International Journal of Computational Methods and Experimental Measurements
https://www.acadlore.com/journals/IJCMEM

Adaptive Multi-Scale Gated Convolution and Context-Aware Attention Network for Accurate Small Object Detection

Jia-Chi Wang[1,2], Min-Po Jung[1*], Shoulin Yin[2*], Hang Li[2*]

[1] Department of Computer and Information Engineering, Graduate School Youngsan University, 612-022 Busan, South Korea
[2] College of Artificial Intelligence, Shenyang Normal University, 110034 Shenyang, China

[*] Correspondence: Min-Po Jung (minpo@ysu.ac.kr); Shoulin Yin (yslin@synu.edu.cn); Hang Li (lihangsoft@163.com)

Received: 08-07-2025 **Revised:** 09-18-2025 **Accepted:** 09-24-2025

Citation: J.-C. Wang, M.-P. Jung, S. L. Yin, and H. Li, "Adaptive multi-scale gated convolution and context-aware attention network for accurate small object detection," *Int. J. Comput. Methods Exp. Meas.*, vol. 13, no. 3, pp. 576–587, 2025. https://doi.org/10.56578/ijcmem130308.

Abstract: Small object detection in complex scenes remains a challenging task due to background clutter, scale variation, and weak feature representation. Conventional deep learning–based detectors are prone to false positives and missed detections when dealing with dense or low-contrast objects. To address these limitations, this paper proposes an Adaptive Multi-Scale Gated Convolution and Context-Aware Attention Network (AGCAN) designed to enhance small object detection accuracy under complex visual conditions. The model introduces an improved Multi-Scale Gated Convolution Module (MGCM) to replace standard U-Net convolutional blocks, enabling comprehensive extraction of fine-grained object features across multiple scales. A Multi-Information Fusion Enhancement Module (MFEM) is incorporated at skip connections by integrating improved dilated convolution and hybrid residual window attention to minimize information loss and optimize cross-layer feature fusion. Furthermore, the Distance-IoU (DIoU) loss replaces the conventional Smooth L1 loss to accelerate model convergence and improve localization precision. Contextual cues are adaptively integrated into region-of-interest classification to strengthen small-object discrimination. Experimental evaluations on the DIOR and NWPU VHR-10 datasets demonstrate that the proposed network achieves superior performance compared with state-of-the-art methods, effectively reducing false detections and improving robustness in complex environments.

Keywords: Deep learning; Small object detection; Adaptive multi-scale gated convolution; Context-aware attention model

1 Introduction

In recent years, with the acceleration of urbanization, the demand for small target detection algorithms in complex scenarios has been increasing day by day. Especially in tasks such as autonomous driving [1, 2], traffic monitoring [3], and unmanned aerial vehicle detection [4, 5], its importance has become increasingly prominent [6]. In the field of optical remote sensing images, Ahmed et al. [7] systematically analyzed the advantages and disadvantages of traditional and deep learning methods, he pointed out the potential of the Transformer algorithm in the fine-grained recognition of ships. The researches show that deep learning methods significantly improve the detection accuracy in complex backgrounds through end-to-end feature learning, but the characterization of small target features and background interference remain core challenges [8, 9]. Factors such as noisy background interference, illumination changes, and occlusion between objects seriously affect the accuracy of small target detection. Therefore, achieving efficient detection of small targets in complex environments has important theoretical and application value [10, 11].

At present, the target detection algorithms based on convolutional neural networks (CNN) are mainly classified into two types: single-stage algorithms and two-stage algorithms. Initially, methods such as R-CNN [12], Fast RCNN [13], Faster R-CNN [14], Mask R-CNN [15], Cascade RCNN [16], etc., which belong to the two-stage series, led the development of target detection technology. The characteristic of the two-stage algorithms is to first generate candidate boxes and then identify the objects within the boxes. However, these methods, due to the need to generate a large number of redundant bounding boxes and the resulting excessive computational load, bring problems such as slow running speed and reduced detection efficiency, making it difficult to meet the real-time requirements brought

about by the rapid development of current technology. Single-stage algorithms, such as RetinaNet, GFL, TOOD, and the YOLO series, enable the model to directly complete end-to-end object detection, reducing the number of calculation steps, addressing the shortcomings of the two-stage algorithms in terms of real-time performance, achieving effective real-time detection of targets, and achieving a more ideal balance between accuracy and speed. It is worth noting that recent researches have made significant progress in lightweight network. Li et al. [17] proposed an optimization scheme based on the attention mechanism for the YOLOv4-Tiny algorithm, providing an effective solution to the problem of balancing accuracy and speed in lightweight models under complex scenarios. Yin and Ding [18] proposed a lightweight model YOLOv8-VSC for the detection of surface defects on steel strips, providing a new approach for small target detection in industrial scenarios. However, both single-stage and two-stage algorithms rely on the Non-Maximum Suppression (NMS) technique to eliminate redundant bounding boxes, which to some extent affects the inference speed and robustness of the model.

To address the issues of missdetection and false detection in multi-scale image small target detection, Zheng et al. [19] introduced the DIoU (Distance Intersection over Union) loss function to make the prediction boxes of YOLOv3 more accurate. Huang et al. [20] incorporated deformable convolution into RetinaNet to enable the algorithm to adaptively adjust the receptive field, thereby improving the detection accuracy in complex environmental backgrounds and when objects are small. Wang and Wang [21] introduced the self-attention mechanism into the YOLOv5 algorithm to address the issue of dense distribution of small targets. Li and Wu [22] introduced frequency channel attention into the YOLOv5 algorithm, thereby improving the detection performance of the algorithm for small targets. Wang et al. [23] introduced the convolutional attention module into the YOLOv5 algorithm, aiming to enhance the target features to improve the detection accuracy of small targets. Su et al. [24] addressed the issue of insufficient pixel and feature information for small targets in images, proposing an algorithm that integrated the attention mechanism, local receptive field, and refined feature pyramid to enhance the contextual information and feature expression ability of effective features, thereby improving the detection effect of small targets. For the problem of missdetection and false detection of multi-scale targets in images, Wang et al. [25] proposed a feature fusion FMSSD (Feature-Merged Single-Shot Detection) algorithm based on the SSD algorithm. By adopting the dilated spatial feature pyramid module, regional weighted cost function, and optimized loss calculation method, etc., it enhanced the detection accuracy of multi-scale targets. Baig et al. [26] addressed the issues of uneven target distribution and drastic size variations, and proposed a rotation target detection algorithm that integrated convolutional channel attention. Bian et al. [27] addressed the problem of poor multi-scale object detection and proposed a cross-scale connection operation to enhance the feature extraction capability of the algorithm. To address the above problems, this paper proposes a novel adaptive multi-scale gated convolution and context-aware attention model for small object detection. The main contributions are as follows.

(1) By replacing the traditional convolution blocks with the multi-scale gate convolution module (MGCM), this approach enables the model to have a larger receptive field through multi-scale feature extraction, which can adapt to lesion regions of different sizes and shapes. At the same time, it combines spatial and channel attention as well as an improved attention gate mechanism to redistribute the weights of spatial and channel features, enhancing the response of lesion features to distinguish the target from the background regions, thereby more accurately locating and segmenting the target regions.

(2) To fully utilize the rich image feature information of the U-shaped network, a multi-information fusion enhancement module (MFEM) is introduced at the jump connection. By improving the dense dilated convolution, the abundant semantic information at the bottom of the U-shaped network is fully extracted, and a mixed-enhanced residual window attention is constructed to capture the local and global features of the current layer of the U-shaped network.

(3) We design a local context enhancement function (LCE), which uses parameter fine-tuning to better focus on the context information required for different categories and scales of small targets, enhance the background information, and reduce false detection.

2 Proposed Method for Small Object Detection

The traditional object segmentation models often suffer from insufficient segmentation accuracy and are prone to be affected by the background region. In this paper, a object segmentation model that integrates multi-scale gated convolution and window attention is proposed. The overall architecture of the network model is shown in Figure 1. The proposed model replaces the traditional convolution blocks with MGCM to fully extract the feature information of objects, and changes the number of channels of the input image to 32. Secondly, the maximum pooling [28, 29] is used for downsampling operation to reduce the information loss at the skip connection and to utilize the rich image semantic information contained in the data at the bottom of the model. A MFEM is introduced at the skip connection to mine the feature information of the current layer of the U-shaped network and to enhance the features using the data at the bottom of the model. Finally, the result of the last decoding layer is input into a 1×1 convolution with a Sigmoid activation function to obtain the final predicted object segmentation image, thereby achieving precise

detection of the object.

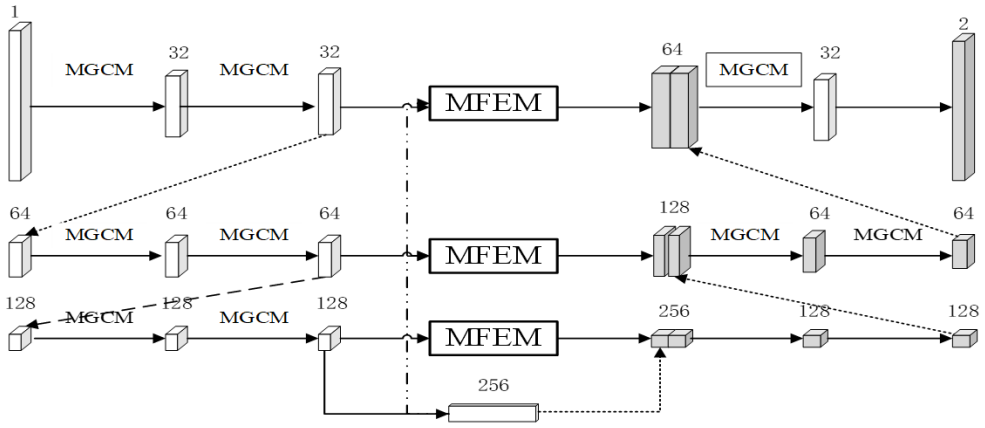

Figure 1. Proposed model

2.1 Multi-Scale Gated Convolution Module (MGCM)

The traditional convolutional blocks have difficulty in specifically training the target pixels during the training process, which may lead to challenges when adapting to targets of different sizes, shapes, and with lower contrast. Inspired by reference [30], this paper designs the MGCM to address these issues. It mainly utilizes the multi-scale gated squeeze-excitation module (MGSM) to achieve more accurate recognition and modeling of targets with different sizes and shapes. At the same time, the double residual attention module (DRAM) and the double attention gate module (DAGM) are adopted to enhance the response of the target features, further distinguishing the target from the surrounding background areas. Compared with traditional networks that capture single-scale features using a single branch, MGSM adopts a parallel branch design consisting of four convolutional blocks of different sizes. At the final stage of the four branches, it references the compressed activation blocks and the attention gate mechanism to design a gate squeeze-excitation block (Gate-SE Block). Multiscale processing can not only increase the receptive field but also capture feature information at different scales, thereby better identifying and covering various shapes, sizes and positions of the lesion areas in the feature map. The compressed activation block dynamically adjusts the weights of the feature maps by learning the correlation dependencies between the channels, so that more important features receive larger weights, which helps to enhance the expressive power of the feature representation. The attention gate can automatically learn and focus on the relevant parts, thereby enhancing the attention to important information. Therefore, in this paper, combining the compression incentive block with the attention gate can enable the network to no longer simply integrate all channel information, but to flexibly emphasize the most significant and important features in the current task and suppress the unimportant ones. At the same time, adding the Gate-SE Block in multiple scales can selectively enhance or weaken the features at specific scales. The network is more likely to learn the universal features and adapt to different sizes and shapes, avoiding the occurrence of overfitting.

In the MGSM module, it is assumed that the input feature map is $A \in R^{C\times H\times W}$, where C represents the channel number of the input image. H and W denote the pixel size of the height and width of the input image respectively. First, the input A is sent to the 4 paths for different-sized convolution operations to generate the feature map $\widehat{A}$ for each path. Then, A and $\widehat{A}$ are input into the Gate-SE block. In the Gate-SE Block module, $\widehat{A}$ is processed through the SE Block and 1×1 convolution, and then a matrix addition and non-linear activation (ReLU) operation is performed with the 1×1 convolution processed A to generate the attention coefficient. Then, the attention coefficient is passed through the SE Block and combined with the attention coefficient that undergoes 1×1 convolution and Sigmoid operation, making the aligned weights in the attention coefficient larger. A is connected with the generated attention coefficient for output. Finally, the feature maps output by the 4 paths through the Gate-SE block are concatenated, and then a 1×1 convolution is used to change the channel dimension to generate a new feature map $B \in R^{C\times H\times W}$.

2.2 Multi-Information Fusion Enhancement Module

Although replacing the traditional convolutional blocks with MGCM can alleviate the problem of gradient vanishing and fully extract the target feature information, during the encoding and decoding process, direct skip connections will inevitably introduce a certain degree of noise interference, thereby causing the phenomenon of missed segmentation. This paper takes into account that the local and global feature information of the current layer in the U-Net model have not been effectively utilized, and the bottom layer integrates spatially and contains

a considerable amount of image semantic information. In order to reduce the phenomenon of missed points, this paper, based on the dense hole convolution block and the Swin Transformer [31], we design the MFEM.

For the detection of small targets at multiple scales, it is also necessary to pay more attention to the contextual information of small targets at different scales. Because for small targets, it is difficult to accurately identify them merely by their appearance. On the contrary, if we want to successfully identify these small targets, it often relies on background information. The surrounding environment can provide valuable clues about the shape, direction and other features of the targets. In the left of Figure 2, due to the appearance characteristics of the target object within the red box, the algorithm might mistakenly identify it as a runway or a basketball court, but in fact, this target is not the object that needs to be detected. In the right side of Figure 2, the true label of the object is an airplane. However, due to the presence of other objects in the yellow box, the algorithm may miss detecting this airplane. The occurrence of these erroneous detections is mainly because the algorithm only considers the limited contextual information near the object. Therefore, a context enhancement function that can better focus on the contextual information of different categories and scales of small objects is needed.

Figure 2. Contextual information diagrams required for different types of targets

In the fine-grained query-aware sparse attention module, the final fine-grained attention employs the LCE function, which can enhance the local context information. The calculation formula of the LCE function can be expressed as shown in Eq. (1).

$$\mathrm{LCE} = \mathrm{DWConv}(\mathrm{V}) \tag{1}$$

Here, V represents the value vector. DWConv denotes the depthwise convolution operation, which is capable of retaining more fine-grained and shallow details. The size of the convolution kernel in this operation is set to 5.

In order to effectively utilize the local and global features of the current layer of the U-shaped network, a better feature extraction is achieved by introducing the Swin Transformer Block (ST Block) and combining it with the channel attention (CAB) and spatial attention (SAB) of the DRAM module. Firstly, the ST Block's conventional window multi-head self-attention (W-MSA) is utilized to divide the input features into local windows, and calculate the self-attention within each small window. Then, the channel attention weights of the input features are calculated through CAB, and the two are added together to enhance the important features of each small window and suppress the non-important ones, thereby capturing the feature information of the local region. However, due to the division of the windows, the information interaction among different windows is impossible. Therefore, the Shifted Window Multi-Head Self-Attention (SW-MSA) is utilized to enable the feature information to be transmitted between adjacent windows. At the same time, SAB is added to retain the feature information in the original space, further enhancing the feature representation of the key areas. Finally, it is connected with the input features of the current layer through a residual connection, thereby better capturing the global structure and local details in the image. The mathematical definition is as follows:

$$\widehat{z}^{l} = \mathrm{W}-\mathrm{MSA}\left(\mathrm{LN}\left(z^{l-1}\right)\right) + \mathrm{CAB}\left(z^{l-1}\right) + z^{l-1} \tag{2}$$

$$z^{l} = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^{l}\right)\right) + \hat{z}^{l} \tag{3}$$

$$\hat{z}^{l+1} = \mathrm{SW}-\mathrm{MSA}\left(\mathrm{LN}\left(z^{l}\right)\right) + \mathrm{SAB}\left(z^{l}\right) + z^{l} \tag{4}$$

$$z^{l+1} = \mathrm{MLP}\left(\mathrm{LN}\left(\hat{z}^{l+1}\right)\right) + \widehat{z}^{l+1} \tag{5}$$

$$x_i = z^{l+1} + z^{l-1} \tag{6}$$

where, $\hat{z}^1$ and z^1 represent the output features of the SW-MSA module and the MLP module respectively. W-MSA and SW-MSA represent the use of conventional and shifted window multi-head self-attention respectively. LN indicates the layer normalization operation. CAB and SAB represent channel attention and spatial attention respectively. z^{l-1} represents the input image at the jump connection. x_i represents the output result after passing through the hybrid-enhanced residual window attention module (HERWM).

Finally, in order to integrate the rich semantic information of the bottom data with the feature information of the current layer, as shown in Figure 3, L is processed through 1×1 convolution, average pooling, and Softmax activation function to generate a spatial weight map. This spatial weight map is then multiplied with the output result of HERWM, and finally added to J to output the feature map $M \in R^{C \times H \times W}$.

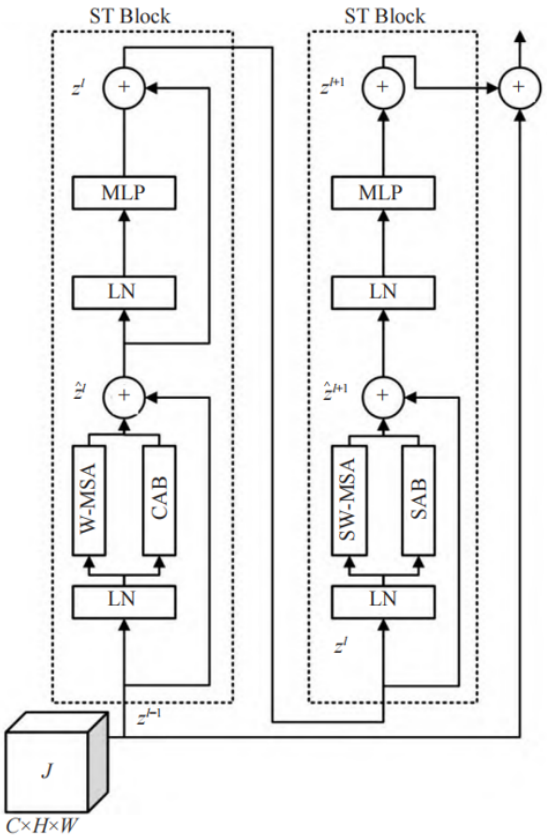

Figure 3. HERWM module

2.3 Loss Function

The original network uses SmoothL1 as the loss function. When multiple detection boxes have the same SmoothL1 loss value, there may be significant differences in their Intersection-over-Union (IoU) with the real boxes. The actual evaluation of the regression indicators of the detection boxes uses IoU. At this time, calculating the SmoothL1 loss value cannot reflect the degree of overlap between the detection box and the real box.

This paper uses the DIoU loss, which is an extension of the IoU loss by adding a penalty term. It minimizes and normalizes the distance between the centers of the two bounding boxes, thereby accelerating the convergence process. The R_{DIoU} is defined as follows:

$$R_{DIoU} = \frac{\rho^2\left(b,\ b^{gt}\right)}{l^2} \tag{7}$$

Here, b and b^{gt} represent the center points of the predicted box and the true box respectively. $\rho^2\left(b, b^{gt}\right)$ represents the Euclidean distance between the two points. l represents the diagonal length of the minimum bounding box that contains the two boxes. Therefore, the loss function of L_{DIoU} can be defined as:

$$L_{DIoU} = 1 - IoU + R_{DIoU} \tag{8}$$

The schematic diagram of L_{DIoU} is shown in Figure 4. Here, $d = \rho\left(b, b^{gt}\right)$ represents the distance between the centers of the two bounding boxes.

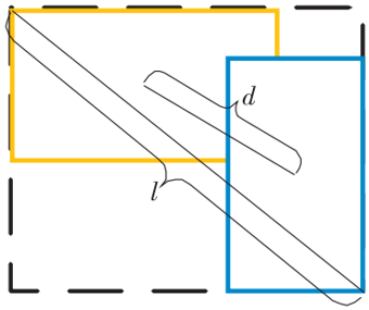

Figure 4. L_{DIoU} calculation principle

Compared to SmoothL1, L_{DIoU} can better minimize the distance between the predicted box and the real box. When there are cases such as containment, perpendicularity, and parallelism between two bounding boxes, L_{DIoU} enables the predicted box to regress more quickly.

3 Results and Discussion

3.1 Dataset

This experiment utilizes four publicly available small object detection datasets: VisDrone2019, KITTI, TinyPerson, and DOTAv1.0. The VisDrone2019 dataset is a large-scale dataset collected by the AISKYEYE team from Tianjin University. It is shot in 14 different cities across China, under various scenarios and with different lighting conditions, covering different environments such as urban and rural areas. The data is highly diverse and complex. It includes ten categories such as pedestrians, cars, trucks, buses, tricycles, bicycles, and motorcycles. The size distribution of the targets shows a distinct small target characteristic, with most targets having small pixel sizes, mainly concentrated within the range of 1501 pixels in size. The training set uses 6471 images, the validation set has 548 images, and the test set employs 1610 images.

The KITTI dataset [32] is jointly created by the Karlsruhe Institute of Technology in Germany and the Toyota Technical Institute in the United States. It is sourced from urban environments and includes scenes with busy traffic and overlapping targets, covering driving data under various weather and urban conditions. The target categories include cars, trucks, pedestrians, cyclists, etc., with most targets being small-sized targets with dimensions concentrated within the range of 30 to 150 pixels. The dataset is divided into training images (5236), validation images (748), and test images (1497).

The TinyPerson dataset [33] is a benchmark dataset specifically focused on small object detection. It contains 1610 labeled images, among which there are 1261 training images, 158 validation images, and 158 test images.

The images in the DOTAv1.0 dataset are sourced from aerial images collected by various sensors and platforms, covering a wide range of real-world scenarios, thus making the dataset comprehensive.

3.2 Experimental Environment

The hardware experiments in this paper are all conducted on the Ubuntu 20.04 operating system, using an RTX 3090 GPU with 24GB of memory and a 14-core Intel(R) Xeon(R) Platinum 8362 CPU operating at 2.80GHz. The experimental environment employs Python 3.8.19 and Pytorch 1.11.0 with CUDA version 11.3. To ensure the fairness of the experiments, the initial learning rate is uniformly set to 0.0001 during the training stage, using the AdamW optimizer with a weight decay value of 0.0001. The batch size and the number of worker processes are both set to 4. Additionally, all experiments do not use pre-trained weights. Depending on the characteristics of different datasets, the number of training epochs is different.

3.3 Evaluation Index

In order to conduct a comprehensive and in-depth assessment of the detection capabilities and performance of the model, this paper carefully selects a series of scientific and effective evaluation indicators. Mean Average Precision (mAP), including mAP@0.5 and mAP@0.5:0.95, is used to evaluate the accuracy of the model in detecting and classifying targets. Precision (P) is used to measure the proportion of positive samples that the model predicts as

positive and are actually positive. Recall (R) reflects the proportion of positive samples that are actually positive and are correctly predicted as positive. Parameters (Params) are a key indicator for measuring the complexity of the model. Computational cost (GFLOPs) can effectively evaluate the computational complexity during the model's operation. The following are the calculation formulas for Recall, Precision, and Average Precision.

$$\mathrm{mAP} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{AP_i} \tag{9}$$

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{10}$$

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{11}$$

TP (True Positive) refers to the number of true positive examples, which is the quantity of positive samples accurately predicted by the model. FP (False Positive) refers to the number of false positive examples, which is the quantity of negative samples misjudged by the model as positive. FN (False Negative) refers to the number of false negative examples, representing the quantity of positive samples misjudged by the model as negative. If T is the time (in seconds) required to process one image, then the FPS calculation formula is:

$$\mathrm{FPS} = 1/\mathrm{T} \tag{12}$$

3.4 Ablation Experiment

This paper employs multiple nodules to improve the proposed model. To verify the effectiveness of the proposed modules, ablation experiments are conducted on the public VisDrone2019 small target dataset. The specific data is shown in Table 1. The experimental results show that the MGCM module demonstrates significant advantages. Compared to the baseline model, it has improved by 1.1% in the mAP@0.5, and by 0.5% in the mAP@0.5:0.95. The parameter count and computational cost have been reduced by 27.1% and 13.2% respectively. The MFEM module has improved by 0.1% in mAP@0.5 and mAP@0.5:0.95, and by 0.7% and 0.8% respectively in mAP@0.5:0.95 compared to the baseline model. LCE has improved by 0.7% and 0.8% in mAP@0.5 and mAP@0.5:0.95 compared to the baseline model.

Table 1. Ablation experiment results on the VisDrone2019 dataset

No.	MGCM	MFEM	LCE	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95/%	GFLOPs	Params	FPS
1	×	×	×	62.3	46.7	48.0	29.4	57.0	19.9	71.3
2	√	×	×	61.7	47.9	49.1	29.9	49.5	14.5	51.3
3	×	√	×	62.2	46.4	48.1	29.7	57.1	20.0	74.5
4	×	×	√	62.2	46.9	48.7	30.2	57.0	19.9	71.7
5	√	√	×	62.6	48.3	49.6	30.5	49.6	14.6	52.2
6	√	×	√	62.4	48.5	50.0	30.9	49.5	14.5	53.2
7	×	√	√	62.0	47.3	49.0	30.4	57.1	20.0	76.2
8	√	√	√	64.3	48.8	50.8	31.7	49.6	14.6	54.6

Then we conduct an ablation experiment on the loss functions. To fully verify the effectiveness of the $\mathrm{L_{DIoU}}$ loss function, an ablation experiment is conducted by comparing it with GIoU and other commonly used IoU (such as CIoU, DloU, InnerloU, MPDIoU) on the VisDrone2019 dataset. The experimental results shown in Table 2 indicate that the mAP@50 and mAP@50:95 of the $\mathrm{L_{DIoU}}$ loss function have reached 48.7% and 30.2% respectively, which are significantly higher than those of the traditional IoU. This loss function has greatly improved the accuracy of boundary positioning. This refined optimization mechanism is particularly suitable for small object detection. Since the proportion of pixels for small objects is low, the impact of boundary offset on IoU is more sensitive. $\mathrm{L_{DIoU}}$ achieves dual enhancements of sensitivity to the boundaries of small targets and consistency of internal features through the multi-point distance penalty of MPD-IoU and the internal region constraint of Inner-IoU. This forces the model to fit the target boundaries more strictly while suppressing background noise. This effect fully validates the superiority and effectiveness of the loss function of the improved model.

Table 2. IoU ablation experiment results

IoU	GIoU	CIoU	DIoU	InnerIoU	MPDIoU	L_{DIoU}
mAP@ 50	48	47.9	47.6	47.6	47.5	48.7
mAP@ 50: 95	29.4	29.5	29.3	29.2	28.8	30.2

Table 3. Test-dev results

Index	Average	SD	95% CI (t-distribution, df = 2)
mAP@0.5	0.507	0.0018	[50.2%, 51.2%]
mAP@0.5: 0.95	0.316	0.0021	[31.1%, 32.1%]

This paper calculated the results of the VisDrone2019 test-dev for three independent experiments as shown in Table 3. Here SD = standard deviation.

Taking VisDrone2019 as an example, this paper conducted a Welch's t-test between the 3-run results and DEIM, RT-DETR, and YOLOv11 as shown in Table 4.

Table 4. Confidence interval comparison with the mainstream models

Model	ΔmAP@0.5/%	95% CI of Δ	*p* Value
DEIM	0.3	$[-0.1, +0.7]$	0.063
RT-DETR	2.8	$[+2.4, +3.2]$	< 0.001
YOLOv11	6.8	$[+6.3, +7.3]$	< 0.001
Proposed	8.8	$[+7.9, +8.5]$	< 0.001

Figure 5 shows the testing error. From the above results, we can know that the statistical results of this method are also superior to those of other methods.

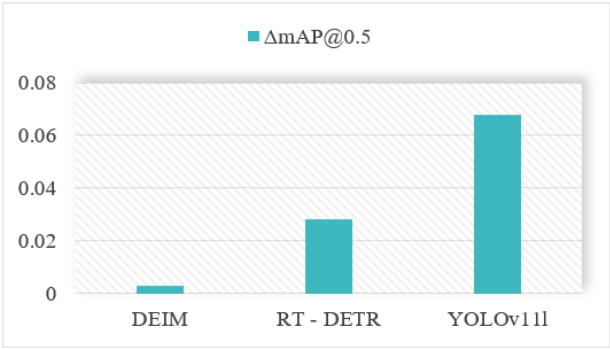

Figure 5. Testing error

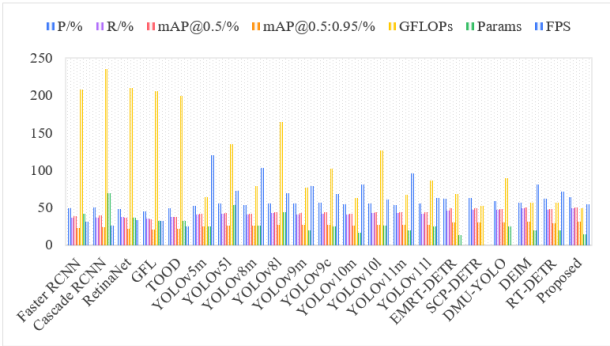

Figure 6. The visualized results of Table 5

3.5 Comparison Experiments

To further verify the superiority of the proposed model in the detection of small targets, we select the VisDrone2019 and KITTI datasets, and compare our algorithm with numerous mainstream baseline algorithms as well as other advanced algorithms (EMRT-DETR [34], SCP-DETR [35], DMU-YOLO [36], DEIM [37], RT-DETR [38]). Results are shown in Table 5, Table 6, Figure 6 and Figure 7.

To conduct a thorough analysis of the detection performance with the proposed model, the small targets detection results obtained using the proposed model in this paper are presented, as shown in Figure 8. As can be seen from the figure, vehicles in dimly lit areas are also successfully detected. This indicates that the proposed method significantly enhances the detection accuracy, especially in complex backgrounds and the identification of small targets. The experiment further verifies that even small targets that are obscured or at a long distance, the proposed model can still successfully detect them, fully demonstrating its effectiveness in detecting small targets under complex scenarios.

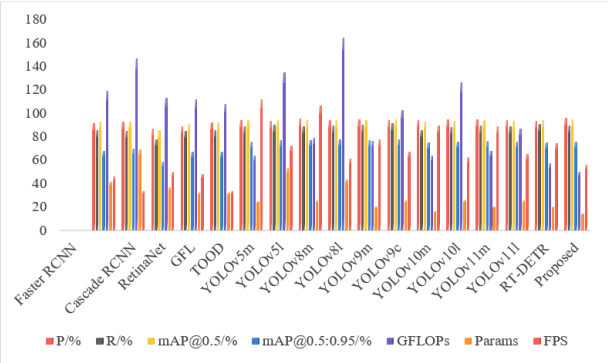

Figure 7. The visualized results of Table 6

Figure 8. Detection results with proposed model

Table 5. Comparison results on VisDrone2019

Model	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95/%	GFLOPs	Params	FPS
Faster RCNN	49.5	36.6	39.1	23.3	208.0	41.4	31.3
Cascade RCNN	50.1	36.5	39.5	23.9	236.0	69.3	25.7
RetinaNet	47.7	38.0	37.0	22.2	210.0	36.5	33.5
GFL	45.0	35.9	34.4	20.3	206.0	32.3	31.9
TOOD	48.8	37.6	37.2	22.1	199.0	32.0	25.0
YOLOv5m	52.1	41.1	41.7	25.2	64.0	25.1	120.4
YOLOv51	55.8	41.7	43.3	26.5	134.7	53.1	72.9
YOLOv8m	53.0	41.1	42.0	25.7	78.7	25.9	103.5
YOLOv81	55.1	42.8	44.1	27.2	164.9	43.6	69.1
YOLOv9m	55.1	41.3	43.4	26.6	76.5	20.0	79.2
YOLOv9c	56.4	42.3	44.1	26.8	102.4	25.3	68.3
YOLOv10m	54.1	40.8	42.0	25.8	63.5	16.5	80.5
YOLOv101	55.2	42.4	44.2	27.4	126.4	25.7	61.3
YOLOv11m	53.7	42.5	43.8	26.8	67.7	20.0	95.4
YOLOv111	55.3	42.3	44.0	27.2	86.6	25.3	63.4
EMRT-DETR	62.0	46.6	48.8	30.5	68.1	13.8	-
SCP-DETR	63.2	47.3	49.4	30.2	52.4	-	-
DMU-YOLO	58.6	47.4	48.1	29.7	89.5	25.4	-
DEIM	57.2	49.1	50.5	30.9	56.4	19.2	80.5
RT-DETR	62.3	46.7	48.0	29.4	57.0	19.9	71.3
Proposed	64.3	48.8	50.8	31.7	49.6	14.6	54.6

Table 6. Comparison results on KITTI dataset

Model	P/%	R/%	mAP@0.5/%	mAP@0.5:0.95%	GFLOPs	Params	FPS
Faster RCNN	91.2	85.7	92.8	67.4	119.0	41.4	45.6
Cascade RCNN	93.0	84.5	92.9	69.4	147.0	69.2	33.3
RetinaNet	86.7	77.7	85.4	58.4	113.0	36.4	50.0
GFL	88.9	85.1	90.9	66.7	112.0	32.3	47.7
TOOD	92.3	85.2	92.1	67.0	108.0	32.0	33.7
YOLOv5m	93.9	88.6	93.7	75.3	64.0	25.0	111.5
YOLOv51	93.5	89.9	94.0	77.0	134.7	53.1	72.0
YOLOv8m	95.1	88.9	93.8	76.6	78.7	25.8	106.5
YOLOv81	94.0	89.3	94.2	77.7	164.8	43.6	61.4
YOLOv9m	94.6	90.0	94.1	77.0	76.5	20.0	77.5
YOLOv9c	93.9	91.1	94.4	77.8	102.3	25.3	67.1
YOLOv10m	93.7	85.7	92.8	74.7	63.4	16.5	89.6
YOLOv101	94.9	88.1	93.5	75.4	126.4	25.7	62.2
YOLOv11m	94.4	89.7	93.9	76.3	67.7	20.0	88.5
YOLOv111	94.1	88.7	93.3	75.7	86.6	25.3	65.1
RT-DETR	93.6	90.7	94.2	75.2	57.0	19.9	74.3
Proposed	95.7	89.6	94.8	75.6	49.6	14.6	56.0

4 Conclusions

To address the problems of severe background interference and insufficient feature expression in small targets in complex scenes, this paper proposes a novel adaptive multi-scale Gated convolution and context-aware attention model for small object detection, which significantly reduces background interference and improves feature expression capabilities. The experimental results show that proposed model has improved the mAP@0.5 and mAP@0.5.0.95 performance on the VisDrone2019 dataset by 2.8% and 2.3% respectively. The parameter quantity and computational cost have been reduced by 26.6% and 13% respectively, making it more advantageous compared to other advanced algorithms. Moreover, it has also been verified for generalization on the KITTI dataset. This further verifies the effectiveness and robustness of the model, and it has significant practical significance and application value.

However, proposed model still has the following limitations. Its performance on small-scale datasets needs improvement, and the model has difficulty converging on small-scale datasets. Real-time detection needs to be enhanced. Although the proposed method achieves a mAP of 50.8% on VisDrone2019, the following limitations still need to be noted. (1) Data scale sensitivity: When the number of training images is less than 1000, the multi-scale gated branch of MGCM suffers from overfitting, and the mAP of the validation set fluctuates by $\pm$ 1.6% (n = 5). (2) Failure in extreme scales: When the short side of the target is less than 8 pixel, the window attention of MFEM degenerates to a single point, and the recall rate drops by 9.2%. Future work may focus on the following directions: optimizing the model architecture to reduce the limitations imposed by data size; combining lightweight design with model compression techniques (such as pruning and knowledge distillation) to further improve the processing efficiency of real-time detection, and potentially providing more efficient solutions for small target detection in complex scenarios.

Funding

This research was funded by Liaoning Provincial Science and Technology Plan Project (Grant No.: 2024JH2/1026 00106).

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] B. Zhou, Z. Zhang, Z. Song, J. Guo, and H. Kong, "Generalizing unsupervised lidar odometry model from normal to snowy weather conditions," *arXiv*, vol. 2509.02011, 2025. https://doi.org/10.48550/arXiv.2509.02011

[2] K. Li, X. Li, Y. Wu, Z. Deng, Y. Wang, Y. Meng, B. Li, X. Su, L. Wang, and X. Wang, "Autonomous dispatch trajectory planning of carrier-based vehicles: An iterative safe dispatch corridor framework," *Def. Tech.*, 2025. https://doi.org/10.1016/j.dt.2025.09.006

[3] R. Pichamuthu, P. Sengodan, S. Matheswaran, and K. Srinivasan, "Energy-efficient hybrid protocol with optimization inference model for WBANs," *J. Appl. Sci. Eng.*, vol. 28, no. 3, pp. 429–440, 2025. https://doi.org/10.6180/jase.202503_28(3).0001

[4] S. L. Yin, L. G. Wang, T. Chen, H. F. Huang, L. Gao, Z. J. L., M. Liu, P. Li, and C. P. Xu, "LKAFormer: A lightweight Kolmogorov-Arnold transformer model for image semantic segmentation," *ACM Trans. Intell. Syst. Technol.*, 2025. https://doi.org/10.1145/3759254

[5] M. Pawe lczyk and M. Wojtyra, "Real world object detection dataset for quadcopter unmanned aerial vehicle detection," *IEEE Access*, vol. 8, pp. 174 394–174 409, 2020. https://doi.org/10.1109/ACCESS.2020.3026192

[6] S. L. Yin, H. Li, A. A. Laghari, L. Teng, T. R. Gadekallu, and A. Almadhor, "FLSN-MVO: Edge computing and privacy protection based on federated learning Siamese network with multi-verse optimization algorithm for Industry 5.0," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 3443–3458, 2024. https://doi.org/10.1109/OJCOMS.2024.3520562

[7] S. F. Ahmed, M. S. B. Alam, M. Hassan, M. R. Rozbu, T. Ishtiak, N. Rafa, M. Mofijur, A. B. M. S. Ali, and A. H. Gandomi, "Deep learning modelling techniques: Current progress, applications, advantages, and challenges," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 13 521–13 617, 2023. https://doi.org/10.1007/s10462-023-10466-8

[8] Y. Wang, "MRCNNAM: Mask region convolutional neural network model based on attention mechanism and Gabor feature for pedestrian detection," *J. Appl. Sci. Eng.*, vol. 26, no. 11, pp. 1555–1561, 2023. https://doi.org/10.6180/jase.202311_26(11).0005

[9] X. Y. Zheng and X. L. Zhao, "Single shot multibox detector-based feature fusion model for building object detection," *J. Appl. Sci. Eng.*, vol. 28, no. 2, pp. 391–398, 2025. https://doi.org/10.6180/jase.202502_28(2).0017

[10] G. Cheng, X. Yuan, X. W. Yao, K. B. Yan, Q. H. Zeng, X. X. Xie, and J. W. Han, "Towards large-scale small object detection: Survey and benchmarks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13 467–13 488, 2023. https://doi.org/10.1109/TPAMI.2023.3290594

[11] W. Wei, Y. Cheng, J. F. He, and X. Y. Zhu, "A review of small object detection based on deep learning," *Neural Comput. Appl.*, vol. 36, no. 12, pp. 6283–6303, 2024. https://doi.org/10.1007/s00521-024-09422-6

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580–587. https://doi.org/10.1109/CVPR.2014.81

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017. https://doi.org/10.1109/TPAMI.2016.2577031

[14] S. L. Yin and H. Li, "Hot region selection based on selective search and modified fuzzy C-means in remote sensing images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5862–5871, 2020. https://doi.org/10.1109/JSTARS.2020.3025582

[15] S. L. Yin, H. Li, L. Teng, A. A. Laghari, A. Almadhor, M. Gregus, and G. A. Sampedro, "Brain CT image classification based on mask RCNN and attention mechanism," *Sci. Rep.*, vol. 14, no. 1, p. 29300, 2024. https://doi.org/10.1038/s41598-024-78566-1

[16] C. Dong, K. Zhang, Z. Y. Xie, and C. J. Shi, "An improved cascade RCNN detection method for key components and defects of transmission lines," *IET Gener. Transm. Distrib.*, vol. 17, no. 19, pp. 4277–4292, 2023. https://doi.org/10.1049/gtd2.12948

[17] P. Li, T. Y. Han, Y. F. Ren, P. Xu, and H. L. Yu, "Improved YOLOv4-tiny based on attention mechanism for skin detection," *PeerJ Comput. Sci.*, vol. 9, p. e1288, 2023. https://doi.org/10.7717/peerj-cs.1288

[18] L. F. Yin and Z. Y. Ding, "Lightweight research on fatigue driving face detection based on YOLOv8," *Recent Adv. Comput. Sci. Commun.*, vol. 19, no. 2, pp. 1–6, 2024. https://doi.org/10.2174/0126662558315127241210053411

[19] Z. H. Zheng, P. Wang, W. Liu, J. Z. Li, R. G. Ye, and D. W. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 993–13 000. https://doi.org/10.1609/aaai.v34i07.6999

[20] B. Huang, S. Wang, L. Li, and X. Fu, "Object detection algorithm based on residual deformable convolution," in *Proceedings of the International Conference on Computer Vision and Pattern Analysis (ICCPA)*, 2024, p. 1325602. https://doi.org/10.1117/12.3037924

[21] C. X. Wang and H. R. Wang, "Cascaded feature fusion with multi-level self-attention mechanism for object detection," *Pattern Recognit.*, vol. 138, p. 109377, 2023. https://doi.org/10.1016/j.patcog.2023.109377

[22] R. Li and Y. P. Wu, "Improved YOLO v5 wheat ear detection algorithm based on attention mechanism,"

Electronics, vol. 11, no. 11, p. 1673, 2022. https://doi.org/10.3390/electronics11111673

[23] Z. Wang, K. Yao, and F. Guo, "Driver attention detection based on improved YOLOv5," *Appl. Sci.*, vol. 13, no. 11, p. 6645, 2023. https://doi.org/10.3390/app13116645

[24] J. Su, Y. Qin, Z. Jia, and B. Liang, "MPE-YOLO: Enhanced small target detection in aerial imaging," *Sci. Rep.*, vol. 14, no. 1, p. 17799, 2024. https://doi.org/10.1038/s41598-024-68934-2

[25] P. J. Wang, X. Sun, W. H. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, 2019. https://doi.org/10.1109/TGRS.2019.2954328

[26] N. A. Baig, M. B. Malik, M. Zeeshan, M. Z. U. Khan, and M. A. Ajaz, "Efficient target detection and joint estimation of target parameters with a two-element rotating antenna," *IEEE Access*, vol. 4, pp. 4442–4451, 2016. https://doi.org/10.1109/ACCESS.2016.2598965

[27] D. Bian, M. Tang, M. Ling, H. Xu, S. Lv, Q. Tang, and J. Hu, "A refined methodology for small object detection: Multi-scale feature extraction and cross-stage feature fusion network," *Digit. Signal Process.*, vol. 164, p. 105297, 2025. https://doi.org/10.1016/j.dsp.2025.105297

[28] S. W. Shi and K. Huang, "Artificial intelligence-based Bayesian optimization and transformer model for tennis motion recognition," *J. Appl. Sci. Eng.*, vol. 29, no. 1, pp. 171–178, 2025. https://doi.org/10.6180/jase.202601_29(1).0017

[29] S. L. Yin, H. Li, A. A. Laghari, T. R. Gadekallu, G. A. Sampedro, and A. Almadhor, "An anomaly detection model based on deep auto-encoder and capsule graph convolution via sparrow search algorithm in 6G internet of everything," *IEEE Internet Things J.*, vol. 11, no. 18, pp. 29 402–29 411, 2024. https://doi.org/10.1109/JIOT.2024.3353337

[30] L. Teng, Y. L. Qiao, and S. L. Yin, "Underwater image denoising based on curved wave filtering and two-dimensional variational mode decomposition," *Comput. Sci. Inf. Syst.*, vol. 21, no. 4, pp. 1765–1781, 2024. https://doi.org/10.2298/CSIS240314057T

[31] Z. Y. Yao, Y. P. Su, H. H. Yang, Y. M. Zhang, and X. J. Wu, "TFSWA-ResUNet: Music source separation with time–frequency sequence and shifted window attention-based ResUNet," *EURASIP J. Adv. Signal Process.*, vol. 2025, no. 1, p. 39, 2025. https://doi.org/10.1186/s13634-025-01249-0

[32] G. Al-Refai and M. Al-Refai, "Road object detection using YOLOv3 and Kitti dataset," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 48–53, 2020. https://dx.doi.org/10.14569/IJACSA.2020.0110807

[33] Y. Yang, X. Han, K. Wang, X. Yu, W. Yu, Z. Wang, G. Li, Z. Han, and J. Jiao, "NRPerson: A non-registered multi-modal benchmark for tiny person detection and localization," *Electronics*, vol. 13, no. 9, p. 1697, 2024. https://doi.org/10.3390/electronics13091697

[34] J. Ferdous, R. Islam, A. Mahboubi, and M. Z. Islam, "A novel technique for ransomware detection using image based dynamic features and transfer learning to address dataset limitations," *Sci. Rep.*, vol. 15, no. 1, p. 32342, 2025. https://doi.org/10.1038/s41598-025-17647-1

[35] E. Q. Liang, D. P. Wei, F. Li, H. M. Lv, and S. T. Li, "Object detection model of vehicle-road cooperative autonomous driving based on improved YOLO11 algorithm," *Sci. Rep.*, vol. 15, no. 1, p. 32348, 2025. https://doi.org/10.1038/s41598-025-18263-9

[36] X. B. Xu, Z. Y. Xing, M. L. Sun, P. R. Zhang, and K. H. Yang, "Enhancing UAV object detection through multi-scale deformable convolutions and adaptive fusion attention," *J. Supercomput.*, vol. 81, no. 14, p. 1301, 2025. https://doi.org/10.1007/s11227-025-07788-5

[37] R. Kıratlı and A. Eroğlu, "Real-time multi-object detection and tracking in UAV systems: Improved YOLOv11-EFAC and optimized tracking algorithms," *J. Real-Time Image Process.*, vol. 22, no. 5, pp. 1–28, 2025. https://doi.org/10.1007/s11554-025-01758-z

[38] D. H. Zhang, C. C. Yu, Z. Li, C. B. Qin, and R. X. Xia, "A lightweight network enhanced by attention-guided cross-scale interaction for underwater object detection," *Appl. Soft Comput.*, vol. 184, p. 113811, 2025. https://doi.org/10.1016/j.asoc.2025.113811