World Scientific Research Journal Volume 11 Issue 10, 2025
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

FAS-YOLO: An Improved Algorithm for Small Target Detection of Unmanned Aerial Vehicles

Xinlong Shui, Wendi Chu

Eighth Oil Production Plant, Daqing Oilfield Co., Ltd. Daqing, Heilongjiang, 163400, China

Abstract

Small target detection from the perspective of unmanned aerial vehicles (UAVs) faces the challenges of low accuracy, high missed detection rate, and high false detection rate. To address these issues, this paper proposes an improved small target detection model based on YOLOv8, named FAS-YOLO (Feature Attention Small Object Detection-YOLO). Firstly, the model replaces the traditional pooling operation by introducing the FPSharedConv module, which can effectively extract fine-grained features and retain the detailed information of small targets. Secondly, based on the improvement of PAFPN, the smallObjectEnhancePyramid feature pyramid structure is proposed: without adding the P2 detection layer, through the fusion of the SPDConv convolution of the P2 feature layer and CSP-OmniKernel, the feature representation ability of small targets is effectively enhanced. In addition, the AFGCAttention mechanism is introduced after the FPSharedConv to further improve the model's attention to key small targets. Finally, the loss function is improved based on WIoUv3, and the detection and positioning accuracy is improved by using a more reasonable aspect ratio measurement. The experimental results show that the precision, recall, mAP50, and mAP50-95 of the improved model on the VisDrone2019 dataset are increased by 9.6%, 8.6%, 10.9%, and 7%, respectively. FAS-YOLO significantly improves the performance of small target detection and provides a new solution for efficient target detection in UAV scenarios.

Keywords

Small target detection, YOLOv8, Feature Pyramid, FPSharedConv, Attention mechanism, AFGCAttention, WIoUv3.

1. INTRODUCTION

In recent years, with the rapid development of unmanned aerial vehicle (UAV) technology, the functions of UAVs have become increasingly diverse, and their roles in various tasks have grown significantly. Applications based on UAV technology have also become more widespread. In the face of complex urban and natural environments, the flexible and lightweight UAVs have huge advantages that traditional visual devices cannot match. [1] However, objects in the field of view of UAVs often have problems such as small size, indistinct features, severe occlusion or overlap, which lead to a series of issues in the detection of small targets by UAVs, such as missed detections, false detections, and low accuracy. [2] Therefore, developing small target detection models suitable for UAVs is of great significance.

In recent years, with the rapid development of deep learning, target detection technology based on deep learning has been increasingly widely applied. Currently, deep learning-based target detection models are mainly divided into one-stage models and two-stage models according to the process. Two-stage target detection models first provide several preselected regions and then select the best prediction regions based on these regions. The detection accuracy of two-stage models is higher than that of one-stage models at the same period, but

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

due to the large computational load of two-stage models, they generally have problems of slow inference speed and low frame rate. For the use of target detection models on UAVs with limited computing power, two-stage target detection models are not a good choice. One-stage target detection models do not need to generate candidate regions and have obvious advantages in inference speed and frame rate while ensuring accuracy, making them a good choice for target detection tasks on UAVs.

One-stage target detection models also have continued to develop, giving rise to a series of representative models, such as RetinaNet [3], SSD (Single Shot Multibox Detector) [4] series, and YOLO (You Only Look Once) [5-7] series. Among them, the YOLO model has been iteratively improved and gradually developed into a widely used target detection model. Especially YOLOv5 and YOLOv8, with their complete network structure and strong generalization ability, are widely applied in target detection tasks and have become the basis for many improved models.

For example, Shen Xueli et al. [8] proposed an improved UAV aerial photography target detection model SFE-YOLO (Shallow Feature Enhancement-YOLO), which has better generalization ability while improving performance. Pan Wei et al. [9] proposed a model for small target detection based on YOLOv8 by integrating multiple attention mechanisms. Although the performance was improved, the number of parameters and the size of the model increased significantly. Liang Xiuman et al. [10] improved the neck network of YOLOv8 by adding shallow detection layers and removing large target detection layers, achieving dual optimization of parameter reduction and performance improvement.

In the field of small target detection, many studies have made progress. For instance, Lei Bangjun et al. [11] proposed the Improved-v8s small target detection algorithm, which enhanced the perception and capture ability of small targets by improving the detection layer and strengthening the fusion of shallow and deep information. They also designed the F_C2f_EMA and SM_SPPCSPC modules to enhance the feature extraction ability of small targets. Wang et al. [12] proposed UAV-YOLOv8, introducing the BiFormer attention mechanism to optimize the backbone network, designing a lightweight feature processing module FFNB, and combining PAFPN to propose two new detection scales. Li et al. [14] addressed the issue of false detection and missed detection of small targets in aerial images by improving the neck network of YOLOv8 with Bi-FPN [13] and introducing the GhostBlockV2 structure based on GhostConv in the backbone network to reduce information loss during transmission and significantly reduce the model's parameter count.

Although the above-mentioned methods have made multi-faceted optimizations for the target detection scenario from the perspective of unmanned aerial vehicles (UAVs), and have improved the performance of small target detection to a certain extent, they still face the following problems: insufficient detection accuracy, slow inference speed, high rates of missed and false detections in complex scenes, large model parameter count, and difficulty in balancing performance and resource consumption. These issues indicate that there is still room for further optimization in small target detection for UAV aerial photography scenarios.

In view of the above problems, the YOLOv8 model is improved from the following aspects.

1) Introduce the FPSharedConv module to replace the traditional pooling operation, and retain more fine-grained feature information of small targets.

2) The SmallObjectEnhancePyramid feature pyramid structure is proposed, and the small object detection capability is enhanced without increasing the P2 detection layer through the fusion of SPDConv and CSP-OmniKernel modules.

3) The AFGCAttention attention mechanism is introduced after feature extraction to adaptively focus on small targets and improve the accuracy and stability of detection.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

4) WIoUv3 is used to improve the loss function, and improve the positioning accuracy and the adaptability of the model to complex backgrounds through a more reasonable aspect ratio measurement method.

Experimental results show that the improved model significantly improves the performance when the model size and parameter quantity are reduced. Among them, mAP50 and mAP50-95 are increased by 10.9 and 7 percentage points, respectively, and can be effectively applied to UAV small target detection tasks.

2. YOLOV8 MODEL

The YOLOv8 model is divided into five versions: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x according to the different network width and depth, and the complexity of the model increases sequentially. Considering the limitation of the computing power of UAV equipment, the lightest YOLOv8n was selected for improvement.

In terms of architectural design, the backbone network of YOLOv8 is based on the CSP-Darknet-53[15] architecture, and the C3 module is replaced by the C2f module to enhance the gradient flow and improve the detection accuracy. The neck network also replaces the C3 module with the C2f module to refine the feature fusion and further optimize the object detection effect. Feature fusion adopts a PAN + FPN structure, which upsamples the smallest feature map generated by the backbone network, gradually scales it to P3 size after fusion with shallow features, and then downsamples it to P5 size. Feature maps for sizes P3 to P5 are entered into the inspection head.

YOLOv8's detection head is a Decoupled Head[16] and uses three branch outputs, each containing two parts, which are responsible for regression and classification tasks. The loss function adopts CIoU (Complete IoU), and adds the calculation of overlap area, center point distance and aspect ratio on the basis of IoU [17], which provides a more comprehensive optimization mechanism than GIoU (Generalized IoU).

Through the above design, YOLOv8 can improve the adaptability to object detection tasks while ensuring the performance of the model, especially in the lightweight model.

3. FAS_YOLO MODEL

Figure 1 shows the network structure of FAS-YOLO. Firstly, in order to solve the problems of feature loss, insufficient receptive field and excessive computational burden in small target detection, the FPSConv module is constructed to replace the SPPF layer, and the accuracy of small target detection is not only improved through multi-scale feature extraction, detail preservation and efficient shared convolutional kernel design, but also takes into account the real-time performance and detection effect of small target detection without significantly increasing the amount of computation. In contrast, the pooling operation of SPPF may lose some detailed information. Then, After feature extraction, the AFGCAttens attention mechanism [18] was introduced to adaptively focus on small targets and improve the accuracy and stability of detection. Then, the OmniKernel module [19] was inserted into the neck, and the CSP idea was introduced to improve the problem of excessive computation, and finally solve the problems of single feature fusion, low efficiency and high computation of the original PAFPN [20]. Finally, based on WIoUv3 [21], the loss function is improved to improve the positioning accuracy and the adaptability of the model to complex backgrounds through a more reasonable aspect ratio measurement method.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

3.1. Convolution

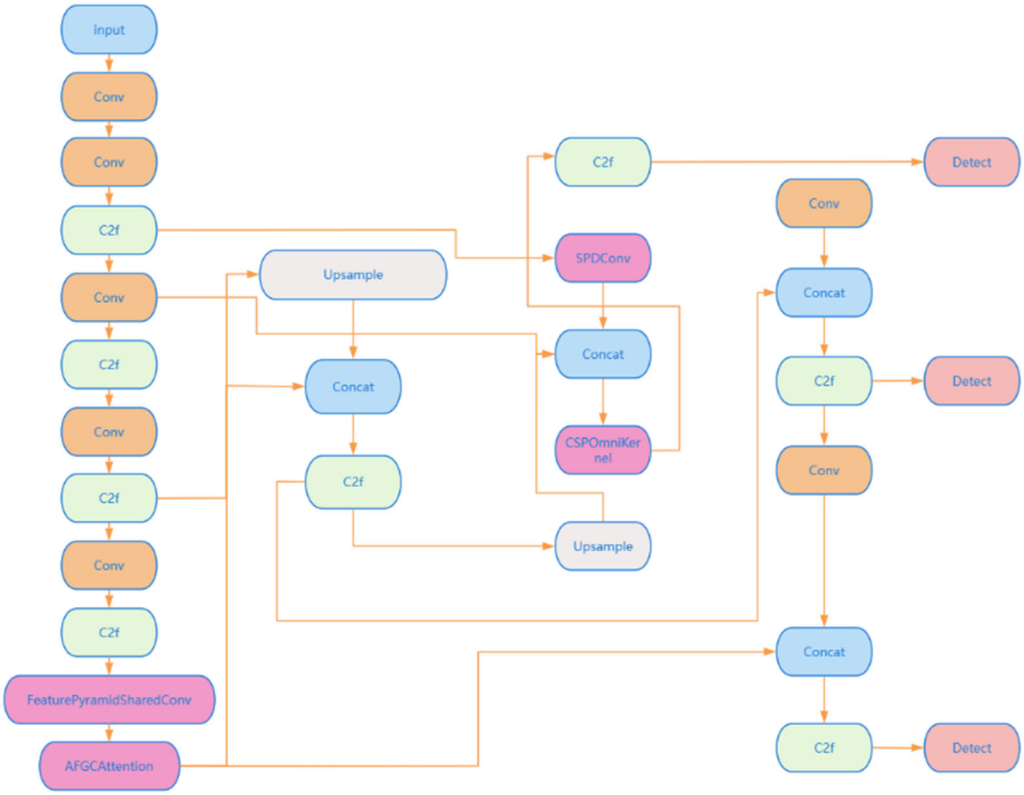


Figure 1. FAS_YOLO network structure diagram

3.1.1 Shared convolution

Shared Convolutional Network (CSN) is a strategy applied in object detection algorithms to improve the efficiency and accuracy of the model by sharing the weight parameters of the same convolutional neural network (CNN) between different steps or components. In traditional object detection algorithms, two independent convolutional neural networks are often used: one to generate candidate regions (e.g., RPNs) and the other to classify and regress candidate regions (e.g., Fast R-CNNs). This design not only increases the complexity of the algorithm, but also wastes computing resources. In order to optimize this problem, the idea of shared convolutional networks came into being. By sharing the weight parameters of the convolutional network among multiple components, redundant computing and memory consumption can be effectively reduced, while improving the speed and accuracy of object detection.

3.1.2 Dilated convolution

Dilated convolution was originally proposed to solve problems in image segmentation. Common image segmentation algorithms usually increase the receptive field through pooling layers and convolutional layers, but in the process, the size of the feature map is reduced, and then the size of the original image is restored by upsampling. However, the process of shrinking and enlarging the feature map will lead to a certain loss of accuracy. Therefore, dilated convolutions have emerged to increase the receptive field without changing the size of the feature map, thus replacing the traditional downsampling and upsampling operations.

Unlike normal convolution, dilated convolution introduces a hyperparameter called the dilation rate, which defines the spacing of the values of the convolution kernel when processing the data. For example, when the size of the convolutional kernel is 3×3, the size of the receptive

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

field can be effectively controlled by adjusting the expansion rate, so that features at different scales can be extracted more flexibly.

In recent years, dilated convolution has been widely used in object detection and other vision tasks. Huang et al. [22] designed a residual roll integration module based on the combination of expansion rates, which used different receptive fields to obtain target features, and improved the target detection performance in complex scenes through feature adaptive fusion. Vishnu Chalavadi et al. [23] proposed a new network, mSODANet, which uses hierarchical dilated convolution for multi-scale object detection in aerial images. Through parallel expansion convolution, the context information of different types of objects at multiple scales and fields of view can be learned to improve the detection ability. Zhen et al. [24] proposed an improved YOLOv4 method in 2023 to introduce an expanded coordinate attention module that combines an expanded convolutional neural network with coordinate attention to improve the accuracy of object detection. At the same time, a multi-scale training strategy is adopted to enhance the detection performance. Zhang et al. [25] proposed a novel conditional generative adversarial network based on extended convolution for infrared small target detection. By designing the generative network, the extended convolution is used to make full use of the context information, retain the semantic information of the infrared small target, enhance the target features, and improve the detection performance.

3.1.3 AFGCA channel attention mechanism

The channel attention mechanism [26] is a common attention mechanism in deep learning, which is widely used in computer vision tasks, and its core idea is to dynamically adjust the weight of each channel according to the importance of each channel in the input feature map, so that the network can pay more attention to the important features and suppress the unimportant features. As shown in Figure 2, firstly, the feature map of each channel is aggregated by global average pooling or global maximum pooling to obtain the global information representation of each channel, and then a small neural network (such as a fully connected layer) is used to process the global information representation of each channel to generate the weight of each channel. Finally, according to the calculated channel weights, each channel of the original feature map is weighted, and then the contribution of each channel to the subsequent processing is adjusted. In this paper, a new adaptive fine-grained channel attention mechanism is used, as shown in Figure 3.

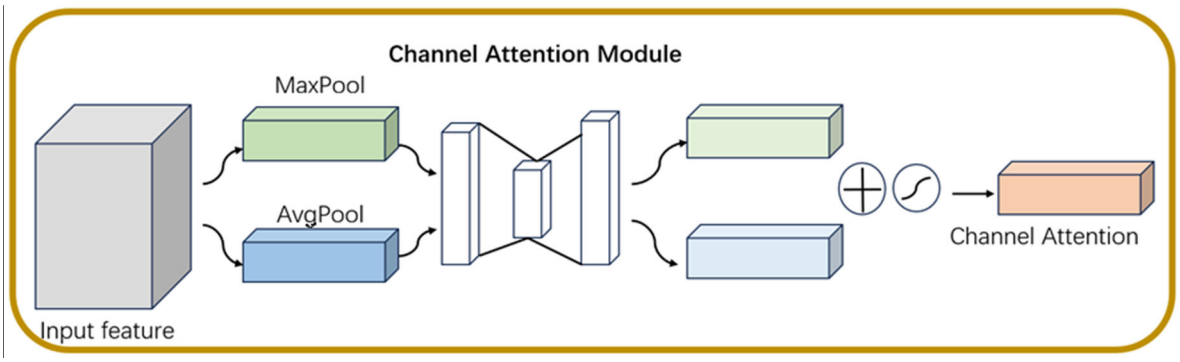


Figure 2. Channel attention structure

Firstly, the input feature map is pooled globally to obtain a channel-level global feature description U with a shape of C×1×1, and then U is divided into two groups: UGC and Ulc, in which UGC is used for global channel features to extract channel information in the global range, and Ulc is used for local channel features, focusing on capturing local frequency information and constructing matrices respectively. Then, the correlation between channels is generated by matrix transformation, the learnable factor θ controls the importance of UGC and Ulc, and the

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

final channel weight W is generated by the Sigmoid function. Finally, the input feature map F is weighted according to the channel weight W, and the output feature F* is generated. The structure of this paper can make the model pay more attention to the characteristics of small and medium-sized targets, and make the model more suitable for small targets, complex backgrounds, or tasks that require high precision.

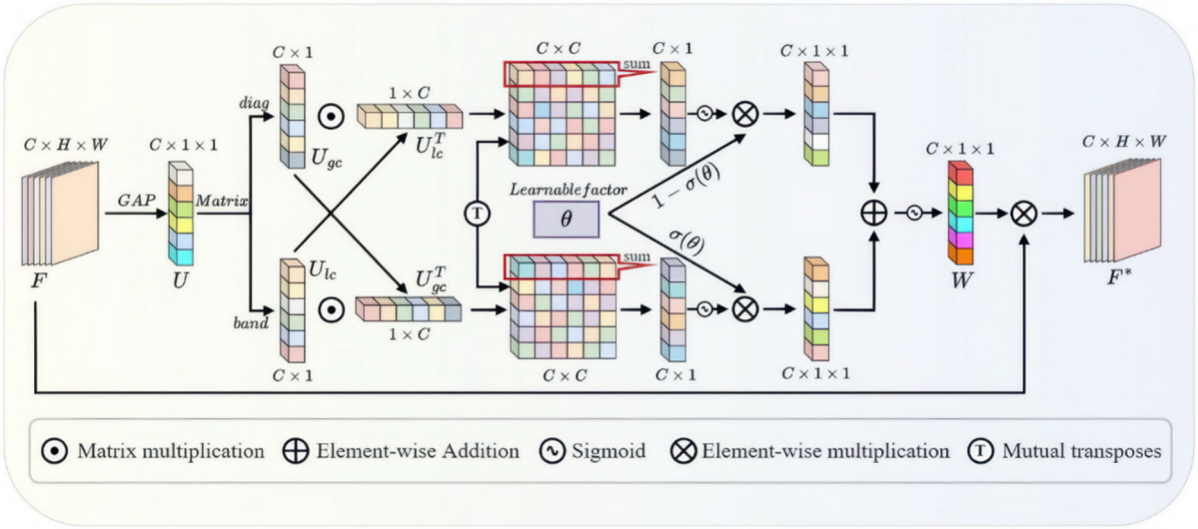


Figure 3. AFGCA Attention Mechanism

Compared with traditional channel attention mechanisms such as SE [27], AFGCA captures multi-scale features through frequency grouping (global and local), which can better deal with small targets and complex backgrounds. SE, on the other hand, only relies on global pooling and ignores local information. Secondly, AFGCA explicitly models the correlation between channels, and introduces learnable factors to dynamically adjust the global and local feature weights, which has stronger adaptability and fine-grained control ability, while the weighting of the SE module is fixed and coarse. In addition, AFGCA has stronger generalization capabilities for complex scenes, and although the computational overhead is slightly higher, it has outstanding performance in improving detail extraction and robustness, making it more suitable for high-performance tasks such as small object detection and medical imaging.

3.1.4 FeaturePyramidSharedConv module

The FeaturePyramidSharedConv module realizes multi-scale feature extraction and detail information retention through the convolution operation of shared convolution kernels and different dilations, and through three shared convolution kernels with expansion rates of 1, 3, and 5, respectively. Extended convolution captures global information by increasing the receptive field, because of the parameter learning ability of the convolution operation, it avoids the loss of details caused by the traditional pooling operation, so as to retain the detail characteristics of small targets. In order to prevent the number of parameters from being too large, the use of shared convolutional kernel reduces the number of parameters and computational complexity, and improves the computational efficiency. In addition, the module can balance the extraction of local and global information, which enhances the detection ability of small targets. The specific structure is shown in Figure 4.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

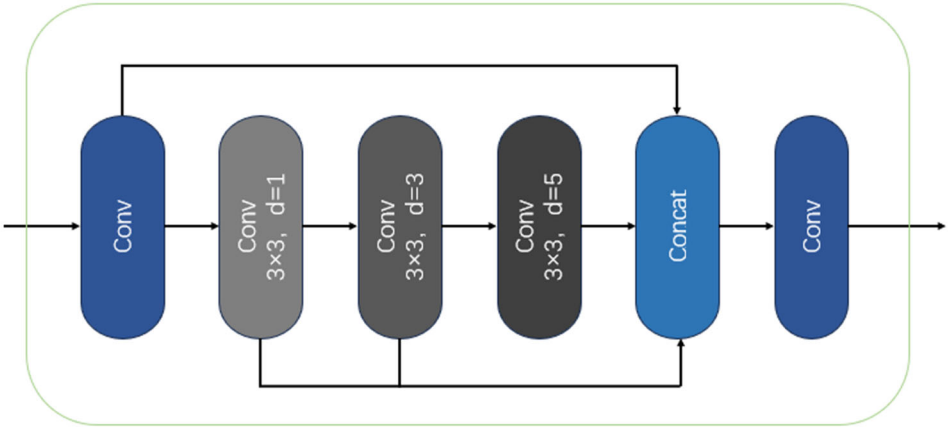


Figure 4. FPSharedConv module

3.2. Small object detection based on multi-scale convolution and CSP structure

On the one hand, because the scale of the corresponding feature map is larger, the receptive field is too large and the detail information is lost more for the small target, and on the other hand, because the P3P4P5 layer is a more representative high-level feature obtained by feature fusion, the low-level features of the small target (such as edge, texture, etc.) are not fully expressed in the high-level features. The traditional approach is to add a P2 detection layer, which can effectively improve the network's perception ability of small targets by improving the features of higher resolution and smaller receptive fields. Chun-Lin Ji et al. [28] added an additional detection layer for small target detection on the basis of YOLOv5 to generate a larger-scale feature map to capture the detailed features of small targets. In addition, the C3CrossConv module and the global attention mechanism are integrated to improve the detection performance of the model. However, it also brings some problems, such as the amount of computation in the network after adding P2 is too large, and the post-processing is more time-consuming. Therefore, based on the improvement of the original PAFPN, the SOEP module is designed, compared with the traditional addition of P2 detection layer, we use the P2 feature layer to obtain features rich in small target information through SPDConv, and then give P3 for fusion, and then use CSP idea and Omnikernel to improve to obtain CSP-Omnikernel feature integration. The Omnikernel module consists of three branches: global branch, large branch, and local branch, in order to effectively learn the feature representation from global to local. The result is an increase in the detection performance of small targets without the problem of excessive computation. The internal structure diagram of Omnikernel and the structure diagram of CSP-Omnikernel are shown in Figure 5 and Figure 6.

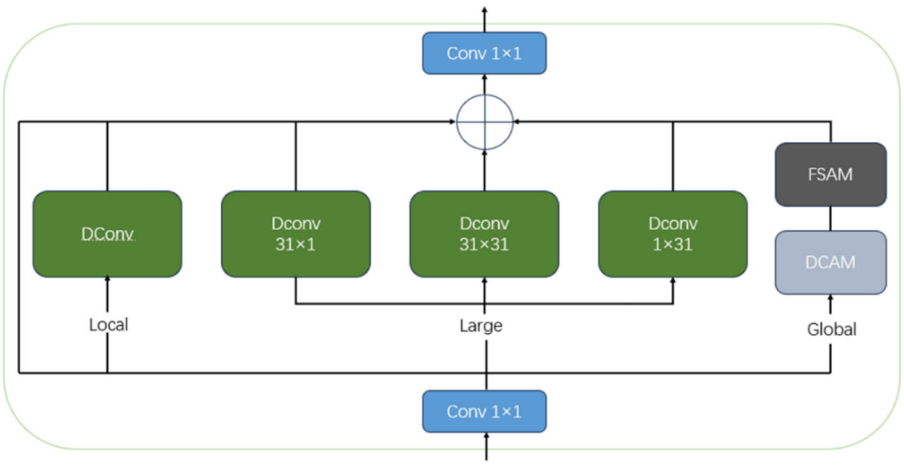


Figure 5. Omnikernel structure

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

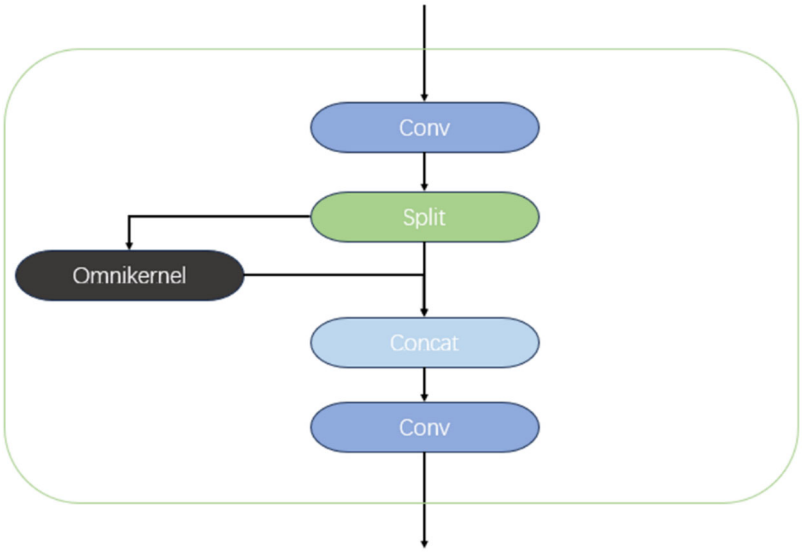


Figure 6. CSP-Omnikernel structure

3.3. Improved loss function

The bounding box regression loss occupies a key position in the overall loss function of the YOLOv8 model, and its reasonable and fine design has a significant effect on improving the overall performance of the model. In recent years, the research work has generally tended to presuppose that the quality of the training samples is high, and based on this, the focus is on optimizing the fitting performance of the bounding box loss function. CIoU [29]LossAs the detection box loss function in YOLOv8, CIoU Loss is designed by default to the high quality of the training data samples, and is committed to strengthening the fitting performance of the model to the bounding box loss function, but it does not fully consider the negative impact of low-quality training data on the model performance. To solve this problem, this chapter introduces Wise-Inner-MPDIoU instead of CIoU Loss based on WIoU[30] Loss. Compared with CIoU Loss, WIoU Loss not only weakens the adaptability of high-quality bounding boxes, but also effectively inhibits the gradient influence caused by low-quality data, so that the model can focus more on processing the anchor boxes of ordinary quality, and comprehensively improve the overall performance of the model. However, WIoU only considers the distance from the center point of the bounding box, ignoring the important contribution of other critical areas of the box, such as edges, midpoints, and vertices, to the positioning accuracy. Therefore, Wise-Inner-MPDIoU is introduced, and the loss function improves the detection performance through fine-grained alignment and dynamic adjustment in complex backgrounds, providing a new technical path for precise positioning. In object detection, the anchor frame is B=[xy wh], and the target box is Bgt=[xgt; ygt wgt hgt]. Figure 7 illustrates the geometric significance of each parameter in an actual IoU operation.

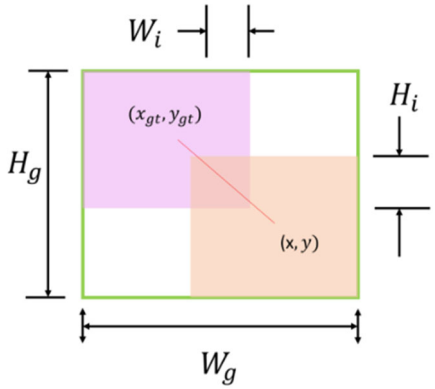


Figure 7. IoU operation

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

$$S_u = wh + w_g th_g t - W_i H_i \tag{1}$$

$$L_{IoU} = 1 - IoU = 1 - \frac{W_i H_i}{S_u} \tag{2}$$

LIoU is used to measure the degree of overlap between the prediction box and the real box, as defined in Eq. (1) and Eq. (2). But IoU has a serious flaw that the backpropagation gradient disappears when there is no overlap between bounding boxes, which results in training not being able to update IoU in the absence of overlapping regions. Therefore, a large number of existing studies have taken into account the set geometry factors related to the bounding box, and constructed penalty terms to solve this problem. DIoU[31] (Distance-IoU) defines the penalty term as the normalized length of the centerpoint connection:

$$\mathcal{R}_{DIoU} = \frac{\left(x - x_{gt}\right)^2 + \left(y - y_{gt}\right)^2}{W_g^2 + H_g^2} \tag{3}$$

In order to further improve the accuracy of bounding box prediction, YOLOv8 introduces CIoU that incorporates the aspect ratio into the bounding box calculation, which improves the detection ability of bounding box regression. When the aspect ratio of the predicted bounding box is the same as that of the real bounding box, the aspect ratio penalty term of the CIoU loss function is always 0, and the convergence process fluctuates greatly. Based on the problem that the CIoU Loss gradient cannot be updated, WIoU Loss additionally solves the impact of abnormal samples (including strong samples and poor samples) on the model. Not only does it reduce the impact of low-quality samples, but it also reduces the possibility of overfitting with strong samples. WIoU is defined in Eq. (4) and Eq. (5).

$$R_{WIoU} = exp\left(\frac{\left(x - x_{gt}\right)^2 + \left(y - y_{gt}\right)^2}{W_g^2 + H_g^2}\right) \tag{4}$$

$$L_{WIoU} = \frac{p}{\delta \alpha^{\beta - \alpha}} R_{WIoU} L_{IoU} \tag{5}$$

$$\beta = \frac{L_{IoU}^*}{L_{IoU}} \tag{6}$$

Among them, α and δ are learning parameters, and β is an indicator to measure the deviation degree of the prediction frame, which is defined by Eq. (6). After WIoULoss is trained, it significantly amplifies the LIoU of the normal quality anchor frame, and focuses on the distance between the center points when the prediction box coincides with the target box. Although WIoU considers the position distance of the center point of the bounding box, it ignores other key areas such as edges and vertices, and the penalty of aspect ratio matching is not enough, which can not constrain the shape of the target box well, so the Wise-Inner-MPDIoU loss function is introduced, and the calculation formula is shown in Eq. (7). The second term is the multi-point distance loss, which improves the fine-grained alignment ability by calculating the matching distance at multiple key points (such as vertex, midpoint, and center point) of the bounding box, as shown in Eq. (8). Among them, Wi is a dynamic weight generated by the

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

embedded intelligence mechanism to emphasize the importance of key points, while D is usually calculated using Euclidean distance.

$$L_{Wise-MPDIoU} = 1 - IoU + \lambda_1 \cdot D_{multi} + \lambda_2 \cdot R \tag{7}$$

In order to optimize the shape matching of the bounding box and improve the prediction accuracy, Wise-Inner-MPDIoU adds the aspect ratio loss term R, as shown in Eq. (9), and the more flexible aspect ratio measurement method effectively suppresses the problem of box morphology, especially in the small target scene, which significantly improves the positioning accuracy of the box.

$$D_{multi} = \sum_i w_i \cdot d\left(P_i(B_{gt}), P_i(B_{pred})\right) \tag{8}$$

$$R = \left(1 - \frac{\min(w_{gt}, w_{pred})}{\max(w_{gt}, w_{pred})}\right)^2 + \left(1 - \frac{\min(h_{gt}, h_{pred})}{\max(h_{gt}, h_{pred})}\right)^2 \tag{9}$$

4. EXPERIMENTS

4.1. Datasets and Metrics

In order to evaluate the performance of the proposed model in the detection of small objects, the VisDrone2019 [31] dataset was selected as the experimental dataset, and detailed comparative experiments and ablation experiments were carried out. The dataset contains 10 types of detection objects, namely people, pedestrians, bicycles, cars, vans, trucks, tricycles, awning tricycles, buses, and motorcycles. The default training set, validation set, and test set are 6471 images, 548 images, and 1610 images, respectively. The dataset is characterized by the majority of small targets and the imbalance of categories, which provides ideal testing conditions for small-scale object detection. The experimental platform is Ubuntu 20.04 focal, Python version is 3.10.14, CUDA version is 12.1, and YOLOv8s dependency library ultralytics version is 8.0.202. Table 1 lists the experimental environment and Table 2.

Table 1. Experimental Environment

configuration	name	Specific information
Hardware environment	GPU	RTX4090
	Video memory	24G
	memory	120G
Software environment	Operating system	Ubuntu
	Python	3.10.14
	Pytorch	2.2.1
	CUDA	12.1

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

Table 2. Experimental parameter settings

Settings	parameters	settings	parameters
epochs	200	Irf	0.01
imgsz	640	patience	20
batch	8	Ir0	0.01
workers	8	momentum	0.937

The size of the input image is 640×640, and the training round is 200 rounds, and the training will be terminated early when the model performance cannot be prompted after more than 20 rounds. The batch input amount of a single image is set to 8.The rest of the experimental parameters are the default settings of the system. Four evaluation indicators were used to measure the performance of the model: precision, recall, m AP50 and m AP50-95.

1) Precision, which indicates the proportion of results that are actually correct in the prediction that is correct.

2) Recall rate, which indicates that the actual result is correct, predicting the correct proportion.

3) mAP50, which refers to the accuracy that the model can achieve when the prediction confidence level reaches 0.5.

4) mAP50-95 refers to the average accuracy that the model can achieve when the prediction confidence is between 0.5 and 0.95.

Where: confidence indicates the confidence level of the model in detecting the target, which is between 0 and 1.

4.2. Comparative experiments on attention mechanisms

In order to verify the effect of the AFGCN module added at the end of the backbone network on the performance of the model, FPSconv is compared with different attention mechanisms on the basis of the improved model, and the experimental results are shown in Table 3. Compared with other attention mechanisms, AFGCN, as a lightweight channel attention mechanism, improves the performance of the model the most when the number of parameters and computations of the model hardly increases.

Table 3. Comparison of attention mechanisms with experimental results

Model	P/%	R/%	mAP0.5/%	mAP0.5:0.95/%	The number of parameters/106	Computational volume/GFlops
FSPConv	45.2	34.4	34.5	19.5	3.24	10.2
+EMA	44.8	34.6	34.2	19.4	3.25	10.4
+SE	45.0	34.5	34.1	19.3	3.24	10.2
+SimAM	44.9	34.9	34.6	19.4	3.24	10.2
+AFGCA	46.5	36.4	36.2	21.1	3.26	10.4

4.3. Improve the loss function comparison experiment

The traditional loss function CIoU and SIoU have the same basic accuracy, and the latter has a slightly higher recall rate, which indicates that SIoU may be better at detecting targets, while CloU is more suitable for tasks requiring higher positioning accuracy, and WiseIoU's mAP50-95 is higher, indicating that it performs better under stricter detection standards, but the precision and recall rate are slightly lower. The improved Inner-MPDIoU recall is close to the final result, but the precision is the lowest, and there may be missed detections2 or false

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

positives. By combining the advantages of Wise and Inner-MPDIoU and adjusting the ratio coefficient to the best effect, it can be seen that the accuracy is greatly improved while maintaining a high recall, and the mAP50-95 is significantly improved (0.201), indicating that it has stronger detection ability for multi-scale targets and achieves a more balanced performance improvement, as shown in Table 4.

Table 4. Experimental results are compared with different loss functions

Loss function	Precision	Recall	mAP50	mAP50-95
CIoU	0.447	0.339	0.334	0.187
SIoU	0.447	0.340	0.330	0.187
WiseIoU	0.441	0.336	0.332	0.191
Inner-MPDIoU	0.438	0.343	0.333	0.188
Wise-Inner-MPDIoU	0.459	0.351	0.342	0.201

4.4. Ablation experiments to improve the model

In order to verify the effectiveness of each improved method, YOLOv8n was used as the benchmark model, and ablation experiments were carried out on each improved module on the benchmark model, FSConv, CSP-Omnikernel, AFGCN attention mechanism and replacement loss function were added as WiseIoU loss function, which were denoted as A, B, C, D in turn, and the obtained network was denoted as MA, MB, MC, MD, and the experimental results are shown in Table 5.

Table 5. Improve model ablation experiments

M	A	B	C	D	The number of parameters/106/106	Computational volume/GFlops	P/%	R/%	mAP0.5/%
YOLOv8n					3.01	8.1	44.7	33.9	33.4
YOLOv8s					11.10	28.5	45.6	34.4	33.0
MA	√				3.24	10.2	45.2	34.4	34.5
MB	√	√			3.45	11.8	46.3	36.2	36.0
MC	√	√	√		3.52	11.8	47.4	37.3	37.3
MD	√	√	√	√	3.52	11.8	48.2	37.9	38.2

The first and second rows are the experimental results of the original YOLOv8n and YOLOv8s models that have not been improved, respectively. MA indicates that the accuracy, recall, and mAP50 of the original model are improved by 0.5, 0.5, and 1.1 percentage points, respectively, compared with the original model, but the multi-parameter characteristics brought by the convolution operation increase the model size and parameter quantity. MB indicates that the CSP-Omnikernel structure is introduced on the basis of MA, and the accuracy, recall and mAP50 of the improved model MB are increased by 1.6, 2.3 and 2.6 percentage points compared with the original YOLOv8n after effective feature fusion. MC represents the model with the AFGCN attention mechanism, and it can be seen that the accuracy, recall, and mAP50 are improved by 1.1, 1.1, and 1.3 percentage points respectively compared with the MC without increasing the number of parameters and the amount of calculation, which significantly improves the performance of the model. MD said that the model with all the improved schemes added, the improved model slightly increased the number of parameters on the basis of YOLOv8n, and the mAP increased by nearly 5%, compared with the deeper model YOLOv8s, the accuracy, recall and mAP50 were greatly improved by 2.6, 3.5 and 5.2 respectively when the number of

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

parameters was less than 30% and the amount of computation was only 41%. This series of optimization measures not only significantly improves the detection accuracy of the model, but also effectively reduces the complexity of the model, making the model more suitable for resource-constrained UAV target detection scenarios.

4.5. FAS-YOLO Comparative Experiment

In order to verify the superiority of the improved YOLOv8n model, some representative YOLO models were selected for comparative experiments, and the network depth and width of the experimental models were unified. Specifically, the network depth and width of all tested models are consistent with YOLOv8n, where the version with the model suffix "n" is one-third the depth and one-quarter width of the original setting. The results of the experiment are shown in Table 6.

In the comparative experiment, the performance of YOLOv3n and YOLOv5n is similar, and both are weaker than YOLOv8n. YOLOv6n is not optimized for small object detection and has the worst performance. Although YOLOv8n is superior to YOLOv5n in terms of model size and number of parameters, it performs better, although the advantage is not significant. The ASF-YOLO [32] model for small-target cell segmentation performs weaker than YOLOv5n in small-target detection tasks because its optimization direction focuses more on target segmentation rather than object detection. The latest YOLOv10 is a significant improvement in lightweighting, with nearly a quarter of the model size and number of parameters of YOLOv8s, but still inferior in accuracy and recall to YOLOv8. The improved FAS-YOLO model exhibits significant performance advantages while remaining lightweight. Although YOLOv8s has a higher network width, a model size of 24.1 MB, and a much higher number of parameters than FAS-YOLO, its precision, recall, mAP50, and mAP50-95 are 2.6, 3.5, 5.2, and 3.5 percentage points lower than FAS-YOLO, respectively. This fully demonstrates the performance advantages of the improved model FAS-YOLO.

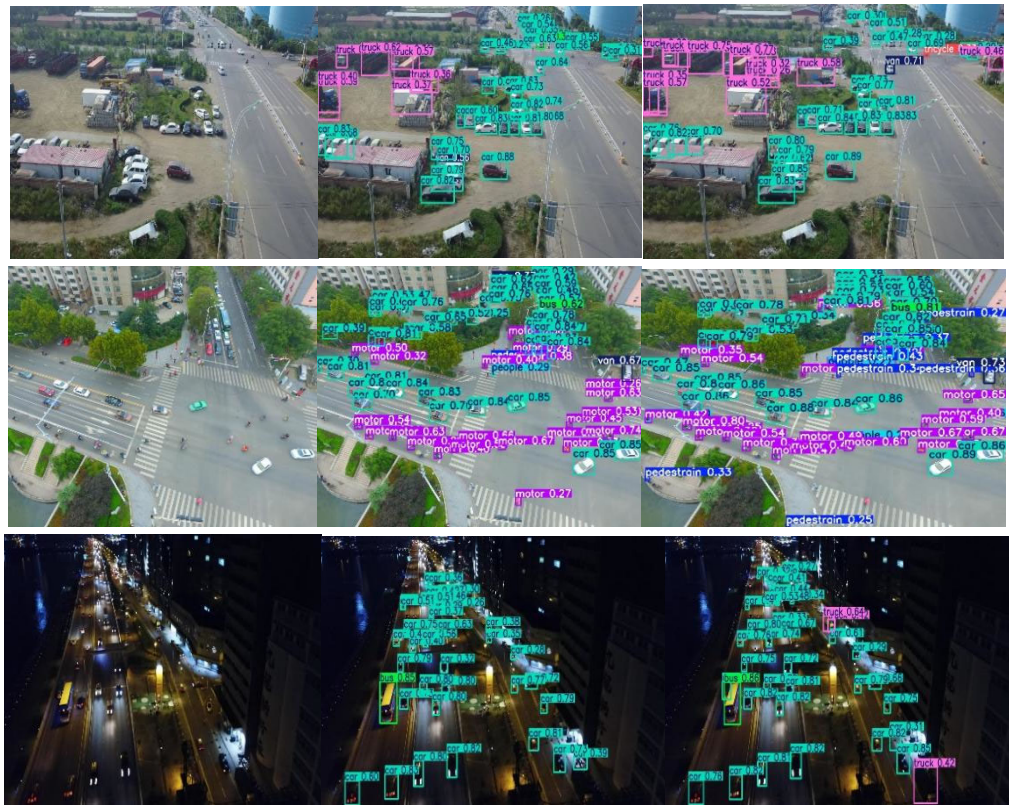
Table 6. Improve model comparison experiments

Model	Model size/mb	The number of parameters/106	Precision	Recall	mAP50	mAP50-95
YOLOv3n	7.95	4.06	0.447	0.333	0.329	0.189
YOLOv5n	5.03	2.50	0.442	0.337	0.332	0.181
YOLOv6n	8.29	4.23	0.404	0.312	0.294	0.167
YOLOv8n	7.5	3.01	0.447	0.339	0.334	0.187
ASF-YOLO	5.08	2.52	0.436	0.341	0.33	0.186
YOLOv8s	22.9	11.10	0.456	0.344	0.330	0.186
YOLOv10n	5.8	2.71	0.434	0.336	0.333	0.191
FAS-YOLOv8n	7.3	3.52	0.482	0.379	0.382	0.221

4.6. Visual comparison of inspection results

By carefully selecting image samples, we performed object detection experiments using the baseline model and the FAS-YOLO model to compare and analyze the detection performance of the two. As shown in Figure 6, FAS-YOLO exhibits higher detection confidence in a variety of scenarios and complex conditions. Specifically, it generates higher confidence scores for the target bounding box, and these scores are highly consistent with the actual target. At the same time, FAS-YOLO has made significant progress in reducing the rate of false detections and missed detections, which can accurately identify the vast majority of targets and effectively avoid misidentification of non-target areas. Even under the conditions of insufficient lighting or large shadow interference, the model still maintains a low missed detection rate, showing excellent robustness and detection performance.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001



(a) ground truth (b) YOLOv8n (c) FAS-YOLOv8n

Figure 8. Visual analytics

5. CONCLUSION

The detection of small targets from the perspective of UAVs usually faces the problems of insufficient accuracy, high missed detection rate and obvious false detection rate. To solve this problem, this paper proposes an improved small object detection model based on YOLOv8, named FAS-YOLO (Feature Attention Small Object Detection-YOLO). The model innovatively uses the FPSharedConv module to replace the traditional pooling method, which effectively retains the detailed information of small targets and enhances the feature extraction ability. Based on PAFPN, the SmallObjectEnhancePyramid feature pyramid structure was designed, and the feature expression of small objects was further strengthened without increasing the P2 detection layer by fusing SPDConv convolution and CSP-OmniKernel. At the same time, the AFGCAttention attention mechanism was introduced to significantly improve the degree of attention to key targets. In addition, the improved Wise-Inner-MPDIoU loss function significantly improves the positioning accuracy of detection by optimizing the aspect ratio measurement.

The experimental results show that FAS-YOLO performs well on the VisDrone2019 dataset and shows strong performance in the field of small target detection, providing an effective solution for efficient target detection in UAV scenarios. Future work will focus on improving the generalization capabilities of the model and optimizing operational efficiency, with a view to deploying it in a wider range of real-world applications.

REFERENCES

[1] SUN S , MO B , XU J ,et al.Multi-YOLOv8: An infrared moving small object detection model based on YOLOv8 for air vehicle[J]. Neurocomputing, 2024, 588. DOI:10. 1016/j.neucom.2024.127685.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

[2] Suthaharan, S. & Suthaharan, S. Support vector machine. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning 207–235 (2016).

[3] LIN T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Shenzhen, China, Piscataway: IEEE, 2017: 29802988.

[4] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector [C]// Proceedings of the 2016 European Conference on Computer Vision. Amsterdam, Netherlands, Springer, 2016: 21-37.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. LasVegas, NV, USA, Piscataway: IEEE, 2016: 779-788.

[6] REDMON J, FARHADI A. YOLO9000: better, faster, stronger [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Hawaii Convention Center, USA, Piscataway: IEEE, 2017: 7263- 7271.

[7] REDMON J, FARHADI A. Yolov3: An incremental improvement [EB/OL]. [2024-07-01]. https://arxiv.org/abs/1804.02767.

[8] SHEN X L, WANG L C. UAV Aerial Photography Target Detection Based on YOLOv8n [J]. Computer Systems Applications, 2024, 33(7): 139-148.

[9] PAN W, WEI C, QIAN C Y, et al. Improved YOLOv8s Model for Small Object Detection from Perspective of Drones [J]. Computer Engineering and Applications, 2024, 60(9): 142-150.

[10] Xiuman Liang, Zihan Jia, Zhendong Liu, et al. A UAV Aerial Image Detection Algorithm Based on Improved YOLOv8n. Electronics Optics & Control. [2024-09-25]. http://kns.cnki.net/kcms/detail/41.1227.TN.20240914.1601.002.html.

[11] LEI B J,YU A,YU K. Small Object Detection Algorithm based on Improved YOLOv8s[J]. Radio Engineering, 2024, 54(04):857-870.s

[12] WANG G, CHEN Y, AN P, et al. UAV-YOLOv8: A smallobject-detection model based on improved YOLOv8 for UAV aerial photography scenarios[J]. Sensors, 2023, 23(16): 7190.

[13] ZHU L, WANG X, KE Z, et al. Biformer: Vision transformer with bi-level routing attention[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 10323-10333.

[14] LI Y, FAN Q, HUANG H, et al. A Modified YOLOv8 Detection Network for UAV Aerial Image Recognition[J]. Drones, 2023, 7(5): 304.

[15] Bochkovskiy A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection [EB/OL]. [2024-07-01]. . https://arxiv.org/pdf/2004.10934.

[16] GE Z, LIU S, WANG F, et al. Yolox: Exceeding yolo series in 2021 [EB/OL]. [2024-07-01]. https://arxiv.org/abs/2107.08430.

[17] JIANG B, LOU R, MAO J, et al. Acquisition of localization confidence for accurate object detection [C]// Proceedings of the 2018 European conference on computer vision (ECCV). Munich, Germany, Springer, 2018: 784-799.

[18] H. Sun, Y. Wen, H. Feng, Y. Zheng, Q. Mei, D. Ren, et al., "Unsupervised bidirectional contrastive reconstruction and adaptive fine-grained channel attention networks for image dehazing", Neural Netw., vol. 176, Aug. 2024.

[19] Cui, Y., Ren, W., & Knoll, A. (2024). Omni-Kernel Network for Image Restoration. Proceedings of the AAAI Conference on Artificial Intelligence, 38(2), 1426-1434. https://doi.org/10.1609/aaai.v38i2.27907

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202510_11(10).0001

[20] LIN T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection [C]// Proceedings of the 2017 IEEE conference on computer vision and pattern recognition. Hawaii Convention Center, USA, Piscataway: IEEE, 2017: 2117-2125.

[21] Tong, Z., Chen, Y., Xu, Z., Yu, R.: Wise-IoU: bounding box regression loss with dynamic focusing mechanism. arXiv:230110051 (2023).

[22] Wenhan Huang, Guodong Yin, Keke Geng, te al. Object detection in complex driving scene based on dilated convolution feature adaptive fusion. Journal of Southeast University (Natural Science Edition), 2021,51(6):1076-1083. DOI:10.3969/j.issn.1001-0505.2021.06.021.

[23] V. Chalavadi, et al.mSODANet: a network for multi-scale object detection in aerial images using hierarchical dilated convolutions[J].Pattern Recognit, 126 (2022), Article 108548.

[24] Yang, Z., Zheng, Y., Shao, J. et al. Improved YOLOv4 based on dilated coordinate attention for object detection. Multimed Tools Appl 83, 56261–56273 (2024). https://doi.org/10.1007/s11042-023-17817-1.

[25] Guodong Zhang, Zhihua Chen, Bin Sheng Infrared Small Target Detection Based on Dilated Convolutional Conditional Generative Adversarial Networks. Computer Science, 2024, 51(2):151-160.DOI:10.11896/jsjkx.221200045.

[26] Guo, MH., Xu, TX., Liu, JJ. et al. Attention mechanisms in computer vision: A survey. Comp. Visual Media 8, 331–368 (2022). https://doi.org/10.1007/s41095-022-0271-y.

[27] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.

[28] Ji, CL., Yu, T., Gao, P. et al. Yolo-tla: An Efficient and Lightweight Small Object Detection Model based on YOLOv5. J Real-Time Image Proc 21, 141 (2024). https://doi.org/10.1007/s11554-024-01519-4.

[29] Zheng, Z., Wang, P., Ren, D., Liu, W., Ye, R., Hu, Q., et al.: Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Trans. Cybern. 52(8), 8574–8586 (2021).

[30] Sun, H., Luo, Z., Ren, D., Hu, W., Du, B., Yang, W., et al.: Partial siamese with multiscale bi-codec networks for remote sensing image haze removal. IEEE Trans. Geosci. Remote Sens. (2023).

[31] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: editors. Distance-IoU loss: faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12993–13000. AAAI Press, Washington, D.C. (2020).

[32] KKANG M, TING C M, TING F F, et al. ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation [EB/OL]. [2024-07-01]. https://arxiv.org/pdf/2312.06458.