Foreign Object Detection Algorithm for Transmission Lines Based on Improved RT-DETR

GE, FuDi <http://orcid.org/0009-0004-6477-5398>, DING, Yunfei <http://orcid.org/0000-0001-5223-7585>, WU, Xingtao <http://orcid.org/0009-0008-1785-3671>, SI, Yuxin <http://orcid.org/0009-0005-8852-7919>, WANG, Lina <http://orcid.org/0000-0003-4919-0645>, DING, Dong, ZHANG, Hongwei <http://orcid.org/0000-0002-7718-021X> and WANG, Xichao

Engineering Research Express

ACCEPTED MANUSCRIPT

Foreign Object Detection Algorithm for Transmission Lines Based on Improved RT-DETR

To cite this article before publication: FuDi Ge *et al* 2025 *Eng. Res. Express* in press https://doi.org/10.1088/2631-8695/ae2c40

Foreign Object Detection Algorithm for Transmission Lines Based on Improved RT-DETR

Fudi Ge[1], Yunfei Ding[1],* Xingtao Wu[1], Yuxin Si[1], Lina Wang[2], Dong Ding[1] , Xichao Wang[1] , Hongwei Zhang[3]

[1] School of Electrical Engineering, Shanghai Dianji University, Shanghai 201306, China

[2] School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

[3] National Centre of Excellence for Food Engineering, Sheffield Hallam University, Sheffield S9 2AA, UK

*Correspondence:yunfeiding@hotmail.com;

Abstract: In order to solve the problems of complex background, variable target scale, and frequent false and missed detections in transmission line foreign object detection, an algorithm based on improved RT-DETR is proposed in this paper. The algorithm enhances the feature extraction capability and background interference suppression by introducing a CRMB module with integrated inverted residual shift module (iRMB) and cascade group attention (CGA). In addition, a SSFF-Slimneck cross-scale feature fusion network is proposed to mitigate the information loss during feature fusion. Focaler-Shape-IoU is adopted as the bounding box loss function to accelerate model convergence, enhance generalisation capability and improve detection performance. The experimental results show that the proposed method improves 3.3% and 2.3% on mAP@50 and mAP@50:95, respectively, while the parameters and computation are reduced by 24.5% and 16.4%, respectively. This indicates that the proposed method achieves higher detection accuracy while reducing the computational complexity, which significantly improves the foreign object detection capability of transmission lines.

Keywords: Transmission line Foreign Object; RT-DETR; CRMB; SSFF-Slimneck; Focaler-Shape-IoU

1.Introduction

As technology advances and urbanization accelerates, the demand for electricity continues to rise, prompting the power grid to evolve for greater intelligence and efficiency. However, this transformation presents significant challenges, particularly regarding the security of power infrastructure [1]. Transmission lines, as the core of the energy network, are vital for maintaining a steady and reliable electricity supply [2]. Various external elements, such as bird nests, kites, balloons, and plastic debris, frequently become entangled in transmission lines and towers for different reasons, posing serious threats to the stable and secure operation of the power grid [3]. Studies suggest that foreign object interference is a major cause of repeated grid failures, compromising the system's stability and safety [4]. In severe situations, such disruptions can trigger extensive blackouts, greatly impacting socio-economic activities and daily life [5]. Consequently, enhancing the efficiency and accuracy of foreign object detection in transmission lines has become a crucial challenge in the power industry.

Traditional target detection approaches primarily depend on feature engineering and classical machine learning methods. Nonetheless, these techniques often face drawbacks, including slow processing, limited generalization capability, reliance on manually designed features, and difficulties in end-to-end training [6]. The emergence of deep learning has led to remarkable progress, significantly boosting detection accuracy, speed, and adaptability, addressing many shortcomings of conventional methods.

Deep learning-based target detection algorithms are primarily divided into two categories. The first category consists of two-stage models like Faster Region-based Convolutional Neural Networks (Faster RCNN) [7] and Mask Region-based Convolutional Neural Networks (Mask RCNN) [8], which boost detection accuracy by generating candidate regions and refining classification. For example, Li et al. [9] applied Faster RCNN for detecting bird nests on transmission lines using UAVs. They leveraged K-means clustering to establish anchor dimensions and enhance coordinate accuracy. Furthermore, they optimized foreground and background balance during training and incorporated the focal loss function in the classification phase of the region proposal network. Despite these advancements, challenges such as intricate backgrounds, small targets, and inconsistent morphological traits remain, resulting in feature loss, reduced accuracy, and increased misclassification rates. The second category includes single-stage approaches like You Only Look Once (YOLO) [10] and Single Shot Multibox Detector (SSD) [11], which perform detection in a single forward pass, making them well-suited for real-time applications. Their simpler network structures also facilitate easier implementation and training [12]. For example, Sun et al. [13] proposed an enhanced YOLOv8-based foreign object detection model, integrating Swin Transformer, Asymptotic Feature Pyramid Network (AFPN), and a novel Focal SIoU loss function to improve accuracy and real-time performance. Similarly, Bin et al. [14] proposed an enhanced YOLOv8n-based model, substituting the Bottleneck in the C2f module with an SC Block structure to minimize computational load. They also integrated a dual-attention mechanism (IEMA) to boost both spatial and channel feature representation. Xu et al. [15] further developed this concept by

embedding the MSDA attention mechanism into YOLOv8n and replacing the CIoU loss function with Focal-EIoU to enhance detection precision. Nonetheless, both two-stage and single-stage models primarily depend on non-maximum suppression (NMS) to eliminate redundant bounding boxes. NMS faces inherent issues, such as inefficiency, challenges in parallel processing, and slower inference speeds. Additionally, determining the optimal NMS threshold to prevent target loss in various contexts continues to be a complex challenge.

While the Transformer [16] is mainly used in the field of natural language processing, the Detection Transformer (DETR) [17] is the first end-to-end target detection model based on the Transformer architecture, proposed by the Facebook team. DETR converts the task of target detection into an ensemble prediction problem. It encodes the input set of images and targets into two sets, and then predicts the class, location, and number of targets by matching these two sets. This conversion not only simplifies the detection process, but also avoids the tedious post-processing steps in traditional methods, creating a new paradigm for target detection. However, its training convergence is slow and requires a large number of training rounds to achieve better performance. And the detection of small targets is poor. The Deformable DETR [18] can well solve the problems of slow convergence and poor performance of small target detection of DETR. Deformable attention is introduced to focus on the local area to accelerate the convergence and improve the performance of small targets. In addition, multi-scale feature fusion can enhance the detection ability of targets at different scales. However, the query initialisation of Deformable DETR lacks explicit semantic information, which causes the decoder to require more iterations to converge to the exact target location. Random or fixed initialisation of queries may lead to insufficient adaptation to complex scenes. In contrast, the DINO [19] combines content queries and anchor box initialisation queries. The query contains both semantic information and positional a priori, which significantly improves the initialisation quality, accelerates convergence and improves accuracy.

RT-DETR [20], based on DETR, integrates Transformer and CNN architectures, enabling the model to directly capture global information from input images and produce final object detection results without the need for traditional NMS algorithms to remove redundant bounding boxes [21]. This approach not only simplifies the detection process but also improves both speed and accuracy. For example, Xue et al. [22] introduced FECI-RTDETR, a lightweight algorithm for UAV-based infrared small target detection. They utilized a lightweight RPConv-Block module to enhance spatial feature extraction and combined an efficient additive attention module with an in-scale feature interaction module, forming the EA-AIFI module, thus improving small target detection accuracy. Similarly, Wang et al. [23] proposed PHSI-RTDETR, another lightweight algorithm for UAV aerial infrared small target detection. They upgraded the backbone feature extraction network with a lightweight RPConv-Block module to better capture small target features and incorporated a HiLo attention mechanism with an in-scale feature interaction module in the hybrid encoder to focus on dense targets. Gao et al. [24] created a lightweight small object detection network for UAV aerial images by substituting convolution operations in RT-DETR with PConv, reducing

computational costs while preserving accuracy. They also introduced structural reparameterization and a multi-scale attention mechanism in the backbone to strengthen feature extraction and enhance small object detection performance. Kong et al. [25] proposed the Drone-DETR model based on RT-DETR, featuring the Effective Small Object Detection Network (ESDNet) to retain detailed small object data and reduce redundant computations. Additionally, they incorporated the Enhanced Dual Path Feature Fusion Attention Module (EDF-FAM) into the neck network to improve multi-scale object detection. Wang et al. [26] proposed a Yolov8-CDD model for concrete defect detection, which improves the feature extraction capability by utilizing a robot transformer module. A convolutional triple state attention module is used to integrate different dimensional features, thus improving the accuracy of the model.

These papers proposed a series of target detection methods, which improved the detection of foreign objects on transmission lines to a certain extent. However, these methods have some limitations, such as easy to be interfered by complex background, easy to miss and misdetect, and poor multi-scale target detection ability. The comparison results of target detection algorithms are shown in Table 1.

In light of the aforementioned limitations, this paper innovatively integrates, for the first time, the CRMB module, the SSFF-Slimneck module, and the Focaler-Shape-IoU module in an effective and synergistic manner, and applies them to the efficient RT-DETR object detection algorithm. The algorithm proposed in this paper significantly enhances feature extraction capabilities in complex backgrounds and improves detection performance for objects of different scales by introducing the CRMB module, which combines the iRMB module with the CGA attention mechanism. Meanwhile, the proposed SSFF-Slimneck cross-scale feature fusion network effectively addresses the issue of information loss during the feature fusion process, further enhancing the model's ability to detect multi-scale objects. Additionally, by adopting Focaler-Shape-IoU as the bounding box loss function, the model not only accelerates convergence but also improves its generalization ability and detection accuracy. The innovation of this paper lies in the meticulous architectural design that achieves organic integration and collaborative operation among various modules, thereby attaining unprecedented performance improvements in the task of detecting foreign objects on transmission lines. This synergistic effect not only enhances detection accuracy but also significantly reduces computational complexity, providing an efficient and robust solution for detecting foreign objects on transmission lines.

Table 1. Comparison results of target detection algorithms

Ref.	Algorithm	Advantages	Disadvantages	originality
[9]	RCNN	k-means clustering to improve accuracy	Low accuracy of small target detection	ROI mining module
[13]	YOLOv8	lightweight construction, Rapid detection	Detection is influenced by the environment	Integrated Swin Transformer
[14]	YOLOv8n	Small model parameters for unmanned and deployed	Affected by complex environments	Dual Attention Mechanism IEMA
[15]	YOLOv8n	Detection of foreign objects at different scales	Slow reasoning	MSDA attention mechanism

[22]	FECI-RTDETR	Lightweight for deployment	High false positive rate	RPConv-Block
[23]	PHSI-RTDETR	Reduced leakage and false detection rates	Insufficient generalisation capacity	AIFI-HiLo module
[24]	RTPR-DETR	Reduced number of parameters in the model	Easy to miss	Partial Convolution
[25]	Drone-DETR	Unaffected by the complexity of the context	Slow detection speed	ESDNet
[26]	YOLOv8-CDD	Improvement of global feature extraction	Susceptible to background interference	Bot-transformer module

2.Methods

2.1RT-DETR Algorithm

RT-DETR is a real-time, end-to-end object detection model based on Transformer architecture. Its core structure consists of three main components: a backbone network, an efficient hybrid encoder, and a detection head-equipped decoder. The backbone network forms the foundation of the RT-DETR model, primarily tasked with extracting relevant feature representations from the input images. The hybrid encoder, which is highly efficient, includes two key modules: the Attention-based Intra-scale Feature Interaction (AIFI) and the CNN-based Cross-Scale Feature Fusion Module (CCFM). These modules collaborate to extract and process multi-scale features from the final three stages (S3, S4, S5) of the backbone network. Lastly, the decoder converts the encoder's output into the final detection results. The complete architecture of the RT-DETR model is shown in Fig. 1.

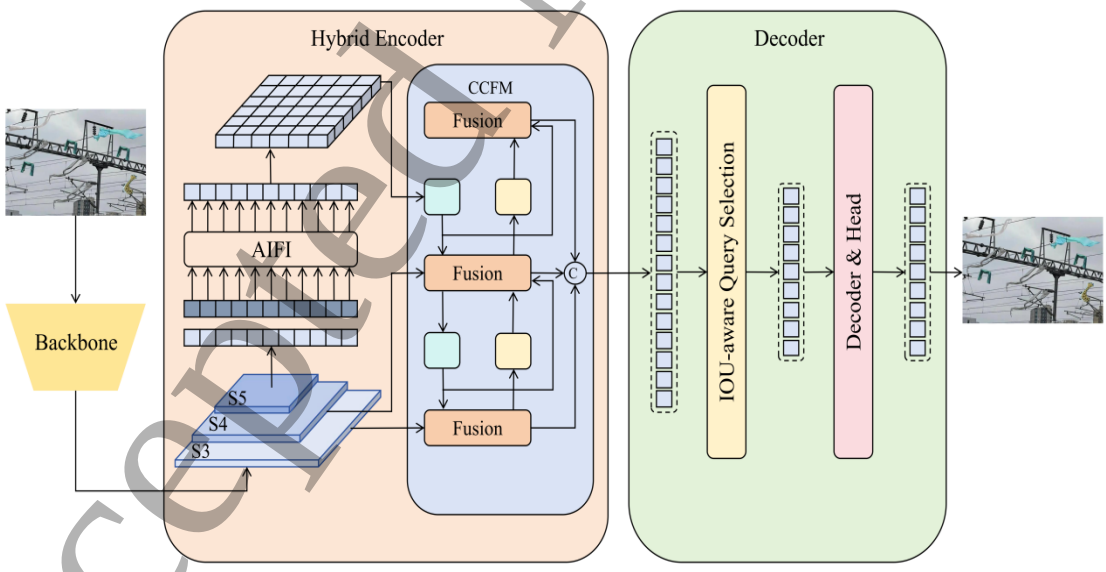


Fig. 1 RT-DETR network structure diagram

The precision of detecting foreign objects on transmission lines is improved through refinements to the RT-DETR model. Initially, the CRMB module substitutes

the original BasicBlock module, greatly enhancing the extraction of features for transmission line foreign objects while efficiently lowering the model's parameter quantity and complexity. Furthermore, the SSFF-Slimneck cross-scale feature fusion network replaces the former CCFM module, aiming to lessen computational burden and inference delay while strengthening the model's capability to detect foreign objects across different scales. Finally, the Focaler-Shape-IoU is utilized as the bounding-box loss function, expediting model convergence and improving its generalization capacity along with detection effectiveness. The layout of the upgraded RT-DETR network is described in Fig. 2.

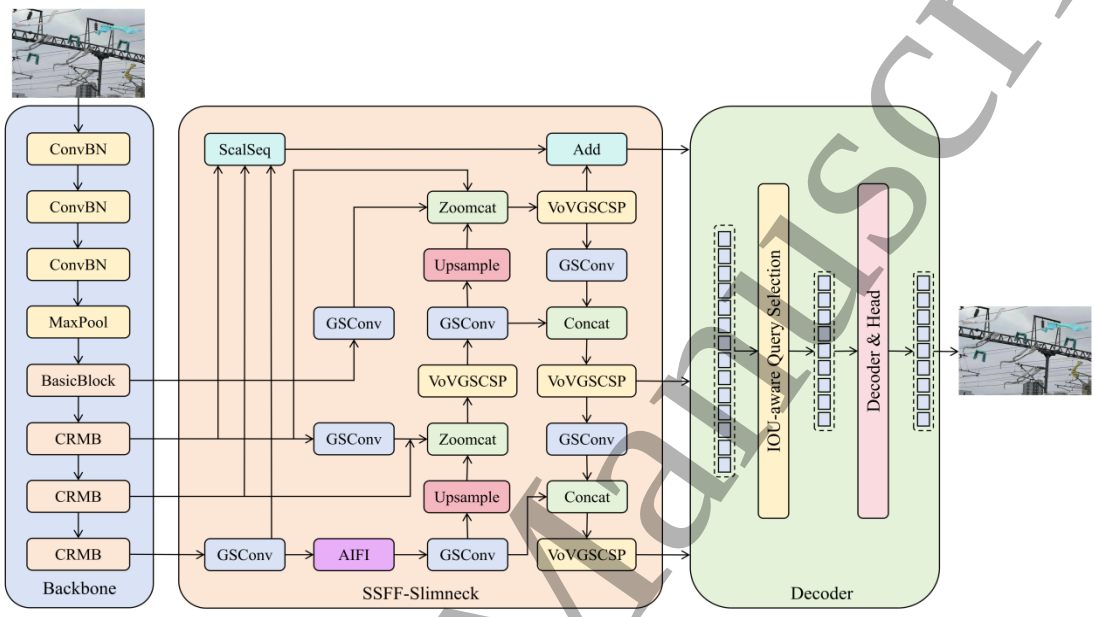


Fig. 2 Structure of the improved RT-DETR network

2.2 Improvement of The Feature Extraction Module

2.2.1 iRMB Module

Due to the challenges posed by complex backgrounds and subtle features in detecting foreign objects on transmission lines, which impact both accuracy and speed, the Inverted Residual Mobile Block (iRMB) [27] is employed for feature extraction. iRMB combines the efficiency of CNN-based local feature modeling with the dynamic modeling capabilities of Transformer. While CNNs excel at extracting local features, Transformers are adept at capturing global context. By alternating between convolutional operations and self-attention mechanisms, the model effectively captures both local and global features, enabling robust feature extraction for foreign object targets. The structure of iRMB is illustrated in Fig. 3. The input undergoes initial compression via a 1x1 convolutional layer, minimizing channel numbers and subsequently decreasing computational demands in later layers. A segment of the output from this first 1x1 convolutional layer contributes to forming the Extended Window Multi-Head Self-Attention Matrix (EW-MHSA) by facilitating interactions between Q and K. This resultant output is then scaled by V to generate an attention-refined feature map. The feature extraction process is further strengthened through a 3x3 depthwise separable convolution. To restore the channel count, another 1x1 convolutional layer

expands the processed output. Ultimately, the final output is derived by merging this layer's output with the module's initial input through a skip connection. iRMB significantly enhances the model's performance by refining its lightweight architecture while simultaneously boosting its robustness and flexibility in detecting foreign objects on transmission lines under complex background conditions.

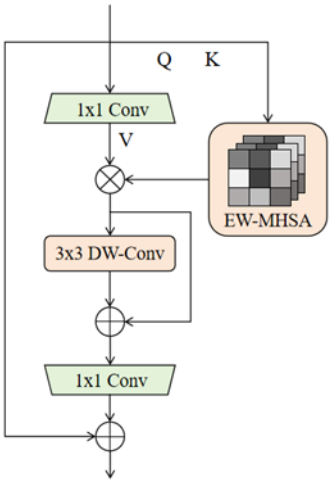


Fig. 3 Structure of iRMB

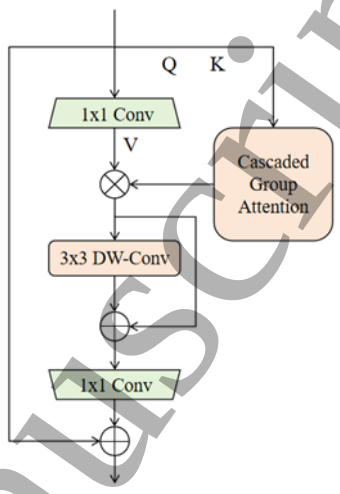


Fig. 4 Structure of CRMB

2.2.2 CGA Module

Although the EW-MHSA in iRMB can capture a larger range of contextual information through the design of the extended window, which enhances the model's feature extraction of transmission line foreign objects in a complex background. However, it is still lacking for the problem of variable scale of foreign objects. In contrast, Cascaded Group Attention (CGA) [28] divides the feature map into multiple groups by grouping and performs attention computation independently within each group. Using multiple cascaded attention layers, each layer captures different levels of feature information, so it is more adaptable to different scales of foreign objects. Therefore, a CRMB module is proposed, by combining iRMB and CGA, which not only captures both local and global features in complex backgrounds, but also improves the detection ability for different scales of foreign objects. The structure of CRMB is shown in Fig. 4.

The workflow of CGA is as follows, firstly, the input feature is partitioned into multiple parts, each part corresponds to an attention head. In Eq. X_i denotes the input feature of the ith block, which is partitioned into multiple attention heads, i.e., $X_i = [X_{i1}, X_{i2}, \ldots, X_{ih}]$, where $1 < j \le h$ and h is the total number of heads. Next, for each attention head j a projection of query (Q), key (K) and value (V) is applied to its corresponding input partition. This typically involves three projection matrices $W_{ij}^{Q}, W_{ij}^{K}, W_{ij}^{V}$, which map the input features into the space of queries, keys and values. Each header performs these projection operations independently. Then, the attention weights are computed and the values (V) are weighted and summed according to the weights to obtain the output of that attention head. The formula for self-attention is shown in equation (1).

$$\tilde{X}_{ij} = Attn\left(X_{ij}W_{ij}^{Q}, X_{ij}W_{ij}^{K}, X_{ij}W_{ij}^{V}\right) \quad (1)$$

where X_{ij} denotes the result of the self-attention head j 's attention on the input segmentation X_{ij} . This step is the core of the self-attention mechanism, which enables each attention head to focus on a different part of the input features. Next, the outputs of all attention heads are cascaded together to form a larger feature vector, which contains information from all the heads and therefore has a stronger representation. This cascaded feature vector is then projected through a linear projection layer to restore its dimensionality to that of the input features. This linear projection layer is usually represented by a weight matrix W_i^P, so the projection operation is calculated as shown in equation (2).

$$\tilde{X}_{i+1} = Concat[\tilde{X}_{ij}]_j =_{1:h} W_i^P \tag{2}$$

Therefore, the final output $\tilde{X}_{i+1}$ is obtained, which contains all the information of the attention header and has the same dimension as the input features. In addition, a recursive formula will be used to further process the input features. The recursive formula is shown in equation (3).

$$X_{ij}^{'} = X_{ij} + \tilde{X}_{i(j-1)} \quad 1 < j \le h \tag{3}$$

denotes the sum of the j th input segmentation X_{ij} and the output $\tilde{X}_{i(j-1)}$ of the previous attention head $(j-1)$. The result of this summation serves as the new input feature for the jth attention head to compute self-attention. The recursive processing can further enhance the representation of the model. The CGA structure is shown in Fig. 5.

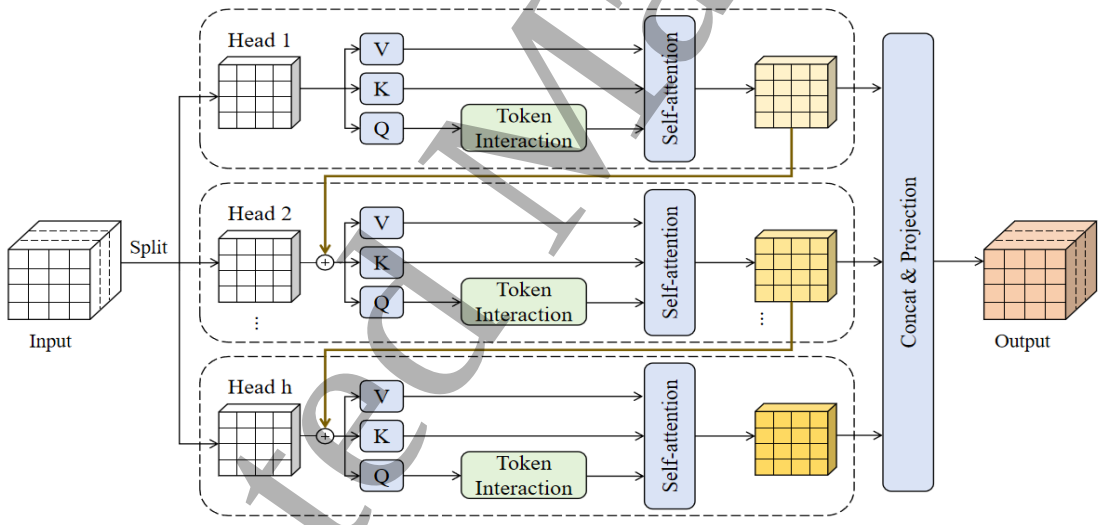


Fig. 5 Structure of CGA

2.3Improvement of Cross-Scale Feature Fusion Module

2.3.1 SSFF Module

In the field of transmission line foreign object detection, conventional single-scale feature fusion techniques often fail to fully capture the varied attributes of foreign objects, which differ in shape, size, and position. To overcome this drawback, Scale Sequence Feature Fusion (SSFF) [29] is utilized, as it efficiently merges feature data across different scales, leading to more accurate detection of foreign objects on transmission lines and improving overall detection precision. The framework of SSFF is shown in Fig. 6. The procedure starts by applying a 1×1 2D convolutional kernel to

the P4 and P5 layers, modifying their channel dimensions to match those of the P3 layer, thereby maintaining scale consistency. Next, the feature map sizes of the P4 and P5 layers are adjusted to align with the P3 layer through nearest-neighbor interpolation. This method ensures smooth multi-scale feature integration, enhancing the model's capability to identify foreign objects with diverse properties. In this process, the dimensions of each feature layer were extended, and from the original three-dimensional tensor form (containing three dimensions: height, width and channel) was converted to a four-dimensional tensor with a new depth dimension. This conversion allows the feature map to have more information in the depth direction. Subsequently, along the dimension of depth, the transformed four-dimensional feature map is stitched together to construct a new three-dimensional feature map. This step lays the foundation for the subsequent 3D convolution operation. Finally, after the feature map stitching was completed, a series of processing means were employed to extract the scalar sequence features. These include the application of the SiLU activation function to increase the nonlinear expressiveness of the model, the use of 3D convolution to further extract the features, and the use of 3D batch normalization to stabilize the training process and improve the generalization ability of the model.

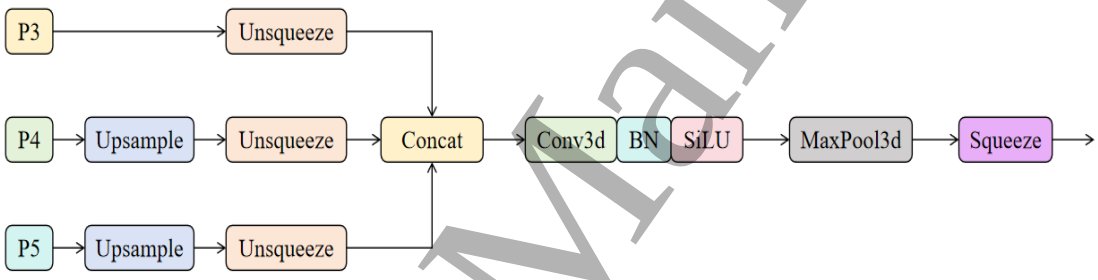
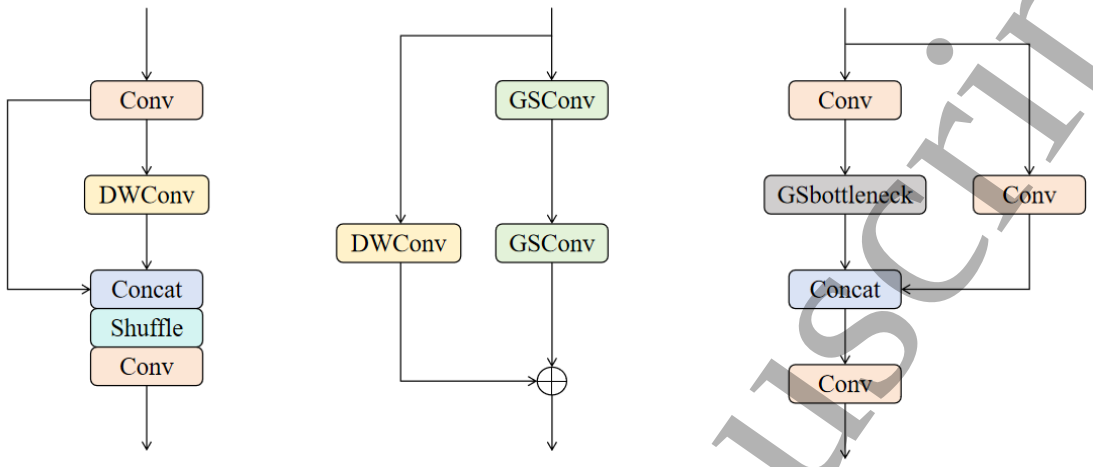


Fig. 6 Structure of SSFF

2.3.2 Slimneck Module

Slimneck [30] follows a lightweight design approach, ensuring high detection accuracy while reducing model parameters and computational overhead. When integrated with Scale Sequence Feature Fusion (SSFF), it not only effectively detects foreign objects in complex backgrounds but also merges multi-scale feature information to improve fusion effectiveness. Furthermore, it minimizes redundant data and enhances computational efficiency. Slimneck consists of three modules: GSConv, GSbottleneck, and VoVGSCSP, as illustrated in Fig. 7. As presented in Fig. 7(A), GSConv initially produces a segment of the feature maps via standard convolution, followed by additional feature maps generated through depthwise separable convolution. These two feature map sets are subsequently combined to ensure uniform information distribution across different feature maps. This method enables GSConv to retain the efficiency of depthwise separable convolution while preserving essential feature details. In Fig. 7(B), GSbottleneck utilizes GSConv as its fundamental element, forming an optimized bottleneck structure by layering multiple GSConvs and depthwise separable convolutions. This configuration lowers model parameters and computational complexity while maintaining strong feature learning performance. As

depicted in Fig. 7(C), VoVGSCSP applies a one-time aggregation technique to create cross-scale feature fusion modules, facilitating effective information fusion across feature maps at various stages. By integrating these components, Slimneck greatly improves the accuracy and efficiency of transmission line foreign object detection while preserving a lightweight and streamlined model design.



(A) GSConv (B) GSbottleneck (C) VoVGSCSP

Fig. 7 Structure of Slimneck

2.4Improvement of The Loss Function

2.4.1 The Loss Function Focaler-IoU

The Focaler-IoU loss function [31] reformulates IoU loss using linear interval mapping, allowing the model to allocate different levels of attention to samples depending on their complexity. This approach helps the model prioritize samples that significantly impact detection accuracy, thereby improving overall performance. Additionally, in transmission line foreign object detection, the distribution of positive and negative samples, as well as easy and difficult samples, is often imbalanced. The Focaler-IoU loss function addresses this imbalance through its unique design, effectively compensating for the limitations of existing bounding box regression methods. This approach strengthens the model's capability to manage imbalanced datasets, thereby enhancing detection accuracy. The definition of the Focaler-IoU loss function can be found in equations (4) and (5).

$$IoU^{\text{focaler}} = \begin{cases} 0, & IoU < d \\ \dfrac{IoU - d}{u - d}, & d \leq IoU \leq u \\ 1, & IoU > u \end{cases} \tag{4}$$

$$L_{\text{Focaler-IoU}} = 1 - IoU^{\text{focaler}} \tag{5}$$

The Focaler-IoU loss function, as outlined in Equation (4), dynamically modifies the loss magnitude in response to the IoU value. If the IoU value drops below the lower threshold d, the loss is reduced to zero. Conversely, when the IoU value surpasses the

upper threshold u, the loss attains its peak value of 1. Within the range between d and u, the loss increases proportionally with the *IoU* value. Equation (5) provides a formalized definition of this loss function. This approach ensures heightened sensitivity of the loss function within a particular *IoU* range, allowing it to concentrate on samples where the predicted bounding box partially coincides with the true bounding box. Such samples present a moderate difficulty level—neither excessively complex nor too simple. Prioritizing these samples aids in minimizing training time and enhancing training efficiency.

2.4.2 The Loss Function Shape-IoU

The Shape-IoU loss function [32] improves the precision of loss computation by integrating the bounding box's shape and size attributes. Conventional bounding box regression techniques primarily emphasize the geometric association between the ground truth (GT) box and the predicted box, computing loss based on their relative position and shape variations. Nonetheless, these approaches often neglect the impact of the bounding box's intrinsic characteristics, such as shape and size, on regression accuracy. In contrast, Shape-IoU incorporates shape and scale elements into its computations, allowing for a more thorough assessment of bounding box regression performance. During processing, Shape-IoU not only accounts for the overlap region between two bounding boxes but also conducts a detailed analysis of their shape and scale discrepancies. This method enhances both the model's detection accuracy and its generalization capability. The framework of Shape-IoU is depicted in Fig. 8, followed by its formula.

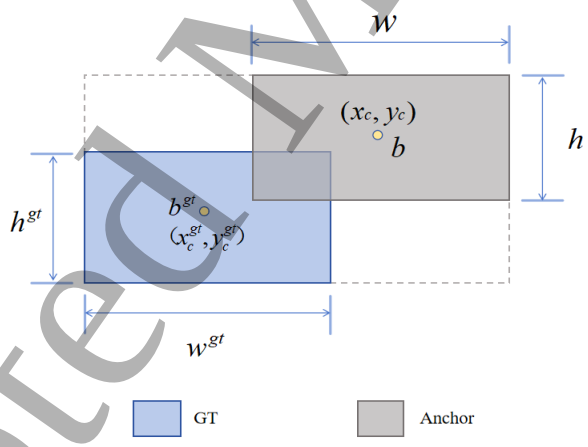


Fig. 8 Structure of Shape-IoU

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{6}$$

$$ww = \frac{2 \times (w^{gt})^{scale}}{(w^{gt})^{scale} + (h^{gt})^{scale}} \tag{7}$$

$$hh = \frac{2 \times (h^{gt})^{scale}}{(w^{gt})^{scale} + (h^{gt})^{scale}} \tag{8}$$

$$distance^{shape} = hh \times \frac{\left(x_c - x_c^{gt}\right)^2}{c^2} + ww \times \frac{\left(y_c - y_c^{gt}\right)^2}{c^2} \tag{9}$$

$$\Omega^{shape} = \sum_{t=w,h} (1 - e^{-\omega_t})^{\theta}, \theta = 4 \tag{10}$$

$$\begin{cases} \omega_w = hh \times \dfrac{|w - w^{gt}|}{\max(w, w^{gt})} \\ \omega_h = ww \times \dfrac{|h - h^{gt}|}{\max(h, h^{gt})} \end{cases} \tag{11}$$

$$L_{\text{Shape-IoU}} = 1 - \text{IoU} + \text{distance}^{shape} + 0.5 \times \Omega^{shape} \tag{12}$$

First, w^{gt} and h^{gt} denote the width and height of the ground truth (GT) box, while w and h indicate the width and height of the predicted box. Equation (6) determines the intersection and union areas of the two bounding boxes, defining *IoU* as their ratio. Equations (7) and (8) introduce a scaling factor that corresponds to the target size within the dataset, where ww and hh serve as weight coefficients for horizontal and vertical directions. These coefficients are derived from the GT box's shape. Equation (9) calculates the center distance by assessing the difference between the x and y coordinates of the bounding boxes' center points. Among them, c in the denominator represents the diagonal length of the smallest rectangle that surrounds the predicted box and the ground truth box. Equation (11) defines two shape cost variables, ω_w and ω_h, which measure the discrepancies in width and height between the bounding boxes. Equation (10) uses these variables to calculate the shape cost, which increases as the differences in width or height grow. Finally, Equation (12) computes the final *IoU* value, incorporating the shape cost and center distance to provide a more accurate assessment of bounding box alignment.

2.5Model improvement analysis

The iRMB module combines the local feature extraction capability of CNN with the global context modeling capability of Transformer. Intended to enhance the adaptability of the model to complex backgrounds while maintaining lightweight structure and reducing computational burden. The CGA module divides the feature map into multiple parts through grouping and cascading attention mechanisms, and performs independent attention calculations for each part to capture feature information at different scales. Thus, it can adaptively enhance the feature representation of targets at different scales, improving the model's detection ability for multi-scale targets. Therefore, a CRMB module is proposed. By combining the iRMB module and CGA attention mechanism, not only can local and global features be captured simultaneously under complex background conditions, but the detection ability of foreign objects at different scales can also be improved. This provides a rich feature information foundation for subsequent feature fusion.

For the multi-scale features extracted by CRMB, a lightweight and efficient feature fusion network is needed to process foreign object information of different

scales. The SSFF-Slimneck proposed in this article combines the multi-scale feature fusion capability of SSFF with the lightweight design advantage of Slimneck. The SSFF module effectively integrates feature information of different scales through sequence feature fusion strategy. It uses 1 × 1 convolution to adjust the channel dimension of the feature map, and adjusts the size of the feature map through nearest neighbor interpolation to ensure smooth fusion of multi-scale features. Slimneck design utilizes lightweight components such as GSConv, GSbleneck, and VoVGSCSP to reduce model parameter and computational complexity while maintaining efficient feature learning capabilities. Therefore, SSFF-Slimneck can effectively integrate feature information from different scales while maintaining model efficiency, improving the model's ability to detect multi-scale targets.

Finally, the fused features of SSFF-Slimneck were optimized using the Focaler-Shape-IoU loss function to make the detection box closer to the real foreign object in terms of shape and scale, reducing localization errors caused by shape mismatch. Shape-IoU introduces a shape penalty term by calculating the difference in width and height between the predicted box and the real box, making the loss function more sensitive to changes in aspect ratio. Focaler-IoU reconstructs the IoU loss through linear interval mapping, allowing the model to give differentiated attention based on the difficulty of different samples, which helps the model focus more on samples of different scales. Therefore, Focaler-Shape-IoU effectively solves the problem of foreign object detection in transmission lines of different scales by introducing shape and scale weights and dynamic gradient allocation mechanism.

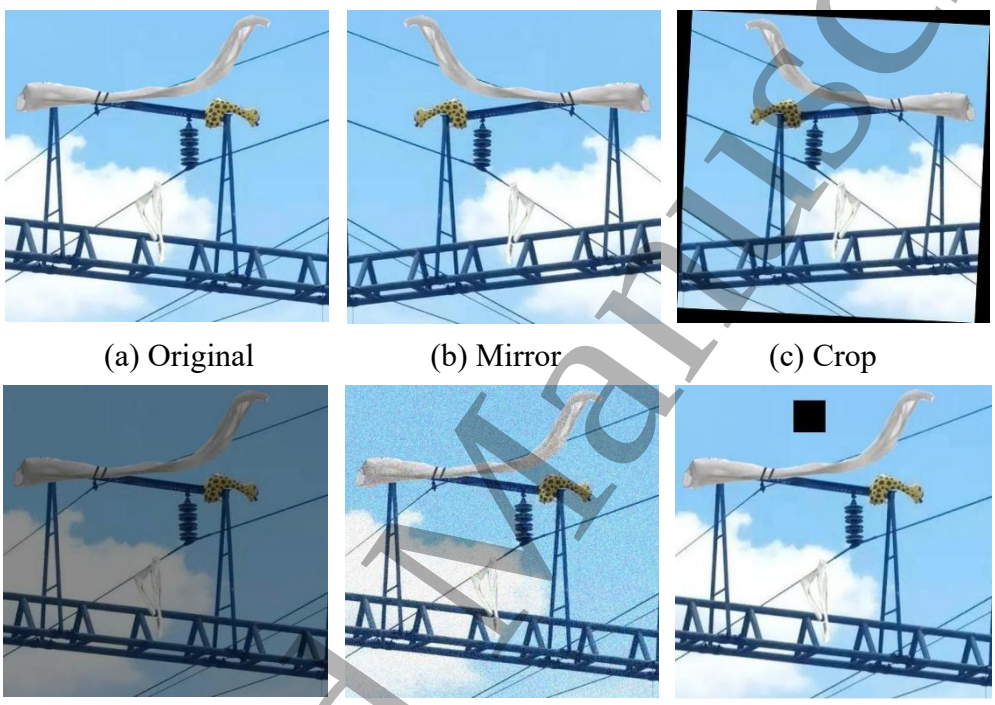
Therefore, by combining CRMB, SSFF-Slimneck, and Focaler-Shape-IoU, an end-to-end efficient detection framework has been formed, which significantly improves the efficiency of foreign object detection in transmission lines while effectively reducing the number of model parameters.

3.Experimental results and analysis

3.1 Datasets

Due to the unique challenges of the power industry, acquiring high-quality data samples is often difficult, resulting in datasets with limited quality. To address this, the dataset used in this study comprises two parts. The first part is based on RailFOD23 [33], a publicly available dataset for foreign object detection on railroad transmission lines. The second part consists of images captured by real UAVs in power grid environments, which include a wide variety of scenarios and foreign object types, thereby enhancing the dataset's diversity. The original dataset contains 822 images categorized into four classes: bird's nest, balloon, kite, and trash. To enable the model to learn more robust features and improve its generalization ability, this paper employs several data augmentation techniques, including (b) mirroring, (c) cropping, (d) grayscaling, (e) adding noise, and (f) adding masks, as illustrated in Fig. 9. After augmentation, the dataset expands to 3,288 images, which are then randomly split into training, validation, and test sets in an 8:1:1 ratio. The dataset is divided into training, validation and test sets according to 8:1:1. 80% of the training set provides sufficient samples for the model to learn the data patterns and reduce the risk of overfitting, while

10% of the validation set assists in the model tuning by evaluating the performance of the model under different hyperparameters to select the optimal combination, and also detects the overfitting in a timely manner. The other 10% of the test set is used to objectively assess the final performance of the model and verify the generalisation ability. This ratio ensures the functionality of each stage while making efficient use of data, balancing the needs of training, validation and testing, and making model development and evaluation more efficient and accurate. In this study, the dataset has been reasonably divided into training, validation and testing sets with moderate proportions. This division has been able to provide an unbiased estimate of the model performance, thus eliminating the need for further use of cross-validation.



(a) Original (b) Mirror (c) Crop

(d) Graying out (e) Adding noise (f) Adding a mask

Fig. 9 Data enhancement methods

3.2 Experimental environment and parameter configuration

The configuration of the experimental environment in this study is outlined in Table 2. In order to ensure the training effect and performance of the model, a series of hyperparameters were carefully selected and set in this study. By observing the change of the loss function during the training process, the initial learning rate was set to 0.0001 to enable the model to converge stably and quickly. Through many experiments, it was found that when the number of iterations reached 200 times, the performance of the model had already stabilised, and continuing to increase the number of iterations had a limited effect on the performance improvement, so the number of iterations was set to 200 times. The number of data loading threads is set to 4 to ensure the efficiency of data loading without overburdening the CPU. The image resolution determines the size of the image for the input model, a higher resolution can provide more image details, but at the same time it will increase the amount of computation, so a resolution of

640X640 is used. The Adam optimiser combines the advantages of the AdaGrad and RMSProp optimisers, and is able to adaptively adjust the learning rate, so the use of the Adam optimiser can result in a faster convergence of the model and better performance. better. In order to make the model fully converged, the early stop training period is set to 50 rounds.

Table 2. Experimental environment

Hardware Name	Model Number
CPU	Intel(R) Xeon(R) Gold 6430
GPU	GeForce RTX 4090
Operating System	Ubuntu20.04
PyTorch	1.11.0
CUDA	11.3
Python	3.8

3.3 Evaluation indicators

To evaluate the improved model's capability in identifying foreign objects on transmission lines, various metrics are employed. These include Precision, Recall, Average Precision (AP), mean Average Precision (mAP), Precision-Recall curve, IoU, Parameters, and GFLOPs. The corresponding calculation approaches are illustrated in the following diagram.

1、Precision

$$P = \frac{TP}{TP + FP} \tag{13}$$

True Positives (TP): indicates the number of samples that the model correctly predicted as positive classes. False Positives (FP): indicates the number of samples that the model incorrectly predicts as positive classes. Precision measures the ratio of correctly identified positive samples to all instances predicted as positive by the model, showcasing its accuracy in recognizing true positives. Its calculation is provided in equation (13).

2、Recall

$$R = \frac{TP}{TP + FN} \tag{14}$$

False Negatives (FN): indicates the number of samples that the model incorrectly predicts as negative classes. Recall indicates the model's ability to correctly detect positive samples among all actual positives, demonstrating its coverage of true instances. Its formula is provided in equation (14).

3、AP

$$AP = \int_0^1 P(r)dr \tag{15}$$

Average Precision obtains a single performance metric by calculating the accuracy

of the model at different recall rates and combining these accuracy values. The higher the AP value, the better the accuracy and robustness of the model in detecting the target. The formula is shown in equation (15).

4、mAP

$$mAP = \frac{1}{n}\sum_{n}^{i=1} AP_i \tag{16}$$

The mAP is derived by computing the mean of average precision (AP) across all categories. This metric integrates both Precision and Recall, enabling the algorithm to identify various target categories effectively. By consolidating these values, it provides a single numerical measure to evaluate the algorithm's detection performance. The corresponding formula is presented in equation (16).

5、Precision-Recall curve

The Precision-Recall curve is an important graphical tool for evaluating the performance of a target detection model by plotting the precision and recall of the model at different confidence thresholds to visualise the model's performance. The curve provides a comprehensive assessment of model performance at different thresholds, especially effective when the distribution of positive and negative samples is not balanced, and can also be used to select the optimal thresholds for balancing precision and recall, or for comparing the performance among multiple models.

6、IoU

$$IoU = \frac{\left|B \cap B^{gt}\right|}{\left|B \cup B^{gt}\right|} \tag{17}$$

B represents the predicted bounding box, and B^{gt} represents the true bounding box. IoU is defined as the ratio of the intersection area between the predicted bounding box and the true bounding box to the union area. This ratio is used to measure the degree of overlap between the predicted bounding box and the true bounding box, in order to evaluate the accuracy of the object detection model. The formula is shown in equation (17).

7、Parameters

Parameters represent the total learnable and optimizable elements in a model, influenced primarily by its network architecture. As the number of layers and model complexity grow, so does the parameter count. Excessive parameters can elevate computational load and memory usage, slowing training and increasing resource demands.

8、GFLOPs

GFLOPs measure a model's computational burden, representing the total floating-point calculations needed for inference or training.

3.4 Ablation Experiment

To assess the efficacy of the proposed model enhancements, ablation studies were performed on each improved module, with findings detailed in Table 3. The mAP@50

comparison of improvement modules is visualized in Fig. 10. Setting RT-DETR-R18 as the reference model, substituting the BasicBlock with the iRMB module in the backbone led to a 17.4% and 13.9% reduction in parameters and GFLOPs, respectively, while boosting mAP@50 and mAP@50:95 by 1%. This indicates that the iRMB module not only reduces model complexity but also strengthens multi-scale feature extraction. Additionally, incorporating the CGA module into the CRMB module improved Recall by 2.7% and further minimized parameters, underscoring the CGA module's role in enhancing feature diversity and computational effectiveness. The Slimneck design concept, when applied to the CCFM cross-scale feature fusion module, resulted in Precision and Recall gains of 3.4% and 2.2%, respectively, while lowering GFLOPs, verifying its ability to decrease computational burden and optimize feature fusion. The combination of Slimneck and the SSFF module contributed to a 1.5% increase in mAP@50, demonstrating the SSFF module's strength in boosting multi-scale information extraction and model resilience. By utilizing Focaler-Shape-IoU as the bounding box loss function, while maintaining the number of parameters unchanged, mAP@50 and mAP@50:95 have increased by 1.8% and 1.5% respectively. This approach not only enhances generalization and detection capabilities but also facilitates faster convergence. The final improved model achieved a 3.3% and 2.3% increase in mAP@50 and mAP@50:95, respectively, while reducing parameters and GFLOPs by 24.5% and 16.4%. These results validate the effectiveness of all proposed improvements, achieving higher accuracy with lower computational complexity, thus enhancing transmission line foreign object detection.

Table 3. Results of ablation experiment

Model	P(%)	R(%)	mAP@ 50(%)	mAP@ 50:95(%)	Param(M)	GFLOPs
RT-DETR-R18	88.7	81.1	86.5	52.1	19.9	57.0
RT-DETR-iRMB	89.6	81.8	87.5	53.1	16.4	49.1
RT-DETR-CRMB	90.7	84.5	88.5	53.3	15.3	46.8
RT-DETR-Slimneck	92.1	83.3	87.9	52.4	19.3	53.3
RT-DETR-SSFF-Slimneck	93.6	82.8	88.6	53.2	19.6	57.6
RT-DETR- Shape-IoU	92.0	82.7	87.4	52.3	19.9	57.0
RT-DETR- Focaler-IoU	91.6	82.1	87.6	53.5	19.9	57.0
RT-DETR- Focaler-Shape-IoU	91.7	83.7	88.3	53.6	19.9	57.0
RT-DETR-CRMB-SSFF-Slimneck	92.7	85.1	89.4	53.5	15.0	47.4
RT-DETR-CRMB-SSFF-Slimneck-Focaler-Shape-IoU	92.4	85.8	89.8	54.4	15.0	47.4

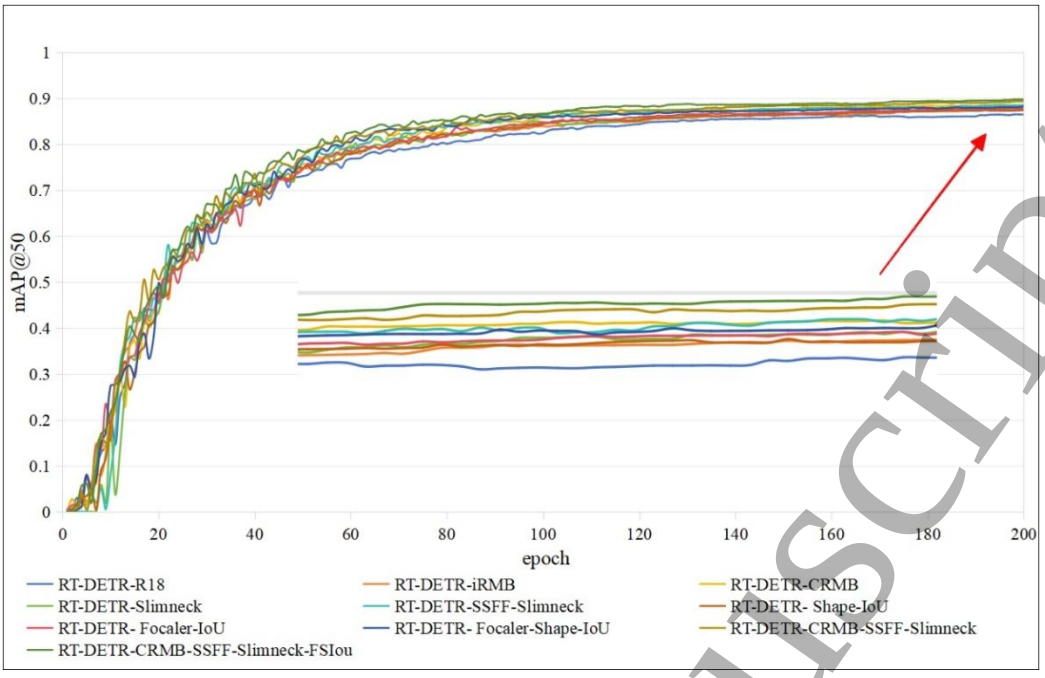


Fig. 10 Comparison of improvement module mAP@50

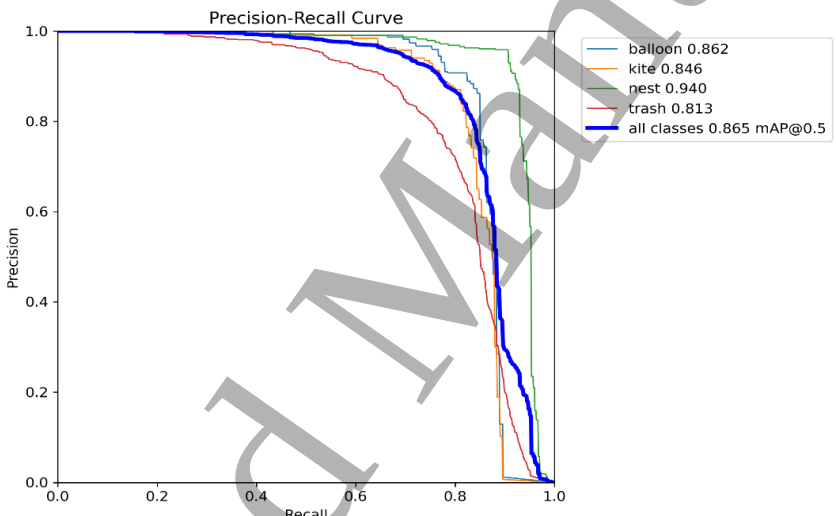


Fig. 11 The effect of PR curve of RT-DETR-R18

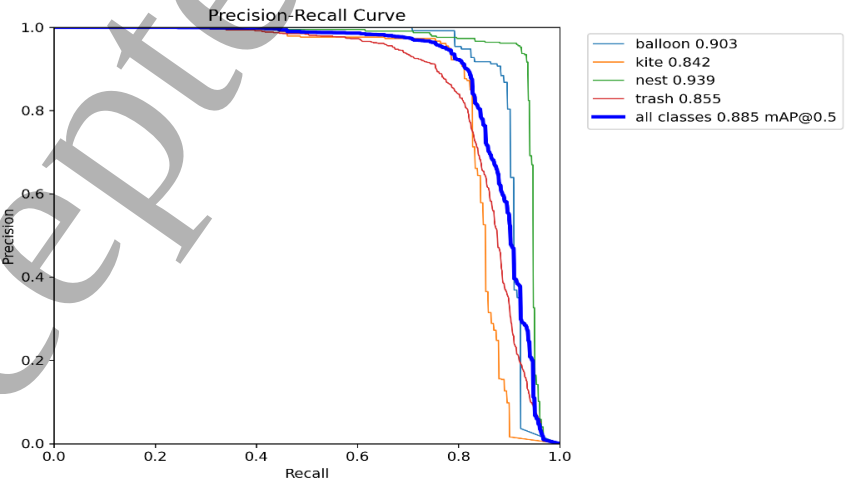


Fig. 12 The effect of PR curve of RT-DETR-CRMB

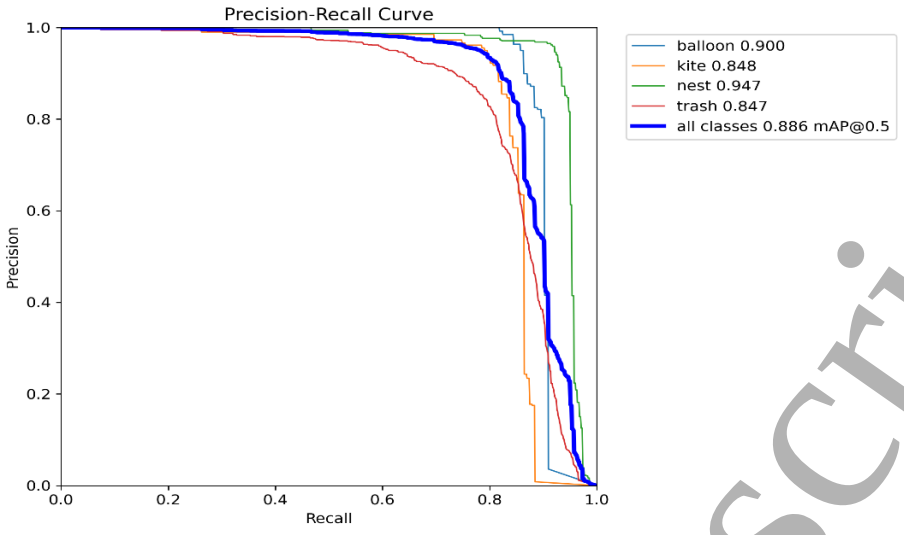


Fig. 13 The effect of PR curve of RT-DETR-SSFF-Slimneck

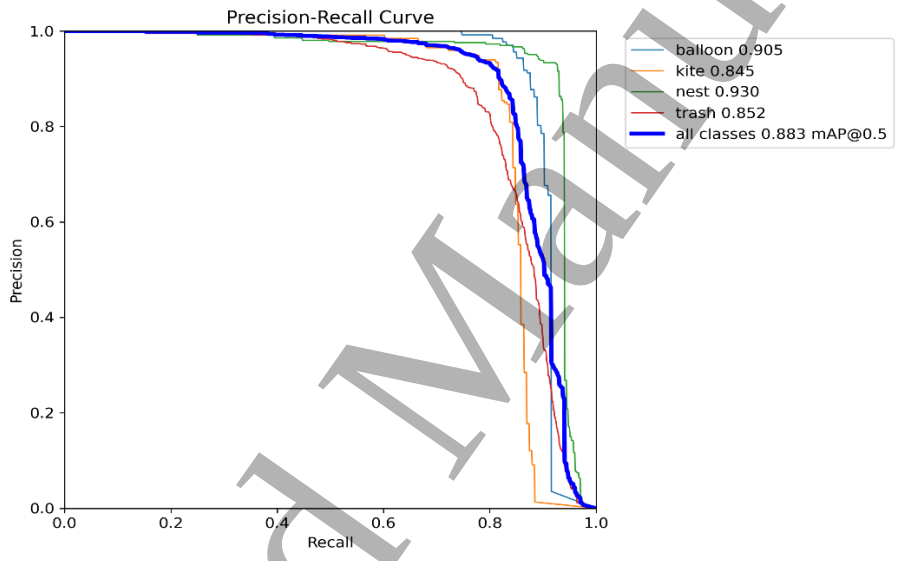


Fig. 14 The effect of PR curve of RT-DETR- Focaler-Shape-IoU

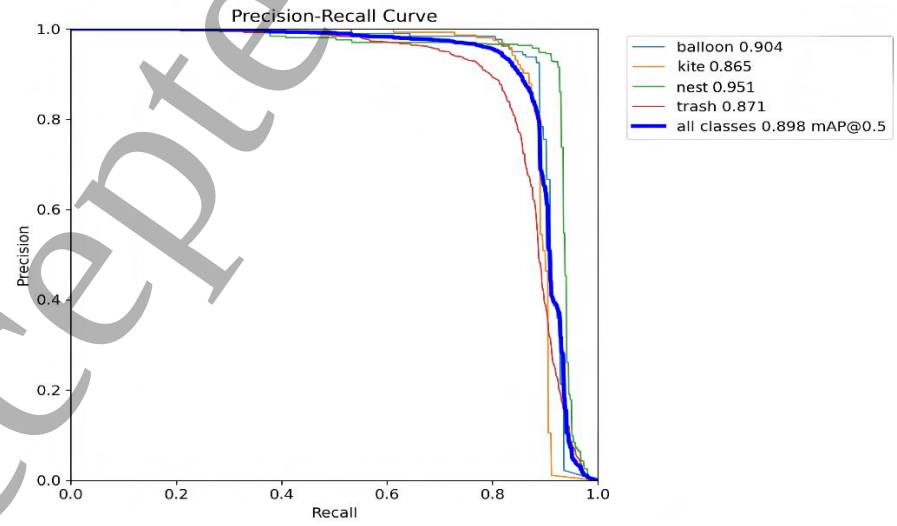


Fig. 15 The effect of PR curve of RT-DETR-CRMB-SSFF-Slimneck-FSIoU

From Figure 12, it can be observed that after incorporating the CRMB module, the detection accuracy for balloons and trash improved by 4.1% and 4.2%, respectively. Therefore, compared to the baseline model in Figure 11, the CRMB module enhances the model's ability to capture both local and global features. From Figure 13, it is evident that after adding the SSFF-Slimneck module, not only did the accuracy for balloons and trash increase, but the recognition accuracy for kites and trash also improved. Thus, the SSFF-Slimneck module not only enhances the effectiveness of multi-scale feature fusion but also further strengthens the model's ability to detect targets of varying scales. From Figure 14, it can be seen that after introducing the Focaler-Shape-IoU loss function, the detection accuracy for balloons and trash increased by 4.3% and 3.9%, respectively. Hence, the Focaler-Shape-IoU accelerates model convergence and improves the model's generalization capability and detection accuracy. From Figure 15, it is clear that the final improved model achieves accuracy enhancements across different categories, significantly boosting the model's ability to detect targets of varying scales in complex backgrounds and enhancing overall performance.

3.5 Comparative Experiment

To assess the efficiency of the proposed algorithms, this research performs a comparative study involving both two-stage and single-stage object detection methods. The evaluated algorithms comprise Faster R-CNN, SSD, YOLOv3-tiny, YOLOv5m, YOLOv6s, YOLOv8m, among others. The corresponding experimental findings are detailed in Table 4.

Table 4. Experimental comparison results of different algorithms

Model	P(%)	R(%)	mAP@ 50(%)	mAP@ 50:95(%)	Param(M)	GFLOPs
Faster R-CNN	85.4	75.6	82.3	47.5	41.1	206.7
SSD	84.9	74.5	80.5	46.6	25.3	87.7
YOLOv3-tiny	88.7	79.2	85.6	48.8	12.1	18.9
YOLOv5m	89.7	75.7	85.2	48.4	20.9	48.0
YOLOv6s	88.0	75.1	84.2	48.8	18.5	45.2
YOLOv8s	87.7	77.6	85.7	49.2	11.3	28.5
YOLOv8m	90.3	77.4	86.1	50.3	25.8	78.7
YOLOv9m	87.3	77.7	85.9	49.7	19.8	75.8
YOLOv10s	86.1	77.5	85.4	49.3	7.5	24.5
YOLOv10m	86.9	78.1	85.9	50.2	16.4	63.4
YOLOv11s	88.1	77.9	85.6	49.5	9.6	23.5
YOLOv11m	88.9	77.1	86.0	50.2	20.0	67.7
YOLOv12s	88.9	75.7	85.1	48.8	9.2	24.3
YOLOv12m	89.6	74.7	85.6	49.2	20.1	67.1
DETR	86.8	73.4	81.7	46.7	41.0	86.2

Deformable DETR	88.5	75.2	83.6	49.3	34.0	78.4
RT-DETR-R18	88.7	81.1	86.5	52.1	19.9	57.0
RT-DETR-R34	90.6	79.4	86.1	51.0	31.2	89.1
RT-DETR-R50	90.1	80.3	86.8	52.5	42.8	134.4
Ours	92.4	85.8	89.8	54.4	15.0	47.4

As indicated in Table 4, the proposed algorithm surpasses the two-stage Faster R-CNN by increasing mAP@50 by 7.5%, while cutting parameters and GFLOPs by 63.5% and 77.1%, respectively, effectively reducing model complexity. Compared to the conventional single-stage SSD, it boosts mAP@50 and mAP@50:95 by 9.3% and 7.8%, respectively, while lowering parameters by 40.7%, ensuring better accuracy with reduced complexity. Regarding the YOLO series, the proposed method demonstrates clear benefits. Compared to YOLOv3-tiny, it enhances mAP@50 and mAP@50:95 by 4.2% and 5.6%, respectively, with comparable parameters. For YOLOv5m and YOLOv6s, it improves mAP@50 by 4.6% and 5.6%, and mAP@50:95 by 6.0% and 5.6%, while reducing parameters by 28.2% and 18.9%, respectively. When tested against YOLOv8m and YOLOv9m, it elevates mAP@50 by 3.7% and 3.9%, and mAP@50:95 by 4.1% and 4.7%, while minimizing parameters by 41.9% and 24.2%, respectively. Compared to YOLOv10m, it increases mAP@50 and mAP@50:95 by 3.9% and 4.2%, respectively, while reducing parameters by 8.5%, maintaining high precision with lower complexity. For YOLOv11m and YOLOv12m, it improves mAP@50 by 3.8% and 4.2%, and mAP@50:95 by 4.2% and 5.2%, respectively, while reducing the parameters by 25.0% and 25.5%, respectively. In addition, by comparing YOLOv8s, YOLOv10s, YOLOv11s, and YOLOv12s with the baseline model, it is evident that the number of parameters and computational complexity are significantly reduced. However, due to their lightweight architectures, a certain degree of accuracy is sacrificed. The mAP@50 and mAP@50:95 decrease by approximately 1% and 3% respectively compared to the baseline model. Their overall performance is inferior to that of the baseline model, which is not conducive to precise object detection. In addition, mAP@50 and mAP@50:95 improved by 8.1% and 7.7 %, respectively, while reducing the parameters by 63.4% compared to DETR. Compared to Deformable DETR, mAP@50 and mAP@50:95 improved by 6.2% and 5.1%, respectively, while reducing the parameters by 55.8%. For RT-DETR-R34 and RT-DETR-R50 it increased mAP@50 by 3.7% and 3.0% and mAP@50:95 by 3.4% and 1.9%, respectively, while reducing parameters by 51.9% and 64.9%, respectively. The baseline model, RT-DETR-R18, outperformed the YOLO series in both mAP@50 and mAP@50:95 and had significantly fewer parameters than RT-DETR-R34 and RT-DETR-R50. RT-DETR-R18 was designated as the baseline model because of its optimised structure and robust performance. In conclusion, the proposed algorithm achieves higher detection accuracy with fewer parameters and GFLOPs than other single-stage and two-stage detectors, which greatly improves the detection efficiency of foreign objects in transmission lines.

3.5.1 Backbone Comparison Experiment

In this study, the lightweight iRMB module replaces the original model's

BasicBlock module. To demonstrate the superiority of iRMB, comparisons are made with Faster-Block, Ortho-Block, DySnake-Block, and DualConv-Block as alternatives to BasicBlock. The experimental results are presented in Table 5.

Table 5. Comparison results of different backbone network experiments

Model	P(%)	R(%)	mAP@ 50(%)	mAP@ 50:95(%)	Param(M)	GFLOPs
BasicBlock	88.7	81.1	86.5	52.1	19.9	57.0
Faster-Block	91.7	80.4	87.1	52.3	16.8	49.5
Ortho-Block	91.0	80.3	86.9	51.6	19.9	57.0
DySnake-Block	90.7	80.0	86.7	52.3	27.8	60.8
DualConv-Block	90.4	80.5	86.7	51.3	15.9	47.3
iRMB-Block	89.6	81.8	87.5	53.1	16.4	49.1

As can be seen from Table 5, compared to the BasicBlock module of the original model, the improvement using the Faster-Block module reduces the parameters and GFLOPs by 15.6% and 13.2%, respectively, while the mAP@50 improves by 0.6%, and the Faster-Block maintains a small increase in the accuracy while significantly reducing the parameters and GFLOPs. Improvements to the Ortho-Block module resulted in a 0.4% improvement in mAP@50, but a 0.5% decrease in mAP@50:95. parameters and GFLOPs were the same as in BasicBlock. Therefore, Ortho-Block has limited improvement in accuracy and is not optimized for computational efficiency. The improvement using the DySnake-Block module resulted in a 0.2% increase in mAP@50, with mAP@50:95 remaining similar. However, parameters increased by 39.7% and GFLOPs increased by 6.7%. Although DySnake-Block has a slight improvement in accuracy, the significant increase in parameters and GFLOPs is not favorable for deployment in resource-constrained devices such as UAVs. After improving the DualConv-Block module, mAP@50 improved by 0.2%, but mAP@50:95 decreased by 0.8%. parameters decreased by 20.1% and GFLOPs decreased by 17%. Despite the reduction of parameters and GFLOPs by the DualConv-Block, the performance in accuracy is not satisfactory, especially the mAP@50:95 has a significant decrease. After improvement with the iRMB-Block module, mAP@50 is improved by 1% and mAP@50:95 by 1%. parameters are reduced by 17.6% and GFLOPs are reduced by 13.9%. It shows that the iRMB-Block module achieves a significant improvement in accuracy while keeping parameters and GFLOPs low, especially in the more stringent mAP@50:95 metric. Comparative experiments show that the proposed iRMB-Block module exhibits more excellent performance and can perform transmission line foreign object detection more efficiently on the basis of lightweight.

3.5.2 Feature Fusion Network Comparison Experiment

The SSFF-Slimneck is employed as the cross-scale feature fusion network. To validate its superiority, comparisons are made with mainstream feature fusion networks, including HSPAN, MAFPN, and BiFPN. The results of these experiments are detailed in Table 6.

Table 6. Comparison results of different feature fusion network experiments

Model	P(%)	R(%)	mAP@ 50(%)	mAP@ 50:95(%)	Param(M)	GFLOPs
Baseline	88.7	81.1	86.5	52.1	19.9	57.0
HSPAN	91.3	83.7	87.5	52.3	20.6	58.1
MAFPN	92.2	82.8	87.6	52.6	22.9	56.3
BiFPN	92.4	81.9	87.8	53.3	20.3	64.3
SSFF-Slimneck	93.6	82.8	88.6	53.2	19.6	57.6

As shown in Table 6, the HSPAN network improves Precision, Recall, and mAP@50 by 2.6%, 2.6%, and 1.0%, respectively, in comparison to the baseline model. Although mAP@50:95 sees only a 0.2% increase, overall performance still experiences an enhancement. Nevertheless, parameters and GFLOPs rise by 3.5% and 1.9%, respectively, suggesting a modest performance improvement with minimal computational overhead. The MAFPN network improves Precision and Recall by 3.5% and 1.7%, respectively, with mAP@50 and mAP@50:95 rising by 1.1% and 0.5%. However, parameters surge by 15.1%, hindering lightweighting despite a slight GFLOPs reduction.The BiFPN network enhances Precision and Recall by 3.7% and 0.8%, respectively, while mAP@50 and mAP@50:95 see gains of 1.3% and 1.2%. Although parameters exhibit only a 2.0% increase, GFLOPs surge by 12.8%, reflecting a balance between performance and computational demand. Conversely, the proposed SSFF-Slimneck network outperforms in all aspects, elevating Precision and Recall by 4.9% and 1.7%, respectively, and boosting mAP@50 and mAP@50:95 by 2.1% and 1.1%. Importantly, parameters experience a slight reduction, and GFLOPs remain unchanged, highlighting its capability to improve transmission line foreign object detection while preserving model efficiency.

3.5.3 Bounding Box Loss Function IoU Comparison Experiment

To demonstrate the advantages of the Focaler-Shape-IoU loss function presented in this study, various mainstream loss functions, including CIoU, GIoU, EIoU, DIoU, and SIoU, are systematically compared and analyzed. The findings from these comparative experiments are detailed in Table 7.

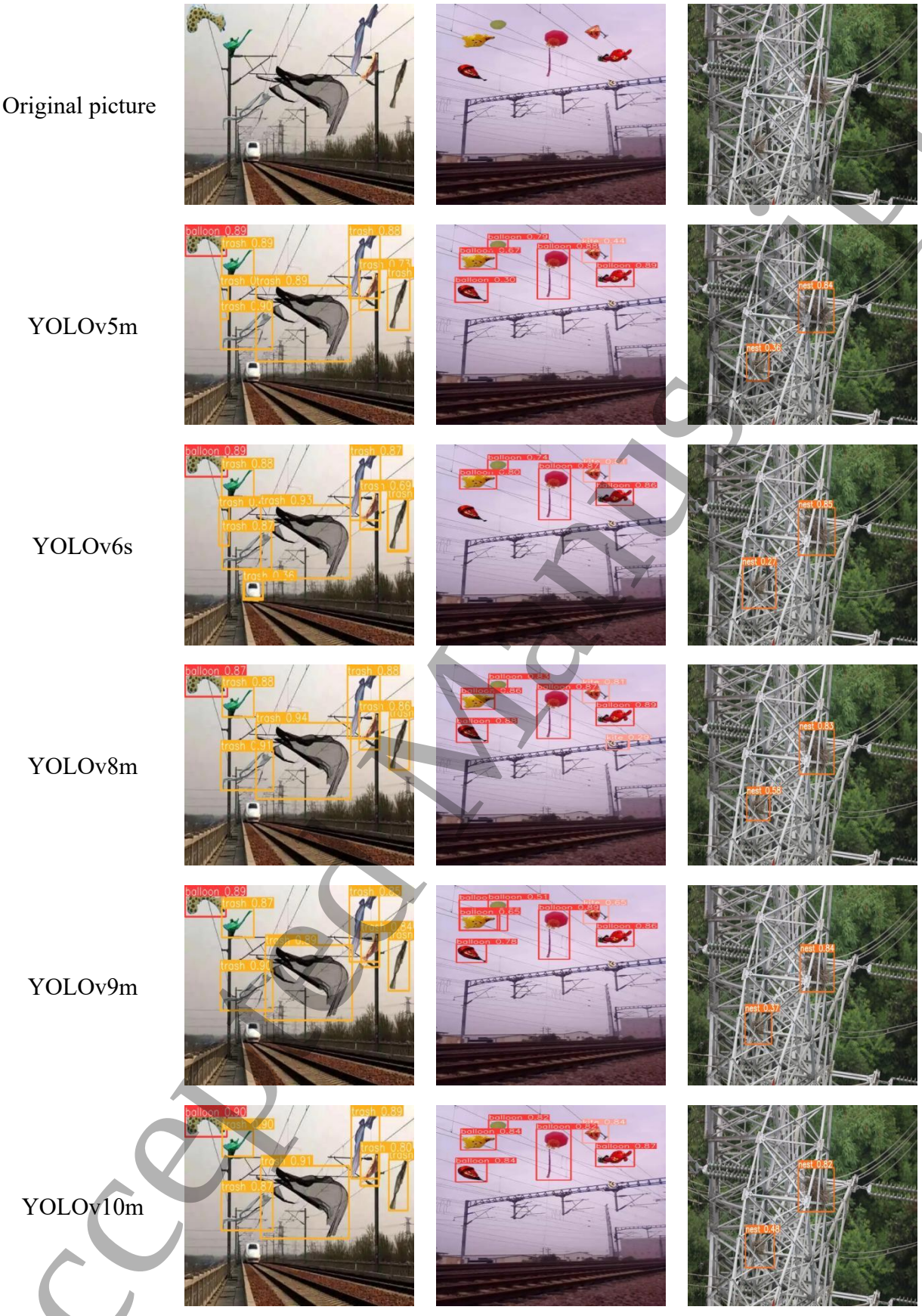
Table 7. Bounding box loss function IoU comparison experiments

Model	P(%)	R(%)	mAP@50(%)	mAP@50:95(%)
GIoU	88.7	81.1	86.5	52.1
DIoU	92.5	80.4	86.7	51.8
CIoU	90.9	81.8	87.1	51.8
EIoU	90.0	81.4	87.2	52.0
SIoU	91.5	82.7	87.4	52.1
Shape-IoU	93.0	82.2	87.7	53.0
Focaler-Shape-IoU	93.3	82.3	87.9	54.1

As indicated in Table 7, when compared to GIoU in the baseline model, DIoU enhances Precision by 3.8% but lowers Recall by 0.7%, leading to a minor 0.2% rise in mAP@50 and a 0.3% drop in mAP@50:95. While DIoU strengthens Precision, the reduction in Recall negatively influences overall performance. CIoU expands on DIoU by incorporating aspect ratio information, increasing Precision by 2.2%, Recall by 0.7%, and mAP@50 by 0.6%, though mAP@50:95 experiences a slight decline. EIoU mitigates CIoU's limitations by adding target and prediction frame aspect details, leading to a 1.3% gain in Precision, a 0.3% rise in Recall, and a 0.7% improvement in mAP@50, while mAP@50:95 remains nearly stable with only a 0.1% reduction. SIoU further strengthens accuracy and robustness through angular loss and soft thresholding, resulting in a 2.8% and 1.6% enhancement in Precision and Recall, respectively, along with a 0.9% rise in mAP@50, while maintaining mAP@50:95. Shape-IoU incorporates shape and scale factors, refining the loss function to better distinguish predicted and actual frames. It elevates Precision and Recall by 4.3% and 1.1%, respectively, with mAP@50 increasing by 1.2% and mAP@50:95 improving by 0.9%. This confirms Shape-IoU's advantage in enhancing detection accuracy while sustaining performance under more rigorous conditions. In contrast, the newly introduced Focaler-Shape-IoU maintains Shape-IoU's strengths while refining the loss function using linear interval mapping, optimizing edge regression. It achieves a 4.6% and 1.2% boost in Precision and Recall, respectively, and increases mAP@50 by 1.4%. Notably, mAP@50:95 improves by 2.0%, marking the highest gain among all models. This underscores Focaler-Shape-IoU's outstanding ability to enhance detection accuracy while conforming to stricter criteria, making it highly suitable for identifying transmission line foreign objects with varying complexities and improving detection precision.

3.6 Visualization and analysis

To more intuitively demonstrate the superiority of the proposed algorithm, this study employs a visualization method to analyze transmission line foreign objects. The visualization results are shown in Fig. 16. In the first column, YOLOv6s misidentifies the train as garbage due to the small size of the locomotive, while the proposed algorithm avoids misdetection and omission, achieving the highest accuracy across all seven foreign objects.In the second column, YOLOv6s fails to detect a balloon. Due to the small size of the kite and occlusion by transmission pylons, YOLOv5m, YOLOv6s, YOLOv9m, and YOLOv10m also miss the target. In contrast, the proposed algorithm accurately identifies all foreign objects with higher precision.In the third column, the bird's nest is small and camouflages with the transmission tower, making detection more challenging. The proposed algorithm not only identifies the bird's nest but also surpasses the YOLO series in accuracy. Overall, this algorithm effectively minimizes both missed detections and false detection rates in complex environments while ensuring improved accuracy. It is particularly suitable for detecting small foreign objects on transmission lines.



Ours

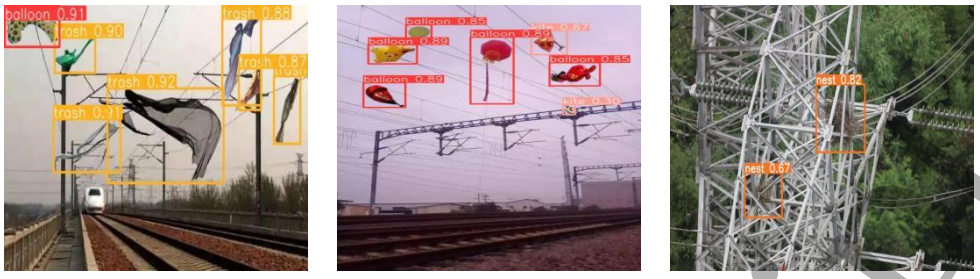
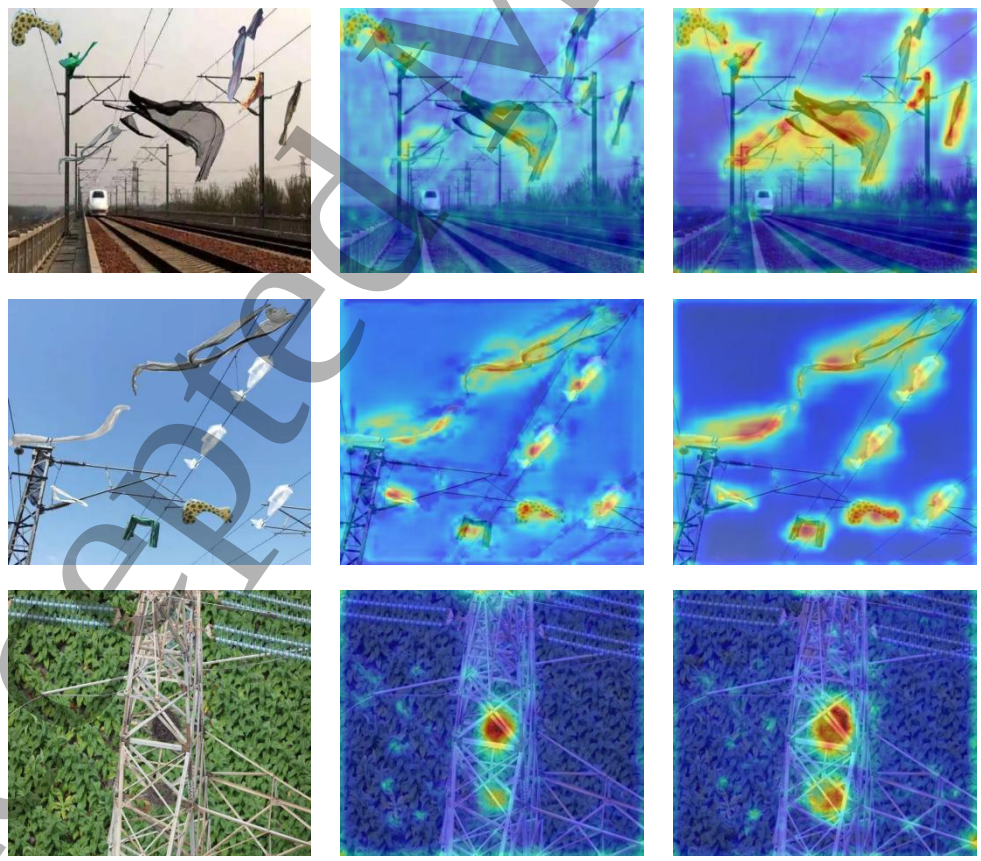
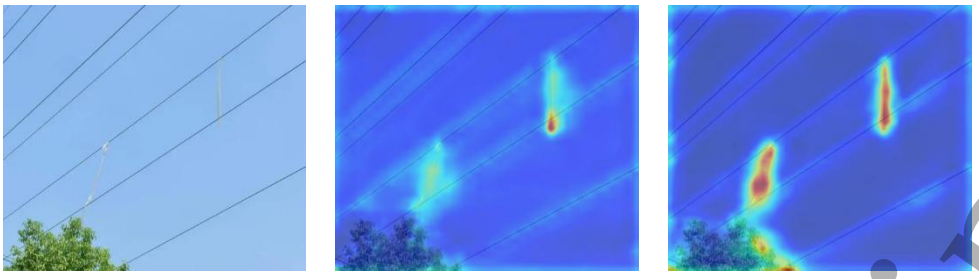


Fig. 16 YOLO series visualization comparison results

A heat map serves as a visualization tool that employs color to depict data size and distribution. In this research, Gradient Weighted Class Activation Mapping (Grad-CAM) [34] is utilized for heat map visualization and analysis. Grad-CAM integrates gradient information with convolutional feature maps to produce heat maps, emphasizing crucial regions. Blue areas indicate lower data values, signifying less focus from the model, whereas red areas correspond to higher values, highlighting regions of greater attention. A comparison of heat maps before and after model enhancement is illustrated in Fig. 17. The first column presents the original image, the second column displays the detection results from the initial model, and the third column showcases the detection outcomes after refinement. It is clear that the heat map before improvement contains numerous scattered hotspots, whereas the refined heat map features more concentrated and precise hotspots. The contours and shapes of the foreign object target regions appear more distinct, signifying that the enhanced model is more effective in recognizing and localizing targets within complex and dense environments.





Original picture Before improvement After improvement

Fig. 17 Comparison of thermograms before and after model improvement

4. Conclusions

In this paper, we propose a transmission line foreign object detection algorithm based on improved RT-DETR to cope with the challenges of complex backgrounds, variable target scales, and frequent false and missed detections. The algorithm integrates several key innovations:

Improvement of the feature extraction module: by introducing the CRMB module, the algorithm combines the inverted residual shift module (iRMB) and cascade group attention (CGA), which significantly enhances the feature extraction capability in complex backgrounds and improves the detection capability of targets at different scales. This improvement not only enhances the robustness of the model, but also makes the model more efficient in dealing with complex scenes.

Optimisation of cross-scale feature fusion network: the SSFF-Slimneck cross-scale feature fusion network is proposed to effectively solve the problem of information loss due to the change of target scale in the feature fusion process. The network improves the effectiveness of feature fusion by combining the Slimneck design paradigm and the Scale Sequence Feature Fusion (SSFF) module, which further enhances the model's ability to detect multi-scale targets.

Innovation of loss function: Focaler-Shape-IoU is adopted as the bounding box loss function, and the algorithm improves the generalisation ability and detection accuracy of the model while accelerating model convergence. This loss function makes the model more sensitive in dealing with samples of different complexity by redesigning the IoU loss, which improves the overall detection performance.

The experimental results show that the proposed algorithm exhibits excellent performance in the transmission line foreign object detection task. Compared with the original RT-DETR algorithm, the improved algorithm achieves significant improvement in both mAP@50 and mAP@50:95, while the parameters and computation amount are also reduced significantly. This not only verifies the effectiveness of the proposed method, but also provides strong support for the practical application of transmission line foreign object detection.

5. Limitations of Algorithms

However, although the proposed algorithm has achieved significant results in transmission line foreign object detection, it still has some limitations. For example, the detection performance of the algorithm may still be affected to some extent under

extreme complex backgrounds. In addition, the detection accuracy and robustness of the algorithm may need to be further improved as the target scale is further reduced or the background complexity increases.

In an image depicting a transmission line traversing a densely forested area, the algorithm failed to accurately detect a plastic bag entangled on the power line. The dense vegetation covered a significant portion of the image background, causing the algorithm to struggle in distinguishing the plastic bag from the surrounding foliage during the feature extraction process. Additionally, the color and shape of the plastic bag closely resembled those of the surrounding environment, further increasing the difficulty of detection.

For future research directions, in-depth exploration is planned in the following aspects: first, further optimisation of the feature extraction and fusion module to improve the detection performance of the algorithm in extremely complex backgrounds. The second is to investigate more efficient loss functions to further improve the detection accuracy and robustness of the algorithm. The third is to explore the application of the proposed algorithm to other related fields, such as UAV inspection and intelligent monitoring, to verify its versatility and scalability. Through these studies, it is expected to make a greater contribution to the development of transmission line foreign object detection and other related fields.

Acknowledgements

The authors wish to express their thanks for the support of the Aeronautical Science Foundation of China under Grant NO.20200001012015. This material is based upon work supported by the Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission with Grant No. 23CGA76. The support of Open Project Funds for the Key Laboratory of Space Photoelectric Detection and Perception, Ministry of Industry and Information Technology under Grant NO.NJ2023029 and ZBA Key Laboratory Open Fund under Grant NO.6142002230102 are also gratefully acknowledged.

Data Availability

Data will be made available on request.

References

[1] Wu Y, Zhao S, Xing Z, et al. Detection of foreign objects intrusion into transmission lines using diverse generation model[J]. IEEE Transactions on Power Delivery, 2023, 38(5): 3551-3560.

[2] Tang C, Dong H, Huang Y, et al. Foreign object detection for transmission lines based on Swin Transformer V2 and YOLOX[J]. The Visual Computer, 2024, 40(5): 3003-3021.

[3] Chen Z, Yang J, Feng Z, et al. RailFOD23: A dataset for foreign object detection on railroad transmission lines[J]. Scientific Data, 2024, 11(1): 72.

[4] Zhang D, Zhang Z, Zhao N, et al. A lightweight modified YOLOv5 network using a swin transformer for transmission-line foreign object detection[J]. Electronics, 2023, 12(18): 3904.
[5] Lv X L, Chiang H D. Visual clustering network-based intelligent power lines inspection system[J]. Engineering Applications of Artificial Intelligence, 2024, 129: 107572.
[6] Han T, Bao M, He T, et al. LW-PV DETR: Lightweight model for photovoltaic panel surface defect detection[J]. Engineering Research Express, 2025.
[7] Ren S, He K M, Girshick R, et al. (2016) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6):1137-1149.
[8] He K M. Gkioxari G, Piotr D, et al. (2017) Mask R-CNN. In: IEEE International Conference on Computer Vision, 2961-2969. IEEE, Venice.
[9] Li J, Yan D, Luan K, et al. Deep learning-based bird's nest detection on transmission lines using UAV imagery[J]. Applied sciences, 2020, 10(18): 6147.
[10] Redmon J. S, Girshick R, et al. (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779-788. IEEE, Las Vegas.
[11] Liu W, Anguelov D, Erhan D, et al. (2016) SSD: single shot multibox detector. In: Computer Vision-ECCV 2016, 21-37. Springer, Amsterdam.
[12] Qin Y D, Li X W, He D, et al. RLGS-YOLO: an improved algorithm for metro station passenger detection based on YOLOv8[J]. Engineering Research Express, 2024, 6(4): 045263.
[13] Sun H, Shen Q, Ke H, et al. Power Transmission Lines Foreign Object Intrusion Detection Method for Drone Aerial Images Based on Improved YOLOv8 Network[J]. Drones, 2024, 8(8): 346.
[14] Bin F, He J, Qiu K, et al. CI-YOLO: A lightweight foreign object detection model for inspecting transmission line[J]. Measurement, 2024: 116193.
[15] Xu W, Xiwen C, Haibin C, et al. Foreign object detection method in transmission lines based on improved yolov8n[C]//2024 10th International Symposium on System Security, Safety, and Reliability (ISSSR). IEEE, 2024: 196-200.
[16] Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.
[17] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.
[18] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. arxiv preprint arxiv:2010.04159, 2020.
[19] Zhang H, Li F, Liu S, et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection[J]. arxiv preprint arxiv:2203.03605, 2022.
[20] Zhao Y, Lv W, Xu S, et al. Detrs beat yolos on real-time object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 16965-16974.
[21] Han Z, Jia D, Zhang L, et al. FNI-DETR: real-time DETR with far and near feature

interaction for small object detection[J]. Engineering Research Express, 2025, 7(1): 015204.

[22] Xue R, Hua S, Xu H. FECI-RTDETR A lightweight unmanned aerial vehicle infrared small target detector Algorithm based on RT-DETR[J]. IEEE Access, 2025.

[23] Wang S, Jiang H, Li Z, et al. PHSI-RTDETR: A Lightweight Infrared Small Target Detection Algorithm Based on UAV Aerial Photography[J]. Drones, 2024, 8(6): 240.

[24] Gao S, Xue F, Tan W, et al. RTPR-DETR: A Real Time Small Object Detection Network Based on Partial-Convolution with structural re-parameterization[C]//2024 36th Chinese Control and Decision Conference (CCDC). IEEE, 2024: 5513-5518.

[25] Kong Y, Shang X, Jia S. Drone-DETR: Efficient small object detection for remote sensing image using enhanced RT-DETR model[J]. Sensors, 2024, 24(17): 5496.

[26] Wang C, Chen B, Li Y, et al. YOLOv8-CDD: an improved concrete defect detection method combined CNN with transformer[J]. Measurement Science and Technology, 2024, 36(1): 015409.

[27] Zhang J, Li X, Li J, et al. Rethinking mobile block for efficient attention-based models[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, 2023: 1389-1400.

[28] Liu X, Peng H, Zheng N, et al. Efficientvit: Memory efficient vision transformer with cascaded group attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14420-14430.

[29] Kang M, Ting C M, Ting F F, et al. ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation[J]. Image and Vision Computing, 2024, 147: 105057.

[30] Li H, Li J, Wei H, et al. Slim-neck by GSConv: a lightweight-design for real-time detector architectures[J]. Journal of Real-Time Image Processing, 2024, 21(3): 62.

[31] Zhang H, Zhang S. Focaler-IoU: More Focused Intersection over Union Loss[J]. arXiv preprint arXiv:2401.10525, 2024.

[32] Zhang H, Zhang S. Shape-iou: More accurate metric considering bounding box shape and scale[J]. arXiv preprint arXiv:2312.17663, 2023.

[33] Chen Z, Yang J, Feng Z, et al. RailFOD23: A dataset for foreign object detection on railroad transmission lines[J]. Scientific Data, 2024, 11(1): 72.

[34] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. International journal of computer vision, 2020, 128: 336-359.