Journal of King Saud University - Computer and Information Sciences 36 (2024) 102113

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University - Computer and Information Sciences

journal homepage: www.sciencedirect.com

Full length article

Improved YOLOv8 algorithms for small object detection in aerial imagery

Fei Feng *, Yu Hu, Weipeng Li, Feiyan Yang

School of Big Data Engineering, Kaili University, Kaili, 556000, China

ARTICLE INFO

Keywords:
Aerial images
YOLOv8
Small target detection
Attention mechanism
Multiscale feature fusion

ABSTRACT

In drone aerial target detection tasks, a high proportion of small targets and complex backgrounds often lead to false positives and missed detections, resulting in low detection accuracy. To improve the accuracy of the detection of small targets, this study proposes two improved models based on YOLOv8s, named IMCMD_YOLOv8_small and IMCMD_YOLOv8_large. Each model accommodates different application scenarios. First, the network structure was optimized by removing the backbone P5 layer used to detect large targets and merging the P4, P3, and P2 layers, which are better suited for detecting medium and small targets; P3 and P2 serve as detection heads to focus more on small targets. Subsequently, the coordinate attention mechanism is integrated into the backbone's C2f, to create a C2f_CA module that enhances the model' s focus on key information and secures a richer flow of gradient information. Subsequently, a multiscale attention feature fusion module was designed to merge the shallow and deep features. Finally, a Dynamic Head was introduced to unify the perception of scale, space, and tasks, further enhancing the detection capability for small targets. Experimental results on the VisDrone2019 dataset demonstrated that, compared with YOLOv8s, IMCMD_YOLOv8_small achieved improvements of 7.7% and 5.1% in mAP@0.5 and mAP@0.5:0.95, respectively, with a 73.0% reduction in the parameter count. The IMCMD_YOLOv8_large model showed even more significant improvements in these metrics, reaching 10.8% and 7.3%, respectively, with a 47.7% reduction in the parameter count, displaying superior performance in small target detection tasks. The improved models not only enhanced the detection accuracy but also achieved model lightweighting, thereby proving the effectiveness of the improvement strategies and showcasing superior performance compared with other classic models.

1. Introduction

As drone technology has significantly advanced, its applications have deeply penetrated various sectors of daily life. Drone aerial photography technology is widely used in industries such as traffic monitoring (Kumar et al., 2021), agricultural reconnaissance (Hafeez et al., 2022), public security patrols (Mohsan et al., 2023), and geological exploration (Asadzadeh et al., 2022), among others. However, this technology presents several challenges. Drones typically operate at high altitudes, resulting in images that contain many small targets against complex and variable backgrounds, which makes target detection from drone-captured images particularly difficult (Shang et al., 2023). To effectively utilize drones for various tasks, it is crucial to enhance the accuracy of drone target detection.

Two predominant approaches are utilized to detect targets within aerial imagery: conventional algorithms and those grounded in deep learning. Conventional techniques are built on manually extracted characteristics such as the histogram of oriented gradients (Dalal and Triggs, 2005) and Haar features (Viola and Jones, 2001). Nevertheless, such approaches tend to demonstrate limited generalizability, particularly in scenarios involving diminutive targets and the intricate and fluctuating nature of aerial imagery, which often results in less-than-ideal outcomes.

The swift advancement of deep-learning technologies has propelled deep-learning-based algorithms for aerial image target detection and showcased exceptional performance. Consequently, these algorithms have emerged as leading technologies in this domain (Hui et al.,

* Corresponding author.
E-mail address: fengfei@kluniv.edu.cn (F. Feng).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2024.102113
Received 15 March 2024; Received in revised form 24 June 2024; Accepted 24 June 2024
Available online 28 June 2024

2024; He et al., 2023; Qi et al., 2023). Target detection algorithms are broadly classified into two groups: one- and two-stage detection methods. Among the one-stage detection approaches, the "you only look once" (YOLO) (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018; Bochkovskiy et al., 2020; Zhu et al., 2021a; Li et al., 2022; Lou et al., 2023) series stands out because of its unique capability to process images a single time to simultaneously determine the bounding box coordinates and probability scores for class regression. Conversely, two-stage detection methods, such as R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2015), R-FCN (Dai et al., 2016), SPP-Net (He et al., 2015), and NAS-FPN (Ghiasi et al., 2019), initiate the detection process by generating potential candidate boxes or regions of interest and then employing convolutional neural networks to classify these specified regions.

Numerous researchers have based their work on the YOLO algorithm and made improvements tailored to various application scenarios to achieve remarkable results. For example, Kisantal et al. (2019) augmented the presence of small objects in their dataset by employing a copy-and-paste technique to amplify their impact on the network. However, this strategy does not optimally leverage the feature information inherent to small objects. On the other hand, Liu et al. (2018) introduced the PAN network architecture, which aimed to boost the model's ability to extract features across multiple scales via the integration of various network layers. Nevertheless, the method that they utilized for feature fusion, which was primarily a straightforward cascading approach, falls short of fully amalgamating feature information from disparate scales. Lim et al. (2021) introduced a method utilizing multiscale feature fusion to leverage the contextual information of targets and an attention mechanism to improve the model's ability to extract small object features. Yang et al. (2022) developed QueryDet to accelerate the inference speed for feature-pyramid-based target detection. Wang et al. (2023b) optimized YOLOv5 to propose a lightweight detection algorithm based on YOLOv5s (MFP-YOLO) and designed a multipath inverse residual module combined with a convolutional block attention module. Wang et al. (2023c) introduced the attention mechanism of BiFormer to optimize the backbone network and designed the FFNB feature processing module to effectively integrate shallow and deep features while adding two detection heads. This method significantly enhances aerial small object detection performance, but the improvements come at the cost of decreased detection speed.

In summary, although existing small object detection algorithms have enhanced detection performance, the accuracy of small object detection in practical applications still fails to meet requirements and leaves substantial room for improvement. This study, based on the YOLOv8s network, targets issues in small object detection tasks and proposes an improved model with enhancements in four key aspects: overall architecture, backbone, neck, and head. Fig. 1 shows an overall structural diagram of the improved network. The main improvements include the following.

(1) Improved Network Structure. In small object detection tasks, the contributions of large object features are relatively limited and may even interfere with the extraction and utilization of small object features. Therefore, to more effectively extract and utilize small object features, we first removed the P5 layer, which is used for detecting large objects, from the backbone network. We then fused the P4, P3, and P2 layers, which are more suitable for detecting the features of small- to medium-sized objects. Finally, we used the P3 and P2 layers as detection heads, focusing on detecting small- to medium-sized objects. This modification not only significantly reduces the number of parameters but also retains more small object feature information, thus improving the accuracy of the model in small object detection tasks.

(2) Design of the C2f_CA Module. To further enhance the model's ability to extract small object features and reduce background noise interference, we introduced a coordinate attention (CA) mechanism into the backbone network and combined it with the C2f module to form the C2f_CA module. To enrich the gradient information flow in the C2f_CA module, we removed the residual structure and performed a concat operation with two CBS modules in the bottleneck and CA modules to enhance the capability of the model to detect multiscale targets.

(3) Design of the Adaptive Multiscale Feature Fusion (AMFF) Module. To further optimize the effective fusion of high-level semantic and low-level detail information, we introduced the AMFF module into the neck of the model. We first used bilinear interpolation and adaptive average pooling techniques to upsample and downsample the features to ensure consistent feature sizes. We then effectively fused these high- and low-level features using the Hadamard product. Finally, we employed the selective kernel attention (SKA) mechanism to further enhance the feature-fusion effect, thereby improving the model's capability for detailed recognition and semantic analysis.

(4) Improved detection head. To focus the model more on densely packed small object areas and effectively extract more small object features, we introduced a Dynamic Head with multiple attention mechanisms and replaced the Decoupled Head of the YOLOv8 network with the Dynamic Head.

Table 1
Parameters corresponding to different sizes of YOLOv8.

Model	Depth	Width	Max channels	Parameters (relative to YOLOv8l)	GFLOPs (relative to YOLOv8l)
YOLOv8n	0.33	0.25	1024	3.0	8.2
YOLOv8s	0.33	0.50	1024	11.1	28.7
YOLOv8m	0.67	0.75	768	25.9	79.1
YOLOv8l	1.00	1.00	512	43.6	165.4
YOLOv8x	1.00	1.25	512	68.2	258.2

Note: The size of each input image is 640 × 640 pixels.

2. YOLOv8 network structure

YOLOv8 represents a newer generation of object detection algorithms and exhibits superior accuracy and performance compared with other mainstream algorithms. To accommodate diverse application scenarios and the need for flexible deployment across different hardware devices, five versions of YOLOv8 have been designed with varying network depths and widths: YOLOv8n, YOLOv8s, YOLOv8 m, YOLOv8l, and YOLOv8x. From the small-scale YOLOv8n to the large-scale YOLOv8x, each version incrementally increases the parameter quantity and resource consumption while also improving the detection performance. The network parameters for the five versions of YOLOv8 are shown in Table 1.

The YOLOv8 network structure is divided into four main parts: the input, backbone, neck, and head. Fig. 2 shows the YOLOv8 network structure. The input section is responsible for image preprocessing, including data augmentation and image scaling, to optimize the image data that are fed into the model. The backbone section, which extracts the image features, consists of the CBS, SPPF, and C2f modules. Among them, the CBS module consists of Conv, BN, and SiLU activation functions. SPPF reduces the computational load and accelerates processing by connecting three 5 × 5 maximum pooling layers. The C2f module employs the concept of gradient diversion by combining the CBS and residual modules to obtain additional gradient flow information, which enhances the feature extraction and learning capabilities of the model. The neck section utilizes an FPN+PAN structure, which effectively enhances the model's ability to detect objects of varying sizes by merging multiscale features through both top-down (Lin et al., 2017a) and bottom-up (Liu et al., 2018) approaches. This structure significantly improves the detection accuracy for images with complex backgrounds and varying object sizes. The head uses a Decoupled Head structure to separately handle the classification and detection tasks. Simultaneously, the Anchor-Free (Hu et al., 2019) idea is used to replace the

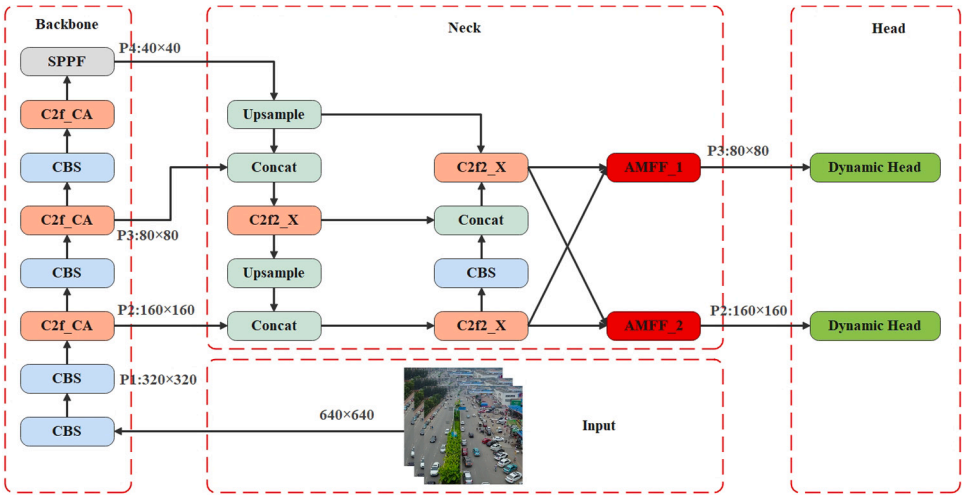

Fig. 1. The overall structural diagram of the improved network.

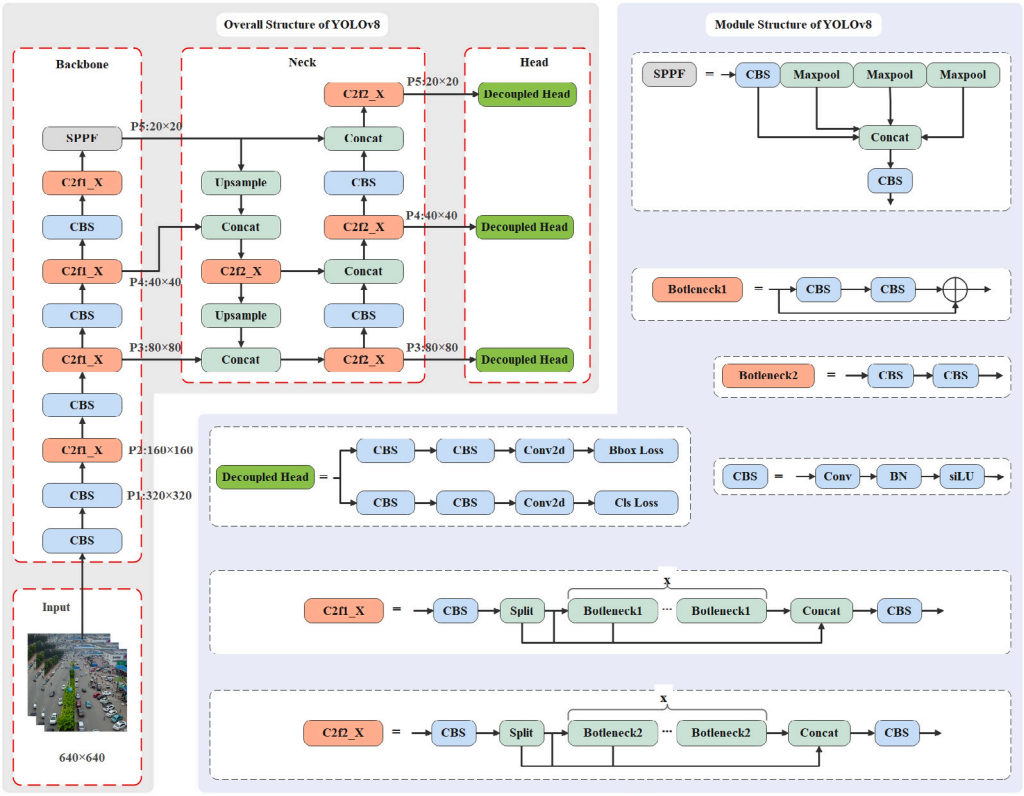

Fig. 2. YOLOv8 network structure.

Anchor-Based (Kong et al., 2020) method. Through these innovative designs and improvements, YOLOv8 achieved a dual breakthrough in terms of accuracy and performance in the field of object detection.

3. Algorithm improvements

3.1. Improved network structure

In the backbone architecture of YOLOv8, the network performs downsampling operations at 2×, 4×, 8×, 16×, and 32× to generate five sets of feature maps (P1–P5) at different scales. Taking an initial image of 640 × 640 pixels as an example, these feature maps were scaled down to the dimensions of 320×320, 160×160, 80×80, 40×40, and 20×20 pixels. Larger, shallower feature maps (such as P1 and P2) excel at recognizing smaller objects, whereas smaller, deeper feature maps (such as P4 and P5) are more suited for recognizing larger objects. By integrating the P3, P4, and P5 layers, YOLOv8 can detect targets with sizes of 8×8, 16×16, and 32 × 32 pixels in the original image. However, the model's ability to capture extremely small targets (less than 8 × 8 pixels) appears to be weaker.

The manually labeled object information in the VisDrone2019 training dataset indicates that the dataset contains many small-area target objects, as shown in Fig. 3. The sub-figures in Fig. 3. are labeled in order from left to right and top to bottom. The first sub-figure displays the number of each type of target in the dataset, indicating that the targets mainly consist of cars and pedestrians. The second sub-figure

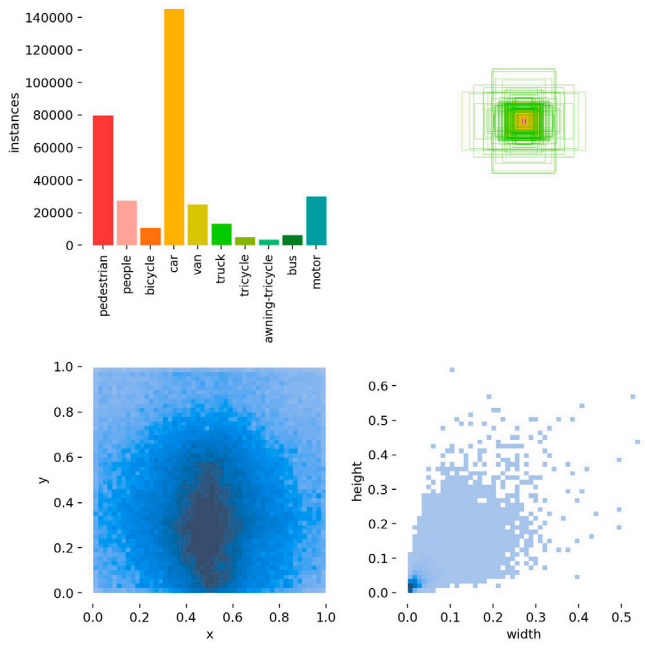

Fig. 3. Information related to manually labeled objects in the VisDrone2019 training data set.

shows the size of the bounding boxes for detected targets in the dataset, with all target box centers fixed at one point, indicating that the dataset contains a large number of targets with small areas. The third sub-figure represents the distribution of the center coordinates of the object bounding boxes, showing that the centers of the objects are mainly concentrated in the middle and lower-right areas of the image data. The fourth sub-figure displays the distribution of the aspect ratio of targets relative to the entire image, with the darkest area located in the lower left corner, further confirming that the dataset primarily consists of small objects. To enhance the ability of the model to detect small objects, some researchers introduced the P2 layer as a small object detection head (Wang et al., 2022; Chen et al., 2023) on top of YOLOv8, which improved the detection capability for small objects to some extent. However, there is scope for further improvements. Without sacrificing the accuracy of the model, the number of parameters can be significantly reduced, and the model can be lightweighted.

In YOLOv8, the extraction of high-level feature maps relies on numerous parameters. Taking the YOLOv8s model as an example, the parameter counts for feature maps P1–P5 in the backbone are listed in Table 2. According to the data in the table, the parameters of the feature information of the P5 and P4 layers account for 64.4% and 27.6% of the parameter count. These two layers are primarily used to improve the detection performance for large and medium-sized objects. However, in tasks focused on detecting small objects, such a high-cost parameter configuration is unnecessary and may even weaken the detection efficiency for small targets. Therefore, this study proposes an optimization strategy that eliminates the P5 layer from the backbone and fuses the features of the P4, P3, and P2 layers, using only the fused feature maps of P3 and P2 as the detection heads. This improvement not only significantly reduces the model's parameter count but also retains more feature information that is useful for small object detection. Hence, the proposed method effectively enhances the model's ability to recognize small objects. Fig. 4 shows the improved network structure.

3.2. C2f_CA module

3.2.1. Design of the C2f_CA module

To overcome the recognition difficulties caused by the high background complexity and small target space proportion in small target detection, this study proposes the integration of the attention mechanism into the model structure. This approach reduces the interference of background information and enhances the ability of the model to detect small-target feature information, as well as the model's extraction capabilities, thereby improving recognition accuracy for small targets. In object detection, squeeze and excitation (SE) (Hu et al., 2018) as well as convolutional block attention modules (CBAM) (Woo et al., 2018) are two commonly used attention mechanisms. The SE mechanism models the dependencies between feature channels to enhance the model's ability to judge the importance of different channel features; however, it overlooks the spatial distribution of features (i.e., their positional information). The CBAM further considers spatial information based on SE, strengthens the model's ability to capture position information through the spatial attention mechanism, and enhances attention to local features. However, while enhancing local feature information, CBAM still faces the limitation of not effectively capturing long-range dependency information, which is particularly critical for detecting small targets against complex backgrounds.

Considering the limitations of the SE and CBAM mechanisms, this study opted for the CA (Hou et al., 2021) mechanism, which can capture both distant spatial and channel information to significantly improve the model's ability to accurately locate and identify target objects against complex backgrounds. Fig. 5 shows the structure of the CA mechanism. Consequently, this study proposes incorporating the CA mechanism into the backbone network of YOLOv8 and combining it with the C2f module to redesign the C2f module as C2f_CA, which can be used to replace the original C2f module in the network. To enable the model to obtain a richer gradient information flow, the bottleneck residual structure in the C2f module was first removed, and this step was followed by embedding the CA module at the backend of the bottleneck. Then, by performing a concat operation with the two CBS modules in the bottleneck and the CA mechanisms, the goal was to strengthen the model's capability to extract target information, which effectively enhanced the detection accuracy for small objects. Fig. 6 shows the structure of the C2f_CA module.

3.2.2. Principle of the CA mechanism

The CA mechanism initially performs global average pooling on the input feature map in two spatial dimensions: height and width. This process allows the model to capture long-distance dependencies in one dimension while retaining precise positional information. This bidirectional global pooling strategy is a core component of the CA mechanism and effectively enhances the sensitivity of the attention mechanism to spatial positional information. This process is realized using the following two equations:

(1) For global average pooling in the direction of the feature map's height, the equation is shown as (1):

$$Z_c^h(h) = \frac{1}{w} \sum_{0 \le i < w} x_c(h, i) \tag{1}$$

where x is the input feature map with size $C \times H \times W$ and Z^h represents the feature map after pooling in the height direction.

(2) For global average pooling in the direction of the feature map's width, the equation is shown as (2):

$$Z_c^w(h) = \frac{1}{h} \sum_{0 \le j < w} x_c(j, w) \tag{2}$$

where x is the input feature map with size $C \times H \times W$, and Z^w represents the feature map after pooling in the width direction.

In the next step of the CA mechanism, the feature maps obtained from the two spatial dimensions are concatenated along the channel direction, and this is followed by a series of transformations on this concatenated feature map, including convolution, batch normalization, and the application of a nonlinear activation function. This process is described by Eq. (3):

$$f = \delta(F_1([Z^h, Z^w])) \tag{3}$$

Table 2
Parameter comparisons of different feature maps.

Feature map	Downsampling rate	Feature map size	Number of channels	Parameters (10^4)
P5	32×	20 × 20	512	183.8
P4	16×	40 × 40	256	78.8
P3	8×	80 × 80	128	19.8
P2	4×	160 × 160	64	2.9
P1	2×	320 × 320	64	1.8

Note: YOLOv8s is the base network, and the size of each input image is 640 × 640 pixels.

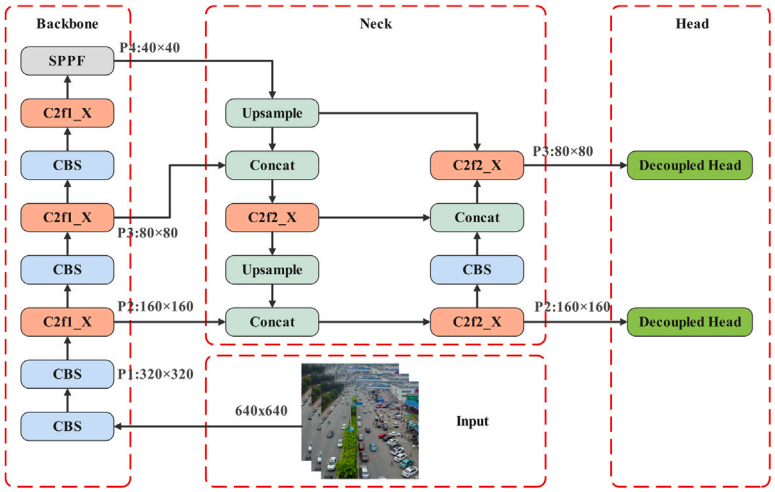

Fig. 4. Improved network structure.

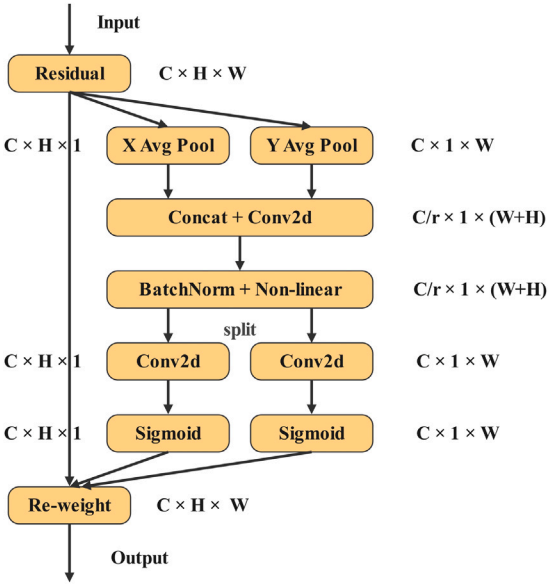

Fig. 5. CA mechanism structural diagram.

where $([Z^h, Z^w])$ represents the splicing operation of Z^h and Z^w along the channel direction, F_1 represents a 1 × 1 convolution transformation function, δ is a nonlinear activation function, and f represents the combined result of the input feature maps in both directions.

Subsequently, the intermediate feature map f is split into two independent feature tensors f^h and f^w along the spatial dimension, which are then transformed via 1 × 1 convolution into feature tensors $F_h(f^h)$ and $F_w(f^w)$ with the same number of channels as the original input x. After processing these two feature tensors with the sigmoid activation function, the attention weights g^h and g^w on the height h and width w are obtained, as expressed in Eqs. (4) and (5):

$$g^h = \sigma(F_h(f^h)) \tag{4}$$

$$g^w = \sigma(F_w(f^w)) \tag{5}$$

where σ represents the sigmoid activation function, and F_h and F_w represent the functions for the 1 × 1 convolution transformation of f^h and f^w, respectively.

Finally, the attention weights obtained by the CA mechanism are applied to the input feature map to achieve weighted feature mapping, thus enhancing feature representation, as expressed in Eq. (6):

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{6}$$

where $y_c(i,j)$ represents the weighted feature map, $x_c(i,j)$ is the original input feature map, and $g_c^h(i)$ and $g_c^w(j)$ are the attention weights obtained along the height and width dimensions, respectively. Using this method, the feature map elements in each channel are weighted according to their importance in the spatial dimensions, which thereby enhances the model's focus on features that are more important at specific spatial locations. This method significantly improves the recognition accuracy and generalization ability of the model, particularly in scenarios where small targets are detected against complex backgrounds.

3.3. AMFF module

3.3.1. Design of the AMFF module

In object detection tasks, the strategy of multiscale feature fusion is crucial for enhancing the detection capabilities of models (Peng et al., 2023). In the feature extraction process, lower-level networks that are closer to the input layer can capture basic information, such as the texture, edges, and location of the image. However, these lower-level networks have relatively low semantic richness. As the network depth increases, the feature maps undergo repeated convolution and pooling operations. Although semantic information is enhanced, this

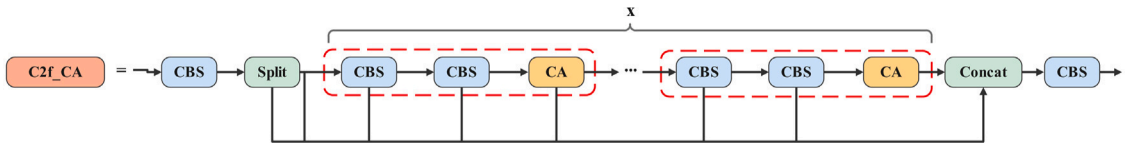

Fig. 6. C2f_CA module structural diagram.

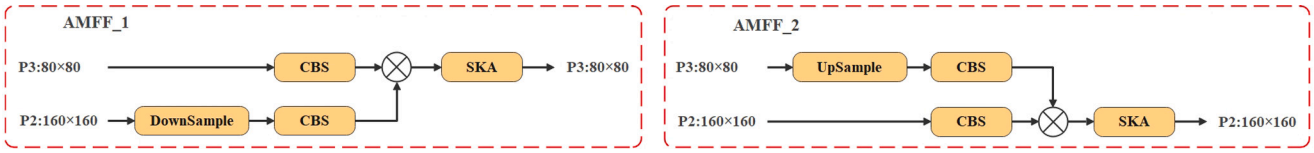

Fig. 7. AMFF_1 and AMFF_2 structural diagram.

also results in a reduction in resolution and the loss of detailed information. This phenomenon is particularly important for the detection of small targets, because the details of small targets are vital for their correct identification. Therefore, effectively merging shallow detail information with deep semantic information not only enhances the model's overall understanding of the image but also compensates for the information loss caused by relying solely on deep features. By integrating features from both deep and shallow levels, the model can achieve high-level semantic understanding and sensitive detail capture capabilities to more accurately locate and identify targets in complex scenes and significantly improve the detection of small targets (Huang et al., 2022; Huan et al., 2021). This multiscale feature-fusion strategy has been widely applied to various object detection frameworks.

YOLOv8 utilizes an FPN+PAN structure to combine high-level semantic information with low-level detail information, which is important for improving the detection accuracy and robustness. However, conventional direct concatenation fusion methods cannot fully exploit the complementarity between the features of different layers, particularly in terms of detail preservation and semantic enhancement. Therefore, this study proposes an innovative AMFF module that aims to further optimize the fusion of high-level semantic information and low-level detail information after the output of YOLOv8's feature pyramid network by effectively fusing feature maps of different scales to improve the detailed capturing and semantic understanding abilities of the model. Fig. 1 shows the location where the AMFF module was added.

The AMFF module first employs adaptive average pooling technology to downsample the input feature maps and bilinear interpolation technology for upsampling. The use of these technologies ensures that feature maps of different scales are scaled to a unified size. In addition, the channel numbers of the multiscale feature maps were adjusted for consistency. Subsequently, the feature maps from each branch are processed using a CBS module to enhance the expressiveness of the scaled features. After the feature maps have undergone upsampling/downsampling, scale adjustment, and CBS module processing, the two branches apply element-wise Hadamard multiplication to achieve deep integration between the feature maps of different scales. This fusion method not only preserves rich semantic information but also maintains detailed information, which is particularly crucial for improving the detection accuracy of small targets. Finally, the AMFF module introduces the SKA mechanism (Li et al., 2019), which focuses on selecting and enhancing features that are more important for the detection task and allows the model to adaptively concentrate on the most useful features for the current task, thereby further enhancing detection accuracy. Fig. 7 shows the structures of AMFF_1 and AMFF_2.

3.3.2. Principle of SKA

The SKA mechanism is an efficient strategy that was designed to mimic the ability of cortical neurons to dynamically adjust the sizes of their receptive fields based on different stimuli. This mechanism operates through three main steps: split, fused, and select-handling convolution paths with different kernel sizes, and it dynamically integrates the results of each convolution kernel to adapt to varying target detection scales.

(1) Split operation: In this step, convolution operations with kernels of different sizes (e.g., 3×3 and 5×5) are applied to the input feature map, to generate two sets of feature maps, namely, U_1 and U_2 at different scales. The purpose of this study was to capture information at different spatial scales.

(2) Fuse operation: This step involves fusing U_1 and U_2 through element-wise addition to produce a new feature map by U. Global average pooling is then performed on U to embed global information and produce channel features, denoted by S. To enhance the processing efficiency, S is then compressed into lower-dimensional channel features denoted Z through a fully connected layer.

(3) Select operation: This step employs a soft attention mechanism across channels to adaptively select information at different spatial scales. This channel attention is multiplied by the corresponding feature maps to obtain the weighted feature maps V_1 and V_2, which are then added together to yield the final feature map V. Branches with different kernel sizes are fused with the corresponding channel information using softmax attention, which allows these branches to receive varying levels of focus to enable the fused-layer neurons to effectively have different receptive field sizes.

The introduction of the SKA mechanism allows the model to adaptively adjust the sizes of its receptive fields based on the characteristics of the input information and effectively capture target information at different scales. This mechanism significantly enhances the model's performance in target detection tasks across various scales, especially in scenarios involving small targets or targets with complex backgrounds, which significantly improves the detection accuracy and efficiency. Fig. 8. illustrates the structure of the SKA.

3.4. Improvements to the Detection Head

In this study, the detection head of YOLOv8 was replaced with a Dynamic Head (Dai et al., 2021) to enhance the capability of the model to detect small targets. In object detection tasks, the challenges faced by the detection head mainly focuses on three aspects: scale, spatial, and task awareness.

(1) Scale awareness capability: A single image can contain multiple objects on different scales. Therefore, the detection head must identify and correctly process targets of various sizes to ensure accurate detection, regardless of the target size.

(2) Spatial awareness capability: Because of the diversity of shooting angles and target positions, the same object may appear in completely different shapes, rotations, and positions under different circumstances. The detection head must have sufficient spatial awareness to adapt to this diversity and accurately identify targets.

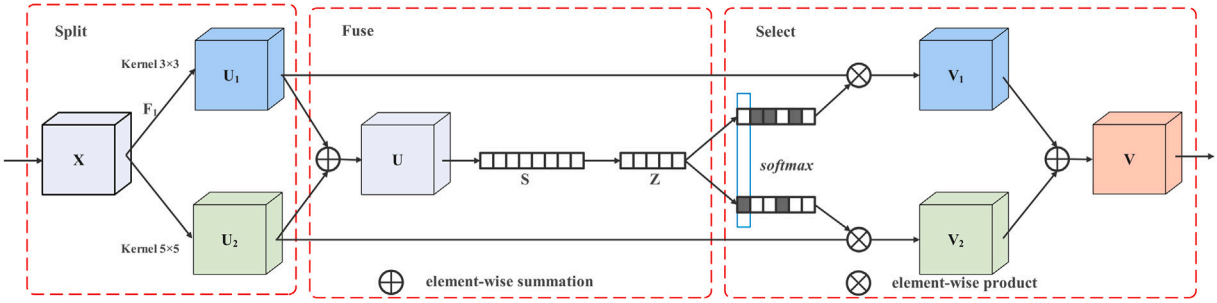

Fig. 8. SKA structural diagram.

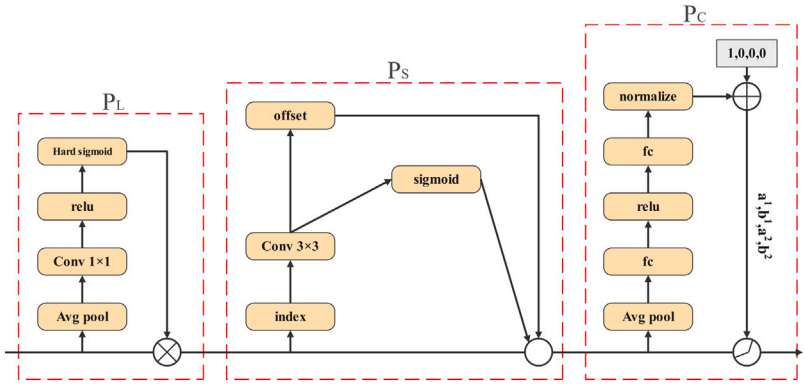

Fig. 9. Dynamic Head structure diagram.

(3) Task awareness capability: In object detection, different tasks may require different target representations, such as bounding boxes, object center points, or corners. The detection head must provide appropriate target representations based on specific task requirements.

The Dynamic Head integrates scale-awareness, spatial-awareness, and task-awareness attention to comprehensively address these challenges. This unified attention mechanism significantly enhances the expressive power and flexibility of the detection head and thereby enables the model to process targets of different scales and shapes more accurately and satisfy the requirements of different detection tasks. Through this Dynamic Head framework, YOLOv8 effectively improves its ability to detect small targets and enhances its recognition and localization capabilities in complex scenes. Fig. 9. illustrates the structure of the Dynamic Head.

The scale-awareness attention mechanism effectively integrates features across different scales by evaluating their semantic importance, as expressed in Eq. (7):

$$\pi_L(F).F = \sigma(f(\frac{1}{SC}\sum_{SC}F)).F \tag{7}$$

where $\pi_L(.)$ denotes scale-awareness attention, $f(.)$ represents a linear function performed by a 1×1 convolution, and $\sigma(x) = \max(0, \min(1, \frac{x+1}{2}))$ is the hard-sigmoid activation function.

The spatial-awareness attention mechanism aims to enhance the discriminative ability of the model with respect to different spatial positions by focusing on areas within the image that are particularly important for the task. This attention mechanism specifically emphasizes adjusting the processing of spatial information by the model to better recognize and locate targets. Employing deformable convolution to achieve the sparsification of attention learning is a key step in this strategy, and this step is followed by cross-scale feature integration to further boost the spatial awareness capabilities of the model, as illustrated in Eq. (8):

$$\pi_S(F).F = \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K}W_{l,k}.F(l;P_k+\Delta P_k;c).\Delta m_k \tag{8}$$

Here, $\pi_S(.)$ signifies spatial-awareness attention, K is the number of sparse sampling locations, $P_k + \Delta P_k$ is the self-learned spatial offset focusing on unique areas, Δm_k is the self-learned importance scalar at location P_k, and all are learned from the input features of the intermediate layer F.

The core concept of the task-awareness attention mechanism is to dynamically adjust the importance of feature channels based on different object-detection tasks and thereby optimize the model's response to specific tasks. By dynamically "switching" the feature channels, this attention mechanism can more effectively assist in completing tasks in various representations, as shown in Eq. (9):

$$\pi_C(F).F = \max(\alpha^1(F).F_C + \beta^1(F), \alpha^2(F).F_C + \beta^2(F)) \tag{9}$$

where $\pi_C(.)$ represents task-awareness attention and $[\alpha^1, \alpha^2, \beta^1, \beta^2]^T = \theta(.)$ is a hyperfunction that is used to control the activation threshold. This process starts with global pooling to reduce dimensions, and it is followed by two fully connected layers and a normalization layer. Finally, the output is normalized to the range [−1,1] by using the sigmoid function.

4. Experiments

4.1. Experimental dataset

The experiments in this study were conducted using the VisDrone2019 dataset (Zhu et al., 2021b), which was meticulously collected by the AISKYEYE team. This dataset was captured using various drone models under different scenarios, weather conditions, and lighting environments, and it has been manually annotated with more than 2.6 million precise object bounding boxes, ensuring diversity in scenes, richness in object categories, and consideration of occlusion situations. The data used for object detection consisted of 10,209 static images that were divided into 6471 training images, 548 validation images, and 3190 test images. The dataset defines 10 categories of detection targets that appear in relatively small sizes in the images,

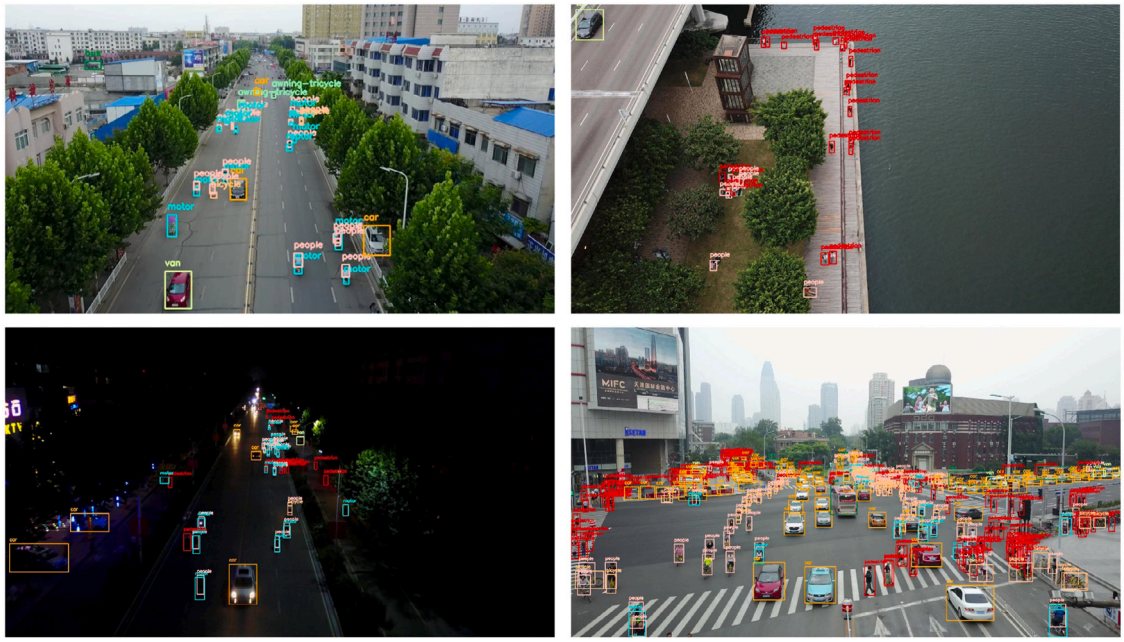

Fig. 10. Images and target annotation boxes from four different scenes in the VisDrone2019 dataset.

Table 3
Experimental environment configuration.

Environment configuration	Version
Operating system	Windows
CPU	Intel i5-13490, 2.5 GHz
GPU	GeForce RTX 3090, 24 GB Graphics Memory
CUDA	11.8
PyTorch	2.0.0
Python	3.11.4

Table 4
Settings for some of the hyperparameters and data augmentation.

Hyperparameters	lr0: 0.01, lrf: 0.1, momentum: 0.937, weight_decay: 0.0005, warmup_momentum: 0.8, (warmup_bias_lr: 0.1, Warmup_epochs: 3.0, batch-size: 8, epochs: 200)
Data augmentation	Scale: 0.5, fliplr: 0.5, mosaic: 1.0, mixup: 0, hsv_h: 0.015, hsv_s: 0.7, hsv_v: 0.4

making it highly suitable for research on small object detection and recognition. The images and target annotation boxes were randomly selected from four scenes in the VisDrone2019 dataset, as shown in Fig. 10.

4.2. Experimental environment and evaluation indicators

Table 3 shows the experimental environment configuration used in this study.

In the experiments conducted in this study, the input image resolution was set as 640 × 640. SGD was used to optimize and adjust the network. Table 4 presents the settings for some of the hyperparameters and data augmentation.

Subsequent comparative experiments were conducted under the same training conditions to accurately measure the performance of the improved model. Additionally, we used various evaluation metrics to assess the model performance, including precision, recall, mAP@0.5, mAP@0.5:0.95, parameters, GFLOPs, and model size. This comprehensive evaluation ensured the reliability and consistency of the results. Precision is the proportion of targets correctly identified by the model, whereas recall is the proportion of true targets correctly identified by the model out of the total number of actual targets. AP denotes the average precision, which is the average accuracy of the model in identifying specific targets. mAP denotes the mean of the AP across all categories, and it is a key indicator for assessing the overall performance of a model; the higher the mAP, the better the comprehensive detection performance of the model across all categories. mAP@0.5 assesses the performance at an intersection over union (IoU) setting of 0.5, primarily by measuring the model performance under more lenient conditions. mAP@0.5:0.95, which is the average mAP as the IoU increases from 0.5 to 0.95 in steps of 0.05, provides a more comprehensive demonstration of the model's detection capabilities under varying levels of strictness. The model parameters refer to the sum of the parameters across all the convolutional layers and filters. Giga floating point operations (GFLOPs) measure the computational complexity of the model (Oneto, 2020). The specific calculations are presented in Eqs. (10)–(13).

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

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

$$AP = \frac{1}{N}\sum_{1}^{N} Precision \tag{12}$$

$$mAP = \frac{1}{N}\sum_{1}^{N} AP \tag{13}$$

When addressing classification tasks, particularly when focusing on the correct classification of a specific category, the task can be viewed as a binary classification problem in which the classification objectives are divided into two types: positive and negative. In such instances, a confusion matrix is commonly used to represent the classification performance of the model. The confusion matrix is a table that illustrates the relationship between the model's predicted outcomes and the actual labels, and it allows for a clear evaluation of the model's performance in both positive and negative cases. Binary classification tasks use a confusion matrix for definition, as shown in Table 5.

Table 5
Confusion matrix.

Confusion matrix		True value	
		Positive	Negative
Predicted value	Positive	TP	FP
	Negative	FN	TN

4.3. Experimental results and analysis

4.3.1. Comparative experiments on network structure improvements

Owing to its good balance between detection accuracy and speed, YOLOv8s is widely used in various object detection tasks. Therefore, this study selected YOLOv8s as the base network to investigate the improvements in small object detection tasks. To evaluate the impact of the backbone network layer and detection head on small object detection, six sets of different structural improvements were designed, and comparative experiments were conducted under the same training conditions. Table 6 lists these improvement strategies, and Table 7 presents the training results for the VisDrone2019 dataset. The experimental results indicate that the improved YOLOv8s_5 achieved the best detection performance. YOLOv8s_5 enhances the extraction and utilization of small object features by removing the P5 layer from the backbone, fusing the features of the P4, P3, and P2 layers, and using the P3 and P2 layers as the detection head. These improvements not only enable the model to focus more on detecting small objects but also simplify the model structure.

Compared with the original YOLOv8s model, the improved YOLOv8s_5 model achieved improvements of 6.0%, 2.9%, 4.5%, and 3.1% in the precision, recall, mAP@0.5, and mAP@0.5:0.95, respectively. These results demonstrate that improvements to the backbone layer and detection head can significantly enhance the ability of the model to detect small objects. In addition to improved detection performance, the updated model also showed reductions of 79.2%, 80.5%, and 4.8% in the number of parameters, model size, and GFLOPs, respectively. Thus, this approach greatly reduced the model's resource requirements, making it more suitable for deployment on resource-limited devices while maintaining high detection accuracy. The improved model, IM_YOLOv8, achieved both lightweight construction and enhanced performance in detecting small objects.

4.3.2. Comparative experiments on different C2f_CA module configurations

(1) Comparative Experiments on Different Combinations of CA and C2f

To ensure optimal performance after integrating the CA mechanism into the C2f module, we designed three different combination methods and conducted comparative experiments. In the first method, the CA mechanism is embedded inside the residual structure of the C2f module, as shown in Fig. 11. The second method places the CA mechanism behind the residual structure of the C2f module, as shown in Fig. 12. The third method adds the CA mechanism after the residual structure, eliminates the residual structure, and performs a concat operation with the two CBS modules within the residual structure and CA mechanism, as illustrated in Fig. 6. In this series of experiments, we used the improved IM_YOLOv8 model as the baseline, replacing the C2f module in the backbone with the C2f_CA module. Table 8 presents the experimental results for the three C2f_CA combinations. The results indicated that the third combination method enabled the IMC_YOLOv8 model to achieve the best detection performance. Compared with the original IM_YOLOv8 model, IMC_YOLOv8 achieved improvements of 0.8% and 0.3% in mAP@0.5 and mAP@0.5:0.95, respectively. This indicates that the third combination method, with the addition of the CA module, enabled the model to more precisely identify key information in the input features. Furthermore, the gradient flow was optimized by eliminating the residual structure, allowing the C2f_CA module to receive a richer gradient information flow. This enhanced the ability of the model to detect multiscale targets in complex scenes, thereby improving its detection performance.

(2) Comparative Experiments Between SE, CBAM, and CA

To validate the superiority of the CA mechanism over the SE and CBAM mechanisms in small object detection tasks, the CA mechanism in the designed C2f_CA was replaced by the SE and CBAM mechanisms, resulting in C2f_SE and C2f_CBAM, respectively. Comparative experiments were conducted under identical experimental conditions, using IM_YOLOv8 as the base network. Table 9 shows the experimental results.

The results indicate that the model with the C2f_CA module performed best on the mAP@0.5 and mAP@0.5:0.95 metrics, confirming the significant advantage of the CA mechanism over the SE and CBAM mechanisms in processing small object detection tasks. The SE mechanism can enhance the recognition of the importance of different channels by recalibrating the channels; however, it neglects spatial information within the feature map, making it unable to fully utilize information in the spatial dimension. CBAM introduces a spatial attention mechanism to enhance the model's capture of positional information; however, it still faces limitations in effectively capturing long-distance dependency information by focusing only on local details. In contrast, the CA mechanism can simultaneously obtain long-distance spatial positional and channel information, which helps the model locate and recognize the objects of interest more accurately.

Fig. 13 shows the variation curves of mAP@0.5 and mAP@0.5:0.95 during the training process for the three attention-mechanism modules. The figure shows that the C2f_CA module exhibited the best detection performance.

4.3.3. Comparisons with YOLOv8

(1) Comparison with YOLOv8s

To demonstrate the improved detection performance of the enhanced model, we conducted a comparative experiment between the final improved model and the baseline model YOLOv8s; the results are shown in Table 10. To satisfy the requirements of the AMFF module and Dynamic Head for consistent input feature channel numbers, we designed two improved models: IMCMD_YOLOv8_small and IMCMD_YOLOv8_large. In the IMC_YOLOv8 model, the channel numbers for P2 and P3 were 64 and 128, respectively, with P2 and P3 serving as input features for the AMFF module and Dynamic Head, respectively, requiring consistent input channel numbers. Therefore, in the IMCMD_YOLOv8_small model, we adjusted the channel number of the P3 layer to 64 to match that of the P2 layer. This adjustment strategy reduces the model's parameter count while maintaining the detection performance. In the IMCMD_YOLOv8_large model, we increased the channel number of P2 layer to 128 to match that of the P3 layer. Although this adjustment increases the model's parameter count, it provides a richer feature representation and further improves detection performance.

According to the results in Table 10, the improved IMCMD_YOLOv8_small model, compared with YOLOv8s, achieved improvements of 8.1%, 5.7%, 7.7%, and 5.1% in precision, recall, mAP@0.5, and mAP@0.5:0.95, respectively, as well as a 73.0% reduction in the model parameter count. This significant performance improvement, along with the substantial reduction in parameters, proves that the improved model is lightweight while enhancing the detection accuracy.

In contrast, the IMCMD_YOLOv8_large model exhibited even more significant improvements in the four performance metrics: 11.7%, 7.8%, 10.8%, and 7.3%, respectively, with a 47.7% reduction in model parameters. These results indicate that by increasing the number of channels to match the P3 layer, the model gained richer feature representations and hence demonstrated superior performance in small object detection tasks.

IMCMD_YOLOv8_small and IMCMD_YOLOv8_large significantly improved the model accuracy and substantially reduced the number of

Table 6
Different improvement strategies for the backbone network layer and detection head.

Models	Delete backbone's P5 layer	Delete backbone's P4 layer	Add P2 detection head	Delete P5 detection head	Delete P4 detection head
YOLOv8s_1	×	×	×	×	×
YOLOv8s_2	×	×	✓	×	×
YOLOv8s_3	×	×	✓	✓	×
YOLOv8s_4	✓	×	✓	✓	×
YOLOv8s_5	✓	×	✓	✓	✓
YOLOv8s_6	✓	✓	✓	✓	✓

Note: "✓" means that this improvement was carried out, and "×" means that this improvement was not carried out.

Table 7
Experimental results of different improvement strategies.

Models	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameters (10^6)	GFLOPs	Model size (M)
YOLOv8s_1	48.3	38.8	39.1	23.5	11.1	28.7	21.4
YOLOv8s_2	53.9	41.6	43.7	26.6	10.6	37.0	20.6
YOLOv8s_3	53.2	42.1	43.3	26.1	7.5	34.8	14.7
YOLOv8s_4	51.8	42.4	43.5	26.5	3.3	30.1	6.6
YOLOv8s_5	54.3	41.7	43.6	26.6	2.3	27.3	4.7
YOLOv8s_6	49.6	39.2	40.2	24.1	0.9	21.4	2.0

Table 8
Comparative experimental results of different combination methods using CA and C2f.

Methods	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
IM_YOLOv8	54.3	41.7	43.6	26.6
+C2f_CA combination method_1	53.6	42.0	43.9	26.5
+C2f_CA combination method_2	54.0	42.0	44.2	26.8
+C2f_CA combination method_3	53.8	42.3	44.4	26.9

Table 9
Experimental results comparing the performance of three attention mechanism modules.

Methods	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	GFLOPs
C2f_SE	52.1	43.0	43.6	26.6	28.3
C2f_CBAM	53.9	41.5	43.9	26.8	28.3
C2f_CA	53.8	42.3	44.4	26.9	28.3

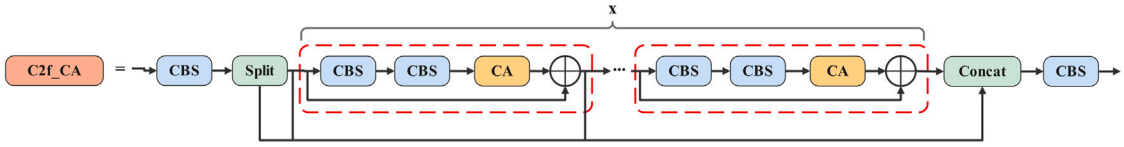

Fig. 11. Combination method 1 using CA and C2f.

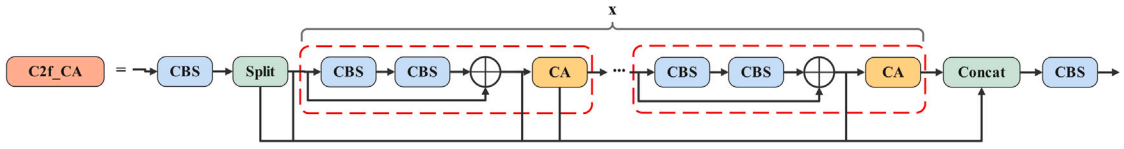

Fig. 12. Combination method 2 using CA and C2f.

model parameters and model size, although there was an increase in the GFLOPs.

To better demonstrate the effectiveness of the improved algorithms presented in this study, the precision–recall curves of the improved models IMCMD_YOLOv8_small and IMCMD_YOLOv8_large were compared with those of the baseline model YOLOv8s during the training process, as shown in Fig. 14. The graph shows that compared to YOLOv8s, both improved models exhibited varying degrees of enhancement for each class. Specifically, IMCMD_YOLOv8_small improved the AP values by more than 10% in the categories of pedestrians, people, and motors, whereas the IMCMD_YOLOv8_large model increased the AP values by more than 10% in the categories of pedestrians, people, bicycles, bicycle, and motors. These substantial performance improvements in the individual categories further validate the effectiveness of the improved models in detecting small objects.

Fig. 15 shows the variation curves for the precision, recall, mAP@0.5, and mAP@0.5:0.95 of the improved models IMCMD_YOLOv8_small and IMCMD_YOLOv8_large compared with the baseline model YOLOv8s during training. The graphs indicate that all three models eventually reached a state of convergence as the number of epochs increased; however, the improved models IMCMD_YOLOv8_small and IMCMD_YOLOv8_large consistently outperformed YOLOv8s across all four metrics. In the early stages of training, the two improved models exhibited higher mAP values, and they maintained relatively stable growth throughout the training process. Compared with YOLOv8s,

Table 10
Comparison between the improved model and YOLOv8s.

Models	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameters (10^6)	GFLOPs	Model size (M)
YOLOv8s	48.3	38.8	39.1	23.5	11.1	28.7	21.4
IMCMD_YOLOv8_small (ours)	56.4	44.5	46.8	28.6	3.0	41.8	6.1
IMCMD_YOLOv8_large (ours)	60.0	46.6	49.9	30.8	5.8	107.6	11.3

Table 11
Comparison between the improved model and other versions of YOLOv8.

Models	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameters (10^6)	GFLOPs	Model size (M)
YOLOv8n	43.1	33.1	32.6	18.9	3.0	8.2	6.0
YOLOv8m	54.2	41.4	42.8	26.0	25.9	79.1	49.6
YOLOv8l	56.8	42.9	44.7	27.5	43.6	165.4	83.5
YOLOv8x	57.5	42.7	45.1	27.9	68.2	258.2	130.0
IMCMD_YOLOv8_small (ours)	56.4	44.5	46.8	28.6	3.0	41.8	6.1
IMCMD_YOLOv8_large (ours)	60.0	46.6	49.9	30.8	5.8	107.6	11.3

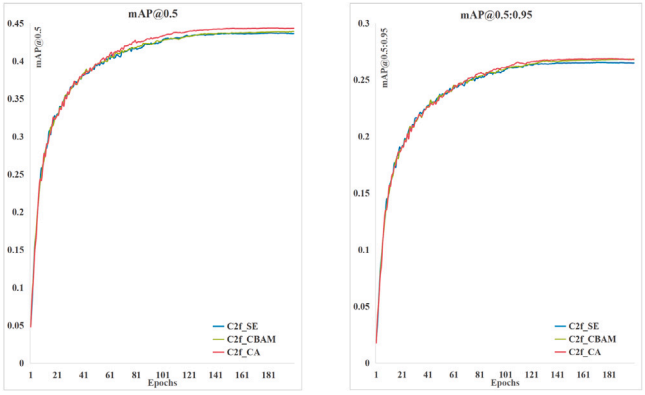

Fig. 13. Training curves of three attention mechanism modules in terms of mAP@0.5 and mAP@0.5:0.95.

our improved models trained faster and delivered better detection performance.

(2) Comparison with Other Versions of YOLOv8

To further demonstrate the effectiveness of the improvement strategy, the improved models were compared with other versions of YOLOv8 (YOLOv8n, YOLOv8 m, YOLOv8l, and YOLOv8x). Table 11 presents the results. The results indicate that, compared with other models, the two model sizes proposed in this study perform best in terms of recall, mAP@0.5, mAP@0.5:0.95, and parameters. Comparing IMCMD_YOLOv8_large, which is our largest model, with YOLOv8x, which is the largest model in the YOLOv8 series, our model showed improvements of 2.5%, 3.9%, 4.8%, and 2.9% in the four core evaluation metrics of precision, recall, mAP@0.5, and mAP@0.5:0.95, respectively. More importantly, in addition to these significant performance improvements, the IMCMD_YOLOv8_large model substantially reduced the parameter count, GFLOPs, and model size, which decreased by 91.5%, 58.3%, and 91.3%, respectively. These results not only highlight the advantages of the improved model in terms of both enhancing detection performance and model lightweighting.

Fig. 16 shows the variation curves of some key evaluation metrics during the training process for the improved models IMCMD_YOLOv8_small and IMCMD_YOLOv8_large, compared with the other versions of YOLOv8. The graph indicates that the improved models IMCMD_YOLOv8_small and IMCMD_YOLOv8_large consistently outperformed the other versions of YOLOv8 in terms of recall, mAP@0.5, and mAP@0.5:0.95. IMCMD_YOLOv8_large consistently performed well throughout the training process, whereas IMCMD_YOLOv8_small started with a mAP value lower than that of YOLOv8x in the early stages of training but surpassed YOLOv8x around epoch 70 and maintained stable growth throughout the subsequent training.

4.3.4. Ablation study

To verify the effectiveness of each improvement strategy proposed in this study, ablation experiments were conducted on the VisDrone2019 dataset. Tables 12 and 13 present the. The results demonstrate that each improvement strategy applied to the baseline model achieved varying degrees of performance enhancement.

(1) Optimized network structure strategy: For small object detection tasks, the removal of the P5 layer from the backbone and the fusion of features from the P4, P3, and P2 layers, with P3 and P2 serving as the detection heads, resulted in increases of 4.5% and 3.1% in mAP@0.5 and mAP@0.5:0.95, respectively. This indicates that simplifying the model structure and focusing on lower-level detail information to capture more small object features can effectively enhance the detection accuracy of small objects.

(2) Design of C2f_CA module strategy: Replacing the C2f module in the backbone with the C2f_CA module led to improvements of 0.8% and 0.3% in the mAP@0.5 and mAP@0.5:0.95, respectively. This result proves that the efficient CA mechanism can better capture key information, and that the removal of the residual structure allows the model to obtain a richer gradient information flow, thereby enhancing the detection accuracy of small objects.

(3) Design of AMFF module strategy: When the channel numbers of P2 and P3 were 64, adding the AMFF module to the neck resulted in increases of 2.0% and 1.2% in mAP@0.5 and mAP@0.5:0.95, respectively. When the channel numbers of P2 and P3 were 128, the increases were 5.3% and 3.5%, respectively. This demonstrates that the improved feature fusion strategy in this study enhances the effective integration of high- and low-level feature information, whereas the SKA mechanism strengthens the focus on features that are crucial for detection tasks. This enables the model to adaptively concentrate on the most useful features for the current task and further improves the model's detection performance.

(4) Introduction of Dynamic Head strategy: Implementing a Dynamic Head with multiple attention mechanisms when the channel numbers of P2 and P3 were 64 led to increases of 1.7% and 1.1% in mAP@0.5 and mAP@0.5:0.95, respectively. When the channel numbers were 128, the increases were 5.2% and 3.4%, respectively. This further confirms that the multiple attention mechanisms of the Dynamic Head improve the focus on densely packed small object areas and therefore extract small object features more effectively.

4.3.5. Comparison with other algorithms

Two sets of comparative experiments were conducted to demonstrate the superiority and effectiveness of the proposed improved algorithm. The first set of experiments compared the proposed models with certain YOLO-series algorithms, as presented in Table 14. The variation curves of some of the key evaluation metrics during the training process are shown in Fig. 17. In the second set, the proposed models were

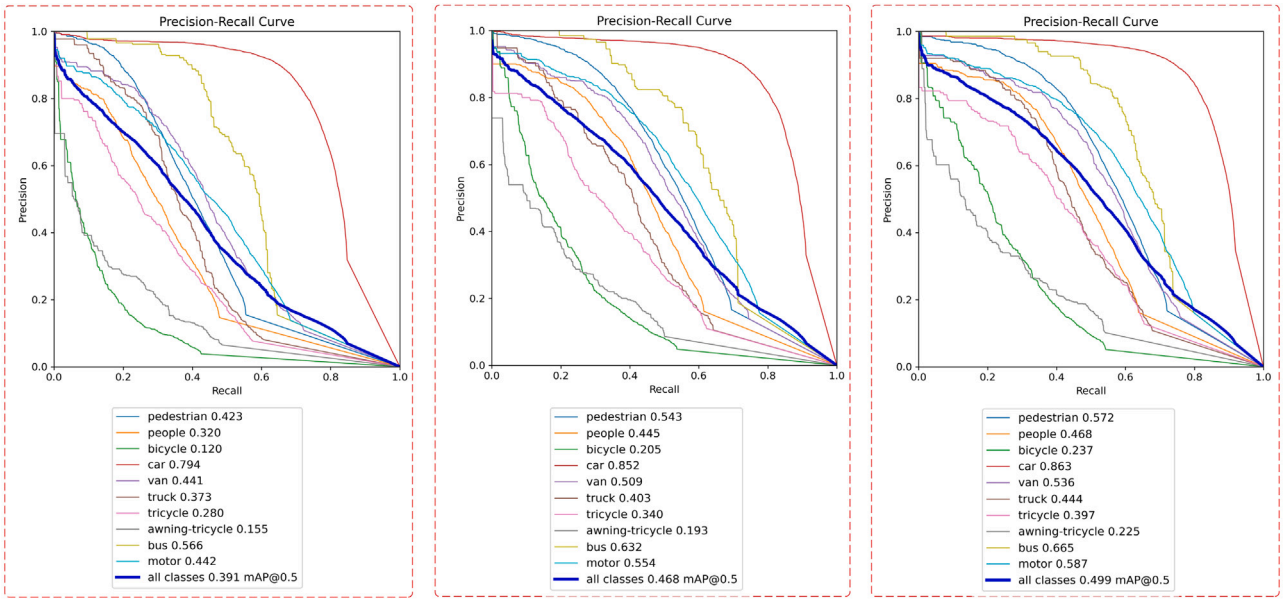

Fig. 14. Precision–Recall curve for YOLOv8s, IMCMD_YOLOv8_small, and IMCMD_YOLOv8_large.

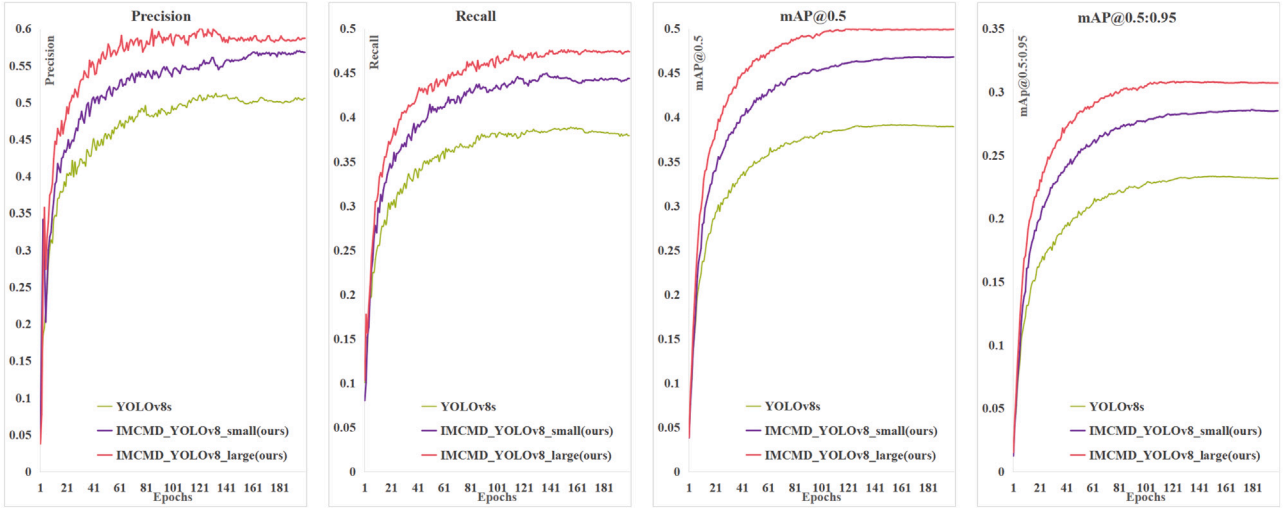

Fig. 15. Training curves for YOLOv8s, IMCMD_YOLOv8_small, and IMCMD_YOLOv8_large in precision, recall, mAP@0.5, and mAP@0.5:0.95.

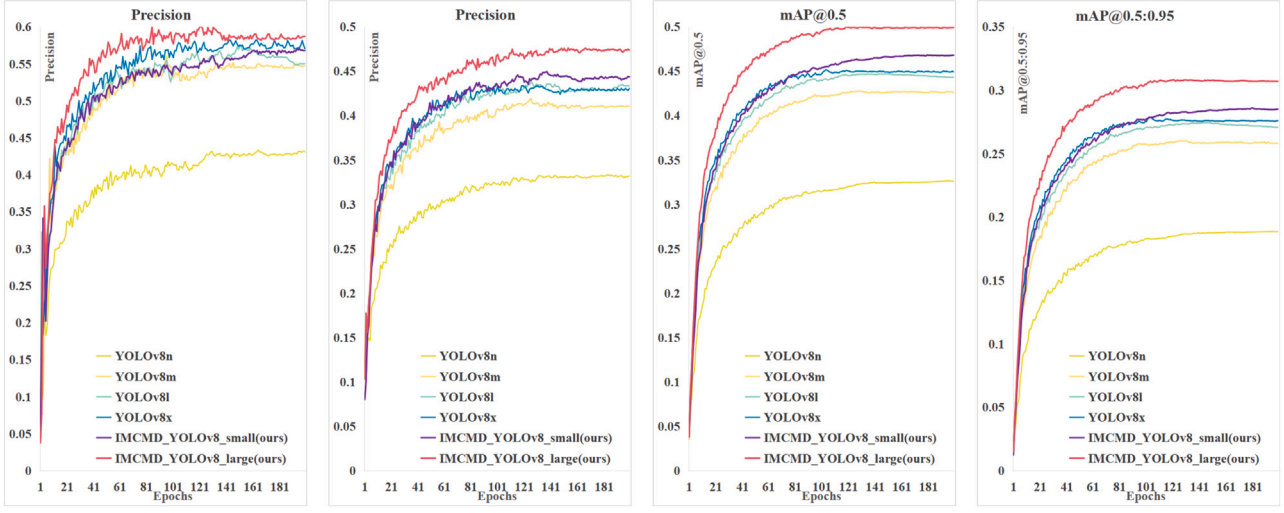

Fig. 16. Training curves for IMCMD_YOLOv8_small, IMCMD_YOLOv8_large, and other versions of YOLOv8 in precision, recall, mAP@0.5, and mAP@0.5:0.95.

Table 12
Different improvement strategies.

Improvement strategy	Improve network structure	Add C2f_CA module	Add AMFF module	Improved dynamic head detection head
YOLOv8s	×	×	×	×
IM_YOLOv8	✓	×	×	×
IMC_YOLOv8	✓	✓	×	×
IMM_YOLOv8_small	✓	×	✓	×
IMM_YOLOv8_large	✓	×	✓	×
IMD_YOLOv8_small	✓	×	×	✓
IMD_YOLOv8_large	✓	×	×	✓
IMCMD_YOLOv8_small	✓	✓	✓	✓
IMCMD_YOLOv8_large	✓	✓	✓	✓

Table 13
Experimental results after introducing different improvement strategies.

Models	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameters (10^6)	GFLOPs	Model size (M)
YOLOv8s	48.3	38.8	39.1	23.5	11.1	28.7	21.4
IM_YOLOv8	54.3	41.7	43.6	26.6	2.3	27.3	4.7
IMC_YOLOv8	53.8	42.3	44.4	26.9	2.5	28.3	5.1
IMM_YOLOv8_small	55.5	42.9	45.6	27.8	2.6	38.7	5.2
IMM_YOLOv8_large	58.9	45.9	48.9	30.1	4.6	102.5	9.1
IMD_YOLOv8_small	54.7	43.1	45.3	27.7	2.4	27.0	4.9
IMD_YOLOv8_large	57.2	46.7	48.8	30.0	3.9	51.8	7.7
IMCMD_YOLOv8_small	56.4	44.5	46.8	28.6	3.0	41.8	6.1
IMCMD_YOLOv8_large	60.0	46.6	49.9	30.8	5.8	107.6	11.3

Table 14
Comparison between the improved model and YOLO series algorithms.

Models	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameters (10^6)	GFLOPs	Model size (M)
YOLOv3	55.5	43.6	44.8	27.5	103.7	283.0	198.0
YOLOv3-Tiny	39.9	24.0	23.7	13.3	12.1	19.1	23.2
YOLOv5s	45.3	33.3	32.6	17.4	7.0	16.0	13.7
YOLOv6s	47.3	35.6	36.3	21.7	16.3	44.2	31.3
IMCMD_YOLOv8_small (ours)	56.4	44.5	46.8	28.6	3.0	41.8	6.1
IMCMD_YOLOv8_large (ours)	60.0	46.6	49.9	30.8	5.8	107.6	11.3

Table 15
Comparisons between the improved and classic models.

Models	mAP@0.5 (%)	mAP@0.5:0.95 (%)
Faster R-CNN	37.2	21.9
RetinaNet	19.1	10.6
Cascade R-CNN	39.1	24.3
CenterNet	33.7	18.8
FSAF	36.5	20.9
ATSS	36.4	22.3
IMCMD_YOLOv8_small (ours)	46.8	28.6
IMCMD_YOLOv8_large (ours)	49.9	30.8

compared with other high-performance models, and these results are presented in Table 15.

(1) Comparisons with YOLO Series Models

Since its introduction, the YOLO series has achieved significant success in the field of object detection with continuous iterations leading to several excellent versions. YOLOv3 introduces the DarkNet-53 network structure, drawing on ResNet's residual structure to enhance the feature extraction capability by stacking more layers. It also implements multi-size inputs and outputs to increase the speed and accuracy of object detection, particularly for small object detection capabilities. YOLOv3-Tiny, which is a lightweight version of YOLOv3, improves the detection speed at the expense of accuracy. YOLOv4 uses the CSP DarkNet-53 (Wang et al., 2020) network architecture and incorporates the SPP-Net structure to accommodate different input sizes. The PANet structure strengthens feature fusion to effectively enhance model performance and adaptability. YOLOv5 made lightweight improvements while maintaining its detection performance. YOLOv6 employs the EfficientRep lightweight network as its base network and uses a feature pyramid network (Rep-PAN) to achieve feature fusion while maintaining good multiscale feature fusion capabilities.

According to the experimental results listed in Table 14, the improved models IMCMD_YOLOv8_small and IMCMD_YOLOv8_large demonstrated significant performance advantages when compared with the other versions of the YOLO series. These improvements include breakthroughs in model lightweighting and substantially enhanced detection accuracy, thereby proving the effectiveness of the proposed methods in optimizing small object detection.

YOLOv3, despite ensuring good detection accuracy, is not suitable for direct deployment in resource-constrained environments, such as drone platforms and mobile devices, owing to its complex structure, high parameter count, and significant computational demand. In contrast, the IMCMD_YOLOv8_small model reduced the parameter count and computational demand by 97.1% and 85.2%, respectively, while achieving a 4.4% increase in mAP@0.5. The IMCMD_YOLOv8_large model decreased its parameters and computations by 94.4% and 70.0%, respectively, with an 11.4% improvement in mAP@0.5.

Although efforts have been made to lighten models, such as YOLOv3-Tiny, YOLOv5s, and YOLOv6s, they generally have lower detection accuracy. In comparison, the IMCMD_YOLOv8_small model reduced the parameters by 75.2%, 57.1%, and 81.6% compared with these models, and it improved the mAP@0.5 by 97.5%, 4.36%, and 28.9%, respectively. The IMCMD_YOLOv8_large model reduced the parameters by 52.0%, 17.1%, and 64.4% while increasing the mAP@0.5 by 110.5%, 53.0%, and 37.1%, respectively. These results demonstrate that the improved models in this study not only achieve lightness but also enhance detection performance.

Fig. 17 shows the variation curves of some key evaluation metrics during the training process for the improved models IMCMD_YOLOv8_small and IMCMD_YOLOv8_large compared with some

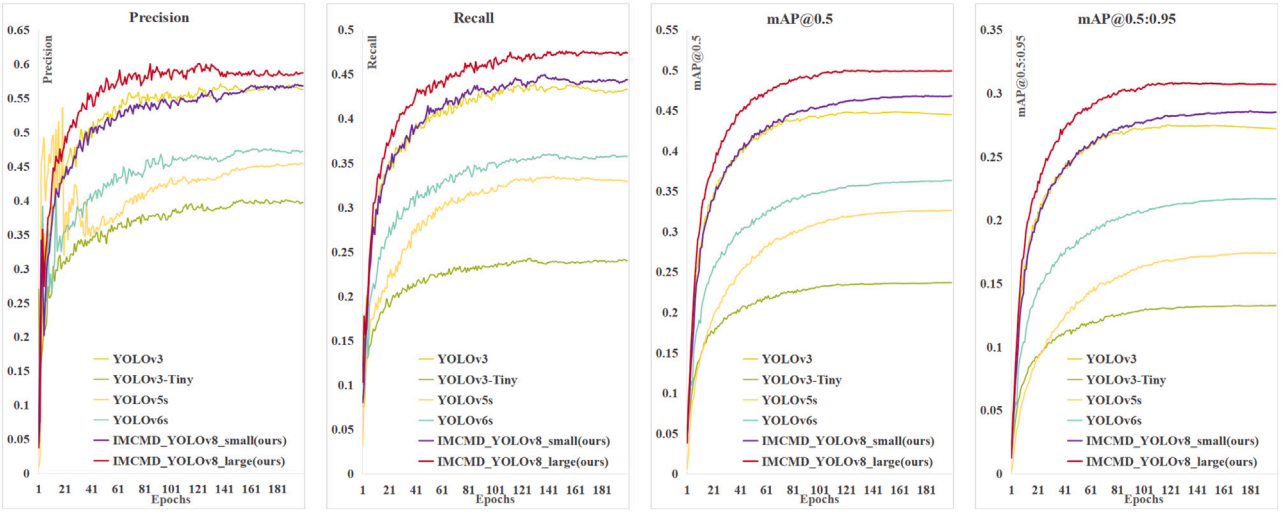

Fig. 17. Training curves for IMCMD_YOLOv8_small, IMCMD_YOLOv8_large, and select YOLO models for precision, recall, mAP@0.5, and mAP@0.5:0.95.

YOLO models. This graph clearly shows that the improved models IMCMD_YOLOv8_small and IMCMD_YOLOv8_large outperform the other YOLO models in terms of precision, recall, mAP@0.5, and mAP@0.5:0.95, maintain stable growth throughout the training process, and exhibit superior performance.

(2) Comparisons with Other Classic Models

The IMCMD_YOLOv8_small and IMCMD_YOLOv8_large models proposed in this study demonstrated significant performance advantages over other classic and efficient object detection algorithms for small object detection tasks. Although these classic models have achieved significant results in specific fields, they still have certain limitations in handling small object detection.

Faster R-CNN, which is a classic and efficient two-stage algorithm, achieves high-precision detection performance by introducing a region proposal network. However, the resolution of the feature maps extracted by the backbone network is relatively low, resulting in poorer accuracy for small-sized feature maps. The core of the RetinaNet algorithm (Lin et al., 2017b) is the introduction of a focal loss function that effectively solves the problem of imbalance between positive and negative samples and improves the detection accuracy of single-stage algorithms. However, the performance of RetinaNet is limited to tasks primarily involving small object detection. Cascade R-CNN (Cai and Vasconcelos, 2018) enhances the detection performance through a multistage detection architecture; however, this also increases the computational complexity and training difficulty. CenterNet (Duan et al., 2019), despite reducing computation by eliminating anchor points, may lead to detection losses in dense scenes and small object occlusions owing to overlapping predicted centroids. FSAF (Zhu et al., 2019) addresses two limitations of anchor-based detection, and, despite using multiscale feature fusion, still has limited detection performance for small objects. ATSS (Zhang et al., 2020) optimizes the performance of anchor-based and anchor-free detectors by automatically selecting positive and negative samples, which results in good versatility.

According to the experimental results shown in Table 15, the improved models proposed in this study surpassed the aforementioned classic models in terms of detection accuracy. IMCMD_YOLOv8_small and IMCMD_YOLOv8_large not only significantly reduced parameter counts in tasks involving small object detection but also showed notable improvements in key performance metrics, such as mAP@0.5, thereby proving the effectiveness of the improvement strategies in enhancing small object detection performance.

Through a comparative analysis of these two sets of experiments, the IMCMD_YOLOv8_small and IMCMD_YOLOv8_large models proposed in this study were shown to significantly outperform other classic object detection models in terms of performance. The improved models not only demonstrated significant improvements in detection accuracy in small object detection tasks but also made important progress in model lightweighting. These experimental results validate the effectiveness of the improved models and demonstrate their potential for application in the field of object detection.

4.3.6. *Visual analysis*

The confusion matrix is an intuitive tool for evaluating the performance of classification models, particularly for multi-category object detection tasks. By comparing the prediction results of the model with the true labels, the confusion matrix clearly shows the prediction accuracy and error situations of the model for each category. In this study, the confusion matrices for the YOLOv8s, IMCMD_YOLOv8_small, and IMCMD_YOLOv8_large models were drawn, as shown in Figs. 18–20. In the confusion matrix, the rows represent the categories predicted by the model, and the columns correspond to the true categories. The values on the diagonal of the matrix indicate instances of correct predictions, whereas those off the diagonal represent cases with incorrect predictions. Observations from the confusion matrices in Figs. 18–20 show that the values in the diagonal area for the IMCMD_YOLOv8_small and IMCMD_YOLOv8_large models were significantly improved, compared with those of the YOLOv8s model. These findings indicate that the improved models exhibited enhanced performance in correctly identifying objects in various categories as well as a corresponding reduction in incorrect predictions.

Through a comparative analysis of the confusion matrices, the effectiveness of the improved models proposed in this study for enhancing the object detection accuracy can be visually verified. This performance improvement is attributed to the optimization of the model structure, introduction of attention mechanisms, feature fusion strategies, and improvements to the detection head. These improvement strategies work together to enhance the capability of the model to capture the features of small objects and improve the classification accuracy.

To visually demonstrate the detection performance of the improved models compared with the baseline model in complex scenes, detection comparisons were conducted on the VisDrone2019 dataset for four different types of scenes: road traffic scenes, aerial shots, dim night scenes, and dense multi-object scenes, as shown in Figs. 21–24, respectively. To clearly assess and compare the performance of the models, detected objects that were incorrectly identified are marked with red ellipses, whereas missed objects are circled in blue in the detection results.

The detection results in Figs. 21–24 show that the improved IMCMD_YOLOv8_small and IMCMD_YOLOv8_large models demonstrate significant advantages in handling these complex scenes. Compared

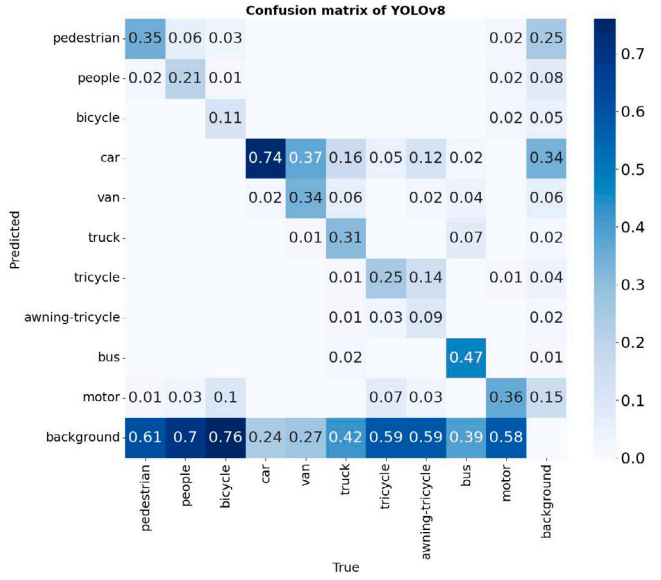

Fig. 18. Confusion matrix plot of YOLOv8s.

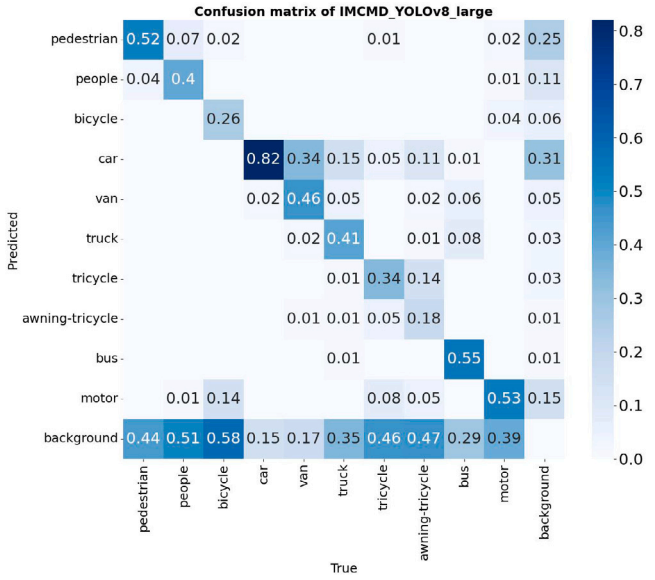

Fig. 20. Confusion matrix plot of IMCMD_YOLOv8_large.

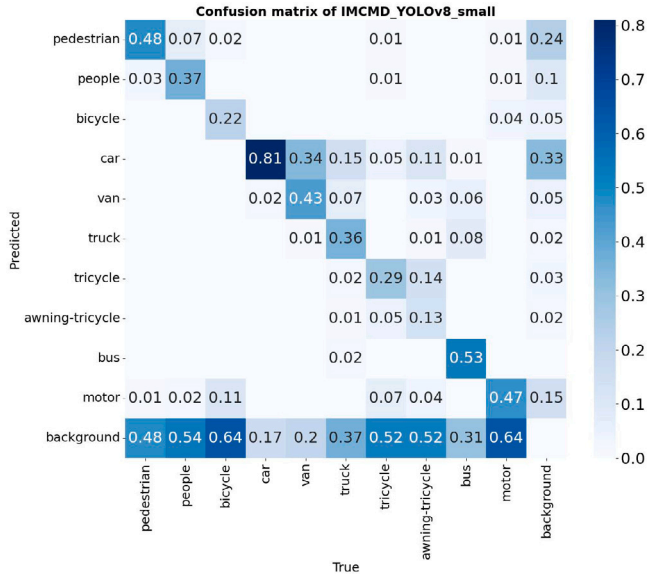

Fig. 19. Confusion matrix plot of IMCMD_YOLOv8_small.

with the baseline model, the improved models can detect smaller and more distant targets, which significantly reduces the rate of false positives and false negatives for occluded and dense small objects, and effectively enhances the detection accuracy and reliability. This comparison illustrates that the improved models can more effectively capture the fine details of small objects and deal with complex backgrounds.

To verify the general applicability of our two improved models, we conducted object detection on the DIOR (Li et al., 2020) mainstream small object dataset for targets in the same categories as those in VisDrone. The detection results are presented in Figs. 25 and 26. The results show that the proposed IMCMD_YOLOv8_small and IMCMD_YOLOv8_large models demonstrate excellent detection performance on this new dataset, proving that our improved models possess a certain level of generalization capability and robustness and that they are capable of effectively completing small object detection tasks.

5. Conclusion

By addressing the issues of poor performance and high miss rates in detecting small objects with general object detection algorithms, we successfully enhanced the precision of small object detection by improving the YOLOv8 algorithm. First, we optimized the network structure by removing the backbone's P5 layer, which is mainly used to detect large objects, and focused on fusing the P4, P3, and P2 layers, which have higher sensitivities to medium and small objects. We also used the P3 and P2 layers as detection heads for small objects to enhance their detection capabilities. Then, by integrating the efficient CA attention mechanism into the backbone's C2f module and designing the C2f_CA module, we strengthened the model's focus on key information in the feature maps, thereby comprehensively improving the detection performance. Additionally, by fusing shallow details and deep semantic features, we significantly reduced the miss rate for small objects and further strengthened the model's ability to detect small targets. Finally, by employing a Dynamic Head that integrates scale, spatial, and task awareness into the head network of the model, we further optimized the overall performance of the model. According to a series of experimental validations, the improved IMCMD_YOLOv8_small model achieved increases of 8.1%, 5.7%, 7.7%, and 5.1% in precision, recall, mAP@0.5, and mAP@0.5:0.95, respectively, as well as a 73.0% reduction in the parameter count, compared with the baseline model YOLOv8s. The IMCMD_YOLOv8_large model showed even more significant improvements in the aforementioned performance metrics: 11.7%, 7.8%, 10.8%, and 7.3%, respectively, with a 47.7% reduction in the parameter count. These significant performance improvements, coupled with the substantial reductions in parameters, demonstrate the effectiveness of the improvement strategies. In tasks involving small object detection, the improved models outperformed several classic models.

Our improved models achieved significantly enhanced accuracy, and the parameter counts and model sizes were substantially reduced; however, both models exhibited varying increases in GFLOPs. In the future, we plan to continue exploring how to further improve the detection accuracy of small objects in aerial images without increasing the GFLOPs.

Fig. 21. Comparison of detection effects of YOLOv8s (left), IMCMD_YOLOv8_small (middle), and IMCMD_YOLOv8_large (right) in traffic road scenes.

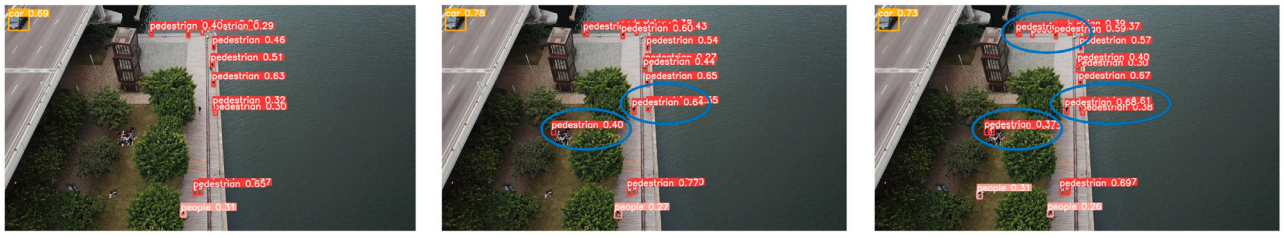

Fig. 22. Comparison of detection effects of YOLOv8s (left), IMCMD_YOLOv8_small (middle), and IMCMD_YOLOv8_large (right) in high-altitude shooting scenes.

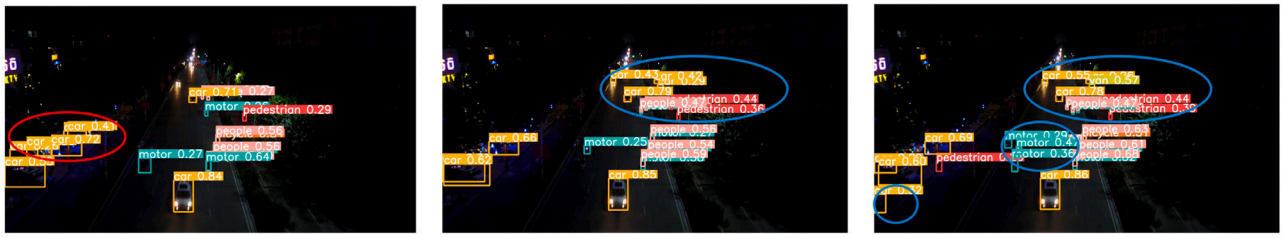

Fig. 23. Comparison of detection effects of YOLOv8s (left), IMCMD_YOLOv8_small (middle), and IMCMD_YOLOv8_large (right) in dim night scenes.

Fig. 24. Comparison of detection effects of YOLOv8s (left), IMCMD_YOLOv8_small (middle), and IMCMD_YOLOv8_large (right) in dense multi-target scenes.

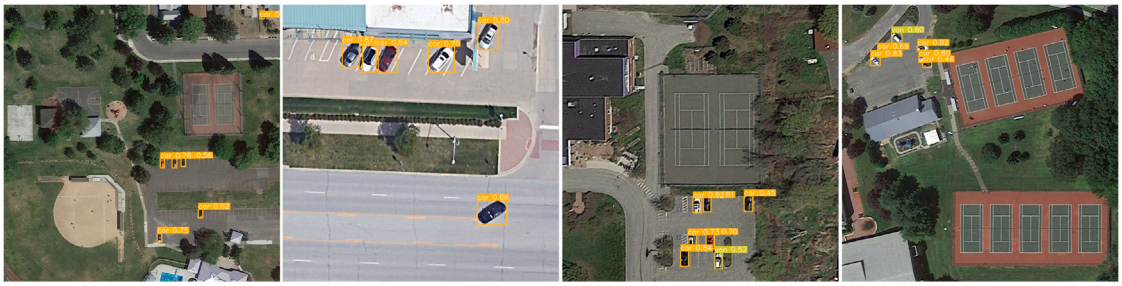

Fig. 25. Detection performance of IMCMD_YOLOv8_small on the DIOR dataset.

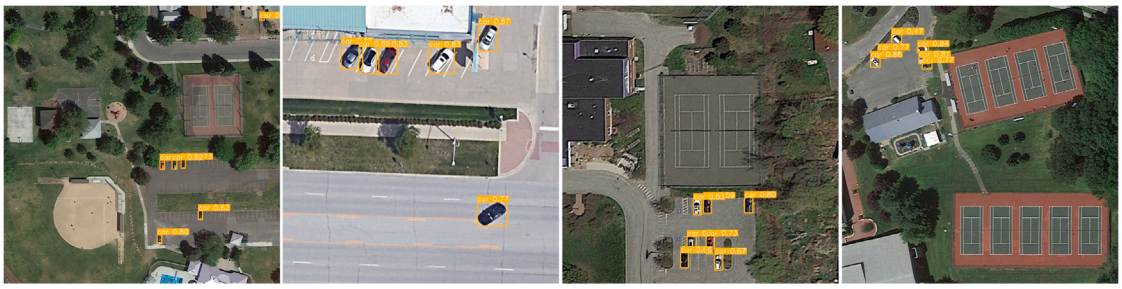

Fig. 26. Detection performance of IMCMD_YOLOv8_large on the DIOR dataset.

CRediT authorship contribution statement

Fei Feng: Writing – original draft, Validation, Methodology, Conceptualization. **Yu Hu:** Validation, Funding acquisition, Conceptualization. **Weipeng Li:** Writing – review & editing, Software, Data curation. **Feiyan Yang:** Writing – review & editing, Project administration, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Guizhou Provincial Department of Education's General Undergraduate College Scientific Research Project (Youth Project) (No: Qian Jiao Ji [2022]366).

References

Asadzadeh, S., de Oliveira, W.J., de Souza Filho, C.R., 2022. UAV-based remote sensing for the petroleum industry and environmental monitoring: State-of-the-art and perspectives. J. Pet. Sci. Eng. 208, 109633.

Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

Cai, Z., Vasconcelos, N., 2018. Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6154–6162.

Chen, W., Jia, X., Zhu, X., Ran, E., Xie, H., 2023. Target detection in UAV aerial images based on DSM-YOLO v5. Comput. Eng. Appl. 59 (18), 226–233.

Dai, X., Chen, Y., Xiao, B., et al., 2021. Dynamic head: Unifying object detection heads with attentions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7373–7382.

Dai, J., Li, Y., He, K., et al., 2016. R-fcn: Object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems. Vol. 29.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR'05, Vol. 1, pp. 886–893.

Duan, K., Bai, S., Xie, L., et al., 2019. Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6569–6578.

Ghiasi, G., Lin, T.Y., Le, Q.V., 2019. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7036–7045.

Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448.

Girshick, R., Donahue, J., Darrell, T., et al., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587.

Hafeez, A., Husain, M.A., Singh, S.P., et al., 2022. Implementation of drone technology for farm monitoring & pesticide spraying: A review. Inf. Process. Agric..

He, J., Shao, L., Li, Y., et al., 2023. Pavement damage identification and evaluation in UAV-captured images using gray level co-occurrence matrix and cloud model. J. King Saud Univ.-Comput. Inf. Sci. 35 (9), 101762.

He, K., Zhang, X., Ren, S., et al., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37 (9), 1904–1916.

Hou, Q., Zhou, D., Feng, J., 2021. Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13713–13722.

Hu, C., He, Y., Savides, M., 2019. Feature selective anchor-free module for single-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 840–849.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Huan, H., Li, P., Zou, N., et al., 2021. End-to-end super-resolution for remote-sensing images using an improved multi-scale residual network. Remote Sens. 13 (4), 666.

Huang, L., Chen, C., Yun, J., et al., 2022. Multi-scale feature fusion convolutional neural network for indoor small target detection. Front. Neurorobot. 16, 881021.

Hui, Y., Wang, J., Li, B., 2024. DSAA-YOLO: UAV remote sensing small target recognition algorithm for YOLOV7 based on dense residual super-resolution and anchor frame adaptive regression strategy. J. King Saud Univ.-Comput. Inf. Sci. 36 (1), 101863.

Kisantal, M., Wojna, Z., Murawski, J., et al., 2019. Augmentation for small object detection. arXiv preprint arXiv:1902.07296.

Kong, T., Sun, F., Liu, H., et al., 2020. Foveabox: Beyound anchor-based object detection. IEEE Trans. Image Process. 29, 7389–7398.

Kumar, A., Krishnamurthi, R., Nayyar, A., et al., 2021. A novel software-defined drone network (SDDN)-based collision avoidance strategies for on-road traffic monitoring and management. Veh. Commun. 28, 100313.

Li, C., Li, L., Jiang, H., et al., 2022. YOLOv6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976.

Li, X., Wang, W., Hu, X., et al., 2019. Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 510–519.

Li, K., et al., 2020. Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS J. Photogramm. Remote Sens. 159, 296–307.

Lim, J.S., Astrid, M., Yoon, H.J., et al., 2021. Small object detection using context and attention. In: 2021 International Conference on Artificial Intelligence in Information and Communication. ICAIIC, pp. 181–186.

Lin, T.Y., Dollár, P., Girshick, R., et al., 2017a. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125.

Lin, T.Y., Goyal, P., Girshick, R., et al., 2017b. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Liu, S., Qi, L., Qin, H., et al., 2018. Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8759–8768.

Lou, H., Duan, X., Guo, J., et al., 2023. DC-YOLOv8: Small-size object detection algorithm based on camera sensor. Electronics 12 (10), 2323.

Mohsan, S.A.H., Othman, N.Q.H., Li, Y., et al., 2023. Unmanned aerial vehicles (UAVs): Practical aspects, applications, open challenges, security issues, and future trends. Intell. Serv. Robot. 16 (1), 109–137.

Oneto, L., 2020. Model Selection and Error Estimation in a Nutshell. Springer International Publishing.

Peng, Y., Sonka, M., Chen, D.Z., 2023. U-net v2: Rethinking the skip connections of U-net for medical image segmentation. arXiv preprint arXiv:2311.17791.

Qi, X., Chai, R., Gao, Y., 2023. Small target detection algorithm with reconstructed SPPCSPC and optimized downsampling. Comput. Eng. Appl. 59 (20), 158–166.

Redmon, J., Divvala, S., Girshick, R., et al., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Redmon, J., Farhadi, A., 2017. YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7263–7271.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ren, S., He, K., Girshick, R., et al., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

Shang, J., Wang, J., Liu, S., et al., 2023. Small target detection algorithm for UAV aerial photography based on improved YOLOv5s. Electronics 12, 2434.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Vol. 1, pp. I–I.

Wang, H., Cao, J., Qiu, C., Liu, Y., 2022. Multi-target detection method in aerial images based on improved YOLOv4. Electro-Opt. Control 29 (05), 23–27.

Wang, G., Chen, Y., An, P., et al., 2023c. UAV-YOLOv8: a small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios. Sensors 23 (16), 7190.

Wang, C.Y., Liao, H.Y.M., Wu, Y.H., et al., 2020. CSPNet: A new backbone that can enhance learning capability of CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 390–391.

Wang, J., Zhang, F., Zhang, Y., et al., 2023b. Lightweight object detection algorithm for UAV aerial imagery. Sensors 23 (13), 5786.

Woo, S., Park, J., Lee, J.Y., et al., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.

Yang, C., Huang, Z., Wang, N., 2022. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13668–13677.

Zhang, S., Chi, C., Yao, Y., et al., 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9759–9768.

Zhu, C., He, Y., Savvides, M., 2019. Feature selective anchor-free module for single-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 840–849.

Zhu, X., Lyu, S., Wang, X., et al., 2021a. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2778–2788.

Zhu, P., Wen, L., Du, D., et al., 2021b. Detection and tracking meet drones challenge. IEEE Trans. Pattern Anal. Mach. Intell. 44 (11), 7380–7399.