}{FP+TN} \tag{17}$$

$$\text{mAP} = \frac{\sum P_{ri}}{K} \tag{18}$$

where TP and FP are respectively true positive and false positive, P_{ri} represents the area under the precise recall curve of a specific category, K is the number of categories, and $K=4$.

The $F1$ is the harmonic mean of precision and recall, which evaluates P and R as a whole. The $F1$ is represented as

$$\text{F1} = \frac{2\times R\times P}{R+P} \tag{19}$$

where P and R are the precision and recall, respectively.

V. EXPERIMENTAL RESULTS AND ANALYSIS

In the YOLOv8s model, we first add different attention mechanism modules at the same location, compare different feature fusion methods and loss functions, and verify the superiority of innovative points. Secondly, the ablation experiments are designed to verify the superiority of each module. Finally, this paper compared the proposed model with different object detection models to verify its superiority.

A. Performance comparison and analysis of feature fusion

This paper redesigns the feature fusion module and propose the MB-FPN module to suppress the false information generated by the upsampling of the original FPN. We compare the MB-FPN module with other modules uch as FPN, Bi-FPN, FPN-PAN, AFPN, and MB-FPN as shown in Table III. Note, the bold numbers in the tables represent the optimal results, while the italicised numbers represent the suboptimal results.

From Table III, it can be seen that compared to the original FPN module, the MB-FPN module has improved precision, recall, and mAP@0.5. Compared with other mainstream feature fusion methods, the mAP@0.5 has also been enhanced to a certain extent, indicating that the MB-FPN module has relatively excellent overall performance in detecting small objects in aerial vehicles. Overall, the MB-FPN module effectively avoids the false information generated by the original FPN when upsampling low resolution feature maps, fully integrates the deep and

shallow information of the backbone network, which can accurately locate the vehicle objects.

B. Comparative analysis of attention mechanisms

In order to solve the problems of background interference and small object feature loss in the feature extraction process, this paper adds mainstream attention mechanisms such as the SENet [30], CBAM, CA [31], SimAM [32], and PAM attention mechanisms to feature maps with downsampling multiples of $\{4,8,16,32\}$ in the backbone network for comparison.

Table IV shows the performance indicators of different attention mechanisms. It can be seen that the PAM module proposed in this paper has the highest precision, recall, and mAP@0.5. Specifically, compared with the original YOLOv5 network, the mAP of the PAM has increased by 1.5%, which can verify the effectiveness of the PAM module.

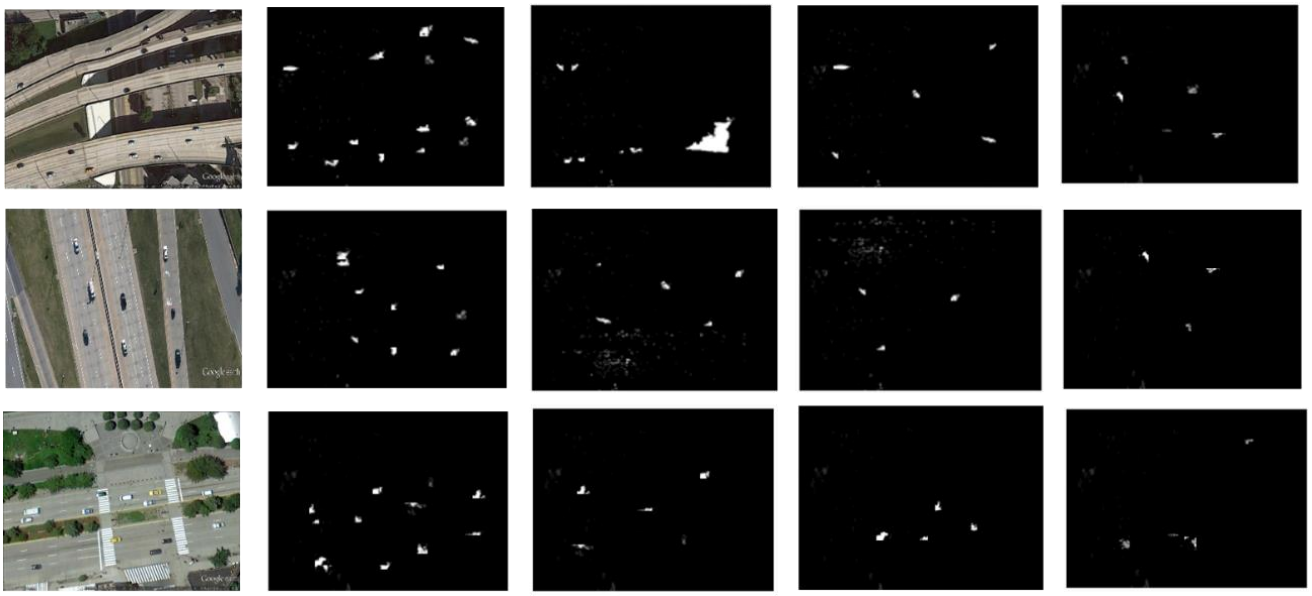
This paper uses visual images to show the attention situation of four different attention mechanisms on small objects of aerial vehicles, which can intuitively illustrate the superiority of the PAM module, as shown in Fig. 7. Due to the image is a high-altitude aerial image with small pixel ratios and significant differences in aspect ratios, we can see from Fig. 7 that the attention mechanisms of the SimAM, CBAM, and SENet all have varying degrees of feature loss. However, the PAM module can accurately focus on the vehicle position information by parallelizing spatial and channel attention mechanisms, adaptively adjusting spatial and channel weights. The module can also highlight vehicle features and suppress background noise interference.

TABLE III
COMPARISON EXPERIMENT OF FEATURE FUSION METHODS

Model \ Index	P/%	R/%	F1	mAP@0.5/%	mAP@0.5:0.95/%
FPN	87.5	87.6	87.5	91.3	65.9
FPN-PAN	92.1	*92.1*	92.0	93.3	68.8
Bi-FPN	**95.2**	90.6	*92.8*	*94.9*	*70.7*
AFPN	*95.1*	90.3	92.6	93.2	68.9
MB-FPN	94.2	**93.4**	**93.7**	**95.2**	**71.8**

TABLE IV
PERFORMANCE INDICATORS OF DIFFERENT ATTENTION MECHANISMS

Model \ Index	P/%	R/%	F1	mAP@0.5/%	mAP@0.5:0.95/%
YOLOv8	92.1	**92.1**	92.0	93.3	68.8
YOLOv8+SENet	93.1	*91.3*	92.1	93.8	67.4
YOLOv8+CBAM	92.5	91.0	91.7	*94.7*	*69.6*
YOLOv8+SimAM	**95.0**	90.6	*92.5*	94.6	68.3
YOLOv8+PAM	*94.6*	90.8	**92.6**	**94.8**	**71.3**



(a) Original (b) PAM (c) SimAM (d) CBAM (e) SENet

Fig. 7 Visualization of different attention mechanisms

C. Performance comparison and analysis of loss functions

This paper compares the loss function of Focaler-CIoU with that of the YOLOv8, as shown in Fig. 8. The overall convergence speed of the Focaler-CIoU loss function is higher than that of CIoU, with smaller loss values and a more stable network. The Focaler-CIoU loss function uses a linear interval mapping method to select the types of samples that are of particular concern, effectively suppressing the adverse effects of sample imbalance on the model.

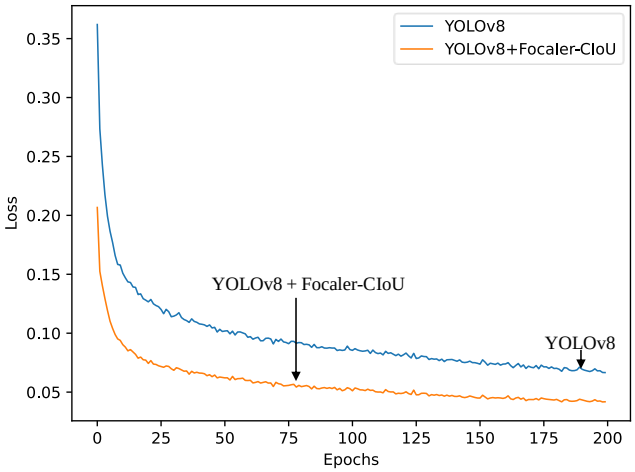


Fig. 8 Comparison of loss functions

Therefore, the Focaler-CIoU loss function has smaller positioning errors, faster and more accurate regression, which improves the detection accuracy of the model.

D. Ablation experiment

We design eight ablation experiments on the aerial datasets using YOLOv8 as the baseline network to verify the improvement of the model performance by PAM, MB-FPN, and Focaler-CIoU.

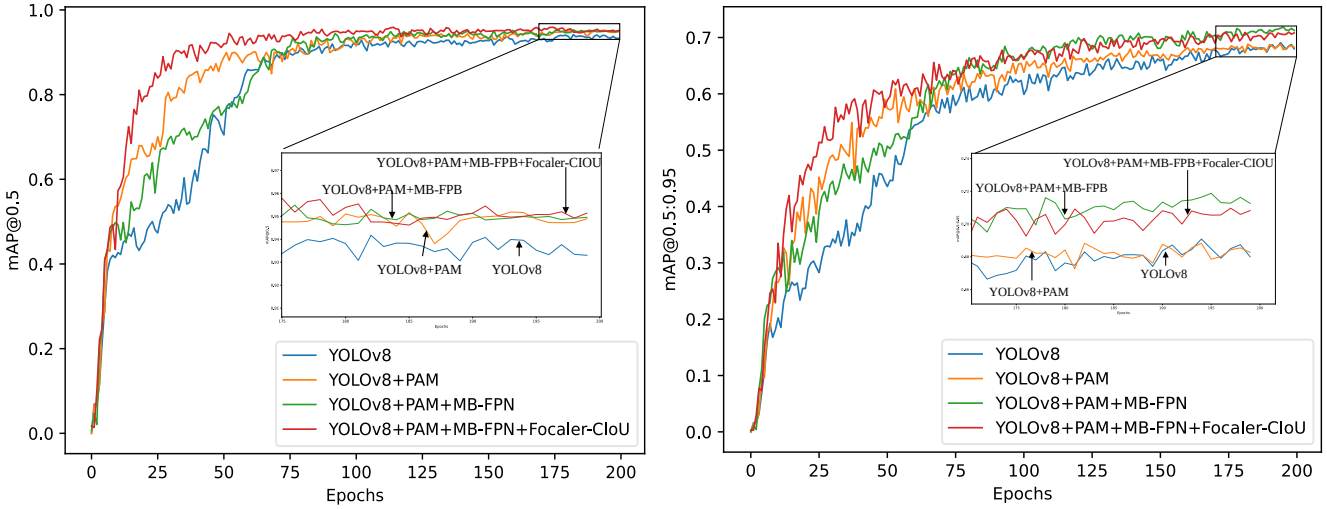
Table V shows the results of the ablation experiment. It can be seen that the introduction of PAM module reduced the detection speed, but mAP also decreased by 5.6 percentage points, and the detection accuracy was significantly improved. Compared to the YOLOv8 network, adding the PAM module increased the precision and mAP@0.5 of the model by 2.5% and 1.5%, respectively. The PAM module effectively enhances the expression of vehicle feature information. The MB-FPN module integrates feature maps of different scales. Compared with the original YOLOv8 network, mAP@0.5 has improved by 1.9% and achieved the highest recall rate, indicating that the MB-FPN network has improved detection accuracy with minimal loss of detection speed. The introduction of the Focaler-CIOU loss function has led to growth in various evaluation indicators, with increases of 1.1%, 0.6%, and 0.7%, respectively. Overall, compared to the YOLOv8 network, the network proposed in this paper improves the precision, recall, and mAP@0.5 by 3.8%, 0.5%, and 2.6%, respectively, with only a small number of parameters added. The results of the ablation experiment prove that all modules proposed in this paper have the effectiveness.

To visually illustrate the impact of each module on model detection performance, Fig. 9 shows the changes in mAP@0.5 values after adding each module. It can be seen that the mAP@0.5 value shows varying degrees of improvement after the successive addition of different modules, reaching stability around 170 rounds. Meanwhile, the model proposed in this paper has better mAP@0.5 value and better convergence of the curve within the iteration period from the locally enlarged graph.

Fig. 10 shows the P-R curves of YOLOv8 before and after improvement. The larger the area enclosed by P-R, and the better the model performs in measuring accuracy and recall, which means that while maintaining the recall rate, the model's prediction accuracy is higher. The P-R area of “vehicle” categories in the YOLOv8 network and MS-YOLO network approaches 1. This is because there are many “vehicle” class objects in the dataset, which allows both networks can effectively extract features. For the “cycle” class with the smallest pixel proportion and small sample size, the P-R area value and the mAP@0.5 value of the network proposed in this paper are higher than those of the original YOLOv8 network, which can demonstrate the superiority of the MS-YOLO network in high-altitude aerial image of small-scale vehicles.

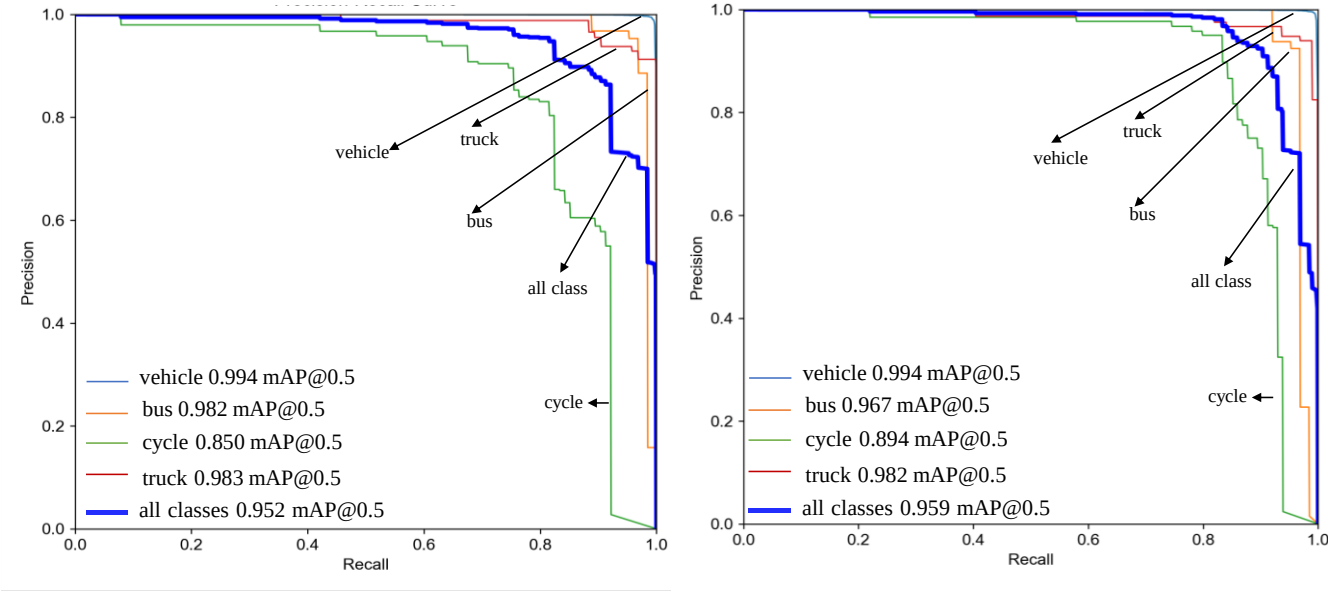
TABLE V
ABLATION RESULT

Index / Model	PAM	MB-FPN	Focaler-CIOU	P/%	R/%	mAP@0.5%	mAP@0.5:0.95%	FPS
1				92.1	92.1	93.3	68.8	**41.5**
2	✓			94.6	90.8	94.8	71.3	35.9
3		✓		94.2	**93.4**	95.2	71.8	39.3
4			✓	93.2	92.7	94.0	69.7	39.6
5	✓	✓		95.0	*93.3*	*95.3*	*72.1*	38.7
6	✓		✓	94.4	88.2	94.0	71.5	37.7
7		✓	✓	*95.6*	92.8	95.1	71.3	*40.1*
8	✓	✓	✓	**95.9**	92.6	**95.9**	**72.3**	35.5



(a) mAP@0.5 (b) mAP@0.5:0.95

Fig. 9 Changes in index in ablation experiments



(a) YOLOv8 network (b) MS-YOLO network

Fig. 10 P-R curve

E. Comparative experiment

To comprehensively evaluate the actual performance of MS-YOLO model, this paper provides the detailed comparison between the MS-YOLO model and six classic algorithms in terms of precision, recall, F1, mAP@0.5, mAP@0.5:0.95, and FPS. Table VI is the comparison of indexes for different models. The MS-YOLO has greater advantages compared to the other object detection methods. Specifically, compared with the original YOLOv8 network, the precision, recall, and mAP@0.5 have improved by 3.8%, 0.5%, and 2.6%, respectively. Compared with the latest YOLOv11model, the recall rate is slightly lower, but other evaluation metrics have slightly improved, which can verify the superiority of MS-YOLO model. There has been a certain improvement in all indicators by compared with the two-stage classical models and other one-stage models. Moreover, the detection speed of the model in this paper is slightly slower than the original YOLOv8 model, but still faster than the two-stage object detection models. Therefore, although the MS-YOLO sacrifices the smaller detection speed, it significantly improves the detection performance of the model, making it more suitable for detecting small object vehicles in the high-altitude aerial images.

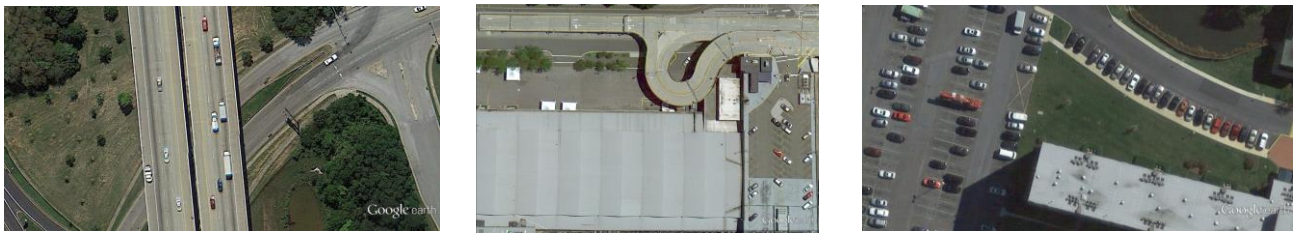
In order to further verify the detection performance of the MS-YOLO model on aerial small target vehicles in different scenarios, this paper selected multiple images from different scenes. Fig. 11 shows the original image, the visualization results of Faster R-CNN, YOLOv5, YOLOX [33], YOLOv8, YOLOv11 and our model, respectively. The yellow boxes in the visualization image represent the missed detections, while the green boxes represent false detections. In Fig. 11(b), the Faster R-CNN model has a large number of missed detections, which cannot meet the actual needs of traffic vehicle detection. The YOLOv5 original model and YOLOX model have improved the missed detections of Faster R-CNN, but there are still missed and false detections when capturing small vehicle objects at high altitudes from Fig. 11(c) and (d). In Fig. 11(e), the YOLOv8 model misclassifies the truck as the bus class and misclassifies the background as the bus, with an amount of false detection. As can be seen from Fig. 11(f), the latest YOLOv11 detection algorithm will still produce false detection when the background is similar to the shape and color of the object,

and there will be missed detection when the object is dense. The MS-YOLO model can effectively distinguish the vehicle categories, significantly reduce the number of missed detections, and has the better confidence.

In summary, the MS-YOLO detection model has superior detection performance, especially in densely populated vehicles. This paper verifies the feasibility of the MS-YOLO model through intuitive data and visualization results in different scenarios. Therefore, the MS-YOLO can be applied to small object detection tasks in traffic scenes.

TABLE VI
COMPARISON OF INDEXES FOR DIFFERENT MODELS

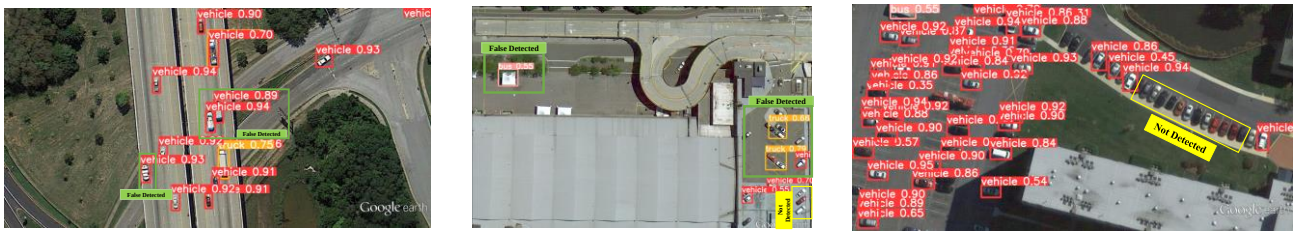
Index / Model	P/%	R/%	F1	mAP@0.5/%	mAP@0.5:0.95/%	FPS
Faster R-CNN	63.1	85.7	73.0	77.4	43.6	7.36
SSD	89.2	81.0	84.9	83.3	51.6	14.3
YOLOv4	81.9	83.2	82.5	86.9	54.9	39.7
YOLOv5	93.2	*91.6*	92.4	92.3	66.4	37.3
YOLOv7	93.1	89.0	90.9	91.1	66.9	*40.6*
YOLOX	93.9	91.6	92.7	94.2	65.3	**43.7**
YOLOv8	92.1	92.1	92.0	93.3	68.8	41.5
YOLOv11	95.3	**93.1**	94.1	94.9	70.3	38.5
MS-YOLO	**95.9**	92.6	**94.5**	**95.9**	**72.3**	35.5



(a) Original images

(b) Faster-RCNN

(c) YOLOv5



(d) YOLOX

(e) YOLOv8

(f) YOLOv11

(g) MS-YOLO

Fig. 11 Visualization results of different object detection network

VI. Conclusion

This paper proposes the new model for detecting small objects in aerial vehicles. Firstly, the PAM module is constructed in the backbone network to parallelize channel and spatial attention mechanisms, intelligently allocate weights between background and target, thereby enhancing the model's feature extraction capability and reducing interference from irrelevant background. Moreover, we propose the MB-FPN module in the neck to fully utilize the feature information of different layers and improve the detection performance of small objects. In order to alleviate the imbalance between background and object, the Focaler-CIoU loss function is introduced, which can selectively focus on simple and difficult samples to improve the detection accuracy of the model. The experimental results on the Aerial dataset show that the MS-YOLO model outperforms other compared models in terms of precision, recall, and mAP evaluation metrics, and can effectively avoid problems such as false positives, missed detections, and duplicate detections. In the future, we will further optimize the model to have better detection ability and faster detection speed when facing complex traffic scenes of vehicle objects.

References

[1] C. J. Kim et al., "End-to-end deep learning-based autonomous driving control for high-speed environment," J. Supercomput. 78, 1961–1982 (2022).

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR), pp. 886-893 (2014).

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vision. 60(2), 91–110 (2004).

[4] H. B. Alwan and K. R. Mahamud, "Cancellable face template algorithm based on speeded-up robust features and winner-takes-all," Multimedia Tools Appl. 79(39-40), 28675-28693 (2020).

[5] Z. Hu et al., "Speech emotion recognition model based on attention CNN Bi-GRU fusing visual information," Engineering Letters. 30(2), 427-434 (2022).

[6] A. S. Girsang and D. Tanjung, "Fast genetic algorithm for long short-term memory optimization," Engineering Letters. 30(2): 528-536 (2022).

[7] Y. Luo et al., "A speaker recognition method based on dynamic convolution with dual attention mechanism," Engineering Letters. 31(2), 825-832 (2023).

[8] S. Ren et al., "Faster R-CNN: towards real-time object detection with region proposal networks," in IEEE Trans. Pattern Anal. Mach. Intell. pp. 1137-1149 (2017).

[9] T .Y. Lin et al., "Feature pyramid networks for object detection," in Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR), pp. 936-944 (2017).

[10] J. Redmon et al., "You only look once: unified, real-time object detection," in Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR), pp. 779-788 (2016).

[11] X. Li et al., "Improved YOLOv4 network using infrared images for personnel detection in coal mines," J. Electron. Imag. 31(1), 013017-013017 (2022).

[12] S. Peng et al., "PS-YOLO: a small object detector based on efficient convolution and multi-scale feature fusion," Multimedia Syst. 30(1), 241-252 (2024).

[13] D. Wang et al., "FS-YOLO: fire-smoke detection based on improved YOLOv7," Multimedia Syst. 30(1), 215-227 (2024).

[14] W. Liu et al. "SSD: single shot multibox detector," in Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 21-37 (2016).

[15] T.Y.Lin et al., "Focal loss for dense object detection," in IEEE Int. Conf. Comput. Vision. (ICCV), pp. 2999-3007 (2017).

[16] S. Liu et al., "Path aggregation network for instance segmentation," in Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR), pp. 8759-8768 (2018).

[17] Z. Song et al., "Deformable YOLOX: detection and rust warning method of transmission line connection fittings based on image

processing technology," IEEE Trans. Instrum. Meas, 72(1), 1-21 (2023).
[18] C. Guo et al., "AugFPN: improving multi-scale feature learning for object detection," in Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR), pp. 12592-12601 (2020).
[19] Y. Song et al., "MEB-YOLO: an efficient vehicle detection method in complex traffic road scenes. Comput," Materials & Continua, 75(3), 5762-5764 (2023).
[20] L. Jiang et al., "MFFSODNet: Multiscale feature fusion small object detection network for UAV aerial images," IEEE. Trans Instrum. Meas, 73(1), 1-14 (2024).
[21] H. Zhang et al., "Accurate detection and tracking of small-scale vehicles in high-altitude unmanned Aerial Vehicle Bird-View Imagery," J. Adv Transp. 2023(1), 538484-538501 (2023).
[22] K. Hao et al., "An insulator defect detection model in aerial images based on multiscale feature pyramid network," IEEE Trans. Instrum. Meas, 71(1), 1-12 (2022).
[23] X. Xiang et al., "Research on infrared small target pedestrian and vehicle detection algorithm based on multi-scale feature fusion," J. Real-Time Image Process., 22(1):31-31 (2024).
[24] V. Rejin, M. Sambath., "YOLOv8: a novel object detection algorithm with enhanced performance and robustness," in Int. Conf. Advances Data Eng. Intell. Comput. Syst. (ADICS), pp. 1-6 (2024).
[25] C. Y. Wang et al., "CSPNet: a new backbone that can enhance learning capability of CNN," in IEEE Conf. Comput. Vision and Pattern Recognit. Workshops. (CVPRW), pp.1571-1580 (2020).
[26] S. Li et al., "Real-time vehicle detection from UAV aerial images based on improved YOLOv5," Sensors, 23(12), 5634-5634 (2023).
[27] B. Mahati et al., "Exploring hardware activation function design: CORDIC architecture in diverse floating formats," in Int. Symp. Qual. Electron. Des. (ISQED), pp. 1-8 (2024).
[28] Z. Zheng et al., "Distance-IOU loss: faster and better learning for bounding box regression," in Proc. AAAI Conf. Artif. Intell. pp. 12993–13000 (2020).
[29] A. M. Ikotun et al., "K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data," Inf. Sciences, 622(1), 178-210 (2023).
[30] J .Hu et al., "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR), pp. 7132-7141 (2018).
[31] Q. Hou, D. Zhou, J. Fen., "Coordinate attention for efficient mobile network design," in Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR), pp. 13708-13717 (2021).
[32] Y. Huang et al., "High-accuracy insulator defect detection for overhead transmission lines based on improved YOLOv5," Appl. Sci. 12(24), 12682-12695 (2022).
[33] Y. Wang et al., "Ships' small target detection based on the CBAM-YOLOX algorithm," J. Mar. Sci. Eng. 10(1), 2013-2031 (2022).