Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.066368

ARTICLE

YOLOv8s-DroneNet: Small Object Detection Algorithm Based on Feature Selection and ISIoU

Jian Peng[1], **Hui He**[2] **and Dengyong Zhang**[2,*]

[1]Elite Engineering School, Changsha University of Science and Technology, Changsha, 410000, China
[2]School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, 410000, China
*Corresponding Author: Dengyong Zhang. Email: zhdy@csust.edu.cn

Received: 07 April 2025; Accepted: 10 June 2025; Published: 30 July 2025

ABSTRACT: Object detection plays a critical role in drone imagery analysis, especially in remote sensing applications where accurate and efficient detection of small objects is essential. Despite significant advancements in drone imagery detection, most models still struggle with small object detection due to challenges such as object size, complex backgrounds. To address these issues, we propose a robust detection model based on You Only Look Once (YOLO) that balances accuracy and efficiency. The model mainly contains several major innovation: feature selection pyramid network, Inner-Shape Intersection over Union (ISIoU) loss function and small object detection head. To overcome the limitations of traditional fusion methods in handling multi-level features, we introduce a Feature Selection Pyramid Network integrated into the Neck component, which preserves shallow feature details critical for detecting small objects. Additionally, recognizing that deep network structures often neglect or degrade small object features, we design a specialized small object detection head in the shallow layers to enhance detection accuracy for these challenging targets. To effectively model both local and global dependencies, we introduce a Conv-Former module that simulates Transformer mechanisms using a convolutional structure, thereby improving feature enhancement. Furthermore, we employ ISIoU to address object imbalance and scale variation This approach accelerates model conver-gence and improves regression accuracy. Experimental results show that, compared to the baseline model, the proposed method significantly improves small object detection performance on the VisDrone2019 dataset, with mAP@50 increasing by 4.9% and mAP@50-95 rising by 6.7%. This model also outperforms other state-of-the-art algorithms, demonstrating its reliability and effectiveness in both small object detection and remote sensing image fusion tasks.

KEYWORDS: Drone imagery; small object detection; feature selection; convolutional attention

1 Introduction

Drone technology has seen rapid advancements in recent years, with these small, versatile devices increasingly deployed for a wide range of domain-specific tasks due to their mobility and adaptability. As a result, drones are being utilized across various applications, including intelligent traffic monitoring [1–4], disaster detection, wildlife monitoring [5], and environmental surveillance [6,7]. With the rise of deep learning, there have been significant advancements in general object detection. However, the detection of small objects within drone imagery remains an ongoing challenge, particularly as small objects are often obscured by noise, occlusions, and the complexity of dynamic environments.

Small Object Detection (SOD) focuses on detecting objects that are smaller in size. Two primary definitions for small objects are commonly used: one based on relative size, as specified by the International

Society for Optical Engineering, where an object is considered small if its size in the image is less than 0.12% of the image; and another way is based on absolute size, it categorizes objects smaller than 32 × 32 pixels as small objects, as defined in MS COCO dataset [8]. The progress in SOD is relatively slower due to the inherent challenges posed by the low resolution and smaller pixel proportions of small objects. These factors make it difficult to extract representation of features, and the downsample operations during feature extraction often lead to critical information loss [9]. Moreover, drone images present unique challenges for small object detection, including varying weather conditions, dense and cluttered environments, occlusion, and differing viewing angles [10].

To address these challenges, this paper proposes the YOLOv8s-DroneNet framework, an efficient model designed to improve the detection accuracy for small objects in Unmanned Aerial Vehicle (UAV) imagery. Fig. 1 shows the overall architecture of the network, which consists of four key parts: feature selection pyramid network, Conv-Former module, a specialized small object detection head and the Inner-Shape IoU. The primary contributions of this paper are summarized as follows:

1. We propose a robust small object detection framework, aiming to improve the accuracy of small object detection in complex environments. The framework includes a feature selection pyramid network that effectively integrates features from different scales and enhances the fusion of features from the network. To enhance the representation capability of small object features, a convolutional attention mechanism is adopted to mimic the Transformer structure, achieving better balance between local and global feature modeling. Additionally, a specialized small object detection head is introduced at the shallow layer of the network architecture, further improving the performance of small object detection.

2. In order to strengthen the sensitivity of small object shapes and scale variation, this paper designs Inner-Shape IoU. Using IS-IoU calculate loss is useful to increase the attention of the model to small targets.

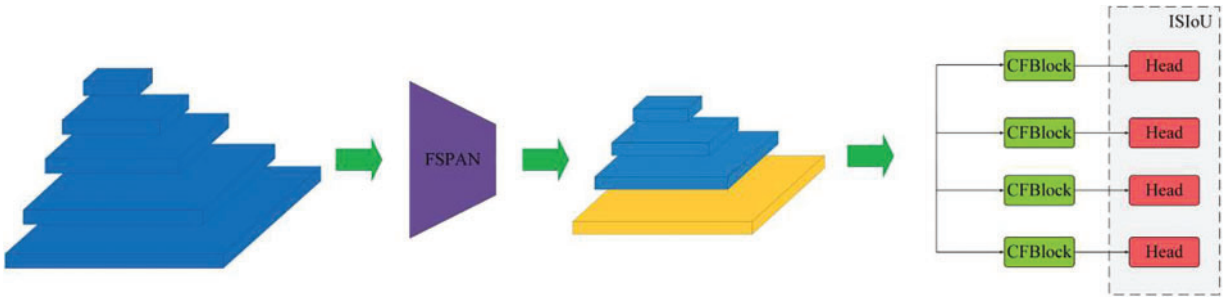

Figure 1: YOLOv8s-DroneNet. This figure illustrates the architecture of the proposed YOLOv8s-DroneNet, designed specifically for small object. The model integrates several advanced components aimed at improving the detection of small objects

2 Related Work

SOD remains a significant challenge in computer vision. Recently, it has been widely applied in various fields, including remote sensing images, medical images, and UAV-based scene analysis. There are two major object detection algorithms: two-stage and one-stage methods. The two-stage approach generates a set of candidate regions and then extracts features for subsequent classification and localization. Representative two-stage detection models include R-CNN [11] and its improved variant, Faster R-CNN [12], which have demonstrated strong detection performance but at the cost of higher computational complexity. Conversely, one-stage methods, such as YOLOv8 [13], YOLOv10 [14], and SSD [15], predict object categories and bounding box coordinates directly in a single forward pass. These models are favoured for their lower

computational cost and real-time performance. However, despite their efficiency, one-stage detectors often face challenges in achieving high detection accuracy, especially when dealing with objects of varying scales. This problem is particularly evident in UAV imagery, where small objects are common, and the multi-scale nature of objects further complicates detection.

2.1 Small Object Detection

Numerous small object detection algorithms have been optimized within the YOLO framework in recent years. Khalili et al. [16] improved the model's feature extraction capability by incorporating an efficient multi-scale attention module, which adaptively redistributes weights to emphasize relevant features. Ni et al. [17] redesigned the feature fusion network and optimized the detection layer architecture. To address the size imbalance issue across datasets, Faraji and Chen [6] devised an enhanced feature pyramid architecture. This architecture leverages attentive feature fusion factors to facilitate the integration of multi-scale features. Further advancements have been made in the design of the backbone and detection head. Chen et al. [18] introduced a custom MFSO structure, which expands objects' feature pixels. Zhang [19] developed a lightweight detection head combined with a Content-Aware Feature Reconstruction module, effectively reconstructing semantically similar feature points to enhance detection accuracy. Similarly, Xiao et al. [20] proposed an enhanced strategy to use inter-layer feature correlation, replacing the conventional FPN-based fusion mechanism. Their approach integrates a Grouped Feature Focus Unit to strengthen contextual feature relevance across different layers, further improving the detection performance.

Recent advancements in Transformer-based architectures have significantly contributed to small object detection. Chen et al. [21] introduced a multi-scale neural network-integrated with a Transformer to enhance the detection of tiny biomedical entities in medical images, achieving promising results. Kumar et al. [5] developed a Transformer-based approach for wildlife monitoring, incorporating a high-frequency feature generator, feature refiner, and query refiner to improve the localization and classification of eight animal species. Liu et al. [22] tackled the challenge of noisy feature fusion by proposing a Trans R-CNN architecture which contains the DN-FPN module. It leverages contrastive learning to mitigate feature noise at each layer of the top-down path.

Super-resolution (SR) is the method that reconstructs high-resolution images, thereby increasing the effective data available for object detection models. Zhang et al. [23] proposed a flexible SR branch that learns high-resolution feature representations, enabling better differentiation of small objects from complex backgrounds in low-resolution inputs. However, their approach did not consider the interaction between the SR network and the detection model. Building upon this, Liu et al. [24] integrated the SR task directly into the detection pipeline by introducing an SR sensing branch and sensing loss, allowing for joint optimization of both SR and detection tasks through SR sensing association. While super-resolution can enhance image quality, it also increases computational overhead, as background pixels–contributing minimally to object detection is also enhanced, leading to inefficient resource usage. The diffusion model, known for its stability over Generative Adversarial Networks (GANs) in image generation, has shown potential to address this issue. Zhang et al. [25] proposed the Bisecting Patch optimization algorithm, which segments an image into patches and selectively applies a conditional diffusion model to reconstruct only high-resolution patches containing objects. This approach significantly boosts detection performance while reducing unnecessary computational costs.

Despite these advancements, existing studies primarily focus on feature fusion at different levels or improving accuracy at the expense of increased computational complexity, highlighting the need for more balanced solutions [26,27].

2.2 Data Augmentation for Small Objects in Remote Sensing

Data augmentation is crucial in improving the robustness of deep learning models, particularly for small object detection in remote sensing imagery. Due to the inherent limitations of small object datasets, such as low object-to-background ratio and imbalanced class distributions–augmentation techniques are essential to enhance generalization and feature diversity. Traditional augmentation methods, such as flipping, rotation, mixup [28], and demosaicking, are often insufficient for small objects, as they do not address the challenges posed by occlusion, scale variation, and environmental factors. Over the past few years, data augmentation techniques have diversified into multiple categories. These span from conventional image processing techniques to more advanced neural network-driven methods, including generative data augmentation and style transfer [29,30]. Nevertheless, there is a relative lack of enhancement strategies specifically designed for small objects. Researchers have explored remote sensing image fusion-based augmentation to overcome these limitations, which synthesizes realistic training samples by integrating multi-sensor data.

Another effective strategy is context-aware copy-paste augmentation, where small objects extracted from one scene are blended into another while maintaining spatial and spectral consistency. In remote sensing applications, this technique is beneficial for enhancing underrepresented classes, such as small ships in maritime surveillance or aircraft in aerial reconnaissance. Faraji and Chen [6] proposed a new copy-paste data augmentation scheme. Kisantal [31] introduced a random copy-paste method with a different strategy for augmentation. However, this method does not increase the number of small objects, and this strategy can lead to context mismatch. Chen et al. [32] proposed adaptive resampling during the pasting stage to sample accurate locations, relies on a pre-trained network.

3 Method

3.1 Improved FPN for Multi-Level Feature Integration

The Feature Pyramid has long been a cornerstone in object detection models, as it allows for the fusion of different scale features effectively. By leveraging features from multiple levels of a feature pyramid, detection heads can identify objects across a wide range of scales. However, many current methods, including those in the YOLO series, rely on deep network architectures to extract features at various scales. While this approach is efficient at capturing high-level semantic information, it often leads to a loss of critical spatial details as the network deepens. In particular, small objects, which are highly dependent on fine-grained spatial information, tend to be underrepresented in the deeper layers of the network, making their detection a significant challenge. Furthermore, the limitations in handling scale differences and feature extraction from feature maps hinder the performance of these models in detecting small targets.

Many feature pyramid models in current research, such as PANet, BiFPN and so on, which are based on the structure of attention mechanisms. PANet is the first model to propose a bottom-up secondary fusion, adding a bottom-up fusion path based on the FPN in Faster RCNN. Many contemporary methods emphasize feature interaction between different layers of the network but fail to address the refinement of features within individual layers. This oversight results in suboptimal utilization of features at various scales, particularly in shallow layers where spatial details are most abundant but are often discarded or diminished through pooling operations.

Considering the above challenges, we propose a novel fusion method within the neck of the YOLOv8 that prioritizes feature selection. This approach is designed to improve the model's capacity for feature extraction and detection of small objects effectively by refining features at each layer before fusion. The core

of this innovation lies in integrating a Channel Attention (CA) mechanism. In our proposed approach, low-level features, rich in spatial details are first filtered through the CA module, which applies attention-based selection to emphasize the most relevant features. The output of this process is then combined with high-level features f_{high} that contain more semantic information but are generally weaker in spatial precision.

To ensure compatibility between the low-level and high-level features, we select a 1×1 convolution to adjust the channel dimensions, making the feature maps from both layers compatible for fusion. Although rich in semantic understanding, the high-level feature maps lack the spatial resolution necessary for detecting small objects accurately. In contrast, while preserving location information, shallow features often fail to capture the complex semantic context. By refining and combining these features using the CA module, our method bridges this gap. This feature fusion mechanism significantly improves the robustness, particularly in UAV-based detection tasks. Moreover, challenging applications where scale variations and complex backgrounds are prevalent.

Traditional approaches usually utilize a direct addition of upsampled high-level features with low-level features, aiming to combine their strengths. However, this method often fails to perform effective feature selection, leading to suboptimal results, especially when the low-level features contain a considerable amount of redundant or irrelevant information. To overcome this limitation, a more selective fusion strategy is used to leverage the high-level feature as a weight to guide the filtering of semantic information from the low-level feature.

The difference of our method is that most of the other FPNs add a bottom-up or more complex branches. The improved method in this section screens each other through features of different scales, which is equivalent to performing multiple self-attention. Fig. 2 shows the fusion process in detail. First, this module applies transposed convolutions (T-Conv) to upsample the high-level feature f_{high} to the same size as the low-level feature. The Channel Attention (CA) module then generates attention weights from the high-level feature to filter out irrelevant parts of the low-level feature. Finally, the filtered low-level feature is fused with the high-level feature, effectively combining semantic information and spatial details. This selective fusion preserves key spatial features while integrating contextual understanding.

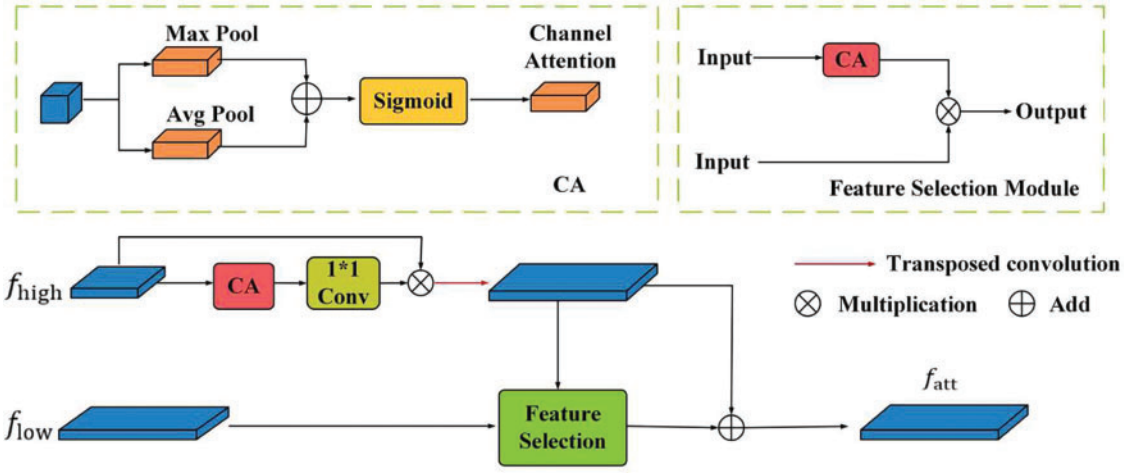

Figure 2: Feature selection pyramid network

3.2 Convolution Feature Enhancement Module Like Transformer Structure

It can be seen from Fig. 3, the convolution feature enhancement module is designed to integrate the strengths of Convolutional Neural Network (CNN) and Transformer, balancing efficiency and performance. While Transformers excel at capturing long-range dependencies through self-attention, their large architectures pose challenges for deployment on UAV platforms with limited computational resources [33,34]. In contrast, CNN efficiently extracts local features but struggles with global context modelling. To address this, Conv-Former mimics the Transformer encoder structure while leveraging lightweight convolutional operations to implement the attention mechanism. Inspired by SCTNet, this hybrid approach enhances feature extraction by incorporating global dependencies while maintaining a compact model size. The Conv-Former framework incorporates a convolution-based self-attention mechanism, capable of effectively gathering spatial and contextual data, and a feed-forward network (FFN) for feature transformation. This architecture enables improved detection accuracy with fewer parameters, making it suitable for real-time UAV applications. The structure of ConvFormer closely resembles the classical Transformer encoder.

$$f = Norm(x + ConvAttention(x)) \tag{1}$$

$$y = Norm(f + FFN(f)) \tag{2}$$

In this module, $Norm$ refers to batch normalization, and x, f and y represent the input, hidden features, and output, respectively. The Feed-Forward Network (FFN) consists of two standard 3 × 3 convolutional layers, effectively capturing local spatial dependencies and enhance feature transformation.

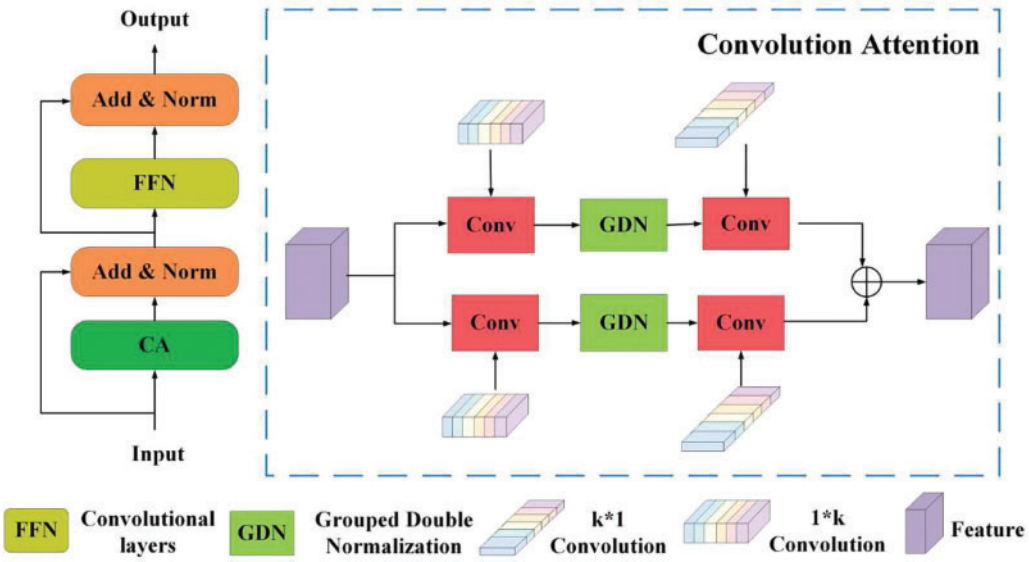

Figure 3: Conv-Former block

To improve efficiency and scalability, we introduce ConvAttention, a GPU-friendly attention mechanism that balances computational cost and feature representation. The process begins with strip convolutions equipped with learnable kernels, which preserve critical spatial details while reducing unnecessary computations. Instead of conventional k × k convolutions, we employ k × 1 and 1 × k convolutions, which approximate the receptive field of standard convolutions while significantly improving efficiency.

Dual-path normalization is used in two key dimensions to further refine attention distribution. First, softmax normalization is performed in the $H \times W$ spatial dimension, ensuring that the sum of similarity scores within the K-region equals 1, thereby enhancing local feature discrimination. Second, Group L_2 normalization is applied in the N-channel dimension, which increases the contrast between different feature points, making it easier for the model to distinguish relevant object regions. This approach effectively improves small object detection by strengthening spatial awareness while maintaining a lightweight and computationally efficient structure. The detailed implementation is illustrated in the accompanying figure.

$$X = GDN(X * K) * V \tag{3}$$

In the context of our model, these dimensions represent the structure of the feature maps as they pass through the network, with C representing the depth, and H and W corresponding to the spatial dimensions. The term N refers to the kernel size of the learnable parameters, which dictates the receptive field of the convolutional operation applied at each layer. The GDN (Grouped Double Normalization) is a normalization technique used to stabilize the learning process by normalizing the feature maps across spatial and channel dimensions. This method ensures that the feature activations are appropriately scaled, facilitating better convergence during training. Finally, the * symbol represents convolution, using the learnable kernels to extract relevant spatial features. The convolution operation is central to feature extraction in deep learning models, enabling the network to capture patterns and structures from the input data at multiple levels of abstraction.

3.3 Detection Heads and Loss Functions

In the feature extraction process, Deep neural networks frequently result in the disappearance of essential details of small objects, while complex background elements can further degrade detection performance. Given the difficulties in identifying small objects, preserving shallow, high-resolution features is critical. To mitigate these issues, we propose the integration of a dedicated small object detection head at the shallow P2 level of the network. This addition significantly enhances the model's performance by focusing on the finer details in the early layers, which contain essential spatial information for accurate localization. Furthermore, to improve object localization, we employ a novel loss function that quantifies the discrepancy between the predicted and ground truth bounding box. This loss is calculated using the Intersection over Union (IoU), which measures the overlap between the predicted and true boxes. While the IoU-based loss function is practical in cases with substantial overlap, its performance deteriorates when the bounding boxes do not overlap, as is often the case with small objects. The CIoU loss used in YOLOv8 considers both the distance and aspect ratio differences between bounding box centres, enhancing its robustness for many cases. However, this approach has limitations when the bounding boxes have identical aspect ratios but differ significantly in size. To address these limitations, various alternatives have been proposed, such as InnerIoU [35], MPDIoU [36], ShapeIoU [37], and NwdIoU [38]. While these methods improve the performance of bounding box regression in general object detection, none are specifically designed with small objects in mind, which are often subject to significant scale variations.

To fill this gap, we introduce ISIoU (Inner and Shape-based IoU), which combines the advantages of both the InnerIoU and ShapeIoU methods. By doing so, our loss function emphasizes the shape and scale during the loss calculation. The process of calculating the ISIoU loss function is detailed in Algorithm 1. In this formula, (x_c^{gt}, y_c^{gt}) and (x_c, y_c) represent the centre points of the ground truth and predicted bounding boxes, respectively. The parameters w^{gt}, h^{gt}, w, h are the widths and heights of the respective bounding boxes, while ratio refers to the scale of the auxiliary box. The scale factor is associated with the object size, and c denotes the distance between the centre points.

Algorithm 1: ISIoU Bounding Box Regression

1 Input:

2 -(x_c^{gt},y_c^{gt}),(x_c,y_c), w^{gt},h^{gt}, w,h, $ratio$, $scale$,c

3 **Output:**

4 -ISIoU

5 **Steps:**

1) Calculate the coordinates of the vertices of the target box and the auxiliary box of the predicted box in the x, y direction.

$b_1^{gt} = x_c^{gt} - (w^{gt} * ratio)/2,\ \ b_2^{gt} = x_c^{gt} + (w^{gt} * ratio)/2$

$b_3^{gt} = y_c^{gt} - (h^{gt} * ratio)/2,\ \ b_4^{gt} = y_c^{gt} + (h^{gt} * ratio)/2$

$b_1 = x_c - (w * ratio)/2,\ \ b_2 = x_c + (w * ratio)/2$

$b_3 = y_c - (h * ratio)/2,\ \ b_4 = y_c + (h * ratio)/2$

2) Calculate IoU^{inner}.

$inter = (\min(b_2^{gt}, b_2) - \max(b_1^{gt}, b_1)) \times (\min(b_4^{gt}, b_4) - \max(b_3^{gt}, b_3))$

$union = (w^{gt} * h^{gt}) * (ratio)^2 + (w * h) * (ratio)^2 - inter$

$IoU^{inner} = inter/union$

3) Calculate the weight coefficients on the horizontal and vertical direction.

$ww = \frac{2\cdot(w^{gt})^{scale}}{(w^{gt})^{scale}+(h^{gt})^{scale}}$, $hh = \frac{2\cdot(h^{gt})^{scale}}{(w^{gt})^{scale}+(h^{gt})^{scale}}$

$distance^{\text{shape}} = hh \cdot \frac{(x_c-x_c^{gt})^2}{c^2} + ww \cdot \frac{(y_c-y_c^{gt})^2}{c^2}$

4) Calculating parameters Ω^{shape}.

$\Omega^{\text{shape}} = \sum_{t=w,h}(1-e^{\omega_t})^{\theta},\quad \theta = 4$

$$\omega_t = \begin{cases} hh \cdot \frac{|\omega-\omega^{gt}|}{\max(\omega,\omega^{gt})}, & t = w \\ ww \cdot \frac{|h-h^{gt}|}{\max(h,h^{gt})}, & t = h \end{cases}$$

5) Calculating parameters L_{ISIoU}.

$L_{ISIoU} = 1 - IoU^{inner} + distance^{shape} + 0.5 * \Omega^{shape}$

By leveraging both Inner-IoU and Shape-IoU, ISIoU provides a comprehensive approach to bounding box regression, placing more emphasis on small objects. A comparison of ISIoU with existing loss functions demonstrates that our method outperforms previous approaches, offering superior accuracy. The results underline the effectiveness of ISIoU in mitigating the challenges.

4 Experiments

4.1 Datasets

To ensure the validity, this paper selects two widely used public datasets, VisDrone2019 [8] and RSOD as the primary sources for evaluation in our experiments. These datasets offer diverse challenges in object detection, making them ideal for assessing the performance of our model.

The VisDrone2019 dataset [8], collected by the AISKYEYE team at the Machine Learning and Data Mining Laboratory of Tianjin University, is designed explicitly for small object detection. This dataset includes 288 video clips, comprising 261,908 frames, along with 10,209 still images captured from various UAV-mounted cameras under various environmental conditions. The pictures of VisDrone2019 cover a variety of scenarios, including diverse weather and lighting conditions, as well as complex backgrounds. It consists of 10 object classes: pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle. Given the prevalence of small objects and the presence of intricate, often occluded backgrounds,

VisDrone2019 is a robust benchmark for assessing object detection algorithms, especially in practical and demanding scenarios. In Fig. 4, we show some dataset examples.

Figure 4: The figure shows scenes captured under different lighting conditions and viewpoints in VisDrone2019. The objects in the image are generally small and numerous, especially the scene in subfigure (d) is particularly complex and challenging

The RSOD dataset, published by Wuhan University, is an open-source collection tailored for object detection in remote sensing imagery. It consists of four object categories: 4993 aircraft images (446 images), 191 playground images (189 images), 180 overhang images (176 images), and 1586 oil barrel images (165 images). Despite the smaller number of object classes, the dataset is dominated by tiny objects, which present significant detection challenges. These small-scale objects are often difficult to detect due to their size. They can be easily missed by detection algorithms, making the RSOD dataset an essential resource for testing the sensitivity and accuracy of models in detecting tiny objects in remote sensing imagery.

VisDrone2019 and RSOD are both small object detection datasets and do have some similarities in terms of sensor viewpoints and categories. However, there are differences. VisDrone2019 is taken from the UAV perspective is a slanted top-down view, while RSOD is a remote sensing image from a vertical view from top to bottom. In terms of categories, VisDrone2019 covers pedestrians to vehicles, etc. In addition, RSOD also contains categories of special scenes, such as playgrounds and overpasses. Both datasets, with their unique characteristics and challenges, provide a comprehensive foundation for testing and refining object detection models, particularly in handling small objects and complex backgrounds.

4.2 Experimental Environment

In this research, YOLOv8s was chosen as the baseline model for analysis and improvement. The implementation was carried out using Python 3.8 and developed within Visual Studio Code, with the PyTorch framework utilized for model construction. The training was performed on an Ubuntu 20.04 system equipped with an NVIDIA RTX3060 GPU (16 GB). The model underwent 200 training epochs, with the learning rate starting at 0.01 and gradually decaying to 0.0001, facilitating stable convergence. A batch size 8 was employed to optimize memory usage and ensure efficient training. This configuration allowed for effective model training while maintaining computational efficiency.

4.3 Evaluation Metrics

The criteria to gauge the model's effectiveness include precision, mean Average Precision (mAP), FPS and model parameters. To conduct a more detailed analysis of the model's performance across various object sizes, we additionally incorporate size-specific Average Precision metrics: AP_s, AP_m, and AP_l. These metrics correspond to small, medium, and large objects, respectively.

4.4 Comparison of Experimental Results

First, we comprehensively compared our proposed model, the baseline YOLOv8s, and other representative models from the YOLO series, specifically evaluating detection performance on the VisDrone2019 dataset. Considering fair and meaningful comparison, the small versions of the YOLO models are used, as they are more suitable for small targets, which are the primary challenge in this study. Table 1 presents the detection accuracy for each category and mAP for all the models tested. The proposed model consistently outperforms all other models across every category, demonstrating superior detection accuracy and confirming its effectiveness in handling small objects. In addition, the confusion matrix exhibited in the Fig. 5 shows that our proposed model outperforms the baseline model.

Table 1: Comparison of YOLO series (The best-performing outcomes are highlighted in bold)

Model	Pedes*	People	Bicycle	Car	Van	Truck	Tricycle	Aw-trc*	Bus	Motor	Parameter	FLOPs
v8s[1]	45.2	38.3	51.2	74.1	75.7	76.9	70.8	71	84.2	55.7	11.12	28.5
v10s[1]	48.1	43.5	56.5	77.7	79.4	79.5	75	73.6	86.3	60.3	8.07	24.5
v11s[1]	47.2	41.6	55.6	77.1	78.6	78.8	73.4	73.3	86.2	59.8	9.41	21.5
v12s[1]	48.5	41.8	56	77.3	78.5	79	74.3	73.4	86.6	60.2	9.23	21.2
SOD[1]	55.1	52.8	60.9	77.8	79.7	**79.4**	75.1	75.9	85.5	64.4	12.14	59.4
Ours	**55.7**	**53.4**	**61.8**	**78.2**	**79.9**	78.6	**75.6**	**76.4**	**86.1**	**64.7**	19.02	48.8

Note: *The special note are respectively abbreviations for the categories of pedestrian and awning-tricycle; [1]The special note indicates different versions of the YOLO series.

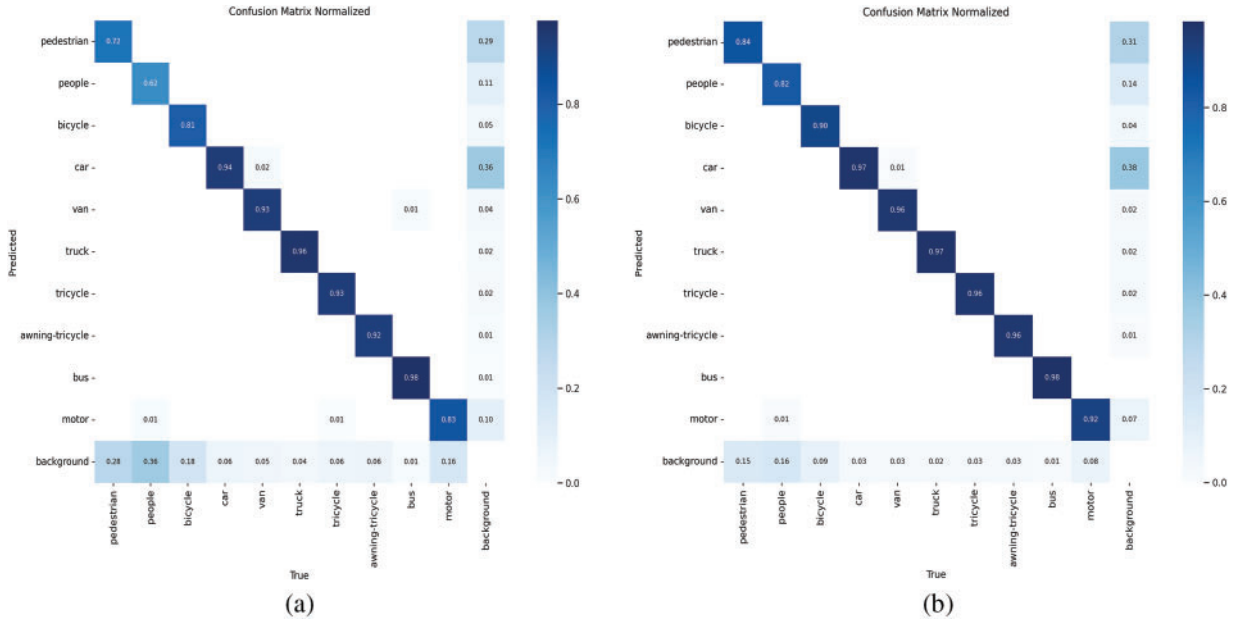

Figure 5: (**a**) and (**b**) represent the confusion matrix of YOLOv8s and the proposed model on the VisDrone2019

Additional comparative experiments, including comparing ISIoU with other commonly used loss functions, are conducted to evaluate the impact of the proposed loss function on the RSOD dataset. The results, shown in Table 2, highlight that ISIoU delivers the best detection performance in terms of mAP@0.5. Specifically, ISIoU improves mAP@0.5 by 0.6%, mAP@0.5:0.95 by 0.3%, precision by 0.7%, and recall by 0.9% compared to the standard CIoU loss. Moreover, ISIoU outperforms alternative loss functions such as MPDIoU, NwdIoU, and ShapeIoU. These results emphasize the efficiency of the ISIoU loss function, which not only enhances detection performance but also simplifies model tuning, making it an effective and efficient method for bounding box regression.

Table 2: Comparison of different bounding box regression methods (The best-performing outcomes are highlighted in bold)

Metrics	mAP50	mAP50-95	Precision	Recall	FPS
CIoU	91.9	64.3	92.8	86.1	169.5
InnerIoU	92	**64.8**	92.8	86.6	185
MPDIoU	91.8	64.2	92.7	86.1	175
ShapeIoU	92	64.5	92.9	86.3	163
NwdIoU	92.3	64.4	92.7	86.9	185
Ours	**92.5**	64.6	**93.5**	**87**	**188**

Many researchers utilize COCO metrics to provide a more detailed detection performance analysis, offering a detailed evaluation across different scales. Table 3 compares the proposed model with several other mainstream models, revealing that the model excels in detecting small objects and shows a significant advantage in detecting medium and large objects. While the Dino model achieves slightly higher APs than our model, its performance gain is marginal—just 0.7%—despite having 2.5 times the number of parameters.

In contrast, our model achieves a 10.6% improvement in APs over the baseline, largely driven by our approach's novel features rather than an increase in model size. The data demonstrate that our method achieves substantial performance gains without the need for excessive model complexity. It maintains a relatively small number of parameters when compared to other models.

Table 3: Comparison in COCO metrics. This table presents a comparative analysis of detection performance across various models using COCO metrics (The best-performing outcomes are highlighted in bold)

Model	AP^{val}_{50-95}	AP^{val}_{50}	AP_s	AP_m	AP_l	Parameter
YOLOv8s	60.0	90.1	39.5	62.6	83.4	11.12
Dino	65.1	94.6	**51.8**	66.9	78.4	47.55
Faster-RCNN	60.8	87.0	45.0	64.2	77.3	41.39
Retinanet	59.6	82.7	36	63.6	79.9	36.51
RTMDet-Tiny	63.5	92.3	47.2	65.1	80.6	4.96
UAV-DETR	65.7	94.5	49.5	67.9	83.4	20
Ours	**65.5**	**94.8**	50.1	**67.5**	**83.1**	19.02

4.5 Ablation Experiment and Visualization

To thoroughly assess the efficacy of each proposed module, we carried out a set of experiments with the baseline model as a benchmark. As shown in Table 4, each optimization brings about a notable enhancement in performance across different evaluation metrics. The visualized heatmaps also emphasize the strength and benefits of our proposed model. In Fig. 6, it presents the heatmap result using three UAV images from different scenes. The heatmaps are overlaid on the original images, with brighter regions indicating areas where the model focuses more attention. The proposed model exhibits a pronounced ability to focus on small objects, critical for improving detection accuracy in complex environments. The red boxes in the visualization mark regions where our model successfully detects objects, demonstrating its superior performance. In contrast, the white boxes highlight areas that the baseline model fails to detect. These visual comparisons effectively demonstrate our approach's enhanced feature representation and attention mechanisms, underscoring its ability to enhance the performance of small and occluded objects in real-world scenarios.

Table 4: This table presents the results of ablation experiments to evaluate the impact of different components

Baseline	ISIoU	FSFPN	CFBlock	mAP_{50}	mAP_{50-95}	Parameter
✓				91.2	64.3	11.12
✓	✓			92.5	64.6	11.12
✓			✓	94.6	68.9	15.47
✓	✓	✓		94.8	67.4	12.47
✓	✓	✓	✓	96.1	71	19.02

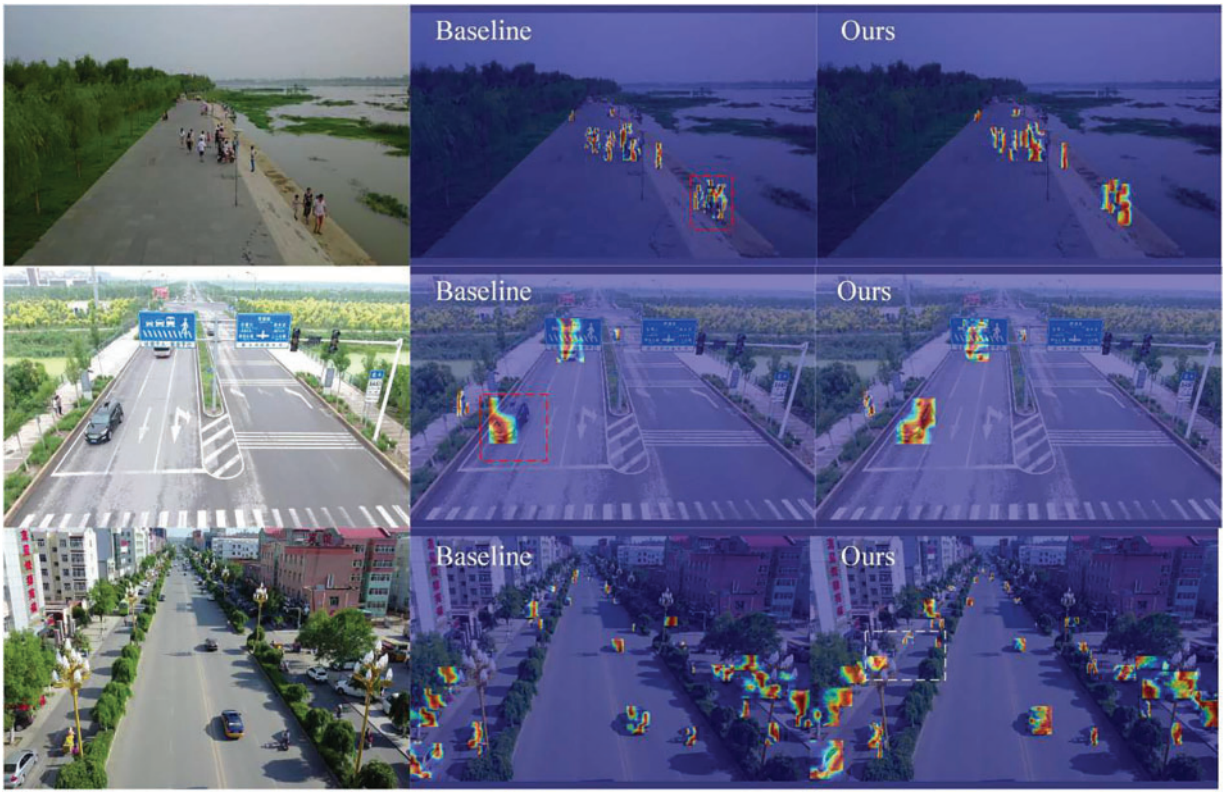

Figure 6: This figure illustrates the heatmap results. From left to right, the image sequence shows: the original input image, the heatmap output from the baseline model and the proposed model

5 Conclusion

This paper presents a robust architecture for detecting small objects in complex scenarios. It incorporates multi-scale feature fusion to enhance detection capabilities. A convolutional attention module inspired by Transformer mechanisms improves the model's feature representation, especially for UAV applications where efficiency is critical. Additionally, introducing a novel loss function that focuses on small objects improves regression accuracy. Although the method significantly enhances small object detection performance, challenges like cluttered backgrounds and low-resolution imagery persist. When the background is complex, objects' severe overlapping and interweaving pose significant challenges. Additionally, low-resolution results in smaller sizes of small objects, thereby reducing detection performance. Further research is needed to refine feature fusion, attention mechanisms, and handling of complex environments, aiming to improve performance and robustness in practical scenarios.

Acknowledgement: Not applicable.

Funding Statement: None.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization and methodology, Dengyong Zhang, Hui He, Jian Peng; data curation and investigation, Hui He; writing—original draft preparation, Hui He, Dengyong Zhang; writing—review and editing, Hui He, Dengyong Zhang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All data included in this study are available upon request by contact with the corresponding author.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Chen X, Ma H, Wan J, Li B, Xia T. Multi-view 3D object detection network for autonomous driving. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA: IEEE; 2017. p. 6526–34. doi:10.1109/CVPR.2017.691.
2. Alqarqaz M, Bani Younes M, Qaddoura R. An object classification approach for autonomous vehicles using machine learning techniques. World Electr Veh J. 2023;14(2):41. doi:10.3390/wevj14020041.
3. Bakirci M. Advanced aerial monitoring and vehicle classification for intelligent transportation systems with YOLOv8 variants. J Netw Comput Appl. 2025;237:104134. doi:10.1016/j.jnca.2025.104134.
4. Bakirci M. Enhancing vehicle detection in intelligent transportation systems via autonomous UAV platform and YOLOv8 integration. Appl Soft Comput. 2024;164:112015. doi:10.1016/j.asoc.2024.112015.
5. Kumar S, Zhang B, Gudavalli C, Levenson C, Hughey L, Stabach JA, et al. WildlifeMapper: aerial image analysis for multi-species detection and identification. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024; Seattle, WA, USA. p. 12594–604.
6. Faraji H, Chen B. Drone-YOLO: improved YOLO for small object detection in UAV. In: 2023 8th International Conference on Image, Vision and Computing (ICIVC); 2023; Dalian, China. p. 93–100.
7. Zhang X, Izquierdo E, Chandramouli K. Dense and small object detection in UAV vision based on cascade network. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW); 2019 Oct 27–28; Seoul, Republic of Korea: IEEE; 2019. p. 27–8. doi:10.1109/iccvw.2019.00020.
8. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: European conference on computer vision. Cham: Springer; 2014. p. 740–55.
9. Du X, Cheng K, Zhang J, Wang Y, Yang F, Zhou W, et al. Infrared small target detection algorithm based on improved dense nested U-Net network. Sensors. 2025;25(3):814. doi:10.3390/s25030814.
10. Zhu P, Wen L, Du D, Bian X, Fan H, Hu Q, et al. Detection and tracking meet drones challenge. IEEE Trans Pattern Anal Mach Intell. 2022;44(11):7380–99. doi:10.1109/tpami.2021.3119563.
11. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA: IEEE; 2014. p. 580–7. doi:10.1109/cvpr.2014.81.
12. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/tpami.2016.2577031.
13. Varghese R, Sambath M. YOLOv8: a novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS); 2024 Apr 18–19; Chennai, India: IEEE; 2024. p. 1–6. doi:10.1109/adics58448.2024.10533619.
14. Ao W, Chen H, Liu LH, Chen K, Lin ZJ, Han JG, et al. YOLOv10: real-time end-to-end object detection. arXiv:2405.14458. 2024.
15. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multibox detector. In: Proceedings of the 14th European Conference on Computer Vision; 2016 Oct 11–14; Amsterdam, The Netherlands. Cham, Switzerland: Springer. p. 21–37.
16. Khalili B, Smyth AW. SOD-YOLOv8-enhancing YOLOv8 for small object detection in aerial imagery and traffic scenes. Sensors. 2024;24(19):6209. doi:10.3390/s24196209.
17. Ni J, Zhu S, Tang G, Ke C, Wang T. A small-object detection model based on improved YOLOv8s for UAV image scenarios. Remote Sens. 2024;16(13):2465. doi:10.3390/rs16132465.

18. Chen D, Lin F, Lu C, Zhuang J, Su H, Zhang D, et al. YOLOv8-MDN-tiny: a lightweight model for multi-scale disease detection of postharvest golden passion fruit. Postharvest Biol Technol. 2025;2019:113281. doi:10.1016/j.postharvbio.2024.113281.
19. Zhang J, Chen Z, Yan G, Wang Y, Hu B. Faster and lightweight: an improved YOLOv5 object detector for remote sensing images. Remote Sens. 2023;15(20):4974. doi:10.3390/rs15204974.
20. Xiao Y, Xu T, Yu X, Fang Y, Li J. A lightweight fusion strategy with enhanced interlayer feature correlation for small object detection. IEEE Trans Geosci Remote Sensing. 2024;62:1–11. doi:10.1109/tgrs.2024.3457155.
21. Chen Z, Lu S. CAF-YOLO: a robust framework for multi-scale lesion detection in biomedical imagery. arXiv:2408.01897. 2024.
22. Liu HI, Tseng YW, Chang KC, Wang PJ, Shuai HH, Cheng WH. A DeNoising FPN with transformer R-CNN for tiny object detection. IEEE Trans Geosci Remote Sens. 2024;62:4704415. doi:10.1109/tgrs.2024.3396489.
23. Zhang J, Lei J, Xie W, Fang Z, Li Y, Du Q. SuperYOLO: super resolution assisted object detection in multimodal remote sensing imagery. IEEE Trans Geosci Remote Sensing. 2023;61:1–15. doi:10.1109/tgrs.2023.3258666.
24. Liu J, Zhang J, Ni Y, Chi W, Qi Z. Small-object detection in remote sensing images with super-resolution perception. IEEE J Sel Top Appl Earth Obs Remote Sensing. 2024;17:15721–34. doi:10.1109/jstars.2024.3452707.
25. Zhang T, Kasichainula K, Zhuo Y, Li B, Seo JS, Cao Y. Patch-based selection and refinement for early object detection. arXiv:2311.02274. 2023.
26. Cheng Y, Wang W, Zhang W, Yang L, Wang J, Ni H, et al. A multi-feature fusion and attention network for multi-scale object detection in remote sensing images. Remote Sens. 2023;15(8):2096. doi:10.3390/rs15082096.
27. Zhang X, Demiris Y. Visible and infrared image fusion using deep learning. IEEE Trans Pattern Anal Mach Intell. 2023;45(8):10535–54. doi:10.1109/tpami.2023.3261282.
28. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. Beyond empirical risk minimization. arXiv:1710. 09412. 2017.
29. Bowles C, Chen L, Guerrero R, Bentley P, Gunn R, Hammers A, et al. GAN augmentation: augmenting training data using generative adversarial networks. arXiv:1810.10863. 2018.
30. Jackson PT, Atapour-Abarghouei A, Bonner S, Breckon T, Obara B. Style augmentation: data augmentation via style randomization. arXiv:1809.05375. 2018.
31. Kisantal M, Wojna Z, Murawski J, Naruniec J, Cho K. Augmentation for small object detection. arXiv:1902.07296. 2019.
32. Chen C, Zhang Y, Lv Q, Wei S, Wang X, Sun X, et al. RRNet: a hybrid detector for object detection in drone-captured images. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW); 2019 Oct 27–28; Seoul, Republic of Korea: IEEE; 2019. p. 100–8. doi:10.1109/iccvw.2019.00018.
33. Xu Z, Wu D, Yu C, Chu X, Sang N, Gao C. SCTNet: single-branch CNN with transformer semantic information for real-time segmentation. arXiv:2312.17071. 2023.
34. Quan Z, Sun J. A feature-enhanced small object detection algorithm based on attention mechanism. Sensors. 2025;25(2):589. doi:10.3390/s25020589.
35. Zhang H, Xu C, Zhang S. Inner-IoU: more effective intersection over union loss with auxiliary bounding box. 2023. doi:10.48550/arXiv.2311.02877.
36. Ma S, Xu Y, Ma S, Xu Y. MPDIoU: a loss for efficient and accurate bounding box regression. arXiv:2307.07662. 2023.
37. Zhang H, Zhang S. More accurate metric considering bounding box shape and scale. arXiv:2312.17663. 2023.
38. Wang J, Xu C, Yang W, Yu L. A normalized Gaussian Wasserstein distance for tiny object detection. arXiv:2110.13389. 2021.