J. Meas. Sci. Instrum., 2026, 17(2): 208-218.
DOI: 10.62756/jmsi.1674-8042.2026018
http://jmsi.publish.founderss.cn
jmsi@nuc.edu.cn

Improved multi-scale feature fusion for infrared small target detection based on YOLOv8

DING Shangsi, YANG Guiqin*, GAN Bingkun

School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

***Corresponding author:** YANG Guiqin (yangguiqin@mail.lzjtu.cn)

Received: May 13, 2025 **Revised:** June 10, 2025 **Accepted:** June 26, 2025

Abstract: Aiming at the problems of low target pixels and intricate background in small target detection in infrared scenes, a target detection model based on multi-scale feature extraction with YOLOv8 was proposed. Firstly, all downsampling convolutions in the network were replaced with the Haar wavelet downsampling (HWD) module to better preserve fine-grained details in infrared imagery during downsampling. Secondly, the spatial pyramid pooling-fast (SPPF) module was improved by introducing separable convolutions, which expanded the receptive field in both horizontal and vertical directions, enabling more comprehensive spatial information capture. Furthermore, a novel C2f_CDWR module was designed using dilated convolutions with varying dilation rates to achieve adaptive feature extraction across multiple receptive fields, thus enhancing detection performance for objects of different sizes. Finally, to improve localization accuracy, the original CIoU loss in YOLOv8 was replaced with Inner-SIoU, which effectively improved bounding box regression accuracy and significantly boosted the model's capability in detecting small infrared targets. The experimental evaluation on the HIT-UAV dataset shows that the precision of the enhanced YOLOv8 model is 90.5%, the recall rate is 75.9%, and the mean average precision is 85.7%. In terms of infrared target detection, its performance was significantly better than that of the baseline YOLOv8 model and other benchmark models.

Key words: infrared image; small object detection; multi-scale feature extraction; dilation convolution; YOLOv8

0 Introduction

Object detection is a fundamental task in computer vision that involves identifying and localizing objects within images or video frames. It plays a crucial role in various applications, including autonomous driving[1], surveillance systems[2], and intelligent security systems. The emergence of deep learning has significantly enhanced the precision and robustness of detection methods. Based on their processing pipeline, deep learning-based object detectors are typically categorized into two groups: two-stage and one-stage approaches. Two-stage models, such as faster R-CNN[3] and mask R-CNN[4], first generate candidate regions and then perform classification and regression. In contrast, one-stage detection detectors perform category prediction and bounding box regression simultaneously in a single forward pass, with YOLO[5] and SSD[6] being the most representative approaches.

Despite rapid advancements driven by deep learning, object detection techniques have subsequently achieved a significant leap forward, resulting in the development of a range of mature methods. However, most existing studies have focused on visible light images, as they provide rich texture and color information. Nevertheless, under challenging conditions such as strong illumination, nighttime environments, or adverse weather (e.g., rain and fog), visible light imaging often degenerates, limiting detection performance[7]. In contrast, infrared imaging technology effectively mitigates the adverse effects of smoke, haze, rain, and snow, ensuring stable imaging. Consequently, it has been widely adopted across various applications. The increasing maturity of infrared technology, along with the availability of cost-effective thermal imaging devices, has further accelerated its adoption in professional domains. Therefore, the primary focus of infrared object detection research is to optimize the model to reduce complexity while maintaining accuracy and enhancing detection efficiency. Achieving robust performance in small target detection, long-range detection, and complex environments remains a critical challenge[8-9].

To enhance the accuracy of infrared object detection while simultaneously reducing algorithmic complexity, Chen et al.[10] proposed an encoder-decoder convolutional

© The Author(s) 2026. Published by Publishing Center of North University of China. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

neural network (CNN) with an attention mechanism to enhance multi-scale feature representation. Their attention module enables better focus on pedestrian targets while suppressing background noise. Ma et al.[11] proposed an edge-assisted multi-scale context network. This approach integrates edge and semantic features through a feature fusion module, leveraging their complementarity to improve the detection of weak and small infrared targets. However, the model exhibits suboptimal performance when detecting large targets. Ma et al.[12] developed an infrared-visible image fusion network based on salient object detection, which enhances object representation by integrating heterogeneous modalities. However, this method primarily focuses on the fusion process and does not leverage the fused images for object detection. To address edge contour recognition challenges in infrared imagery, Li et al. [13] introduced the YOLO-FIRI network by incorporating residual blocks and a fusion attention mechanism into the YOLOv5 framework.

Although this approach enhances edge feature learning, it significantly increases the number of parameters and computational cost, limiting training and deployment efficiency. However, infrared detection algorithms still face certain challenges. 1) Infrared images lack rich texture and color information, and their image quality is poor, making it difficult for existing algorithms to extract features. 2) Existing algorithms struggle to balance infrared targets that vary significantly in size, shape, and texture. 3) Infrared images typically have low contrast, making it difficult for existing algorithms to effectively distinguish between targets and backgrounds[14]. An enhanced infrared target object detection algorithm YOLOv8-based with enhanced multi-scale feature extraction is studied in this paper.

1 YOLOv8 network

YOLOv8[15] is a newer generation of single-stage object detection algorithms developed by Ultralytics. Building upon the strengths of the YOLO series, YOLOv8 offers improved performance and flexibility, achieving a well-balanced trade-off between detection velocity and precision. Its architecture incorporates a novel backbone network, an anchor-free detection head, and an optimized loss function, thereby enabling state-of-the-art efficiency, accuracy, and robustness across a wide range of computer vision tasks.

The backbone of YOLOv8 adopts the C2f module, an evolution of the C3 module used in YOLOv5. This design reduces the number of parameters while enhancing feature extraction capabilities, thereby improving feature fusion and detection performance. The neck of YOLOv8 employs an FPN+PAN structure, which effectively integrates deep-semantic information with low-level spatial features, further enhancing detection performance for multi-scale targets[16]. For classification, YOLOv8 employs varifocal loss[17], while distribution focal loss and CIoU loss are applied to bounding box regression, which enhances the precision of boundary localization.

2 Improved YOLOv8 network

Although YOLOv8 demonstrates notable improvements in detection capability compared to its predecessors, infrared images inherently suffer from reduced sensitivity to brightness variations and limited clarity, often containing a high proportion of low-resolution small targets[24]. Therefore, we introduced several key modifications to enhance YOLOv8.

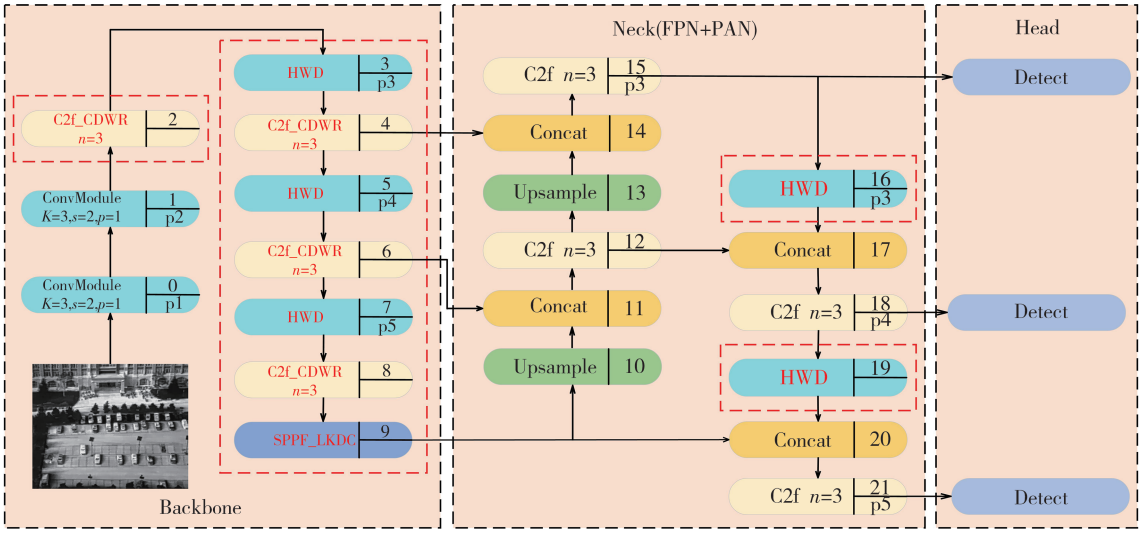


Fig. 1 Improved YOLOv8 network architecture

Firstly, large-kernel convolutions were integrated into the SPPF module to enhance its ability to capture long-range dependencies. Secondly, HWD was used to replace stride convolution-based downsampling, effectively preserving fine-grained features and minimizing information loss. Thirdly, the original C2f module was redesigned using the DWR module, which expanded the receptive field and enhanced deep feature extraction. Ultimately, the baseline CIoU loss was replaced by the internal SIoU. The internal SIoU can take into account the directional relationship between the predicted results and the actual discovery results in terms of bounding boxes, and the loss calculation used other bounding boxes, thereby accelerating the convergence speed of the model. The enhanced structure is shown in Fig.1.

2.1 SPPF_LKDC module

To overcome the limitations of conventional small-kernel convolutions in capturing long-range feature dependencies, numerous studies have introduced large-kernel convolutions or attention mechanisms to enhance the global receptive field. Although attention mechanisms such as transformers have demonstrated notable success, they often introduce substantial computational overhead, making them less suitable for lightweight applications. In contrast, large kernel depthwise convolutions provide comparable long-range modeling capabilities at a significantly lower computational cost, rendering them particularly suitable for infrared detection tasks that demand both high inference efficiency and accuracy.

In this study, 11×11 large kernel depthwise convolutions are integrated into the spatial pyramid pooling-fast (SPPF) structure. To further reduce parameter count and computational complexity, a factorized design using sequential $1\times k$ and $k\times 1$ convolutions is adopted. This design enhances the model's ability to capture long-range dependencies in both horizontal and vertical directions. The resulting architecture remains compact while exhibiting stronger global contextual modeling capability, which is especially effective in complex infrared scenarios characterized by diverse target sizes and limited texture information. This approach effectively expands the receptive field while maintaining a low computational burden, thereby enhancing the network's capacity to capture spatial relationships over longer distances. The proposed module first applies max pooling to the inputs, followed by concatenation to generate multi-dimensional feature vectors, shown as

$$X_1=\mathrm{Concat}\left(X_1,\left\{\mathrm{MaxPool}_{5\times 5}^{(n)}\left(\mathrm{Conv}_{1\times 1}(X)\right)\right\}_{n=1}^{3}\right). \quad (1)$$

Subsequently, depthwise grouped convolutions with 1×3 and 3×1 kernels are utilized, followed by depthwise dilated convolutions with 1×5 and 5×1 kernels to extract spatial information across a broader range and in multiple directions. They are

$$X_2=\mathrm{DWGConv}_{3\times 1}\left(\mathrm{DWGConv}_{1\times 3}(X_1)\right), \quad (2)$$

$$X_3=\mathrm{DWGDConv}_{3\times 1}\left(\mathrm{DWGDConv}_{1\times 3}(X_2)\right), \quad (3)$$

where DWGConv and DWGDConv denote depthwise grouped convolution and depthwise grouped dilated convolution, respectively.

Ultimately, a 1×1 convolution is employed to fuse channel-wise information, shown as

$$Y=X_1\otimes\left(\mathrm{Conv}_{1\times 1}(X_3)\right), \quad (4)$$

where $\otimes$ denote the Hadamard product (element-wise multiplication).

The improved model effectively captures long-range dependencies, thereby enhancing model performance. The structure of the enhanced SPPF with large kernel depthwise separable convolution (SPPF_LKDC) model is depicted in Fig.2.

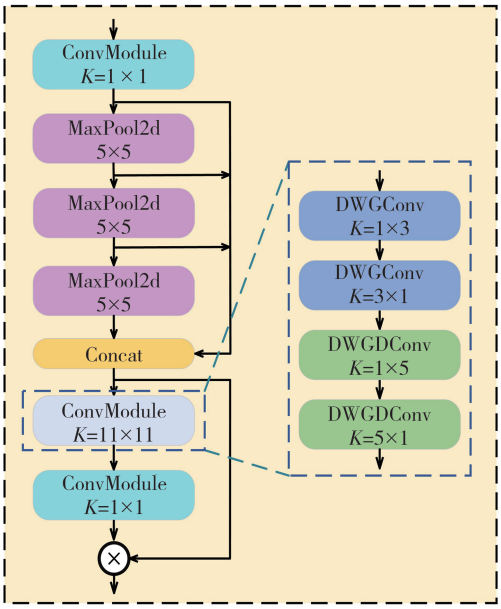


Fig. 2 Structural diagram of improved SPPF_LKDC model

2.2 C2f_CDWR module

Multi-scale feature extraction is crucial for addressing scale variations and weak textures in infrared imagery. A common strategy involves enlarging convolutional kernels. However, this often incurs significant computational overhead, limiting their applicability in lightweight detection models. Dilated convolutions provide a more efficient alternative by expanding the receptive field without

increasing the number of parameters, making them well-suited for multi-scale modeling in resource-constrained environments.

Therefore, a multi-receptive-field feature extraction module, C2f_CDWR, was designed, which was based on an improved dilation-wise residual (DWR) module. The DWR module leverages multi-rate dilated convolutions to enhance multi-scale perception and capture richer feature information. Moreover, the dilation rate and convolutional capacity were meticulously designed to ensure comprehensive feature extraction across various spatial regions. Nevertheless, a single DWR module has limited capacity to fuse shallow features with deep semantic information. To overcome this limitation, an enhanced DWR module was introduced, comprising two branches. One branch focused on deep feature extraction through stacking operations, while the other processed features via variable convolutional kernels. The features from both branches were subsequently spliced and further fused through variable convolution. This design not only preserved the lightweight nature of the module but also significantly enhanced its capability to extract deep features. The enhanced structure is shown in Fig.3. $2c$ denotes the number of channels, d indicates the dilation rate used in the dilated convolution, and RR and SR refer to regional residualization and semantic residualization, respectively.

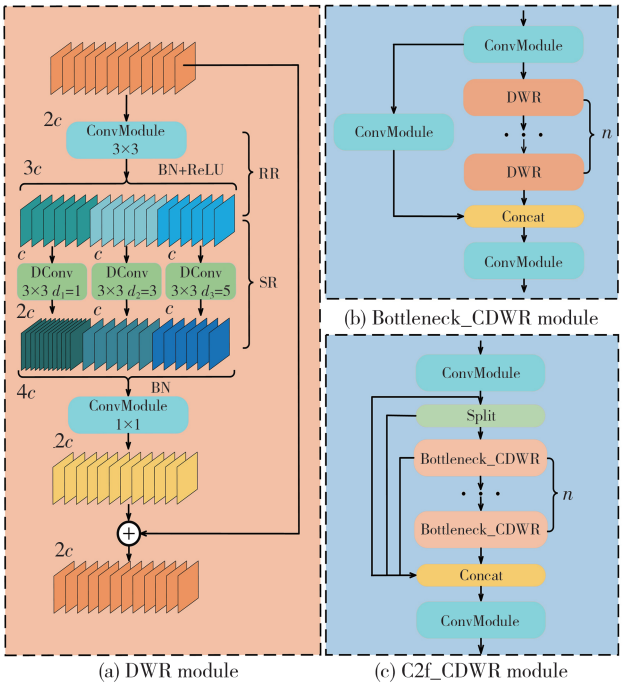


Fig. 3 C2f_CDWR feature extraction module

The formulas of the original DWR module are

$$F_i = \mathrm{Conv}_{3\times 3}^{d_i}\left(\mathrm{Conv}_{3\times 3}(X)\right), \tag{5}$$

$$\mathrm{DWR}(X) = X + \mathrm{Conv}_{1\times 1}\left(\mathrm{Concat}(F_1, F_2, F_3)\right), \tag{6}$$

where d_i denotes the dilation rate, and F_i represents the output after convolution with a dilation rate of d_i.

The enhanced Bottleneck_CDWR module exhibits improved capability in extracting features from targets of varying sizes, significantly enhancing deep feature extraction. The extraction process is detailed as

$$X_1, X_2^0 = \mathrm{Conv}_{k\times k}(X), \tag{7}$$

$$X_2^i = \mathrm{DWR}\left(X_2^{i-1}\right), i = 1, 2, \cdots, n, \tag{8}$$

$$\mathrm{Bottleneck\_CDWR}(X) = \mathrm{Conv}_{k\times k}\left(\mathrm{Concat}\left(X_1, X_2^n\right)\right), \tag{9}$$

where X_2^i denotes the output after the i-th DWR module.

The original network's C2f module is ultimately refined with the improved module, thereby enhancing the network's capability to extract and integrate multi-scale features and deep semantic representations. The process is

$$X_\mathrm{d}, X_\mathrm{b}^0 = \mathrm{Split}(X), \tag{10}$$

$$Y = \mathrm{Conv}_{1\times 1}\left(\mathrm{Concat}\left(X_\mathrm{d}, \left\{\mathrm{Bottleneck\_CDWR}\left(X_\mathrm{b}^i\right)\right\}_{i=1}^{n}\right)\right), \tag{11}$$

where X_b^i denotes the output after the i-th Bottleneck_CDWR module.

The improved module enhances the receptive field by integrating variable convolutional kernels and a channel separation strategy, further refining the DWR module. The model's capacity for encompassing a more extensive array of contextual information is thereby enabled, significantly strengthening feature extraction capabilities, particularly in complex scenarios and multi-scale detection tasks. The traditional Bottleneck module is replaced with Bottleneck_CDWR, making the improved C2f_CDWR more flexible in convolutional kernel design and better suited for diverse detection requirements.

2.3 HWD downsampling module

CNN has shown superior capability in object detection. However, as highlighted by Sunkara R et al.[18], stride convolution and pooling operations have limitations. While modern datasets frequently feature high resolution and moderate-sized targets, thereby enabling networks to bypass redundant information through stride convolution and pooling to enhance learning efficiency, this approach is less effective for infrared datasets. The low contrast, noise interference, and lack of detailed target features in infrared images result in poor image resolution and indistinct target characteristics, thereby posing significant challenges for feature extraction. In this case, due to the scarcity of redundant information, the stride convolution and pooling

operations will inevitably lead to the loss of details, thereby significantly reducing the detection performance for small or occluded objects.

To address this issue, this paper proposes to replace the traditional stride convolutions in YOLOv8 with HWD. By leveraging wavelet transformation, HWD effectively captures richer feature representations in both the spatial and frequency domains. This approach not only preserves fine-grained target features but also mitigates the information loss caused by conventional downsampling methods. Particularly under low-quality infrared imaging environments, HWD enables more stable and robust feature extraction, thereby enhancing the overall detection accuracy and adaptability in complex scenes. The structural configuration of HWD is illustrated in Fig.4. Firstly, the resolution $H \times W$ feature map is decomposed by means of a low-pass filter $\boldsymbol{h}_0$ and a high-pass filter $\boldsymbol{h}_1$, which are utilized to extract approximation and high-frequency features, respectively. Subsequently, both the approximation and detail components undergo downsampling. The wavelet transform produces four components: the low-frequency, herein designated LL, and the high-frequency components, including the horizontal LH, vertical HL, and diagonal HH components, defined as

$$\mathrm{LL} = X * \boldsymbol{h}_0 \bullet \boldsymbol{h}_0 \tag{12}$$

$$\mathrm{LH} = X * \boldsymbol{h}_0 \bullet \boldsymbol{h}_1, \tag{13}$$

$$\mathrm{HL} = X * \boldsymbol{h}_1 \bullet \boldsymbol{h}_0, \tag{14}$$

$$\mathrm{HH} = X * \boldsymbol{h}_1 \bullet \boldsymbol{h}_1, \tag{15}$$

where $*$ denotes the convolution operation, and $\bullet$ represents separable filtering, indicating that the convolution is performed along both the row and column directions.

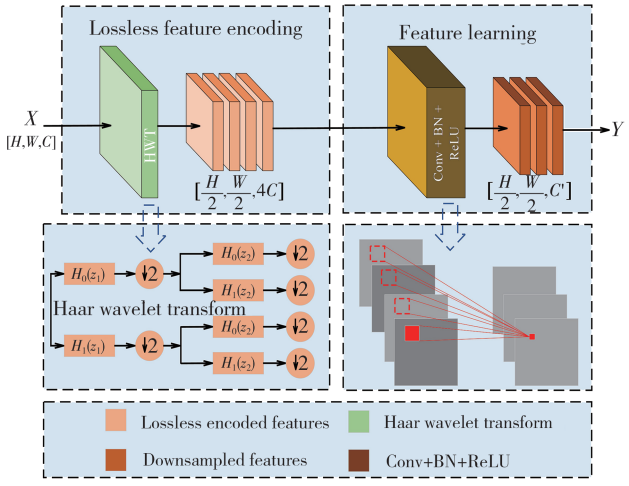


Fig. 4 HWD downsampling module

Then, the extracted features are concatenated to obtain a feature map with half the spatial resolution and four times the channel depth. This transformation effectively encodes spatial information into the channel dimension while preserving critical details. Finally, a convolution-BatchNorm-ReLU (CBR) layer is applied to extract representative features with reduced spatial resolution from the transformed feature map, described as

$$Y = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}\left(\mathrm{cat}\left(\mathrm{LL},\mathrm{LH},\mathrm{HL},\mathrm{HH}\right)\right)\right)\right). \tag{16}$$

The proposed method not only effectively overcomes the limitations posed by low-resolution and noisy infrared images but also enhances the capacity for the identification of diminutive and obscured targets. Moreover, HWD successfully reduces computational complexity while retaining critical features, thereby improving detection accuracy and enhancing model robustness in infrared object detection tasks.

2.4 Inner-SIoU loss function

CIoU demonstrates good boundary regression capabilities in a variety of object detection tasks. However, in the context of infrared imaging, the low contrast, indistinct target-background separation, and susceptibility to noise interference pose significant challenges. CIoU-based bounding box regression is prone to background interference, leading to reduced detection accuracy. Moreover, CIoU exhibits heightened sensitivity to noise, especially when target edges are blurred or when noise levels are high, thereby resulting in unstable Bounding-box regression. Although some studies have introduced advanced loss functions such as GIoU and DIoU, these methods still fall short in addressing the challenge of distinguishing targets from complex backgrounds. To minimize these limitations, this paper puts forward the concept of inner-SIoU, replacing the traditional CIoU.

SIoU loss enhances the model's stability in low-contrast and noisy environments through a smoother optimization process and stronger geometric constraints. By comprehensively considering factors such as angle, distance, and aspect ratio, SIoU effectively adapts to variations in target shapes, thereby improving detection precision. Inner-IoU[19] calculates the overlap within the internal regions of the predictive and real boxes, effectively reducing the interference from external regions and focusing on the core target area. By integrating SIoU's smooth optimization and geometric constraints, Inner-SIoU achieves more accurate bounding box regression, particularly in scenarios with blurred target edges or partial occlusion. This improvement in detection performance serves as a key contribution of this study.

As illustrated in Fig.5, Inner-IoU defines the GT box

and the anchor box, which are labeled as b^{gt} and b. The centre point of the GT box is denoted as (x_c^{gt}, y_c^{gt}). The anchor box is represented as (x_c, y_c). The width and height of the GT frame are expressed as w^{gt} and h^{gt}, while those of the anchor frame are denoted by w and h. The scaling factor, denoted as scale, ranges between grades 0.5 and 1.5.

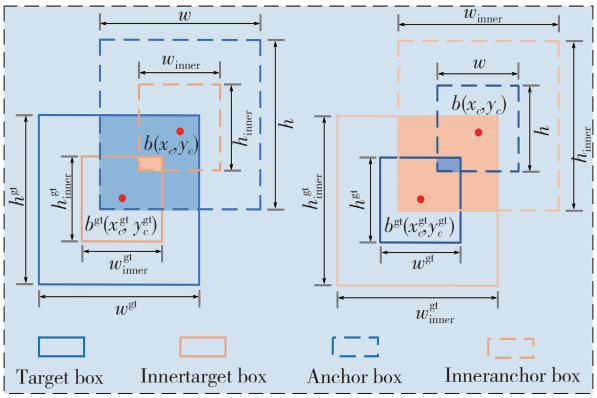


Fig. 5 Description of Inner-IoU

The Inner-SIoU metric integrates the anti-interference capability of Inner-IoU with the optimization properties of SIoU, thereby offering enhanced stability and accuracy in infrared detection. The Inner-SIoU is calculated as follows.

1) The location of the auxiliary ground truth box is determined by the centre point (x_c^{gt}, y_c^{gt}), width w^{gt}, and height h^{gt} of both the anchor frame *scale* and the GT frame. This involves determining the upper-left vertex (b_l^{gt}, b_b^{gt}) and the lower-right vertex (b_r^{gt}, b_t^{gt}), shown as

$$b_l^{gt} = x_c^{gt} - \frac{w^{gt} * \text{scale}}{2}, \tag{17}$$

$$b_r^{gt} = x_c^{gt} + \frac{w^{gt} * \text{scale}}{2}, \tag{18}$$

$$b_t^{gt} = y_c^{gt} - \frac{h^{gt} * \text{scale}}{2}, \tag{19}$$

$$b_b^{gt} = y_c^{gt} + \frac{h^{gt} * \text{scale}}{2}. \tag{20}$$

2) Using the center point (x_c, y_c), width w, and height h of both scale and the GT frame, the location of an auxiliary ground truth frame is computed. This step involves determining the upper-left vertex (b_l, b_b) and the lower-right vertex (b_r, b_t), formulated as

$$b_l = x_c - \frac{w * \text{scale}}{2}, \tag{21}$$

$$b_r = x_c + \frac{w * \text{scale}}{2}, \tag{22}$$

$$b_t = y_c - \frac{h * \text{scale}}{2}, \tag{23}$$

$$b_b = y_c + \frac{h * \text{scale}}{2}. \tag{24}$$

3) The intersection in and the union un between the auxiliary ground truth box and the auxiliary predicted box are then computed, followed by the calculation of the IoU ratio $\text{IoU}^{\text{inner}}$, defined as

$$\text{in} = \left(\min(b_r^{gt}, b_r) - \max(b_l^{gt}, b_l)\right) * \left(\min(b_b^{gt}, b_b) - \max(b_t^{gt}, b_t)\right), \tag{25}$$

$$\text{un} = (w^{gt} * h^{gt}) * (\text{scale})^2 + (w * h) * (\text{scale})^2 - \text{in}, \tag{26}$$

$$\text{IoU}^{\text{inner}} = \frac{\text{in}}{\text{un}}. \tag{27}$$

4) The IoU is computed based on the areas of the predicted box and the ground truth box, and both IoU and $\text{IoU}^{\text{inner}}$ are incorporated into the SIoU. The final Inner-SIoU is computed by

$$L_{\text{Inner-SIoU}} = L_{\text{SIoU}} + \text{IoU} - \text{IoU}^{\text{inner}}. \tag{28}$$

3 Experiments

3.1 Datasets

In this experiment, the HIT-UAV infrared dataset[20] is employed for the detection of infrared small targets in UAV scenarios. The dataset under consideration consists of 2 898 infrared images, extracted from a total of 43 470 video frames, covering diverse environments, including schools, parking areas, roads, and playgrounds. It contains five object categories: person, bicycle, car, other vehicle, and Dont-Care. Compared to traditional ground-based imaging methods, the UAV perspective provides a wider field of view, allowing for the capture of a larger number of targets. However, the objects are significantly smaller, particularly pedestrians and distant objects, which are small in size and densely annotated, making this dataset highly suitable for research on small infrared target detection.

3.2 Experimental environment

The experimental environment was delineated as such: operating system: Windows 10; CPU: 12th Gen Intel® Core™ i5-12600KF; memory: 32 GB; GPU: NVIDIA GeForce RTX 4060Ti 16 G; deep learning framework: PyTorch 2.2.2; CUDA version: 11.8; programming language: Python 3.9; development environment: PyCharm.

Given that different versions of the YOLOv8 framework can yield varying training results, this study adopted the stable YOLOv8.1.9 version. The training settings were as follows: number of epochs: 200; initial LR: 0.01; batch size: 16; momentum: 0.937; weight decay coefficient: 0.000 5; Opt: SGD; other hyperparameters: default values provided by the system. To ensure fairness, no pre-trained weights were used in this experiment, and training was configured to utilise the same hyper-parameter settings.

3.3 Evaluation indicators

To assess the experimental performance, several key

metrics are considered, which are P, R, and the mean average precision (mAP), parameters, model size, and computational complexity measured in GFLOPs.

P reflects the proportion of correctly identified positive instances out of all predicted positives. It is calculated by

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{29}$$

R is used to measure the extent to which the model captures actual positive samples in all real data instances.

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{30}$$

where TP stands for true positives, FP denotes false positives, and FN for false negatives.

The mAP serves as a comprehensive indicator for object detection, representing the average performance across different target categories. It is typically computed by integrating the area under the $P-R$ curve, The average precision (AP) is computed by

$$\mathrm{AP}=\int_0^1 P(r)\mathrm{d}r. \tag{31}$$

Overall detection capability is then derived by averaging the AP scores of all categories. The mAP is computed by

$$\mathrm{mAP}=\frac{1}{N}\sum_{i=1}^{N}\mathrm{AP}_i. \tag{32}$$

3.4 Ablation experiment

Ablation experiments were performed on the YOLOv8n model to show the contribution of each of the four improved modules introduced in this study: the SPPF_LKDC module, the HWD downsampling module, the Inner-SIoU, and the C2f_CDWR multi-receptive field feature extraction module. As shown in Table 1, A represents the SPPF_LKDC module with the addition of large kernel separation convolution, B represents the addition of the HWD downsampling module, C represents the addition of the Inner-SIoU loss function, and D represents the addition of the multiple receptive field feature extraction module C2f_CDWR. The "√" in the table indicates that the module represented by the corresponding letter has been added.

The experimental results are presented in Table 1. After integrating the large kernel decoupled convolution module, the model exhibits an increase of 3.8 percentage points in precision, an improvement of 2.9 percentage points in recall, and a rise of 1.6 percentage points in mAP50. However, this enhancement comes at the cost of increased model size and computational complexity. The introduction of the HWD downsampling module leads to a marginal accuracy reduction but improves recall by 5.5 percentage points and mAP50 by 2.2 percentage points, effectively mitigating missed detections while reducing the parameter size. Replacing CIoU with the Inner-SIoU loss function further enhances performance, yielding a gain of 2.2 percentage points in precision, an increase of 2.7 percentage points in recall, and an improvement of 2.3 percentage points in mAP50, particularly for low-contrast and low-detail infrared targets. Finally, integrating the C2f_CDWR module into the backbone network enhances the feature extraction capability, resulting in an increase of 3.8 percentage points in precision, 3.3 percentage points in recall rate, 3.3 percentage points in mAP50 value, while reducing the parameters by 15.2% and the computational cost by 13.6%. These results demonstrate that the enhanced model outperforms the original YOLOv8n across all metrics, delivering superior detection performance while maintaining a lightweight architecture, thereby validating the efficacy of the proposed design.

Table 1 Results of ablation experiments

A	B	C	D	P/%	R/%	mAP50/%	Parameters/M	GFLOPs
				86.7	72.6	82.4	3.0	8.1
√				90.5	75.5	84.0	3.3	8.3
√	√			84.0	78.1	84.6	3.0	7.8
√	√	√		88.9	75.3	84.7	3.0	7.8
√	√	√	√	90.5	75.9	85.7	2.6	7.0

3.5 Comparison experiment of different Inner-SIoU ratio factors

To determine the optimal scale factor for the proposed Inner-SIoU loss function, a series of ablation experiments were conducted on the enhanced model. The scale hyper parameter, varied from 0.5 to 1.5, modulates the geometry and position of the auxiliary bounding box, thereby influencing the accuracy of bounding box regression. As presented in Table 2, when the scale is set to 1.0, Inner-SIoU degenerates into the standard SIoU, yielding the lowest performance. Within the interval [1.0, 1.5], only marginal gains in detection accuracy are observed.

Table 2 Comparison of experimental results with different proportion factors

Scale	P/%	R/%	mAP50/%	mAP50-95/%
0.5	81.2	74.0	82.1	55.9
0.6	83.2	74.7	82.3	57.2
0.7	**90.5**	**75.9**	**85.7**	**58.2**
0.8	86.4	75.1	83.4	57.5
0.9	88.6	76.0	83.6	57.8
1.0	81.1	75.8	80.2	55.0
1.1	82.4	73.2	81.1	55.8
1.2	82.4	72.6	82.4	55.5
1.3	81.8	72.6	81.0	55.2
1.4	85.5	74.6	82.5	55.9
1.5	81.2	73.9	81.7	55.3

In contrast, a substantial improvement is achieved when the scale lies between 0.6 and 0.9, with the best performance obtained at a scale factor of 0.7. Therefore,

this study adopts a scale of 0.7 for Inner-SIoU as the final bounding box regression loss.

To validate the effectiveness of the Inner-SIoU loss function, comparative experiments were conducted between the proposed method (scale=0.7) and several widely used loss functions. As presented in Table 3, the Inner-SIoU demonstrates superior performance over EIoU across all evaluation metrics, except for a slight decrease in recall. Furthermore, when compared to PIoUv2 and CIoU, the proposed loss function consistently achieves better results in all performance indicators. These findings confirm that Inner-SIoU effectively reduces missed detections and false positives, thereby significantly enhancing overall detection performance.

Table 3 Loss function comparison experiment results

Module	P/%	R/%	mAP50/%	mAP50-95/%
YOLOv8n (EIoU)	80.9	74.7	81.9	56.7
YOLOv8n (PIoUv2)	81.9	73.3	81.7	55.9
YOLOv8n (CIoU)	86.7	72.6	82.4	57.0
YOLOv8n (Inner-SIoU)	89.6	73.9	82.9	57.6

3.6 Comparative experiment

To further validate the advantages of the aforesaid model, a series of comparative experiments are conducted against several mainstream algorithms, as presented in Table 4.

Table 4 Comparative test results

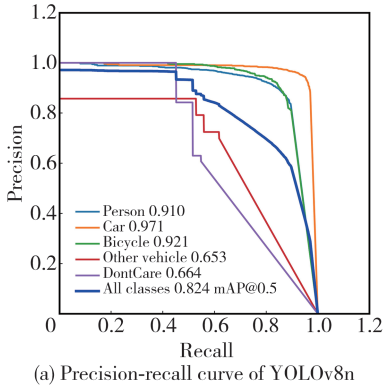
Model	P/%	R/%	mAP50/%	Parameters/M	GFLOPs
MASK-RCNN	76.0	77.4	78.1	44.0	216.1
RT-DETR-l	76.2	66.3	70.6	2.9	100.6
YOLOv5n	80.9	62.8	75.1	2.5	7.1
YOLO-FIRI[13]	80.3	70.8	78.1	20.2	13.1
YOLOv8n	86.7	72.6	82.4	3.0	8.1
EAMCNet[11]	85.4	73.9	82.8	21.2	6.2
YOLOv9t	84.2	71.9	80.9	2.1	8.5
YOLOv10n	80.6	74.9	81.4	2.7	8.2
YOLOv11n	86.6	76.3	83.1	2.6	6.3
YOLOv12n	82.4	73.5	80.6	2.6	6.5
DINO-4scale	86.1	73.7	81.1	47.6	221.5
DDQ-DETR	83.8	81.9	83.3	41.1	173.2
Ours	90.5	75.9	85.7	2.6	7.0

Compared with conventional methods, including MASK-RCNN, YOLOv5n, YOLOv9t, and YOLOv10n, the proposed approach achieves notable improvements in mAP by 7.6 percentage points, 10.6 percentage points, 4.8 percentage points, and 4.3 percentage points, respectively. Moreover, it demonstrates significant advantages in precision, recall, parameter count, and computational complexity, surpassing these baseline algorithms by a considerable margin. Compared with YOLOv11n and YOLOv12n[25], the proposed approach still achieves superior results in terms of precision, mAP, and parameter size. Although it exhibits a slightly lower recall than YOLOv11n, the difference is marginal and does not compromise overall performance. Furthermore, when compared to newly emerged DETR-based models such as RT-DETR-l[21], DINO-4scale[22], and DDQ-DETR[23], the proposed model outperforms these methods with mAP improvements of 15.1 percentage points, 4.6 percentage points, and 2.4 percentage points, respectively. It also significantly reduces both parameter count and computational burden, underscoring its efficiency and suitability for resource-constrained environments. Compared with YOLO-FIRI, the proposed algorithm demonstrates clear advantages in terms of accuracy, recall, mAP, parameter count, and computational efficiency. While the proposed method exhibits a slight increase in computational cost relative to the EAMCNet algorithm, it outperforms EAMCNet across other evaluation metrics.

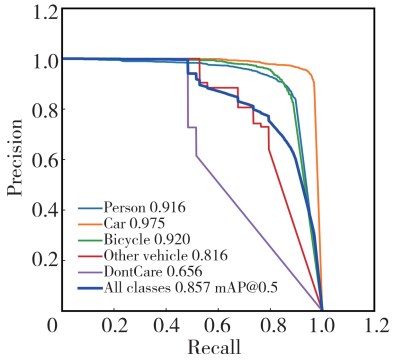
In conclusion, while the proposed model has a marginally lower recall than YOLOv11n and DDQ-DETR, and a slightly higher computational complexity compared to YOLOv11n and YOLOv12n, these differences are negligible. Overall, the proposed approach demonstrates superior performance across nearly all evaluation dimensions. The experimental results affirm that the proposed model not only achieves excellent detection accuracy but also maintains a lightweight and efficient architecture, thus validating the effectiveness and robustness of the design.

3.7 Visualisation of test results

As illustrated in Fig. 6, the precision-recall ($P-R$) curve before and after incorporating the proposed enhancements reveals a substantial improvement of 16.3% in the detection of the other vehicle category, along with an overall increase of 3.3% across all categories. This notable advancement effectively addresses the original model's deficiency in detecting the other vehicle class and significantly enhances the overall detection capability. These findings further validate the effectiveness of the proposed algorithm in improving infrared target detection performance.



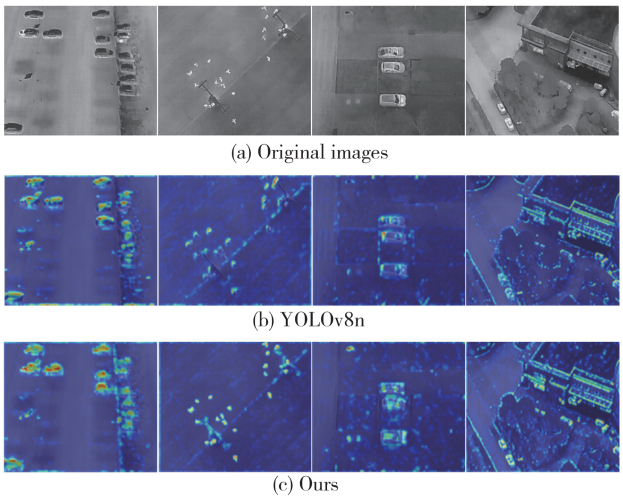
(a) Precision-recall curve of YOLOv8n



(b) Precision-recall curve of ours

Fig. 6 Comparison of *P*–*R* curves

The visual heat map is shown in Fig.7, further accentuating the advantages of the proposed method. In scenarios where targets are sparsely distributed throughout the scene, although both the original and the improved models can detect all targets, the enhanced model demonstrates superior robustness and localization precision. In regions characterized by densely packed small targets, the baseline model frequently misidentifies multiple adjacent objects as a single entity. In contrast, the proposed model accurately detects and distinguishes each individual target. Moreover, in scenarios involving significant variations in target scale, the original model tends to overlook small targets and inadequately focus on large objects. The proposed approach effectively captures both small-scale targets and enhances attention to large-scale ones. In complex occlusion environments where target-background separation is particularly challenging, the original model often misclassifies targets as background clutter. Conversely, the improved model demonstrates the capacity to distinguish occluded targets with higher precision, yielding enhanced detection performance in such challenging contexts.



(a) Original images

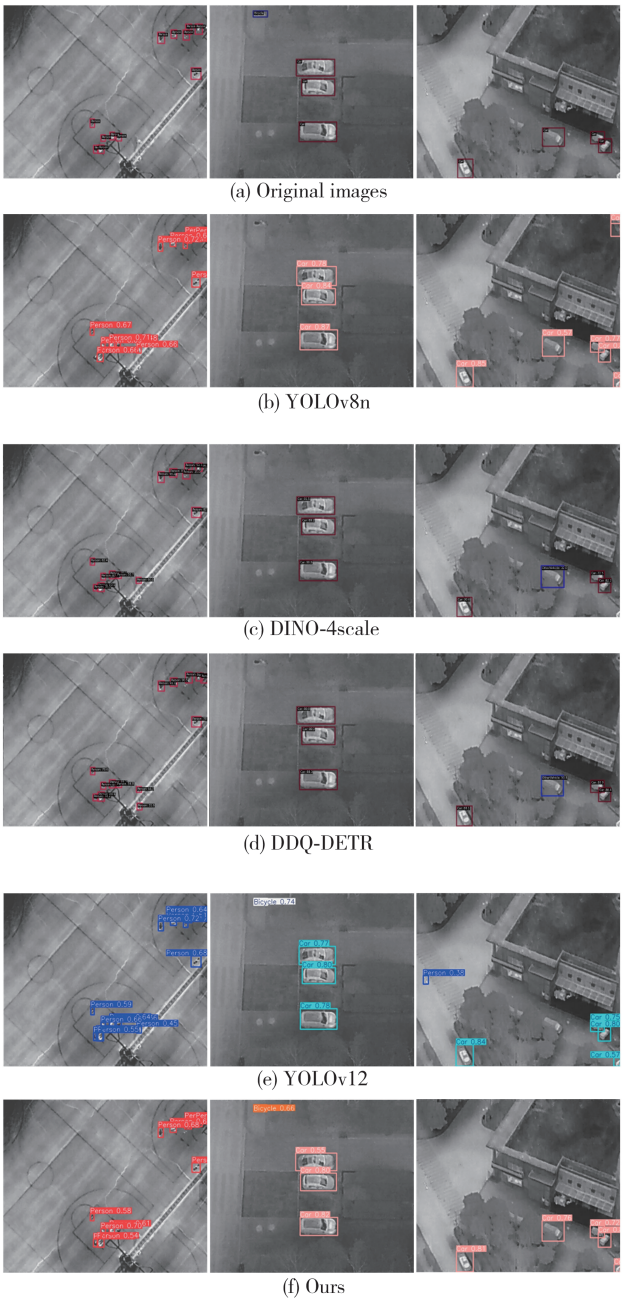
(b) YOLOv8n

(c) Ours

Fig. 7 Heat map comparison of YOLOv8n with proposed algorithm

In summary, the proposed model demonstrates remarkable detection capability, effectively capturing fine-grained details in infrared images, precisely distinguishing the boundaries of small objects, and accurately identifying targets in occluded scenarios. Consequently, the improved algorithm achieves superior results in dense small target detection, multi-scale target detection, and complex occlusion scenarios.

Fig.8 presents the visualization of detection results of various algorithms. Three representative scenarios are selected for evaluation performance: dense small target clustering, significant variation in target sizes, and complex occlusion environments.



(a) Original images

(b) YOLOv8n

(c) DINO-4scale

(d) DDQ-DETR

(e) YOLOv12

(f) Ours

Fig. 8 Visualization of detection effect of different algorithms

In the dense small target clustering scenario, algorithms such as YOLOv8n, DINO-4scale, DDQ-DETR, and YOLOv12 exhibit false detection, erroneously identifying non-target regions such as stadium fences as human figures. In contrast, the proposed algorithm accurately detects all valid targets without producing any false positives, demonstrating superior discriminative capability in cluttered scenes. In the large-scale variation scenario, most algorithms fail to detect certain targets, only YOLOv12n and the proposed algorithm successfully identify the bicycle. In the complex occlusion scenario, the majority of the comparison algorithms encounter issues such as false detection, misidentification, or missed detection, while the proposed algorithm successfully detects all targets without any errors.

These experimental results clearly demonstrate that the designed method outperforms several algorithms. It not only achieves accurate target identification in infrared imagery but also yields higher detection confidence. Therefore, the designed method shows outstanding performance in infrared target detection tasks, demonstrating enhanced practicality and robustness.

4 Conclusions

To mitigate the intrinsic challenges posed by small-object recognition in infrared scenes, including low target pixel resolution, complex backgrounds, and difficulty in extracting effective features, we presented an enhanced detection algorithm based on YOLOv8. Firstly, a large kernel separable convolution was introduced into the SPPF module, enabling the aggregation of feature information from a broader receptive field through multiple large kernel pooling operations, thereby improving the model's capacity for spatial context modeling and boosting detection accuracy. Secondly, the HWD downsampling module replaced the stride convolution downsampling module. By leveraging wavelet transform, the HWD module captured more informative representations across both spatial and frequency domains, effectively preserving fine-grained target details and mitigating the information loss typically caused by traditional downsampling techniques. To better account for the spatial alignment between predicted and actual bounding boxes and thus improve the convergence speed, the Inner-SIoU was introduced. Finally, a novel module named C2f_CDWR was constructed to enhance feature representation by incorporating diverse receptive field patterns, enabling more effective extraction of deep and multi-scale information. Experimental findings confirmed that the designed method substantially enhanced the detection performance of small infrared targets while maintaining model efficiency.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No.62361034)

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] LU A D, QIAN C, LI C L, et al. Duality-gated mutual condition network for RGBT tracking. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(3): 4118-4131.

[2] LI C L, XIANG Z Q, TANG J, et al. RGBT tracking *via* noise-robust cross-modal ranking. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(9): 5019-5031.

[3] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[4] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 386-397.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 779-788.

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 21-37.

[7] LIU Y, SUN P, WERGELES N, et al. A survey and performance evaluation of deep learning methods for small object detection. Expert Systems with Applications, 2021, 172: 114602.

[8] TANG G Y, NI J J, ZHAO Y H, et al. A survey of object detection for UAVs based on deep learning. Remote Sensing, 2023, 16(1): 149.

[9] DAI J S, WU L, WANG P K. Overview of UAV target detection algorithms based on deep learning//2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence, December 17-19, 2021, Chongqing, China. New York: IEEE, 2022: 736-745.

[10] CHEN Y F, SHIN H. Pedestrian detection at night in infrared images using an attention-guided encoder-decoder

convolutional neural network. Applied Sciences, 2020, 10(3): 809.
[11] MA X Y, LI Y. Edge-aided multiscale context network for infrared small target detection. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 7001405.
[12] MA J Y, TANG L F, XU M L, et al. STDFusionNet: an infrared and visible image fusion network based on salient target detection. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 5009513.
[13] LI S S, LI Y J, LI Y, et al. YOLO-FIRI: improved YOLOv5 for infrared image object detection. IEEE Access, 2021, 9: 141861-141875.
[14] LAU K W, PO L M, REHMAN Y A U. Large separable kernel attention: rethinking the large kernel attention design in CNN. Expert Systems with Applications, 2024, 236: 121352.
[15] VARGHESE R, M S. YOLOv8: a novel object detection algorithm with enhanced performance and robustness//2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems, April 18-19, 2024, Chennai, India. New York: IEEE, 2024: 1-6.
[16] WANG X Q, GAO H B, JIA Z M, et al. BL-YOLOv8: an improved road defect detection model based on YOLOv8. Sensors, 2023, 23(20): 8361.
[17] FU A M, ZHANG X L, XIONG N X, et al. VFL: a verifiable federated learning with privacy-preserving for big data in industrial IoT. IEEE Transactions on Industrial Informatics, 2022, 18(5): 3316-3326.
[18] SUNKARA R, LUO T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. September. 19-23, 2022, Cham: Springer Nature Switzerland, 2022: 443-459.
[19] ZHANG H, XU C, ZHANG S J. Inner-IoU: more effective intersection over union loss with auxiliary bounding box. 2023: arXiv: 2311.02877.
[20] SUO J S, WANG T Y, ZHANG X Z, et al. HIT-UAV: a high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection. Scientific Data, 2023, 10: 227.
[21] ZHAO Y A, LV W Y, XU S L, et al. DETRs beat YOLOs on real-time object detection//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 16-22, 2024, Seattle, WA, USA. New York: IEEE, 2024: 16965-16974.
[22] ZHANG H, LI F, LIU S L, et al. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection. 2022: arXiv: 2203.03605.
[23] ZHANG S L, WANG X J, WANG J Q, et al. Dense distinct query for end-to-end object detection//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 17-24, 2023, Vancouver, BC, Canada. New York: IEEE, 2023: 7329-7338.
[24] ZHANG Y, MA R G, LIANG C. Road target detection algorithm based on improved YOLOv5 in UAV images. Journal of Measurement Science and Instrumentation, 2024, 15(1): 128-139.
[25] TIAN Y J, YE Q X, DOERMANN D. YOLOv12: attention-centric real-time object detectors. 2025: arXiv: 2502.12524.

改进YOLOv8的多尺度特征融合红外小目标检测算法

丁尚思，杨桂芹*，甘炳坤

兰州交通大学 电子与信息工程学院，甘肃 兰州 730070

摘　要：针对红外场景下小目标检测目标像素低、背景复杂等问题，本文提出了一种基于YOLOv8的多尺度特征提取的目标检测模型，首先，将网络所有的下采样卷积替换为小波下采样模块（Haar wavelet downsampling），在下采样过程中有效地保留了红外图像中的细粒度细节。为了进一步加强特征提取，对SPPF模块进行了修改，加入了可分离卷积，可双向（即水平和垂直）扩展感受野，从而在更大范围内捕捉更全面的空间细节。然后，设计了一个C2f_CDWR模块，利用不同速率的扩张卷积实现不同感受野的自适应特征提取，从而提高对不同大小物体的检测性能。最后，为了提高检测性能，将YOLOv8中原有的CIoU损失函数替换为Inner-SIoU，从而有效提高了边界框回归精度，并显著增强了模型检测小型红外目标的能力。实验结果表明，在HIT-UAV数据集上进行的实验评估表明，增强型YOLOv8模型的精确度达到90.5%，召回率达到75.9%，平均精度mAP达到85.7%。这些结果表明，在红外目标检测方面，YOLOv8基线模型和基准模型都有了明显改善。

关键词：红外图像；小目标检测；多尺度特征提取；扩张卷积；YOLOv8

引用格式：DING Shangsi, YANG Guiqin, GAN Bingkun. Improved multi-scale feature fusion for infrared small target detection based on YOLOv8. Journal of Measurement Science and Instrumentation, 2026, 17(2): 208-218. DOI: 10.62756/jmsi.1674-8042.2026018