



OPEN A foreign object detection dataset and network for electrified railway catenary systems

Fengqi Li, Jinhao Cao, Haolin Yang, Li Diao, Yanhua Wang, Shiyu Yu, Xiaohong Yan & Fengqiang Xu

Intrusion of foreign objects into the Electrified Railway Catenary System can lead to power failures, train service interruptions, and even casualties, making accurate detection essential for safe operation. Due to the scarcity of railway datasets, this study constructs a Railway Catenary Foreign Object Dataset to support model training and evaluation. Existing detection methods often struggle with complex railway environments, diverse object morphologies, and varying scales. To address these challenges, we propose a Railway Catenary Foreign Object Detection Network. It leverages the hierarchical architecture and window-based attention mechanism of Swin Transformer for multi-scale semantic feature extraction and global relational modeling, effectively distinguishing foreground from background. A Multi-branch Fusion Feature Pyramid Network is designed to deeply fuse low- and high-level features across scales, improving detection of objects of different sizes. Additionally, a Regional Receptive Field-Enhanced Edge Module expands the receptive field and enhances edge extraction for elongated foreign objects. Extensive experiments on the constructed dataset demonstrate the effectiveness of the proposed approach, achieving an Average Precision of 60.2%, with 53.8% for small object detection.

Keywords Railway catenary, Dataset, Foreign object detection, Swin transformer, Feature fusion, Dilated convolution

As a critical component of the electrified railway system, the catenary system constitutes a high-voltage power supply network, comprising key elements such as supporting structures, contact wires, cables, and suspension ropes, in addition to auxiliary facilities associated with power transmission, distribution, and protection. During long-term operation, the catenary system is inevitably affected by environmental factors and human activities. Foreign objects such as bird nests, kites, balloons, and plastic films represent significant operational risks. These intrusions may cause localized damage to the catenary, potentially leading to electric shocks, power outages, or train operation incidents, thereby posing severe threats to train safety and passenger life and property.

Figure 1 presents the overall workflow for the preventive maintenance of foreign objects in Electrified Railway Catenary Systems (ERCS). Serving as an important component of railway maintenance strategies, this workflow emphasizes the necessity of preventive monitoring. Currently, in addressing foreign object interference, railway authorities primarily rely on manual measures, arranging personnel to conduct regular patrols along the lines for inspection. In recent years, camera-based video surveillance systems have gradually become a cost-effective solution. By capturing images and videos of the railway catenary, these systems can directly reflect the operational status on site. However, after acquiring large volumes of visual data, video surveillance systems still depend on trained technicians to perform frame-by-frame inspection. This process is not only time-consuming and labor-intensive but also highly reliant on human expertise, making it challenging to achieve efficient and comprehensive monitoring of railway lines. Therefore, realizing automated detection of foreign objects in railway catenary systems has become a pressing technical challenge in the field of railway maintenance.

Railway intrusion poses a serious threat to operational safety and may lead to catastrophic consequences, including severe casualties and economic losses. Ensuring timely and reliable detection of foreign objects and intrusions along railway infrastructure is therefore a critical requirement for modern railway systems. Traditional railway catenary inspection systems primarily rely on physical sensing devices mounted on inspection vehicles to monitor the operational condition of pantograph–catenary interactions. For instance, Fiber Bragg Grating (FBG) sensors are widely used to measure contact force and vertical acceleration of pantograph heads, while

School of Railway Intelligent Engineering, Dalian Jiaotong University, Dalian 116028, China. ✉email: xfq@djtu.edu.cn

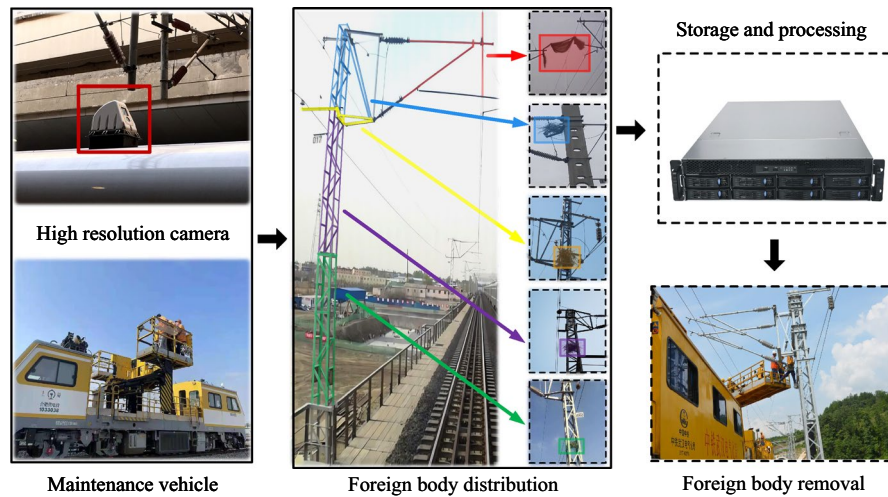


Figure 1. Preventive maintenance process of foreign objects in ERCS.

optical radar technologies are employed to record key geometric and dynamic parameters of the catenary for condition assessment^{1,2}. However, such physical inspection systems mainly focus on structural health and operational status monitoring and are inherently unsuitable for detecting foreign objects intruding into the catenary system.

With the increasing complexity of railway operating environments, intrusion detection has gradually evolved from traditional physical sensing toward intelligent perception based on machine vision. Vision-based approaches offer clear advantages, including non-contact measurement, rich semantic representation, and strong scalability. A recent comprehensive survey by Cao et al.³ systematically reviewed railway intrusion detection methods based on machine vision over the past two decades, covering ground-based monitoring, on-board inspection, and unmanned aerial vehicle (UAV) inspection. The survey highlights that vision-based methods have become a dominant paradigm in railway intrusion detection. At the same time, it identifies several key challenges that hinder practical deployment, such as the scarcity of intrusion samples caused by the low frequency of fault events, complex and cluttered backgrounds, multi-scale and elongated object characteristics, and performance degradation under diverse environmental conditions.

Early vision-based studies mainly relied on handcrafted image features and conventional image processing techniques. For example, foreign object detection in Electrified Railway Catenary Systems (ERCS) was achieved through image analysis and matching strategies⁴. Although such methods demonstrated effectiveness in detecting specific targets, such as bird nests, their applicability was largely restricted to single-scene and single-object settings. Moreover, their performance depended heavily on carefully designed assumptions and controlled environments, limiting their robustness in real-world railway scenarios.

In recent years, the rapid advancement of artificial intelligence (AI) has significantly promoted the application of deep learning-based methods in railway inspection. Convolutional neural network (CNN)-based detectors, such as YOLO and Faster R-CNN, have been widely adopted for railway component inspection and defect detection. For instance, YOLOv8-FAM⁵ enhances the feature extraction capability of YOLOv8 for defective fastener detection, while improved Faster R-CNN variants⁶ have been used for precise localization and condition assessment of railway fasteners. Despite their effectiveness, CNN-based methods suffer from inherent locality in convolution operations, which limits their ability to model long-range dependencies. This limitation becomes particularly problematic in railway catenary environments characterized by high-altitude installation, complex structural layouts, and highly cluttered backgrounds.

Beyond 2D image-based approaches, recent research has explored 3D perception techniques to improve spatial understanding and long-distance obstacle detection in rail transit scenarios. Lian et al. proposed RVSA-3D⁷, a voxel-based fully sparse attention framework that enhances feature representation for distant and sparse obstacles through sparse attention mechanisms and multi-scale information fusion. Similarly, the RAE3D model⁸ integrates multi-scale voxel aggregation and bird's-eye-view (BEV) auxiliary learning to alleviate information loss caused by point cloud sparsity. Although these 3D-based methods achieve state-of-the-art performance on datasets such as KITTI and Rail3D, they typically require expensive sensors and substantial computational resources, which limits their practicality for large-scale deployment in railway catenary inspection systems.

To address the specific challenges of foreign object detection in railway catenary systems, Chen et al.⁹ proposed a lightweight perception framework combined with a scarce image generation strategy. Their work emphasizes two critical issues in catenary foreign object detection: the severe scarcity of fault samples and the inherent trade-off between lightweight network design and global dependency modeling. Although their approach effectively reduces computational complexity and improves detection efficiency, balancing local feature extraction with global contextual modeling remains challenging, especially for small-scale and elongated foreign objects commonly observed in catenary environments.

The emergence of large-scale models has further advanced visual perception tasks, with the Transformer architecture¹⁰ demonstrating remarkable success in capturing long-range dependencies. Vision Transformer (ViT)¹¹ was the first to introduce Transformer architectures into computer vision and has since been increasingly adopted in railway-related applications, such as multi-scale obstacle detection¹². However, while ViT exhibits strong global modeling capability, its limited ability to capture fine-grained local features constrains its performance in complex railway scenes. The Swin Transformer (ST)¹³ addresses this limitation by introducing hierarchical feature representation and shifted window attention, enabling effective joint modeling of local details and global context. Therefore, this study adopts the Swin Transformer as the backbone of the proposed detection network to extract discriminative image features for railway catenary foreign object detection.

At present, the number of publicly available datasets for foreign object detection (FOD) in railway catenary systems remains limited, and the field is still at an early stage of development. Due to stringent requirements on railway operational safety and data confidentiality, real-world inspection data are difficult to acquire and release at scale, resulting in a long-standing scarcity of abnormal samples. Among existing studies, RailFOD23¹⁴ is one of the more representative public datasets. It constructs a relatively large-scale FOD benchmark by leveraging manual synthesis, data augmentation, and AI-generated content (AIGC), which to some extent alleviates the shortage of real abnormal samples. However, the foreign object instances in this dataset are predominantly synthetic, with a relatively small proportion of anomalies collected from real-world scenarios. As a result, discrepancies remain between the synthetic objects and real operating environments in terms of physical consistency and spatial relationships with the background. In addition, some studies utilize real railway scene datasets, such as RailSem19¹⁵, as background imagery and construct foreign object samples through post-processing and compositing. Although this strategy improves scene realism to a certain extent, the diversity of object morphology, category coverage, and risk representation remains limited, making it difficult to comprehensively reflect the complex and diverse foreign object types and operational risks encountered in real railway catenary systems.

Due to the particularity of railway scenarios, publicly available datasets are extremely limited, while most existing ones are either inaccessible or small in scale, making it difficult to meet the demand of deep learning models for large-scale, high-quality data. To validate the effectiveness of the proposed model, we constructed a scenario-specific dataset the Railway Catenary Power Supply System Foreign Object Dataset (RailCatFOD-DS). Images of the ERCS exhibit significant variability, including changes in illumination, lighting conditions, occlusion, weather effects, and diverse foreground–background characteristics. Moreover, foreign objects vary widely in shape and are distributed across different locations of the ERCS, with some partially occluded by the catenary itself. In addition, identifying foreign objects within the ERCS becomes particularly challenging in complex backgrounds such as mountains, trees, and buildings.

In railway environments, conventional CNN models often struggle to perform effectively. These models rely on stacked convolutional kernels to capture contextual information, but their limited receptive fields restrict the integration of global information, making precise foreground-background discrimination difficult. Furthermore, complex weather conditions frequently degrade image quality, requiring the model to possess robust feature extraction capabilities. To address these challenges, we propose a method that combines a powerful Transformer encoder with a Feature Pyramid Network (FPN) decoder. The self-attention mechanism of the Transformer enables global information modeling, overcoming the local receptive field limitations inherent to CNNs. Meanwhile, objects exhibit significant scale variations due to viewpoint differences, and traditional FPN with large-stride convolutions often lose critical details, reducing detection performance. To mitigate this, we developed the Multi-branch Fusion Feature Pyramid Network (MFFPN) to decode and integrate finer-grained features. Railway foreign object detection also faces challenges from varying object shapes and sizes, particularly elongated foreign objects. Previous models typically fuse features from backbone outputs within the FPN, but the limited regional receptive fields of Transformer encoder outputs hinder effective handling of object boundary information. To address this, we design the Regional Receptive Field-Enhanced Edge Module (RRFEM), which expands the regional receptive field, allowing the MFFPN decoder to more effectively integrate contextual information across different regions. Extensive experiments conducted on the constructed dataset demonstrate that the model achieves an average precision (AP) of 60.2%, with an AP of 53.8% for small object detection, outperforming existing object detection models in railway scenarios.

The main contributions of this study are summarized as follows:

- **Dataset Construction:** We construct a dedicated RailCatFOD-DS, which provides a solid benchmark for foreign object detection in ERCS.
- **Detection Network:** We propose RailCatFOD-Net, a refined detection framework tailored for complex railway scenarios, effectively addressing the challenges posed by diverse object morphologies, varying scales, and cluttered backgrounds.
- **Multi-branch Fusion Feature Pyramid Network (MFFPN):** We design MFFPN to deeply fuse shallow and high-level features across multiple scales, detecting objects of different scales.
- **Regional Receptive Field-Enhanced Edge Module (RRFEM):** We design RRFEM to expand the receptive field and strengthen boundary handling, particularly improving detection of elongated foreign objects.

Railway catenary foreign object dataset

In deep learning-based object detection, access to suitable datasets is essential. Yet, due to task specificity, publicly available datasets are scarce, and existing ones are often restricted or small in scale, falling short of the demand for large, high-quality data. This study constructs a dedicated image dataset for foreign object detection in ERCS, supporting subsequent model training and optimization. This dataset is divided into two main categories:

Number of pictures	source
1300	https://aistudio.baidu.com/datasetdetail/234072
166	https://download.csdn.net/download/qq_44042678/87949346
500	https://pan.baidu.com/s/1SPX025S9BmyrULDogcwqIA
4517	https://download.csdn.net/download/weixin_45941288/87446746
155	Internet pictures
1000	Self-collected pictures
6228	Data augmentation

Table 1. Primary source of RCSFO dataset.

Object type	Number of pictures
Fog	3286
Changes in lighting	1985
Occlusion	3248

Table 2. Data augmentation quantity distribution.

1. Bird nests: Bird nests are among the most common foreign objects in railway catenary systems. Nests built by different bird species vary significantly in size, shape, color, and material. Some are woven from branches, forming relatively regular spherical or bowl-like structures.
2. Light debris: This category includes items such as plastic bags, kites, and plastic films that are blown onto or entangled with the catenary. These objects are typically lightweight and highly susceptible to wind, exhibiting considerable variation in shape and size. Many of them appear elongated or irregular.

After determining the object types for railway catenary system foreign object detection, we have collected a total of 7,638 images from the internet and other sources, including 166 synthetic images. The sources of these images are listed in Table 1. We used the Labelling tool to annotate the images, labeling the object types, positions, and sizes to ensure the annotation data adheres to the COCO dataset format, with additional support for the YOLO format. Furthermore, to enhance the model's generalization ability and increase the diversity of the dataset, a series of data augmentation techniques were employed, including common methods such as rotation, translation, and mirroring. Considering the specific characteristics of the railway environment and the challenges of image acquisition under complex weather conditions, more specialized augmentations were also introduced, such as dynamic rain simulation, image degradation, and illumination enhancement, aiming to replicate various real-world weather and environmental factors. Detailed distribution quantities are shown in Table 2. Partial training sample images from the dataset are shown in Fig. 2.

The rain layers produced by current deep learning-based methods are synthetic and exhibit discrepancies compared to real data, leading to a decline in image quality in railway catenary system scenarios. To address this issue, we employ the S2VD¹⁶ model, which utilizes a dynamic rain generator to simulate more realistic rain layers. The results are shown in Fig. 3a. The Transition Model is responsible for updating the hidden state s_t , which reflects the dynamic properties of rain, such as velocity and acceleration, and provides the input state for the Emission Model at the current time step. The Emission Model then generates the observed rain image R_t based on this state.

The transition model is used to describe the dynamic changes of the rain layer, and the formula is as follows:

$$s_t = F(s_{t-1}, z_t, m; \alpha) \quad (1)$$

The generation formula for rain streaks is realized through the emission model:

$$R_t = H(s_t; \beta) \quad (2)$$

where R_t represents the rain streak image at the t -th frame. s_t is the hidden state variable, representing the rain layer features at the t -th frame (such as direction, density, etc.). H is the mapping function. β is the parameter of the emission model. $z_t \sim N(0, I)$ is the random noise, used to simulate external influences (such as wind speed and changes in raindrop angle). $m \sim N(0, I)$ is the rain variation vector. α is the parameter of the transition model.

In this study, Gaussian noise was added to the images to effectively simulate blurring and color distortion under environmental conditions such as haze and dense fog. The results are shown in the Fig. 3b. A random variable $v \sim Uniform(0, 0.2)$ is generated within the range $[0, 0.2]$. For each image pixel $I(x, y)$, Gaussian noise $\varepsilon(x, y) \sim N(0, v)$ is generated, where the mean is 0 and the variance is v . Add Gaussian noise to the image to get the new pixel value, as shown in the following formula:

$$I'(x, y) = I(x, y) + \varepsilon(x, y) \quad (3)$$

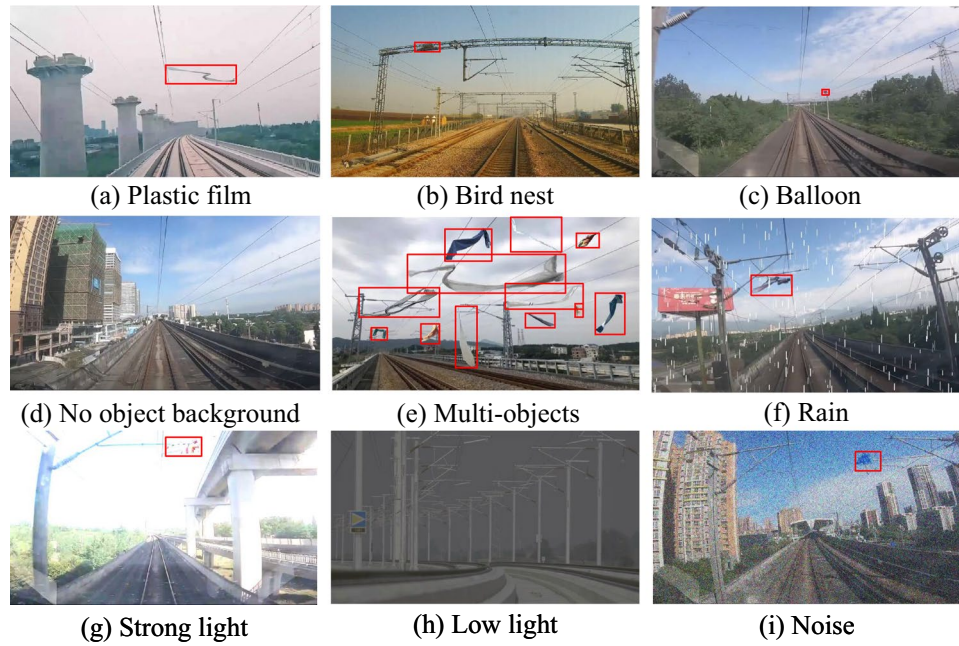


Figure 2. Partial training sample images from the dataset.



Figure 3. Special treatment effect diagram.

We also employed the Cutout¹⁷ data augmentation technique, which randomly masks portions of the image to simulate occlusion scenarios in real-world environments. This forces the model to learn effective recognition under partial information loss. Cutout can be described by the following formula:

$$I'(x, y, c) = \begin{cases} 0 & \text{if } x_1 \leq x \leq x_2 \text{ and } y_1 \leq y \leq y_2, \\ I(x, y, c) & \text{otherwise} \end{cases} \quad (4)$$

where $I'(x, y, c)$ is the original image and $I(x, y, c)$ is the image with Cutout.

In the railway catenary foreign object detection scenario, lighting conditions vary with time, weather, and seasons (e.g., strong sunlight or low light at night). Brightness is randomly adjusted using OpenCV. The results are shown in Fig. 3c,d. The formula is as follows:

$$I'(x) = \alpha I(x) + \beta \quad (5)$$

where $I(x)$ is the original pixel intensity, $I'(x)$ is the adjusted pixel intensity, α is the brightness scaling factor, and β is the brightness offset.

These augmentation techniques have increased the final dataset to a total of 13,866 images. The object type, number of objects, and number of images in the RailCatFOD-DS are summarized in Table 3. The dataset was stratified and randomly divided into training, validation, and test sets in a 7:2:1 ratio to facilitate subsequent

Object type	Object quantity	Number of pictures
Bird nest	8420	8336
Light debris	6378	4748
No object background image	0	800
Total	14,798	13,866

Table 3. Statistics of object types, number of objects and number of images in the data set.

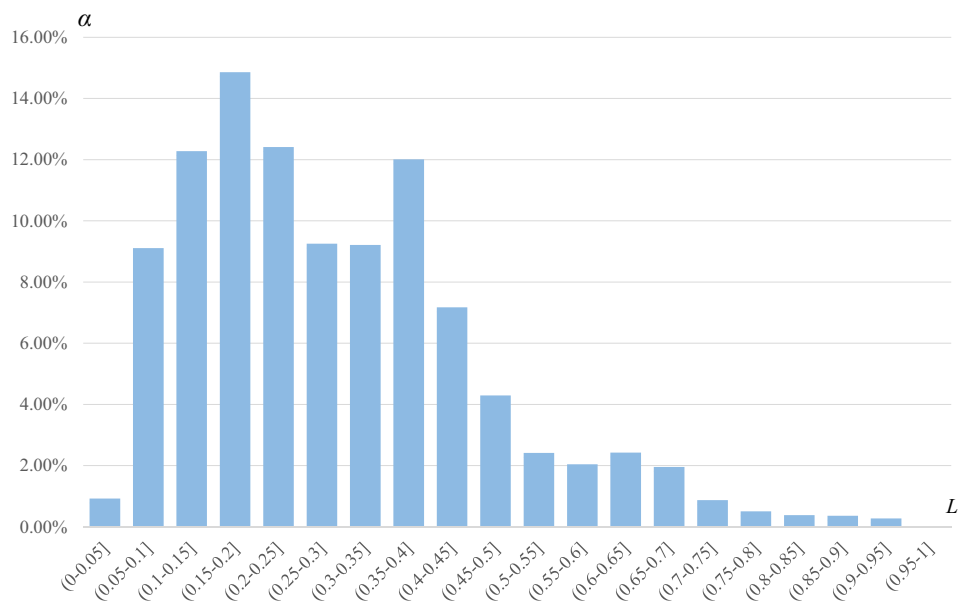


Figure 4. Object size distribution diagram in the RailCatFOD-DS.

training and evaluation. Finally, we visualize the annotations on the images and verify the accuracy of the annotations.

Since all images are scaled to one or more fixed dimensions during training or testing of the model, formula L is selected as the size distribution of the object in the judgment standard dataset.

The formula for L is as follows:

$$L = \max(w'/w, h'/h) \quad (6)$$

where w' and h' denote the width and height of the object's bounding box, respectively, and w and h represent the width and height of the corresponding image.

Figure 4 represents the scale distribution of objects in RailCatFOD-DS. The horizontal axis is labeled as L . The vertical axis is labeled as α and represents the ratio of the number of objects in that region to the total number of objects. According to the distribution depicted in Fig. 4, it contains a large number of small objects. Therefore object detection on it is more challenging.

Proposed method

Overview framework

The overall architecture we propose is illustrated in Fig. 5, which comprises four main components: Encoder, Bottleneck, Decoder, and Cascade Detector¹⁸. In the encoder, we use the Swin Transformer-tiny, which consists of four stages with 2, 2, 6, and 2 Swin Transformer blocks at each stage, respectively. The output of each stage is not only passed to the bottleneck layer but also combined with the output of the bottleneck layer to serve as input to the next stage. Additionally, the bottleneck output is passed to the decoder.

In the bottleneck layer, we propose the RRFEM, which consists of three layers of DBS with dilation rates of {1, 2, 5}. This design aims to effectively expand the regional receptive field without altering the size and dimension of the feature maps, thereby better perceiving regional contextual information.

In the decoder, we employ a Multi-branch fusion feature pyramid network. This enhances adaptability to object scale variations by fusing features from different branches. Finally, the obtained feature maps are fed into a Region Proposal Network (RPN) to produce candidate regions. Subsequently, these candidate regions are standardized into fixed-size Region of Interest (ROI) feature maps through the ROI Align operation, which are then used for subsequent object detection tasks.

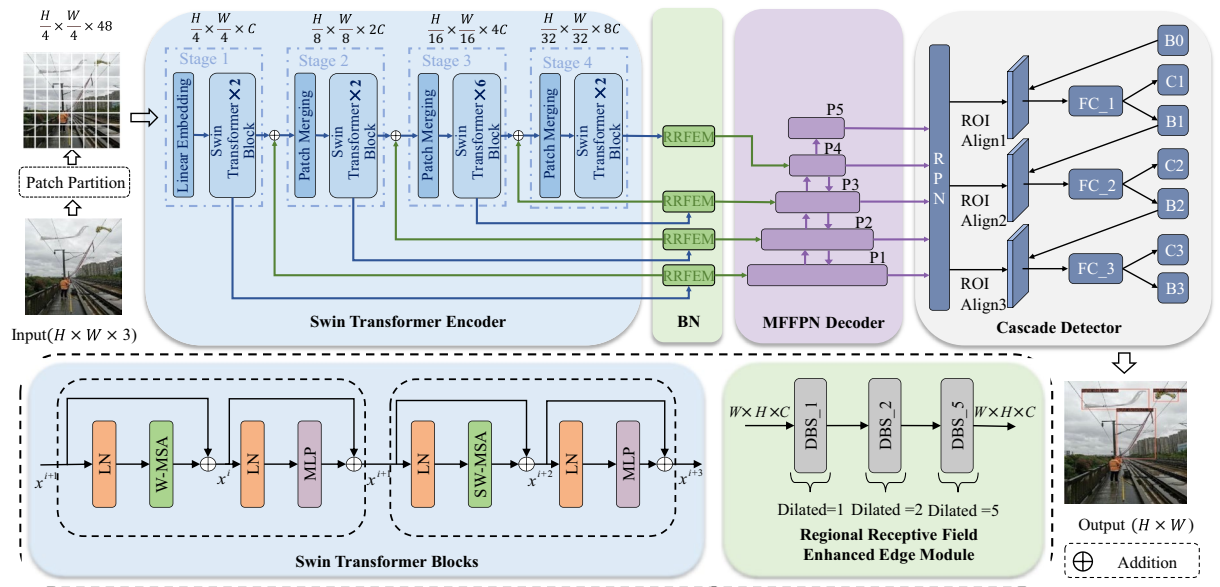


Figure 5. Overall framework diagram of the mode. Illustration includes the general framework of our model, Swin Transformer Blocks, and Regional Receptive Field-Enhanced Edge Module.

Transformer encoder to extract features

Swin transformer

In our model, ST is selected as the core component of the encoder for extracting image features. It learns various feature representations in images during training, as shown in Fig. 5. Each stage of the encoder consists of a Patch Merging Layer and several Swin Transformer Blocks. Figure 5 clearly demonstrates the characteristics of Swin Transformer Blocks, namely, they consistently occur in pairs, with each block internally incorporating Layer Normalization (LN) and Multilayer Perceptron (MLP). In these blocks, the first one implements Window-based Multi-head Self-Attention (W-MSA), which focuses on feature correlations within the window. The second one applies Shifted Window-based Multi-head Self-Attention (SW-MSA).

The algorithm flow of Swin Transformer Blocks is as follows:

$$x^i = W-MSA(LN(x^{i-1})) + x^{i-1} \tag{7}$$

$$x^{i+1} = MLP(LN(x^i)) + x^i \tag{8}$$

$$x^{i+2} = SW-MSA(LN(x^{i+1})) + x^{i+1} \tag{9}$$

$$x^{i+3} = MLP(LN(x^{i+2})) + x^{i+2} \tag{10}$$

where x is the feature mapping of each layer after processing.

W-MSA

The self-attention mechanism is a crucial factor for the outstanding performance of Transformer models in handling computer vision tasks. As a model originally designed for sequence-to-sequence tasks, the Transformer adopts a unique approach when applied to image processing.

It typically segments the image into a series of fixed-sized patches, which are then treated as elements in a sequence for processing. Self-attention processes information across the entire sequence of the image by utilizing three learnable matrices: Queries (W^Q), Keys (W^K) and Values (W^V). The input sequence I is initially multiplied by these weight matrices, resulting in $Q = IW^Q$, $K = IW^K$, and $V = IW^V$.

We can obtain the attention score by the following equation:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_q}}\right)V \tag{11}$$

where $\frac{1}{\sqrt{d_q}}$ is the scaling factor.

The self-attention mechanism computes the dot product between Q and all K . It then applies the softmax operator for normalization to obtain attention scores. Every output becomes a weighted sum of all outputs in the sequence, with the attention scores providing the weights.

The traditional self-attention mechanism computes attention scores across the entire image. In this study, a window-based multi-head self-attention mechanism is employed, as depicted in Fig. 6a, to compute attention

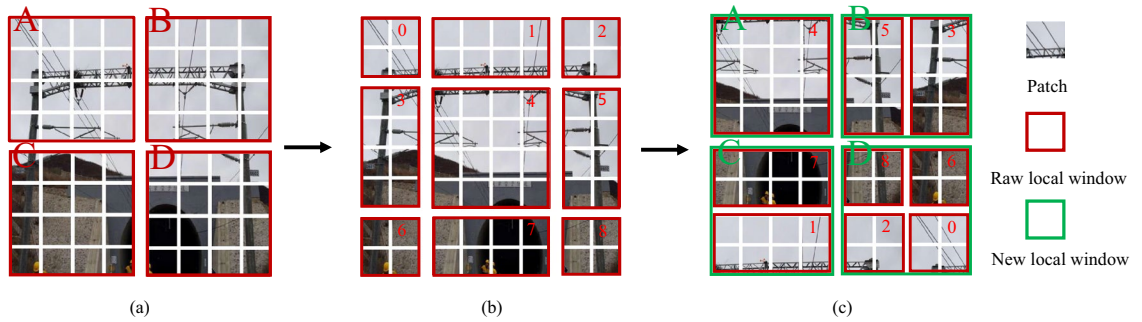


Figure 6. Schematic diagram of SW-MSA window shift. W-MSA only performs the steps in (a) and performs attention calculations in the Raw local window. SW-MSA performs attention calculation in the New local window through the transformation from (b) to (c).

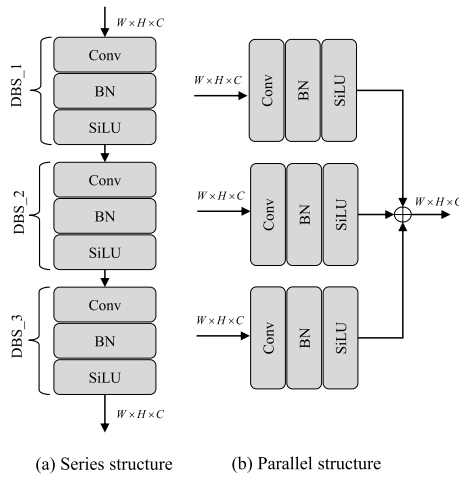


Figure 7. (a) The serial structure of DBS, which forms the RRFEM. (b) The parallel structure of DBS.

scores within each window. Similarly, the image is initially divided into n patches, with adjacent patches merged into a window. The advantage of this approach is that the computational intensity of calculating attention scores within a window is much lower than calculating attention scores across the entire image.

The self-attention mechanism that introduces relative position bias in this study is calculated as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_q}} + B\right)V \tag{12}$$

where B is the introduced positional bias.

SW-MSA

The ST also incorporates shifted window-based self-attention to better interact with information from other windows. This design is retained in our work. Firstly, as shown in Fig. 6, the image is divided into nine regions as illustrated in Fig. 6b. Next, each region is shifted to the corresponding position as shown in Fig. 6c. Finally, as depicted in Fig. 6c, the rearranged image is divided into four parts for computing local attention.

Regional receptive field-enhanced edge module

In the ERCS, foreign objects such as bird nests and plastic films exhibit significant shape variations. Conventional feature pyramid network-based fusion of feature maps output by the backbone often fails to effectively capture the fine-grained local edge information of these objects, leading to less informative representations and limited receptive fields. To overcome these limitations, this study designs the RRFEM and the DBS, as illustrated in Fig. 5.

This serial structure applies convolution operations layer by layer, with each convolution based on the output of the previous one, thereby gradually expanding the receptive field and enabling seamless integration of low-level local features with high-level global information, as illustrated in Fig. 7a. In contrast, in a parallel structure, each branch operates independently, which prevents progressive cross-level feature integration and contextual relationship construction, as illustrated in Fig. 7b. In ERCS, where background elements are often similar or

Next, the multi-scale context is integrated. M_3 is upsampled by a factor of 2 to match the spatial dimensions of F_2 , while the channel number is adjusted using a 1×1 convolution, resulting in M'_3 :

$$M'_3 = \text{Up}_{2 \times}(M_3) \quad (17)$$

The adjusted M'_3 is fused with F_2 :

$$M_2 = \text{Conv}_{1 \times 1}(F_2) + M'_3 \quad (18)$$

A 3×3 convolution operation is applied to F'_2 to obtain the final feature map P_2 :

$$P_2 = \text{Conv}_{3 \times 3}(M_2) \quad (19)$$

The fusion feature at any level P_i can be abstracted as the following formula:

$$F_i = \text{Conv}_{2 \times 2}(C_{i-1}) + C_i \quad (20)$$

$$M_i = \text{Conv}_{1 \times 1}(F_i) + \text{Up}_{2 \times}(M_{i+1}) \quad (21)$$

$$P_i = \text{Conv}_{3 \times 3}(M_i) \quad (22)$$

where C_i represents the basic feature map of the i -th layer. M_i is the intermediate multi-scale feature map of the i -th layer. P_i represents the final fused feature map of the i -th layer.

Experiments

Extensive experiments are conducted on the RailCatFOD-DS to demonstrate the effectiveness of the proposed method. The experiment is divided into the following six parts: Comparative experiment, Ablation experiment results and analysis, Experimental analysis of dilation rates in RRFEM, Comprehensive impact experimental analysis, RRFEM series Parallel structure impact experiment and Generalization experiments on the RailFOD23 dataset.

Detailed experimental settings

Experimental setup

The model training was conducted using Ubuntu operating system, NVIDIA PCIe A100 GPU, and the MMDetection toolbox. The proposed method utilizes the AdamW optimizer with an initial learning rate of 0.0001 and weight decay of 0.005. The betas are set to (0.9, 0.999), and the batch size is set to 14. The model is trained for 60 epochs. For multi-scale training, the long edge of the images is fixed at 1333 pixels, while the short edge ranges from 480 to 800 pixels.

Datasets

To accurately evaluate the performance of the proposed model, the training and validation splits of RailCatFOD-DS are used for model training and validation, while the test set is strictly reserved for performance evaluation. To further assess the generalization capability of the proposed approach, training and validation are additionally conducted on the public RailFOD23 dataset.

Evaluation metrics

We follow the COCO format for annotating the RailCatFOD-DS. To evaluate our results, we primarily use average precision (AP), along with AP50 (AP when the IoU threshold exceeds 0.5), AP75 (AP when the IoU threshold exceeds 0.75), APS (AP for objects smaller than 32×32 pixels), APM (AP for objects between 32×32 pixels and 96×96 pixels), and APL (AP for objects larger than 96×96 pixels). In object detection and instance segmentation, the overall AP is approximately equivalent to the area under the Precision-Recall curve. Precision and Recall are calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (23)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (24)$$

where TP represents the number of correctly identified positive samples, FP indicates the number of negative samples mistakenly identified as positive, and FN denotes the number of positive samples incorrectly identified as negative. In COCO's evaluation metrics, AP is calculated as the mean of AP values across different IoU thresholds (ranging from 0.50 to 0.95 with a step size of 0.05). The IoU is formulated as follows:

$$\text{IoU} = \frac{A \cap B}{A \cup B} \quad (25)$$

where A denotes the prediction, B represents the ground truth, and IoU is the ratio of their intersection to their union.

Comparative experiment

To evaluate the performance of RailCatFOD-Net, this study compares it with existing models, including DETR²¹, Dab-detr²², Dino (ResNet50)²³, Dino (ST-large), Sparse r-cnn²⁴, Cascade r-cnn¹⁸, YOLOx-L²⁵, YOLOv8-L²⁶, DiffusionDet²⁷, YOLOv10-L²⁸ and RT-detr²⁹. These models have not been tested on the RailCatFOD-DS, so we experiment with the default settings of these models. The experimental results are shown in Table 4.

Compared with DINO equipped with the largest backbone in the ST series (Swin-Large), our model adopts a significantly smaller backbone while still achieving a higher AP by 1.2 percentage points (60.2% vs 59.0%). Moreover, the number of parameters in our model is substantially lower than that of DINO (Swin-Large) (78.05 M vs 218.23 M). Although our model and Cascade R-CNN (ResNet-50) are built upon the same detection framework, our approach consistently outperforms it across all evaluation metrics, while maintaining a comparable parameter scale (78.05 M vs. 69.17 M). In particular, a remarkable improvement of 18.0 percentage points is observed for small object detection (53.8% vs 35.8%). Furthermore, when compared with the latest YOLOv10 model, our method also achieves a higher AP by 4.4 percentage points (60.2% vs. 58.8%).

Specifically, for the object occlusion scene in Fig. 9a, our model accurately detects the bird's nest, which is partially occluded by the railway catenary's rigid horizontal beam, with only a small portion visible. In contrast, Dino, YOLOv10 and RT-detr fail to detect it. YOLOx mistakenly detects the connection between the beam and the steel pillar as the bird's nest. In the multi-object scenario shown in Fig. 9b, our model successfully detects all targets, whereas the other methods suffer from missed detections to varying degrees. In contrast, RT-detr additionally produces false detections. In the rainy scene Fig. 9c, our model successfully detects elongated foreign objects in its entirety, while Dino misidentifies the object as two separate objects. DETR and YOLOv10-L fail to fully recognize it. This is because our RRFEM effectively enhances the edge features of the object, improving the detection of elongated foreign objects. In the noisy environment Fig. 9d, both our model and DINO (Swin-L) exhibit strong performance, as the introduction of ST effectively distinguishes the object from the background. In contrast, DETR suffer from false detections. In the no-object background Fig. 9e and low-light scene Fig. 9f, all other models experience false detections. In the strong light environment Fig. 9g, although Dino uses ST as the backbone, the object is unclear due to the strong light, and the model fails to effectively integrate the object's detailed information. In contrast, our model benefits from MFFPN, which effectively fuses detailed features, improving detection performance. The detection precision and detection results verified that the overall detection performance of our model for objects in images was relatively balanced, and the detection of objects was more stable under different scenarios.

Ablation experiment results and analysis

Our model uses Cascade r-cnn as the baseline. The paper conducted ablation experiments on the RailCatFOD-DS to validate the effectiveness of the ST encoder, RRFEM, and MFFPN decoder. Table 5 provides comparative results of the ablation experiments.

The paper first conduct experiments on training and testing using the baseline model. While keeping the other structures of baseline unchanged, replace the network backbone with ST as the Transformer encoder, and add the MFFPN decoder and RRFEM module respectively and the results after adding them simultaneously are shown in the last few rows of the table.

As observed, by simply adding the ST, the model's AP shows significant improvements, with increases of 3.6% (58.1% vs 54.5%). This validates that integrating the ST into baseline model effectively consolidates global information and captures long-range dependencies. Based on M1, M2 incorporates the RRFEM module, leading to improvements in AP, APS, and APM by 1.5% (59.6% vs 58.1%), 12.3% (52.6% vs 40.3%), and 6.2% (59.3% vs 53.1%), respectively. This enhancement further improves the model's local receptive field and edge feature extraction capability. Building upon M1, M3 incorporates the MFFPN decoder, achieving improvements of 1.7% (59.8% vs 58.1%), 10.3% (50.6% vs 40.3%), and 6.6% (59.7% vs 53.1%) in AP, APS, and APM, respectively. The model performs more refined feature fusion, leading to further performance enhancements. Finally, the RRFEM module is utilized as a bridge between the ST encoder and the MFFPN decoder. Compared to the baseline model, our model achieves significant performance improvements.

Method	Backbone	AP	AP50	AP75	APS	APM	APL	Params (M)
DETR	ResNet50	0.576	0.913	0.648	0.436	0.564	0.597	41.56
Dab-detr	ResNet50	0.553	0.928	0.604	0.385	0.565	0.571	43.72
Dino	ResNet50	0.553	0.911	0.598	0.413	0.564	0.569	47.55
Dino	Swin-Large	0.590	0.941	0.653	0.497	0.590	0.607	218.23
Sparse r-cnn	ResNet101	0.542	0.907	0.581	0.409	0.553	0.559	124.94
Cascade r-cnn	ResNet50	0.545	0.889	0.597	0.358	0.553	0.562	69.16
YOLOx-L	CSPDarknet	0.573	0.903	0.631	0.409	0.566	0.591	54.21
YOLOv8-L	CSPDarknet	0.582	0.889	0.662	0.406	0.526	0.609	43.70
YOLOv10-L	CSPDarknet	0.588	0.896	0.659	0.401	0.558	0.612	24.42
DiffusionDet	ResNet101	0.541	0.898	0.582	0.353	0.559	0.559	62.12
RT-detr	ResNet50	0.595	0.918	0.656	0.450	0.576	0.623	41.97
Our model	Swin-Tiny	0.602	0.943	0.656	0.538	0.592	0.619	78.05

Table 4. The impact of different methods on detection precision.



Figure 9. Comparison of the visualization of detection results between this method and other mainstream object detection methods on the RailCatFOD-DS Dataset.

Model	ST-tiny	RRFEM	MFFPN	AP	AP50	AP75	APS	APM	APL
Baseline				0.545	0.889	0.597	0.358	0.553	0.562
M1	✓			0.581	0.914	0.647	0.403	0.531	0.608
M2	✓	✓		0.596	0.946	0.648	0.526	0.593	0.611
M3	✓		✓	0.598	0.943	0.661	0.506	0.597	0.614
Ours	✓	✓	✓	0.602	0.943	0.656	0.538	0.592	0.619

Table 5. Ablation experiment results.

The ablation experiment results indicate that the addition of each module leads to a certain improvement in detection precision. Figure 10 compares the visual results of detection examples from the ablation experiments. In the first column of Fig. 10, the baseline model fails to detect the occluded bird’s nest, resulting in a missed detection. However, when the backbone is replaced with ST, the M1 model accurately detects all objects. In the third column of Fig. 10, the M2 model, incorporating both ST and RRFEM, successfully detects elongated foreign objects. The RRFEM enhances the edge extraction capabilities of M1. In the fourth column of Fig. 10, the M3 model, with ST and MFFPN, enriches detailed features, further improving object precision. However, in

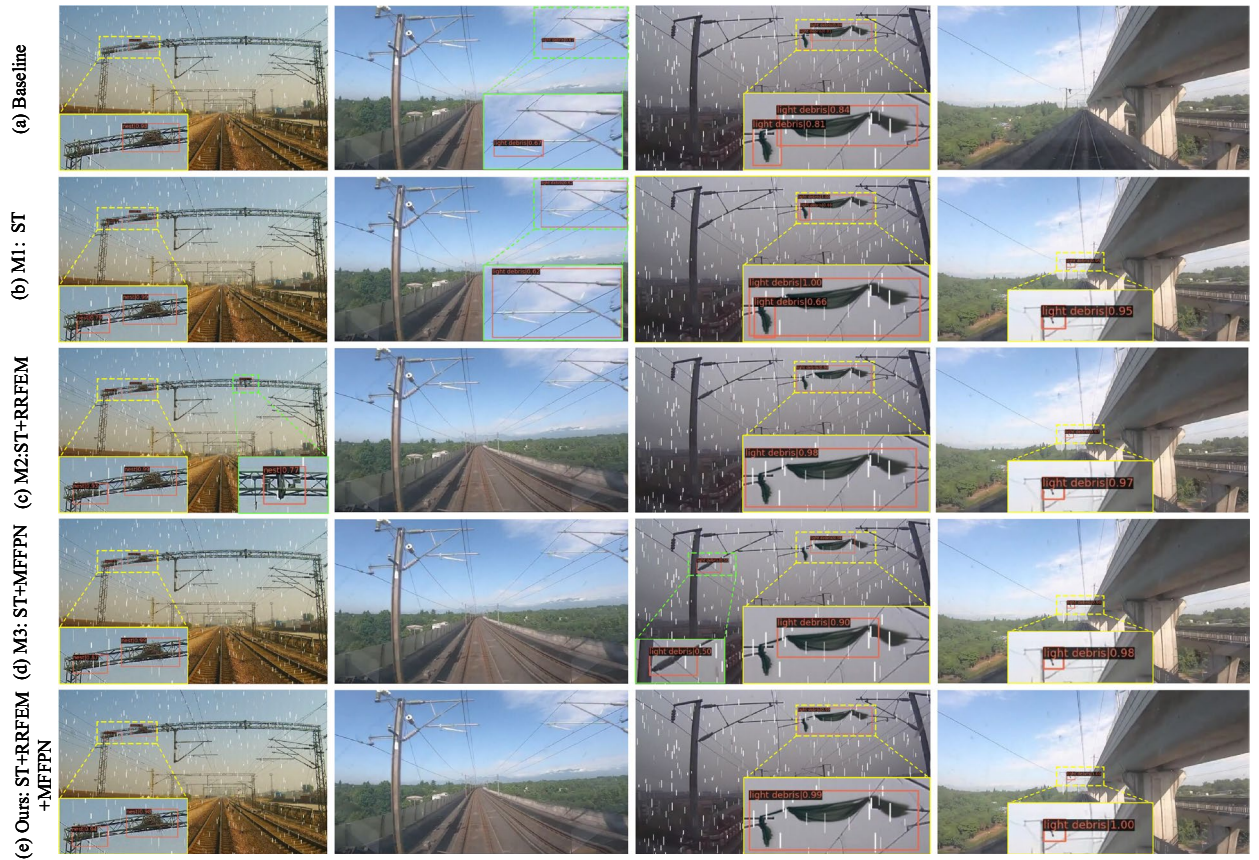


Figure 10. Visual comparison of the detection results of the ablation experiment on the RailCatFOD-DS dataset. (a) shows the baseline experiment, (b) shows the experiment with the addition of Swin Transformer, (c) shows Swin Transformer + RRFEM, (d) shows Swin Transformer + MFFPN, and (e) shows Swin Transformer + RRFEM + MFFPN (ours).

Dilation rates	AP	AP50	AP75	APS	APM	APL
1, 2, 3	0.600	0.939	0.657	0.465	0.584	0.617
1, 2, 5	0.602	0.943	0.656	0.538	0.592	0.619
1, 3, 5	0.595	0.938	0.655	0.512	0.597	0.609
1, 3, 9	0.593	0.940	0.655	0.473	0.601	0.609

Table 6. Effect of different dilation rates on detection precision.

M1, M2, and M3, varying degrees of false detections occur within the green boxes. When RRFEM is utilized as a bridge between the ST encoder and MFFPN decoder, our model accurately detects all objects without missed or false detections.

Experimental analysis of dilation rates in RRFEM

The RRFEM expands the model’s regional receptive field, enabling it to capture local information from a broader area. To evaluate the impact of different dilation rates on object detection precision, experiments were conducted with dilation rates set to {1, 2, 3}, {1, 2, 5}, {1, 3, 5}, and {1, 3, 9}. The results are summarized in Table 6.

The findings demonstrate that RRFEM significantly enhances object detection precision, effectively increasing the network’s focus on objects during training. However, as the regional receptive field grows, precision for small and medium objects decreases, leading to an overall decline. This occurs because excessively large dilation rates extend the receptive field beyond a practical range, impairing the network’s ability to learn small and medium objects effectively. Hence, the improvement in detection precision from RRFEM has an upper limit.

The experiments indicate that RRFEM achieves the best performance with dilation rates of {1, 2, 5}, which is used as the default setting throughout this study unless otherwise noted. Figure 11 visualizes feature map heatmaps with different dilation rates, further highlighting the impact of this configuration.

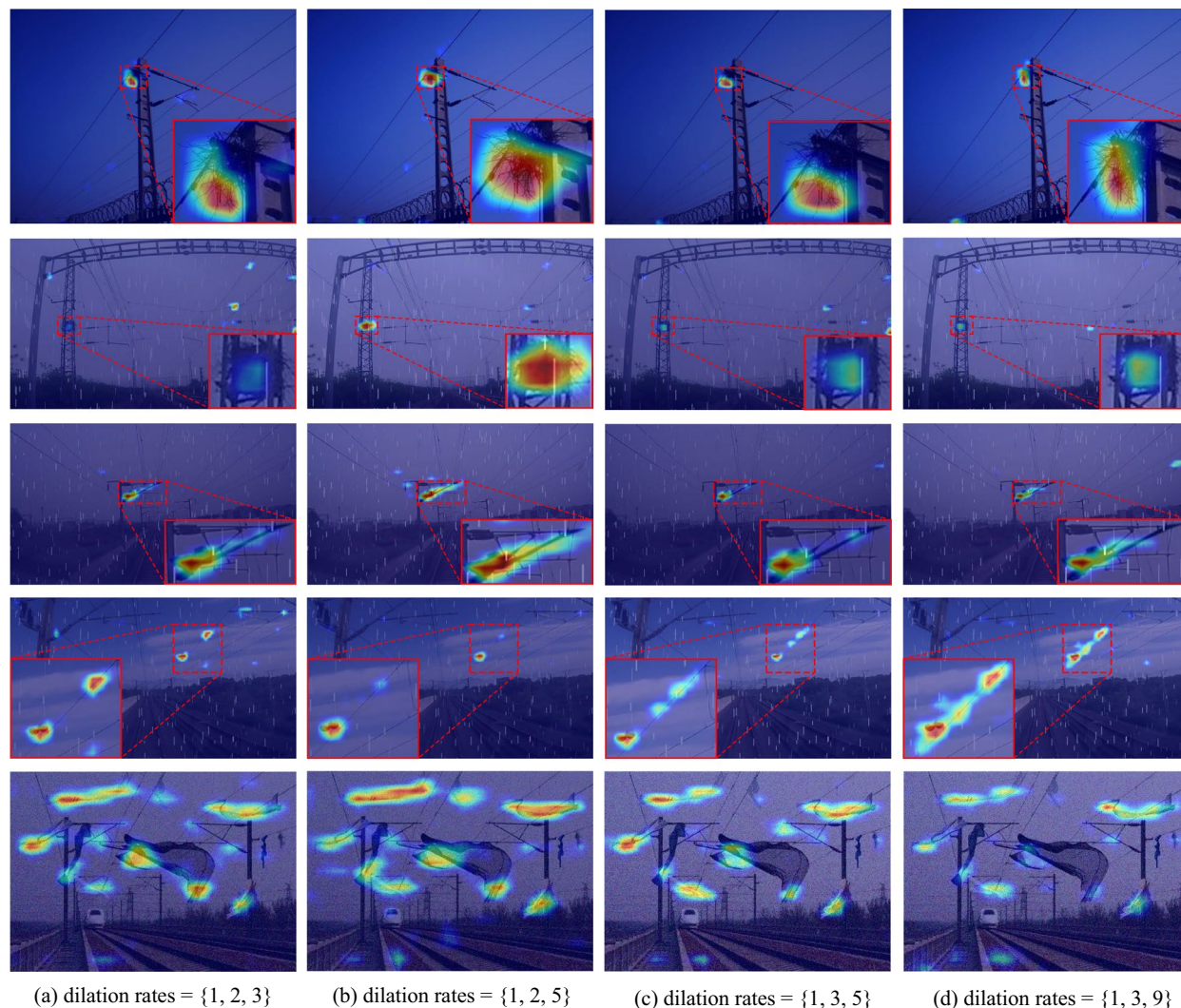


Figure 11. Heatmap visualization comparison of feature maps with different dilation parameters. When the dilation rates are set to {1, 2, 5}, the detection model can more accurately focus on the objects, and the range of captured feature areas can encompass more complete objects.

Structure	Dilation rates	AP	AP50	AP75	APS	APM	APL
Parallel structure	1, 2, 5	0.598	0.930	0.661	0.472	0.572	0.618
Series structure	1, 2, 5	0.602	0.943	0.656	0.538	0.592	0.619

Table 7. Effects of series and parallel structures.

RRFEM series and parallel structure impact experiment

To validate the impact of different RRFEM structures on detection performance, we conducted experiments to compare the effectiveness of these structures. The experimental results are summarized in Table 7. As shown in the results, the series structure demonstrates certain advantages in the task of foreign object detection in railway catenary systems.

Figure 12 presents the visualization of heatmaps for the two structures. In the first and third columns of Fig. 12, the parallel structure RRFEM shows redundancy due to multiple branches processing similar features, which leads to excessive attention to irrelevant information and fails to fully focus on the target objects.

Comprehensive impact experimental analysis

To verify the combined impact of adding RRFEM and MFFPN simultaneously on object detection performance, we set the dilation rates of RRFEM to {1, 2, 5} and conducted experiments using ST with different depths as the backbone network. The numbers of Swin Transformer Block in the four stages of ST-tiny, ST-small and ST-base are {2, 2, 6, 2}, {2, 2, 18, 2}, and {2, 2, 18, 2}. The number of channels in the first-stage hidden layer for each of

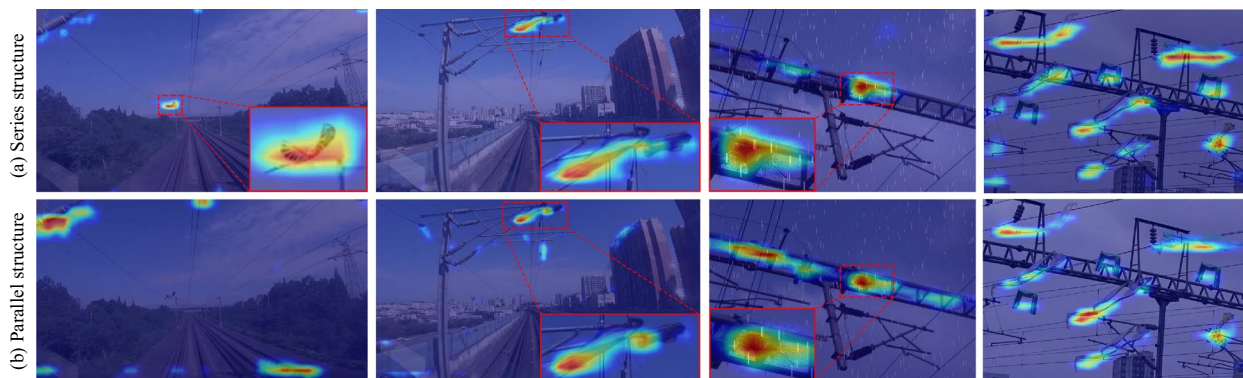


Figure 12. Visualization of different structures of RRFEM.

Model	Stage 1	Stage 2	Stage 3	Stage 4
Swin-T	96 dim 3 heads 2 blocks	192 dim 6 heads 2 blocks	384 dim 12 heads 6 blocks	768 dim 24 heads 2 blocks
Swin-S	96 dim 3 heads 2 blocks	192 dim 6 heads 2 blocks	384 dim 12 heads 18 blocks	768 dim 24 heads 2 blocks
Swin-B	128 dim 4 heads 2 blocks	256 dim 8 heads 2 blocks	512 dim 16 heads 18 blocks	1024 dim 32 heads 2 blocks

Table 8. Comparison of Swin-T, Swin-S and Swin-B architectures across four stages.

Backbone	AP	AP50	AP75	APS	APM	APL
ST-tiny	0.602	0.943	0.656	0.538	0.592	0.619
ST-small	0.604	0.949	0.671	0.502	0.601	0.620
ST-base	0.607	0.939	0.677	0.498	0.583	0.624

Table 9. The impact of different depths on detection precision.

Dataset	Plastic bag	Fluttering object	Nest	Balloon	Total number of labels	Total images
AIGC-based	1675	2762	6992	923	12,352	5000
PS-based	96	279	193	62	630	412
AUG-based	2790	11,573	8610	4586	27,559	9203
Total (RailFOD23)	4561	14,614	15,795	5571	40,541	14,615

Table 10. RailFOD23 distribution of labeled samples. AIGC-based, PS-based, and AUG-based represent data generated by AIGC, PS software, and image enhancement methods on Railsem, respectively.

them is 96, 128 and 192 respectively. Table 8 shows the detailed parameters. Table 9 shows that for ST models of different depths, there are varying degrees of improvement in detection precision. The deeper the network, the more significant the improvement, but there is a decrease in detection precision for small objects as the network depth increases. This experimental result indicates that deeper networks are less favorable for detecting small objects.

Generalization experiments on the RailFOD23 dataset

RailFOD23 is a publicly available dataset designed for foreign object detection in railway traction power supply systems, and it is primarily intended to investigate robust detection performance under complex railway scenarios. The dataset focuses on the traction catenary and its surrounding regions, and it represents one of the few specialized benchmarks specifically constructed for foreign object detection in railway catenary environments.

RailFOD23 mainly covers four typical categories of railway foreign objects: plastic bags, fluttering objects, bird's nests, and balloons. As summarized in Table 10, the dataset construction integrates AIGC-generated data,

Photoshop-based synthesis, and data augmentation. Through multi-source data fusion, RailFOD23 significantly enhances scene diversity and appearance variations, resulting in substantial uncertainty in target appearance, background distribution, and imaging conditions. These characteristics make RailFOD23 particularly suitable for evaluating model generalization capability and cross-scene robustness. Therefore, in this study, the RailFOD23 dataset is adopted as the benchmark for generalization experiments to validate the detection performance and adaptability of the proposed model under complex railway environments and diverse foreign object morphologies.

To evaluate the generalization capability of RailCatFOD-Net, generalization experiments are conducted on the public RailFOD23 dataset, where the proposed method is compared with DETR, DAB-DETR, DINO, Cascade R-CNN, YOLOv5, YOLOX, and YOLOv11. The experimental results are reported in Table 11. Our model consistently outperforms existing methods in terms of detection precision. Specifically, compared with DETR, the AP is increased from 70.4% to 84.8%, yielding an absolute improvement of 14.4 percentage points. When compared with YOLOX, our approach achieves an AP gain of 8.4 percentage points. To further evaluate the visual detection performance, several representative images are selected from the test set for qualitative comparison, as illustrated in the Fig. 13. Although the YOLO-series methods are capable of detecting foreign objects, their detection accuracy consistently remains inferior to that of the proposed approach. In particular, in the third column of the images, YOLOX incorrectly classifies background regions as fluttering objects and misidentifies plastic bags as fluttering objects and bird's nests. Overall, the proposed model demonstrates higher accuracy and stronger robustness in complex scenarios, thereby validating its effectiveness for practical foreign object detection tasks.

Training loss and accuracy analysis

The loss curve reflects the model's learning progress during training, ideally decreasing as training progresses. The accuracy curve demonstrates the model's performance on a validation or test set, ideally increasing with training epochs.

As shown in Fig. 14a, the total loss decreased and converged further after reducing the learning rate at 10k iterations. After 20k iterations, when the learning rate was lowered again, the loss curve showed no significant change, indicating that the model had converged. The classification loss (Fig. 14c) and bounding box loss (Fig. 14b) followed a similar trend.

Figure 14d depicts the accuracy curve. After 10k iterations and the initial learning rate reduction, accuracy improved. However, following the second reduction at 20k iterations, the accuracy curve stabilized, indicating performance had plateaued. Observing all four curves, the model achieved convergence around 25k iterations.

Conclusion

In this study, we address the challenging problem of foreign object detection in electrified railway catenary systems (ERCS), where complex structural layouts, severe environmental variations, and diverse foreign object characteristics significantly hinder reliable detection. To this end, we construct a dedicated dataset, RailCatFOD-DS, and propose a customized detection framework termed RailCatFOD-Net. By integrating the hierarchical feature representation of the Swin Transformer, the refined multi-scale fusion strategy of the MFFPN, and the receptive field enhancement of the RRFEM, the proposed framework achieves superior detection performance, particularly for small-scale and elongated foreign objects, while maintaining robustness under occlusion, illumination variations, and adverse weather conditions. Despite these advantages, several limitations of the proposed method remain. First, the incorporation of transformer-based backbones and multi-branch feature fusion inevitably increases computational complexity, which may restrict direct deployment on resource-constrained embedded platforms. Second, the current framework relies solely on visual information, making it vulnerable in extreme scenarios such as heavy fog, nighttime glare, or severe occlusion, where visual cues become unreliable. Third, although RailCatFOD-DS covers multiple typical railway scenarios, the diversity of real-world operating environments is still broader, and domain shifts may affect generalization performance. To address these limitations, future work will focus on lightweight architectural optimization, including model pruning, knowledge distillation, and efficient transformer variants, to reduce computational overhead while preserving detection accuracy. In addition, multi-sensor data fusion, such as combining vision with infrared or LiDAR signals, will be explored to enhance detection reliability under challenging environmental conditions.

Method	Backbone	AP	AP50	AP75	APS	APM	APL	Params(M)
DETR	ResNet50	0.704	0.919	0.81	–	0.574	0.725	41.56
Dab-detr	ResNet50	0.770	0.929	0.835	–	0.681	0.776	43.72
Dino	ResNet50	0.825	0.927	0.864	–	0.771	0.824	47.55
Cascade r-cnn	ResNet50	0.824	0.930	0.864	–	0.789	0.819	69.16
YOLOx-L	CSPDarknet	0.764	0.927	0.844	–	0.708	0.767	54.21
YOLOv5-m	CSPDarknet	0.819	0.934	0.876	–	0.763	0.820	25.05
YOLOv8-m	CSPDarknet	0.829	0.934	0.880	–	0.780	0.829	25.84
YOLOv11-m	–	0.814	0.930	0.863	–	0.775	0.821	20.03
Our model	Swin-Tiny	0.848	0.938	0.882	–	0.844	0.841	78.05

Table 11. Generalization experiments of different models on RailFOD23.



Figure 13. Comparison of the visualization of detection results between this method and other mainstream object detection methods on the RailFOD23 dataset.

Finally, large-scale dataset expansion and cross-domain adaptation techniques will be investigated to further improve the generalization capability of the model across diverse railway lines and operational environments. In summary, this study provides both a dedicated dataset and an effective detection framework for foreign object intrusion detection in railway catenary systems, offering a foundation for future research.

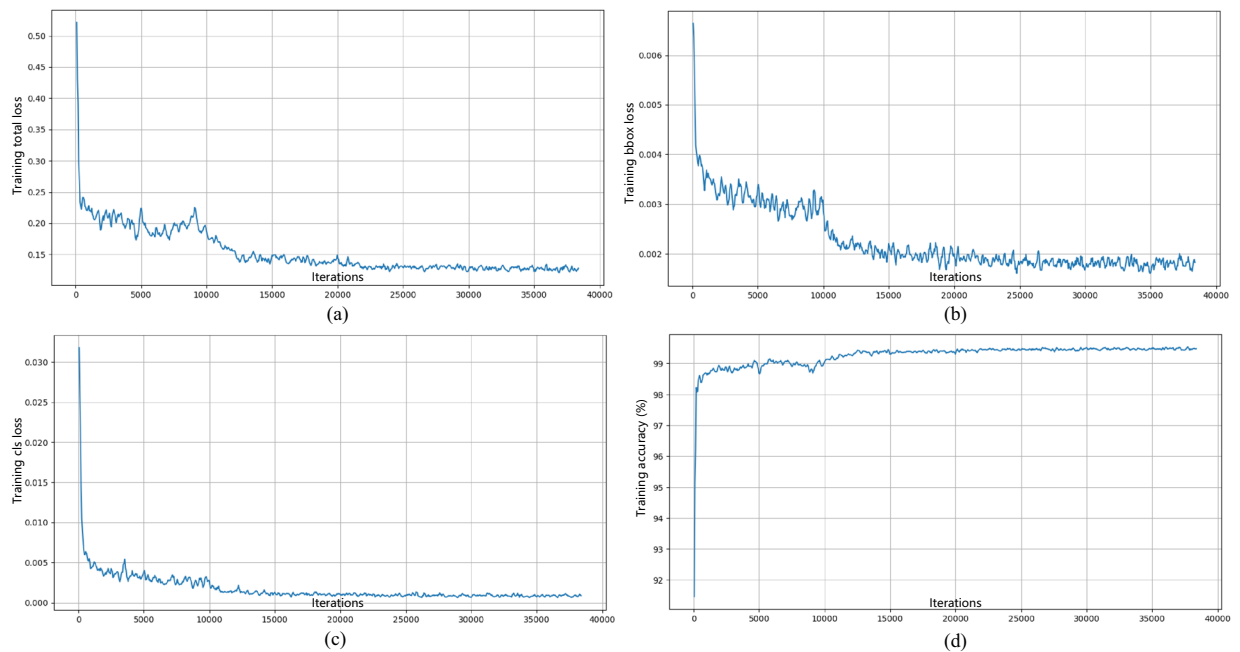


Figure 14. Loss and accuracy curves. (a) Illustrates the total loss curve during model training. (b) Illustrates the bounding box loss curve during training. (c) Illustrates the classification loss curve during training. (d) Illustrates the accuracy curve during model training.

Data availability

The constructed Railway Catenary Foreign Object Dataset is available on GitHub at <https://github.com/dashgfuiddng/CPSSFOdataset.git>

Received: 10 November 2025; Accepted: 3 February 2026

Published online: 14 February 2026

References

- Bocciolone, M., Bucca, G., Collina, A. & Comolli, L. Pantograph–catenary monitoring by means of fibre bragg grating sensors: Results from tests in an underground line. *Mech. Syst. Signal Process.* **41**, 226–238. <https://doi.org/10.1016/j.ymssp.2013.06.030> (2013).
- Hofler, H., Dambacher, M., Dimopoulos, N. & Jetter, V. Monitoring and inspecting overhead wires and supporting structures. In *IEEE Intelligent Vehicles Symposium* 512–517, 2004. <https://doi.org/10.1109/IVS.2004.1336436> (IEEE 2004).
- Cao, Z. et al. Railway intrusion detection based on machine vision: A survey, challenges, and perspectives. *IEEE Trans. Intell. Transp. Syst.* **25**, 6427–6448 (2024).
- Wu, X., Yuan, P., Peng, Q., Ngo, C.-W. & He, J.-Y. Detection of bird nests in overhead catenary system images for high-speed rail. *Pattern Recogn.* **51**, 242–254. <https://doi.org/10.1016/j.patcog.2015.09.010> (2016).
- Qiu, S. et al. Automated detection of railway defective fasteners based on yolov8-fam and synthetic data using style transfer. *Autom. Constr.* **162**, 105363. <https://doi.org/10.1016/j.autcon.2024.105363> (2024).
- Bai, T., Yang, J., Xu, G. & Yao, D. An optimized railway fastener detection method based on modified faster r-cnn. *Measurement* **182**, 109742. <https://doi.org/10.1016/j.measurement.2021.109742> (2021).
- Lian, L. et al. Rvsa-3d: voxel-based fully sparse attention 3d object detection for rail transit obstacle perception. *Pattern Recogn.* 112324 (2025).
- Lian, L. et al. Rae3d: Multiscale aggregation-enhanced 3d object detection for rail transit obstacle perception. *IEEE Trans. Ind. Inf.* (2025).
- Chen, Z. et al. Foreign object detection method for railway catenary based on a scarce image generation model and lightweight perception architecture. *IEEE Trans. Circ. Syst. Video Technol.* (2025).
- Vaswani, A. et al. Attention is all you need. In Guyon, I. et al. (eds.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017).
- Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).
- Zhao, Z., Kang, J., Sun, Z., Ye, T. & Wu, B. A real-time and high-accuracy railway obstacle detection method using lightweight cnn and improved transformer. *Measurement* **238**, 115380. <https://doi.org/10.1016/j.measurement.2024.115380> (2024).
- Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022 (2021).
- Chen, Z., Yang, J., Feng, Z. & Zhu, H. Railfod23: A dataset for foreign object detection on railroad transmission lines. *Sci. Data* **11**, 72 (2024).
- Zendel, O. et al. Railsem19: A dataset for semantic rail scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0 (2019).
- Yue, Z., Xie, J., Zhao, Q. & Meng, D. Semi-supervised video deraining with dynamical rain generator. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 642–652. <https://doi.org/10.1109/CVPR46437.2021.00070> (2021).
- DeVries, T. Improved regularization of convolutional neural networks with cutout. arXiv preprint [arXiv:1708.04552](https://arxiv.org/abs/1708.04552) (2017).

18. Cai, Z. & Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6154–6162. <https://doi.org/10.1109/CVPR.2018.00644> (2018).
19. Wang, P. et al. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1451–1460. <https://doi.org/10.1109/WACV.2018.00163> (2018).
20. Lin, T.-Y. et al. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944. <https://doi.org/10.1109/CVPR.2017.106> (2017).
21. Carion, N. et al. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229 (Springer, 2020).
22. Liu, S. et al. Dab-detr: Dynamic anchor boxes are better queries for detr. In *The Twelfth International Conference on Learning Representations (ICLR)* (2022).
23. SZhang, H. et al. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Twelfth International Conference on Learning Representations (ICLR)* (2023).
24. Sun, P. et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14449–14458. <https://doi.org/10.1109/CVPR46437.2021.01422> (2021).
25. Ge, Z., Liu, S., Wang, F., Li, Z. & Sun, J. Yolox: Exceeding yolo series in 2021. *ArXiv* **abs/2107.08430** (2021).
26. Varghese, R. & M., S. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 1–6. <https://doi.org/10.1109/ADICS5844.8.2024.10533619> (2024).
27. Chen, S., Sun, P., Song, Y. & Luo, P. Diffusiondet: Diffusion model for object detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 19773–19786. <https://doi.org/10.1109/ICCV51070.2023.01816> (2023).
28. Wang, A. et al. Yolov10: Real-time end-to-end object detection. In *NeurIPS 2024* (2024).
29. Lv, Y., Chen, W., Chen, B., Zhang, Z. & Wang, J. Rt-detr: Dets beat yolos on real-time object detection. In *Proc. of the IEEE/CVF Int. Conf. on Computer Vision Workshops (ICCVW)*, 2420–2430 (2023).

Acknowledgements

Thanks for the hard work of the editors and the constructive suggestions of the anonymous reviewers.

Author contributions

Li contributed to the conceptualization, methodology, project administration, and writing—original draft. Cao was responsible for investigation, methodology, software, visualization, and writing—original draft. Yang contributed to data curation, software, and visualization. Diao performed formal analysis, investigation, and validation. Wang contributed to formal analysis, resources, and validation. Yu was responsible for data curation, resources, and writing—review & editing. Yan contributed to data curation and writing—review & editing. Xu provided funding acquisition, supervision, and writing—review & editing. All authors reviewed the manuscript.

Funding

This research was supported in part by the National Natural Science Foundation of China (No.62276042), by the Artificial Intelligence Scientific and Technological Innovation Program of Liaoning Province (No. 2023JH26/10100008), and by the Science and Technology Plan Joint Program of Liaoning Province (No.2025-MSLH-138, No.2025 BSLH-100, No.2025-BSLH-101).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026