scientific reports

Check for updates

OPEN

A railway surface defect detection model based on topology-enhanced feature association

Qike Wu[1,2], Sharafiz bin Abdul Rahim[2]✉, SaiHong Tang[2], Muhammad Azim bin Azizi[2] & Jiapei Wei[2]

Rail defect detection is essential for ensuring railway safety and enabling reliable maintenance. However, existing methods often struggle in complex environments, where feature ambiguity and occlusion significantly degrade detection performance, leading to frequent missed and false detections. To overcome these limitations, we propose a topology-enhanced relational modeling framework that jointly captures local, global, and high-order dependencies. Specifically, a semantic association graph is constructed to enable dynamic multi-order feature extraction, allowing the model to learn structured relationships among spatial regions and effectively enhance the representation of subtle defects. To further improve global perception, a multi-directional state-space modeling module is introduced to capture long-range dependencies and enhance spatial sensitivity. Moreover, a hypergraph-based interaction mechanism is designed to model complex high-order relationships, where hypergraph convolution facilitates deep feature fusion and improves discrimination of heterogeneous defect patterns. Experimental results demonstrate that the proposed method consistently outperforms state-of-the-art approaches, achieving higher accuracy and stronger robustness under challenging conditions.

Keywords Rail defect detection, Topology-enhanced modeling, Graph neural networks, State space models, Hypergraph learning

Rail defect detection plays a crucial role in ensuring railway safety, preventing accidents, and improving maintenance efficiency[1]. Undetected rail defects may compromise train operation safety and even lead to catastrophic accidents[2,3]. Therefore, the timely and accurate detection of rail surface defects is essential for maintaining railway integrity and promoting the sustainable development of transportation systems[4].

With the advancement of autonomous inspection vehicles, onboard camera systems, and related sensing technologies, artificial intelligence (AI) has been increasingly integrated into rail defect detection[5,6]. AI-based automated inspection systems enable early identification of potential hazards and reduce the reliance on manual inspection. However, accurate defect detection remains challenging due to complex environmental noise, varying illumination conditions, and limited defect samples, all of which hinder the robustness and generalization of detection models[7,8]. Related studies on anomaly analysis have shown that cross-modal and target-focused modeling can improve robustness in complex environments, especially when appearance cues are ambiguous or partially degraded[9].

In recent years, the rapid development of deep learning–based object detection algorithms has provided new opportunities for automated rail defect inspection. Detection frameworks based on convolutional neural networks (CNNs) are generally classified into two categories: two-stage and one-stage methods. Two-stage detectors, such as Fast R-CNN[10], first generate region proposals and then perform classification and bounding-box regression to achieve high accuracy. In contrast, one-stage detectors, such as YOLO[11] and SSD[12], directly predict object categories and locations in a single step and were developed to improve detection speed. Recent advances have further strengthened the trade-off between detection accuracy and computational efficiency in real-time object detection[13,14]. Although these methods have achieved remarkable success in general computer vision tasks, their direct application to railway track defect detection remains limited due to the small size, low contrast, and diverse appearance of rail defects, which often degrade detection performance.

To address these issues, researchers have proposed various methods to enhance rail surface defect detection. Reference[15] proposed the Pixel-aware Frequency Conversion Network (PFCNet), which integrates frequency-

[1]School of Foreign Languages, Hainan Tropical Ocean University, No. 1 Yucai Road, Sanya 572022, Hainan, China. [2]Mechanical Engineering Department, Universiti Putra Malaysia, Jalan UPM, Serdang 43400, Selangor, Malaysia. ✉email: sharafiz@upm.edu.my

domain information with attention mechanisms to improve feature representation and detection accuracy. Reference[16] developed a multitask learning framework that integrates multiple detectors to handle diverse defect types, image variations, and limited training samples, achieving superior accuracy. Reference[17] tackled the challenge of limited samples by reformulating defect detection as a sequential learning task and proposed two line-level models—OC-IAN and OC-TD—based on one-dimensional CNNs and LSTMs, which achieved state-of-the-art results on the RSDDs dataset. Reference[18] developed the Dense Attention-Guided Cascaded Network (DACNet) for defect detection under complex industrial conditions, effectively capturing fine-grained defect features and outperforming prior methods across multiple public datasets. Furthermore, Unsupervised track anomaly detection has also attracted increasing attention. Reference[19] introduced an unsupervised stereoscopic saliency detection method based on a binocular line-scanning system capable of simultaneously acquiring high-resolution images and 3D profiles. By combining low-rank nonnegative reconstruction, geometric outlier detection, and depth-saliency fusion, the authors achieved robust performance and established the RSDDS dataset. Reference[20] proposed a segmented-aggregated framework with an internal-external constraint mechanism to model track anomalies without relying on dense manual annotation. In addition to image-based defect analysis, semisupervised railway track anomaly detection combined with vision-LiDAR decision fusion has also been explored to improve robustness under complex inspection conditions[21].

Overall, existing research has achieved considerable progress in rail defect detection, demonstrating the effectiveness of deep learning-based approaches in complex environments. However, accurately detecting small-scale or low-contrast defects remains challenging, often resulting in missed detections and reduced reliability.

More importantly, most existing methods focus on optimizing individual components, while lacking a unified framework to jointly model local details, global context, and complex feature relationships. This limitation restricts their ability to capture the intrinsic structural and textural characteristics of rail defects. In addition, many approaches are directly adapted from general-purpose vision architectures, which are not specifically designed for rail inspection scenarios, leading to suboptimal performance in real-world applications.

Furthermore, while recent studies have explored relational modeling and graph-based methods, their theoretical advantages over conventional convolutional or standard graph convolution approaches in defect detection tasks remain insufficiently analyzed. At the same time, the practical applicability of these methods, including computational complexity, model size, and inference efficiency, is often underreported, making it difficult to assess their deployment potential in real-world railway systems. From a broader industrial monitoring perspective, robustness to nonstationary conditions and process-level variation has also been recognized as critical for reliable deployment[22,23].

Therefore, instead of introducing entirely new architectures, a more effective direction is to integrate multiple complementary modeling strategies into a unified and efficient framework. Such integration not only enhances multi-level feature representation but also provides a more practical balance between detection accuracy, robustness, and computational cost, which is critical for real-world rail defect detection systems.

To address these challenges, we develop a rail defect detection framework based on topology-enhanced relational modeling. The main contributions of this work are summarized as follows:

1) We introduce a dynamic multi-order feature extraction module based on a semantic association graph. This module constructs spatial-semantic topological relationships among feature regions and utilizes a routing mechanism to model contextual dependencies. Attention operations are further applied to highly correlated regions, which helps improve the representation of defect-related features.

2) We incorporate a state-space modeling module to capture long-range dependencies. To alleviate its limited sensitivity to spatial structures in image data, a random cross-scanning strategy is adopted, allowing the model to better aggregate contextual information and enhance feature representations, particularly in the presence of occlusion.

3) We design a hypergraph-guided feature fusion pyramid to model higher-order relationships among features. By introducing hypergraph convolution, the proposed method facilitates multi-level feature interaction and improves the integration of semantic information across different scales.

Methodology

As illustrated in Fig. 1, we develop an object detection framework based on spatial–semantic relational modeling, built upon the RT-DETR[24] architecture. The proposed framework is intended to improve defect detection performance under complex railway inspection conditions, where challenges such as occlusion, low contrast, and irregular defect patterns are commonly encountered. By incorporating relational modeling mechanisms, the framework aims to better capture dependencies among feature regions and enhance feature representation.

RT-DETR follows an end-to-end detection paradigm consisting of a backbone, a neck, and a transformer-based decoder. Given an input image, hierarchical multi-scale features are first extracted by the backbone network and then progressively aggregated by the neck to facilitate feature fusion across different resolutions. The transformer-based decoder subsequently performs global reasoning over the fused features and directly generates the final detection results, eliminating the need for handcrafted post-processing steps.

Overall, the proposed framework follows a hierarchical design that integrates three complementary components: (1) local spatial-semantic relational modeling in the backbone, (2) global context modeling via a state-space mechanism, and (3) multi-scale feature fusion with high-order relational reasoning in the neck. These components jointly enhance feature representation from local, global, and cross-scale perspectives.

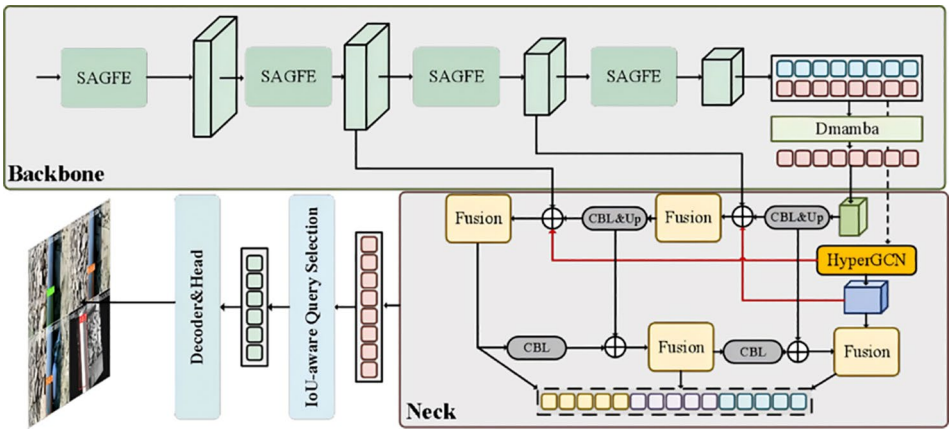

Figure 1. Overall architecture of the proposed model.

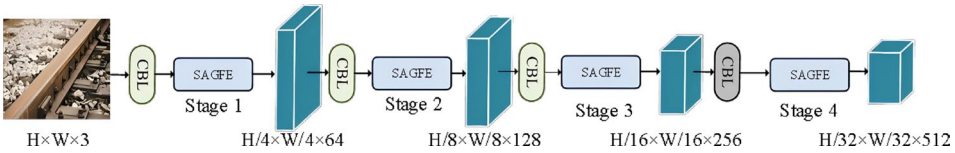

Figure 2. Overall architecture of the backbone.

Dynamic multi-order feature extraction method based on semantic association graph

The overall architecture of the backbone network is shown in Fig. 2. which consists of stacked semantic association graph-based dynamic multi-order feature extraction (SAGFE) modules and convolutional downsampling blocks (CBL), organized into four stages.

In this design, the SAGFE modules are responsible for modeling spatial-semantic relationships and enhancing feature representation, while the downsampling blocks progressively reduce spatial resolution and expand the receptive field, enabling hierarchical feature extraction across different scales.

Dynamic multi-order feature extraction module based on semantic association graph (SAGFE)

The SAGFE module serves as the core component for feature extraction in the backbone network. It is motivated by the limitation of conventional convolutional and graph-based methods in modeling complex spatial dependencies, particularly for small or occluded defects. Inspired by the regional reasoning strategy in[25], the module is designed to facilitate adaptive aggregation of contextual information through a relational modeling framework. The overall structure of SAGFE is illustrated in Fig. 3.

Specifically, the input features are first encoded to generate adaptive contextual representations, enabling the model to capture local variations in feature distributions.

Subsequently, a sparse attention mechanism is introduced to selectively model interactions among feature regions. By focusing on highly relevant regions and reducing redundant connections, this mechanism improves the efficiency of contextual information aggregation while suppressing background interference.

Based on the routing relationships derived from sparse attention, a semantic association graph is then constructed to model spatial-semantic dependencies among regions. Unlike conventional graph convolution networks that rely on fixed or dense connectivity, the proposed approach adopts a dynamic and sparse graph structure, allowing more flexible modeling of feature relationships. Finally, dynamic graph convolution is applied to aggregate information across regions, which contributes to more informative and structured feature representations.

Dynamic context information extraction

Following[26], convolutional operations are employed to extract contextual information. Compared with fully connected layers, convolutional layers are more suitable for modeling visual patterns such as shape, size, and texture, while preserving the spatial structure of the input.

Moreover, the proposed design introduces limited additional computational overhead, as dynamic context generation is implemented through the concatenation of the input feature x during key construction, followed by a linear projection. This formulation enables input-conditioned feature transformation without significantly increasing model complexity.

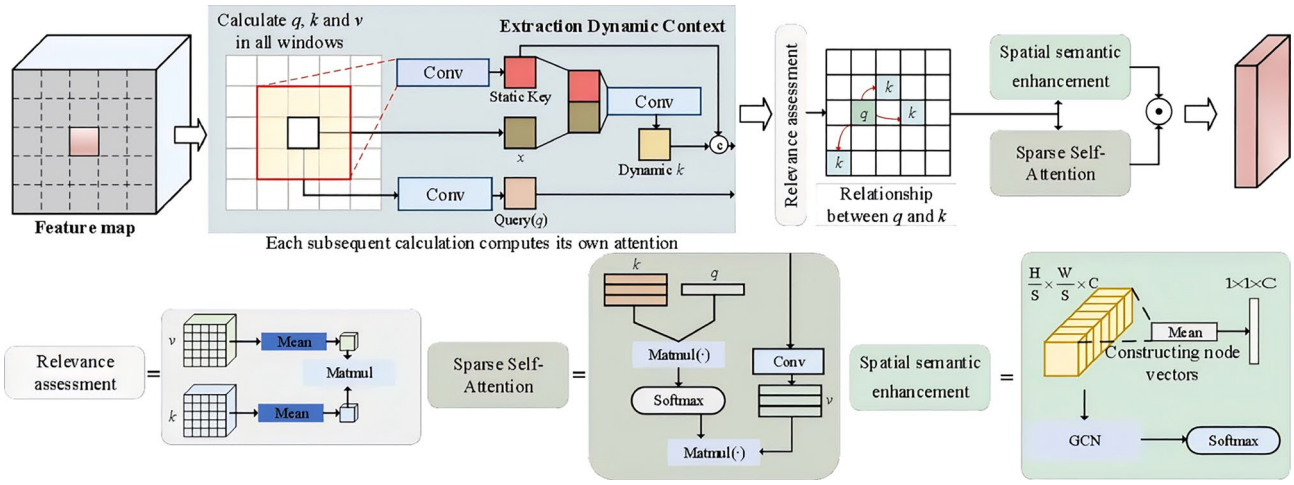

Figure 3. Overall architecture of the SAGFE module.

Accordingly, the input features are first encoded to produce the static *(k)* and *(q)* representations, which can be expressed as follows:

$$\begin{aligned} k &= \mathrm{Conv}_{1\times1}(x), \\ q &= \mathrm{Conv}_{1\times1}(x), \\ \mathrm{Conv}(x) &= x \cdot W + b \end{aligned} \tag{1}$$

Here, *W* and *b* denote the learnable parameters and bias term, respectively. The static *(k)* and *(q)* are generated using convolutional layers with a kernel size of 1. The 1×1 convolution is functionally equivalent to the linear projection commonly used in Transformer architectures, while preserving spatial structure through its convolutional formulation and introducing limited computational overhead. Subsequently, the input feature *x* is concatenated with the static *k* along the channel dimension and processed by a convolutional operation. This design allows the key representation to be conditioned on the input features, introducing adaptive variations that reflect local contextual differences. The operation is formulated as follows:

$$k_d = \mathrm{Conv}_{1\times1}\big(\mathrm{concat}([x, k],\ \mathrm{dim} = c)\big) \tag{2}$$

Construction of the spatial-semantic association graph

After obtaining the q and dynamic k_d, the feature map is partitioned into $s \times s$ local windows to facilitate region-level modeling. Following[27], a directed graph is constructed to represent dependency relationships among spatial regions. To obtain compact node representations, mean pooling is applied over the spatial dimensions within each local window, producing a channel-wise feature vector for each region. This operation reduces local variations while preserving the overall semantic distribution, and can be formulated as follows:

$$q^r = \mathrm{Mean}(q), \quad k^r = \mathrm{Mean}(k_d) \tag{3}$$

Here, q^r, $k^r \in \mathbb{R}^{s^2\times c}$, and Mean() denotes the mean operation. Subsequently, the matrix multiplication between q^r and the transpose of k^r derives the correlation matrix $A^r \in \mathbb{R}^{s^2\times s^2}$, which represents the dependency relationships among regions. The *i*-th row A^r corresponds to the correlation between the query q in the *i*-th region and the *k* in all other regions. Therefore, by pruning the matrix A^r, we can obtain the routing index that links each query q to its most relevant *k* within the corresponding region, expressed as follows:

$$I^r = \mathrm{TopIndex}(A^r) \tag{4}$$

Based on the matrix $I^r \in \mathbb{R}^{s^2\times k}$, we can further derive the adjacency matrix, which explicitly defines the edges of the topological graph structure. For each query q, we retain the Top-k *k* with the highest correlation weights, setting the corresponding entries to 1, while assigning 0 to all others. In this way, the adjacency matrix $M \in \mathbb{R}^{s^2\times s^2}$ is constructed. This sparsification strategy also helps control computational complexity, making the graph construction feasible for high-resolution feature maps. The node representations are obtained by performing mean pooling over the spatial dimensions within each local window, resulting in a feature vector of channel dimension for each window. The operation can be expressed as follows:

$$x^r = \mathrm{Mean}(x, \mathrm{dim} = c) \tag{5}$$

Here, x^r effectively aggregates the local contextual information within each window. Therefore, the spatial-semantic topological graph of the features can be represented as $G = (x^r, M)$. Subsequently, a sparse attention

mechanism is applied to enhance the spatial semantic correlations within the features by leveraging long-range dependencies among targets in the graph structure, thereby producing a hybrid dual-attention representation.

Sparse self-attention mechanism
Based on the routing relationships between the query and key, a sparse attention mechanism is constructed to restrict computation to the most relevant regions. This design reduces redundant interactions and improves computational efficiency. Specifically, the value (v) is computed within the selected regions as follows:

$$\tilde{v} = \mathrm{Conv}_{1\times1}\left(\mathrm{gather}(x^r, I^r)\right) \tag{6}$$

Here, the gather() function is used to collect tensors from the effective regions based on the routing indices. Subsequently, the key is gathered as follows:

$$\tilde{k} = \mathrm{gather}(k_d, I^r) \tag{7}$$

Then, the attention computation is performed, which can be formulated as follows:

$$O_1 = \mathrm{Attention}(q,k,v) = \mathrm{Softmax}\left(\frac{q\tilde{k}^T}{\sqrt{C}}\right)\tilde{v} \tag{8}$$

Semantic feature enhancement
Based on the constructed spatial-semantic topological graph, graph convolution is employed to aggregate information across different regions and model their spatial dependencies. Unlike conventional graph convolution that relies on fixed adjacency structures, a dynamic graph convolution mechanism is introduced, where the adjacency matrix is adaptively updated according to feature similarity. This design enables more flexible modeling of relationships among regions, including interactions between defect areas and their surrounding context.

To ensure stable training, the input features are first normalized to control the scale of activations and mitigate potential gradient instability. The normalization process can be expressed as follows:

$$\hat{x}^r = \frac{x^r - \mathrm{E}(x^r)}{\sqrt{\mathrm{Var}(x^r) + \mathcal{S}}}, \breve{x} = \hat{x}^r \cdot \lambda + \gamma \tag{9}$$

Here, $\mathrm{E}()$ and $\mathrm{Var}()$ denote the mean and variance of the input node features, respectively. By normalizing the input data, the training process can be effectively stabilized. The term $\mathcal{S}$ is introduced to prevent division by zero, while γ, λ represent the learnable parameters. Therefore, the node feature matrix can be formulated as follows:

$$x_p = \breve{x}p \tag{10}$$

Here, p denotes the attention weight parameters computed from the input features. Subsequently, the feature similarity matrix is calculated as follows:

$$W = \sigma(x_p \cdot x_p^T) \tag{11}$$

Here, $\sigma()$ represents the sigmoid activation function. After obtaining the similarity matrix W, the correlation coefficients are computed by normalizing the neighboring nodes of each feature node. The adjacency matrix M is subsequently updated according to these correlation values, allowing it to be represented in a dynamic form as follows:

$$M_d = \mathrm{Softmax}\left(\mathrm{diag}(0,\infty)\right) \cdot (MW + E) \tag{12}$$

Here, $\mathrm{diag}(0,\infty)$ denotes a diagonal matrix used to eliminate self-loops in the generated adjacency matrix, while E represents the identity matrix. After obtaining the dynamic adjacency matrix, it is first normalized to prevent feature vanishing or explosion. The normalization process can be expressed as follows:

$$L = D^{-\frac{1}{2}}(D - M_d)D^{\frac{1}{2}} \tag{13}$$

Here, the matrix D denotes the degree matrix. Subsequently, the graph convolution operation is performed and can be formulated as follows:

$$\hat{y} = \sigma(Lx_pW_d) \tag{14}$$

The output $\hat{y} \in \mathbb{R}^{s^2\times c}$ obtained after the graph convolution operation integrates both spatial and semantic information, thereby enriching the feature representations. We then apply the output $\hat{y}$ of the sparse attention mechanism is then applied to perform channel-wise weighting within each window, resulting in a representation that integrates both spatial and channel-wise mixed weighting.

$$O = \hat{y} \odot O_1 \tag{15}$$

Finally, O represents the output of the module; $\odot$ denotes matrix element-wise multiplication.

It is worth noting that directly applying conventional graph convolution networks (GCNs) to image data is not straightforward. Unlike graph-structured data, images are represented as regular grid-structured tensors, where explicit adjacency relationships are not naturally defined. As a result, constructing meaningful graph connectivity for visual features remains a non-trivial problem.

In addition, the relationships among visual regions are often content-dependent and difficult to predefine using fixed adjacency structures. Modeling each pixel as an individual node would further introduce prohibitive computational complexity, especially for high-resolution inputs.

Therefore, a key consideration is how to incorporate graph-based relational modeling in a manner that is both computationally feasible and structurally meaningful. In this work, we build upon a region-level and sparsity-constrained strategy with task-oriented adaptations, providing a practical way to incorporate graph-based relational modeling into defect detection.

Dynamic scanning state space model

The Dynamic Scanning State Space Model (DMamba) is incorporated into the final stage of the backbone network. Similar to the SPPF module, its role is to enlarge the effective receptive field and enhance the modeling of global contextual information. Recent studies, such as S4ND[28], have shown that state space models provide a promising alternative for visual representation learning. In essence, the state space formulation describes the evolution of latent states conditioned on previous states and external inputs[29], making it suitable for modeling long-range dependencies.

In this work, the state space mechanism is introduced to complement convolution-based feature extraction by improving global context modeling in the later stage of the backbone. This design is particularly relevant for defect detection, where subtle targets may depend not only on local texture cues but also on broader structural context. Rather than relying solely on local receptive fields, the introduced module provides an additional pathway for capturing long-range interactions in a relatively efficient manner. More broadly, recent studies in sequential representation learning have suggested that attention-based sequence modeling is effective for capturing complex dependency structures, which supports the rationale of reorganizing visual features into ordered sequences for global modeling[30].

The detailed structure of the DMamba module is illustrated in Fig. 4. Since the Mamba module operates on sequential inputs, the feature map is first partitioned into patches and then flattened into a one-dimensional sequence. To preserve positional information during this transformation, a learnable positional embedding is introduced, which can be formulated as follows:

$$T = \mathrm{Flatten}(x) + \mathrm{P} \tag{16}$$

Here, $T \in \mathbb{R}^{L\times D}$, $\mathrm{P} \in \mathbb{R}^{L\times D}$ represent the feature map flattened into L sequences, each with a length of D, L denotes the number of patches and D denotes the number of pixels within each patch. The flattened features are subsequently arranged along four directions of the image, resulting in four spatially ordered sequences, which can be formulated as follows:

$$T^k = \mathrm{CSM}(T) \tag{17}$$

Here, k denotes k ordering methods, and CSM() represents the sorting function[31]. In addition, following the approach in[32], an additional random ordering method is introduced. The five sequences are subsequently fed in parallel into five spatial state estimation modules to perform global information modeling. The spatial state estimation process for a single module can be formulated as follows:

$$y = \phi_{ssm}(T) \tag{18}$$

Here, y represents the output of the spatial state estimation computation, while $\phi_{ssm}()$ denotes the multi-path dynamic activation spatial state estimation module, which consisting of N blocks at each stage. Since $\phi_{ssm}()$

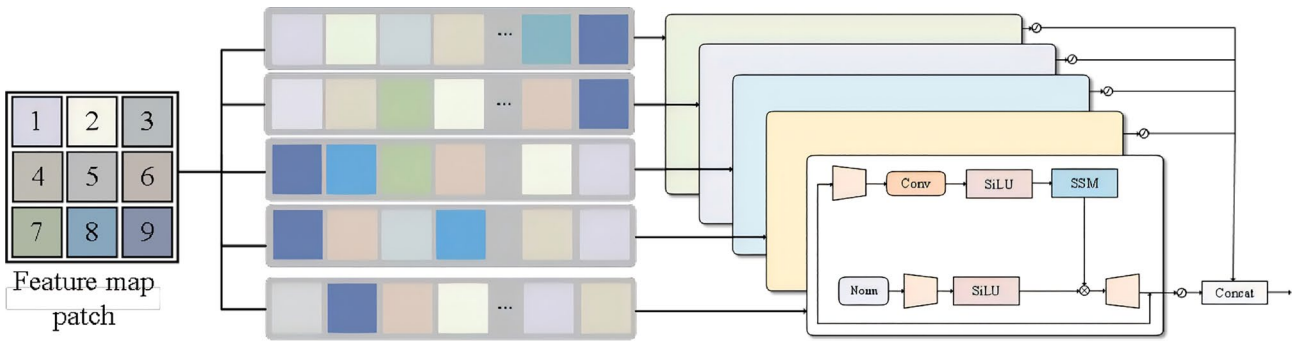

Figure 4. Detailed structure of the DMamba module (for clearer visualization, some fully connected layers and normalization operations within the Mamba module are omitted).

was originally designed for 1-D causal sequence modeling, it lacks the capability to capture spatial positional information, thereby limiting its expressive power for representing image features. To address this limitation, the sequences sorted along five different directions are input into five parallel spatial state estimation modules, which can be formulated as follows:

$$y^k = \phi_{ssm}^k(T^k) \tag{19}$$

Next, the outputs of the parallel modules are gated and aggregated to fuse the information from the five ordered sequences, producing the final output, which can be formulated as follows:

$$\begin{aligned} \hat{y} &= \sum_{k=0}^{4} y^k \cdot g_k, \\ g_k &= \mathrm{Softmax}\left(\phi_{\mathrm{pro}}\left(\phi_{\mathrm{mean}}\left(\phi_{\mathrm{concat}}(y^k)\right)\right)\right) \end{aligned} \tag{20}$$

Here, $\phi_{\mathrm{concat}}()$ denotes the concatenation of the outputs from the five parallel spatial state estimation modules. $\phi_{\mathrm{mean}}()$ denotes the mean pooling operation along the length dimension. $\phi_{\mathrm{pro}}()$ denotes the projection of the averaged features into a five-dimensional vector. The Softmax function serves as the activation mechanism and is applied to obtain the final normalized weights for each parallel module. In summary, DMamba feeds five groups of spatially reordered features into five parallel Mamba modules for multi-branch global modeling. It then estimates the importance of each spatial ordering and fuses the five reordered feature representations according to their respective significance, thereby capturing more comprehensive global information.

Hypergraph-inference-based feature fusion pyramid

The Hypergraph-Inference-Based Feature Fusion Pyramid (HG-FFP) is designed as an enhanced feature fusion module based on the neck structure of RT-DETR, as illustrated in Fig. 5. Specifically, three modifications are incorporated: (1) a bidirectional fusion pathway, where the top-down branch enhances high-level semantic information while the bottom-up branch preserves low-level spatial details; (2) a hypergraph-based inference module integrated at the high-level feature stage to model higher-order relationships among features and provide structured semantic guidance during feature propagation; and (3) the introduction of a C2f module with SCConv-based self-calibrated convolution at fusion nodes to refine feature responses and facilitate more effective multi-scale feature integration.

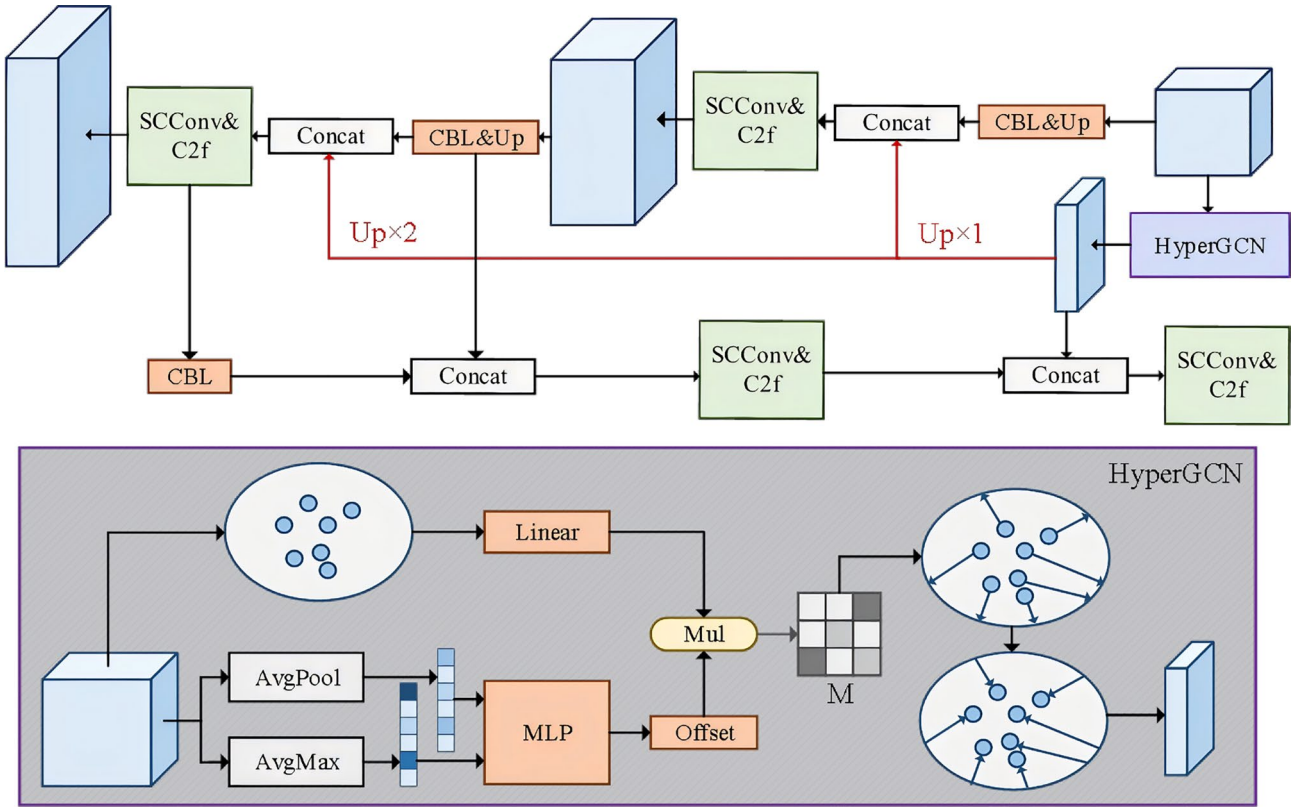

Figure 5. Overall structure of the HG-FPN.

Through these modifications, the feature fusion process is made more structured and adaptive, which is beneficial for handling complex visual patterns in defect detection scenarios.

Hypergraph convolution

Recent studies have explored hypergraph-based relational modeling in visual recognition and detection tasks, showing its potential for capturing high-order interactions that are difficult to represent with conventional convolution or pairwise attention mechanisms[33]. CNNs are effective at modeling local patterns within limited receptive fields, while attention mechanisms can capture long-range dependencies but may still be insufficient for representing complex many-to-many relationships among feature regions. In contrast, hypergraph-based modeling provides a more flexible way to characterize higher-order associations between objects and contextual information.

Therefore, HyperGCN is introduced as a semantic guidance module in our model, as illustrated by the blue components in Fig. 5. For low-level semantic features, each pixel in the feature map corresponds to a grid point in the original image, representing a specific spatial region. Accordingly, each pixel is treated as a node to model the correlations among regions and objects. First, the feature map is reshaped by flattening its spatial dimensions, denoted as $x \in \mathbb{R}^{N \times C}$, where N represents the total number of pixels. Then, both max pooling and average pooling are performed along the N dimension to obtain a globally aggregated node feature vector, denoted as $z \in \mathbb{R}^{2C}$. Subsequently, a linear projection is applied to derive the node offsets, represented as ΔP. Meanwhile, a trainable parameter is randomly initialized to obtain the global hyperedge prototype P_{global}. By adding the node offsets to this prototype, the hyperedge matrix is generated, which effectively models the latent visual correlations among regions and can be formulated as follows:

$$P = \Delta P + P_{\text{global}}, \quad \Delta P = z_{*c}, \quad P \in \mathbb{R}^{M \times C} \tag{21}$$

Here, *c represents the project operator; M denotes the number of hyperedges. The contribution of each node is computed through an inner product operation similar to that employed in attention mechanisms. Specifically, the node features x are first projected, followed by an inner product with P to generate the final hyperedge matrix, which can be formulated as follows:

$$A = \text{Softmax}\left(\frac{x_{*c} \cdot P^{\mathrm{T}}}{\sqrt{h}}\right) \tag{22}$$

Here, A denotes the hyperedge mapping matrix, which serves as the basis for performing hypergraph convolution. Specifically, unlike conventional graph convolution, hypergraph convolution first compresses node features into hyperedges, formulated as follows:

$$x_e = W_e D_e^{-1} A^{\mathrm{T}} x_v \tag{23}$$

Here, D_e^{-1} denotes the degree matrix of hyperedges, x_v denotes the set of nodes, where each node corresponds to a pixel in the feature representation, and W_e denotes the projection parameter. The above process can be interpreted as a feature compression operation that captures relational dependencies through the hyperedge matrix. Subsequently, the hyperedge features are projected back into the node space, formulated as follows:

$$\hat{x}_v = W_v D_v^{-1} A x_e \tag{24}$$

The above formulation projects the hyperedge features back into the node domain, thereby constructing relational representations that effectively strengthen semantic associations among regions.

Self-calibrating convolution

The detailed architecture of the Self-Calibrated Convolution (SCConv) module[34] is illustrated in Fig. 6. Conventional convolutional operations tend to learn highly similar feature patterns, leading to limited feature discriminability. Moreover, the fixed receptive field of each convolutional layer constrains its ability to capture contextual information. Therefore, SCConv is introduced to perform neighborhood feature fusion, facilitating the effective integration of fine-grained details and semantic information.

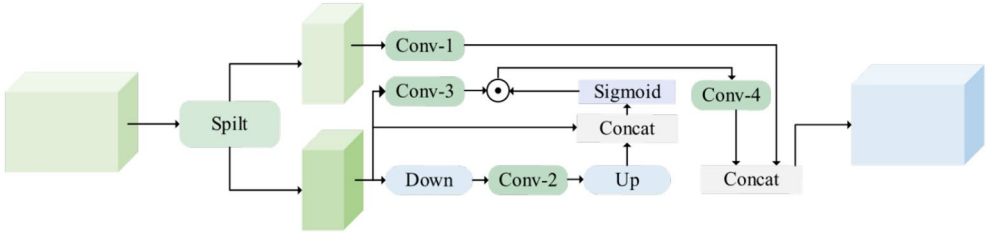

Figure 6. Architecture of the SCConv module.

Specifically, the input x is equally divided into two parts along the channel dimension. These two parts are processed by two separate branches to model local information with a small receptive field and contextual information with a large receptive field, respectively. The first branch further consists of two sub-branches. The first sub-branch downsamples the original-resolution features to expand the receptive field and extract contextual information, formulated as follows:

$$\hat{x}_1 = \mathrm{Up}\big(\mathrm{DWConv}(\mathrm{Augpool}(x_1))\big) \tag{25}$$

Here, $\mathrm{Augpool}()$ denotes the pooling operator for downsampling, $\mathrm{DWConv}()$ represents a depthwise separable convolution with a kernel size of 3 that captures contextual information from low-resolution features, and $\mathrm{Up}()$ denotes the bilinear interpolation operator used to restore the original spatial resolution. Subsequently, the self-calibration operation is performed as follows:

$$y_1 = \mathrm{DWConv}[\mathrm{DWConv}(x_1) \cdot \sigma(\hat{x}_1 + x_1)] \tag{26}$$

Here, $\sigma()$ denotes the Sigmoid activation function, and x_1 serves as a residual connection to generate the calibrated weights. Finally, a calibrated convolutional mapping is applied to produce the final output of the first branch.

In the second branch, only a simple convolution operation is applied to produce the output $y_2 = \mathrm{DWConv}(x_2)$, aiming to preserve the original spatial information. Finally, the two outputs $[y_1, y_2]$, are concatenated to form the final output of the module.

Experiment

Construction of the dataset

In this study, a publicly available railway track image dataset was utilized, comprising 2,324 defect images categorized into eight defect types. The number of annotations and the corresponding label information for each category are presented in Fig. 7.

From the label distribution patterns, it can be observed that the targets in this dataset exhibit a distinct concentration in both spatial location and geometric shape. The pairwise relation map shows that target centers are relatively dispersed horizontally but clearly concentrated in the middle region vertically. Most targets extend vertically across a large portion of the image height while maintaining a narrow width, resulting in a slender appearance. The bar chart further indicates that the category distribution is somewhat imbalanced, with most samples concentrated in a few dominant classes. Overall, the dataset is characterized by vertically elongated targets with concentrated spatial positions and a head-heavy class distribution, suggesting that the model should focus on enhancing its detection capability for slender targets during training. This also highlights the importance of robustness to distribution variation, which is closely related to the broader problem of transferable out-of-distribution generalization[35]. To improve the model's generalization ability for objects appearing at different positions, horizontal shifting, random cropping, and Mosaic augmentation are employed.

As shown in Fig. 8, the railway defect data exhibit high inter-class similarity and substantial intra-class variability, while variations in viewing angles further influence the visual appearance of defects. Therefore, it is

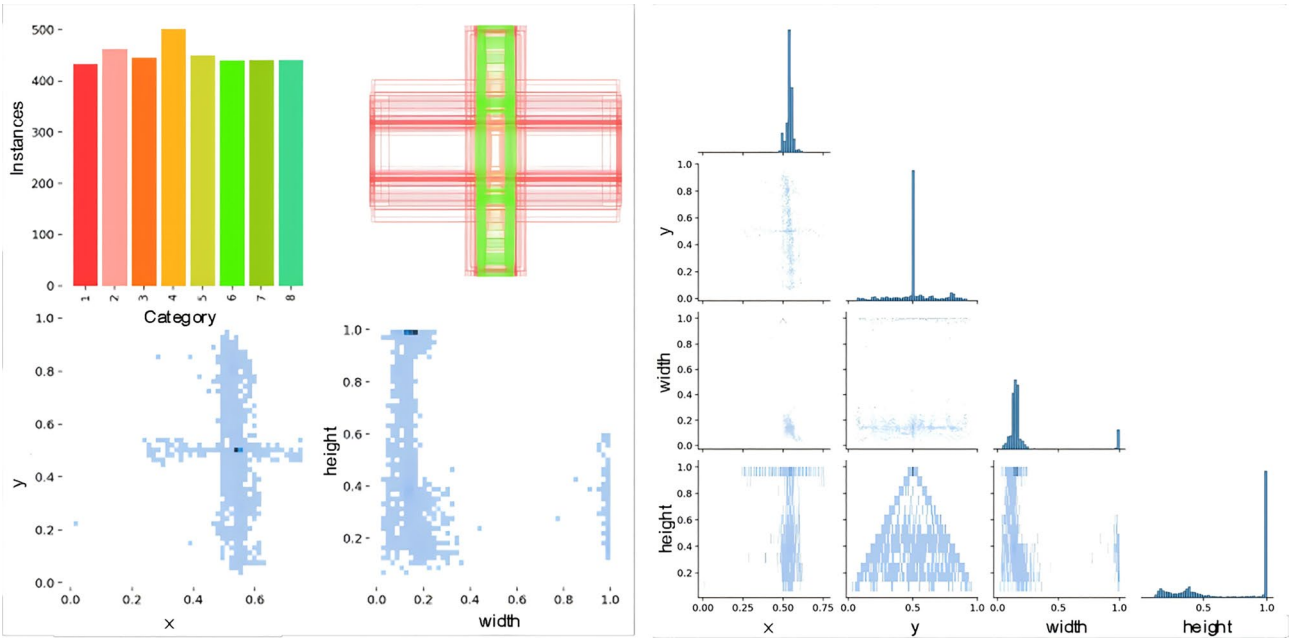

Figure 7. Distribution of labels in the dataset.

Figure 8. Detailed visualization of the dataset.

Hyperparameter	Setting
Batch size	32
Number of epochs	300
Initial learning rate	0.01
Optimizer	SGD
Input image resolution	640 × 640

Table 1. Hyperparameter settings.

crucial for the model to exhibit strong generalization ability to ensure robust performance and enable practical engineering deployment.

Hyperparameter settings

The experiments were conducted on a workstation equipped with an Intel i7-11700K CPU, an NVIDIA GTX 4090 GPU, and 64 GB of RAM. From the constructed dataset, 80% of the samples were used for training, 10% for testing, and 10% for validation. A stratified random sampling strategy was adopted to ensure that the class distribution in the training and testing sets remained consistent with that of the original dataset. Under this experimental setup, the model was trained, and the detailed hyperparameter configurations are summarized in Table 1.

Experimental results and analysis

Training curve analysis

To verify the training stability and convergence performance of the proposed model, a comparative analysis was conducted between the baseline and the improved models. As shown in Fig. 9, both models exhibit a steadily decreasing trend in training loss as the number of epochs increases. The improved model converges more rapidly and demonstrates smaller fluctuations in later stages, indicating higher fitting efficiency. Meanwhile, the validation loss follows a consistent convergence trend with the training loss and eventually stabilizes, suggesting that the improved model achieves enhanced training performance while maintaining strong generalization ability. Furthermore, examination of the evaluation metrics reveals that all indicators increase steadily and ultimately converge as training progresses. Overall, the improved model achieves consistently higher performance curves across different stages, confirming its effectiveness in enhancing detection accuracy.

As illustrated in Fig. 10, the proposed improved model achieves a better overall Precision-Recall trade-off than the baseline model. The PR curves show that our method maintains higher precision and recall for most

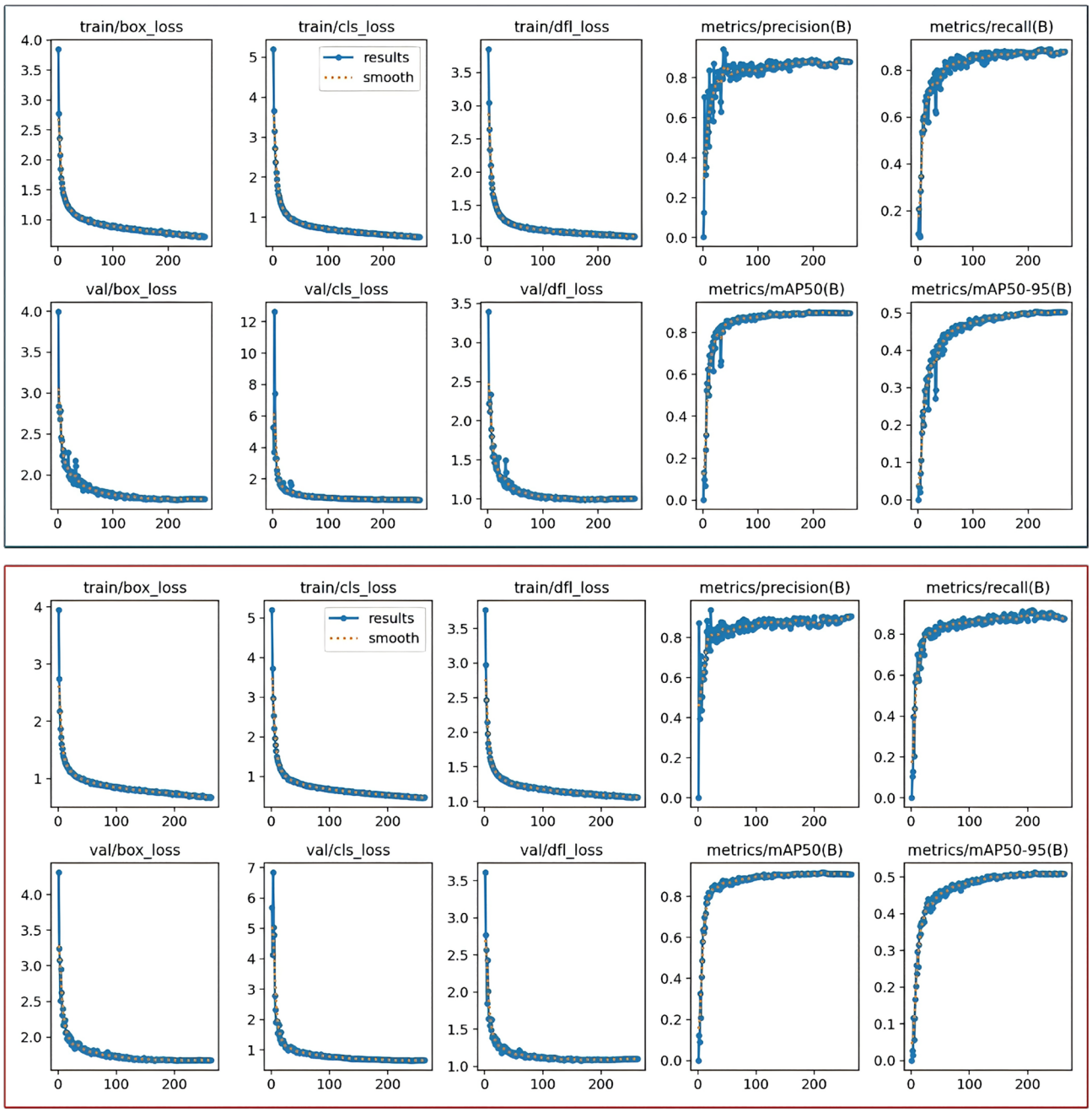

Figure 9. Convergence curves of the baseline model and the improved model (the red box represents the baseline model RT-DETR, while the blue box represents our improved model).

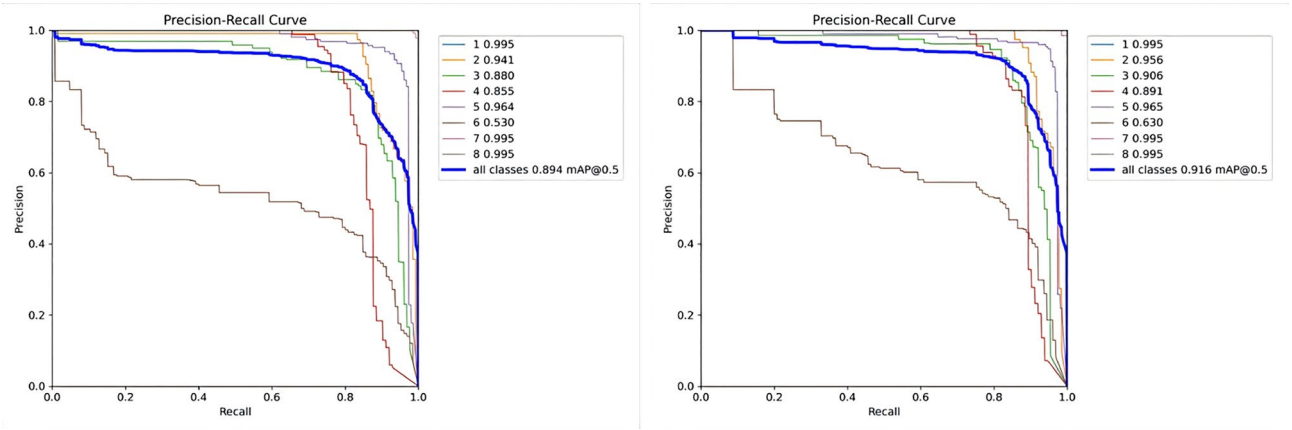

Figure 10. mAP curves of the model for each object category (the left figure represents the baseline model RT-DETR, while the right figure represents our improved model).

defect categories, and the overall mAP@0.5 is also improved, indicating stronger detection capability for railway defects under complex conditions.

As shown in Fig. 11, the Recall-Confidence curve and the Precision-Confidence curve indicate that, compared with the baseline model RT-DETR, the proposed model demonstrates higher stability and broader applicability. As the confidence threshold increases, the recall of the proposed model remains at a relatively high level for a longer range before gradually declining, whereas the baseline model shows a relatively faster downward trend. Meanwhile, the precision curve of the proposed model consistently maintains an advantage over the baseline across the entire confidence range, further confirming its effectiveness.

Meanwhile, as shown in Fig. 12, the confusion matrix indicates that the proposed method achieves comparable or superior classification accuracy across all categories compared to the baseline approach.

Overall, these results demonstrate that the topology-enhanced model achieves more stable and reliable detection performance in complex railway defect scenarios.

Visualization of experimental results

As shown in Fig. 13, the visualization of the model's detection results demonstrates that the proposed model performs well in identifying clear and visually distinct defects, achieving high confidence scores with an average confidence above 0.5. This indicates that the model is effective in handling well-defined targets. However, for complex or blurred defect regions, the detection accuracy still requires further optimization to enhance robustness and precision.

(a) Visualization and validation of the SAGFE module As shown in Fig. 14, the feature maps extracted by the SAGFE module are more prominent and exhibit broader feature coverage, demonstrating its superior capability in feature extraction.

(b) Visualization and validation of the DMamba module As shown in Fig. 15, the output feature maps of the DMamba and SPPF modules are visualized to analyze their roles in multi-level feature extraction. The right images display the output feature maps of the DMamba module, where the responses are concentrated in localized regions and the distribution of strong activation areas is relatively sparse. This indicates that the DMamba module, through dynamic spatial state modeling, can more precisely capture fine-grained and subtle target features. Such capability makes it particularly suitable for tasks sensitive to local information, such as small object detection and fine detail extraction.

In contrast, the SPPF module produces feature maps with more uniformly distributed responses, where multiple regions exhibit strong activations. This suggests that the SPPF module, leveraging multi-scale pooling, effectively captures features at different scales and integrates hierarchical information. Such feature representations are advantageous for large object detection or scenes with complex backgrounds, enhancing the model's ability to perceive and process global contextual information.

Through comparison, it can be observed that the DMamba module focuses more on capturing local details and fine-grained features—characteristics well aligned with the nature of railway defect detection—whereas the SPPF module emphasizes multi-scale global feature fusion, making it more suitable for general-purpose vision tasks.

(c) Visualization and Validation of the HyperGCN Module As shown in Fig. 16, several feature maps generated by the HyperGCN module are visualized. The results indicate that HyperGCN is not limited to pixel-level or locally receptive convolutional representations; instead, it models high-order relational patterns by mapping the latent correlations between different regions through hyperedge structures. In these feature maps, bright regions correspond to strong correlation responses between nodes, while dark regions indicate weak or insignificant connections. This relational modeling enables the network to connect semantically related yet spatially distant regions, forming a non-Euclidean structured feature representation.

Compared with conventional convolutional features, the output of HyperGCN exhibits a distinct clustering pattern: some channels show multiple bright, clustered regions, implying that the module captures high-order

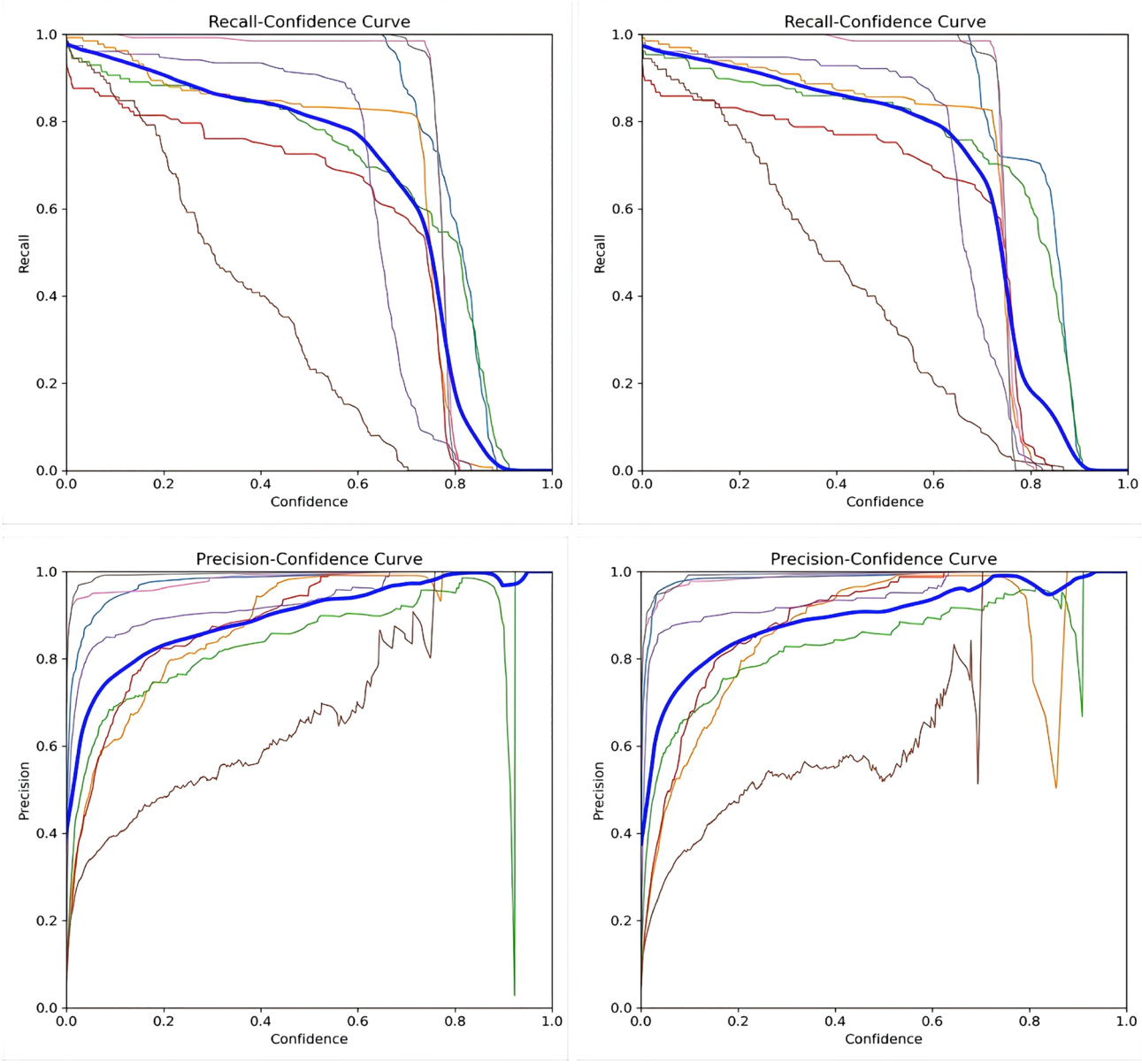

Figure 11. Precision-confidence curve and recall-confidence curve (the right figure represents the baseline model RT-DETR, while the left figure represents our improved model).

dependencies among different object parts or semantic regions; other channels display sparse activations, emphasizing selective associations between key areas. Overall, the relational patterns learned by HyperGCN not only enhance the discriminative power of local features but also integrate cross-regional contextual information into node representations through hypergraph reasoning, effectively improving the model's robustness in complex detection scenarios.

(d) Visualization and Validation of the SCConv Module As shown in Fig. 17, the SCConv module is employed to integrate neighborhood features and perform guided refinement based on the relational features output by the HyperGCN module, thereby obtaining high-level representations with strong attention to defect regions. It is evident that the resulting features exhibit a highly distinct separation between foreground and background areas, clearly demonstrating the effectiveness of the proposed approach.

Ablation study and comparative study

The ablation study aims to systematically remove or replace key components of the model to analyze the specific contribution of each part to the overall performance. This approach provides an intuitive validation of the model's design rationality and necessity, while highlighting the importance of individual modules within the architecture.

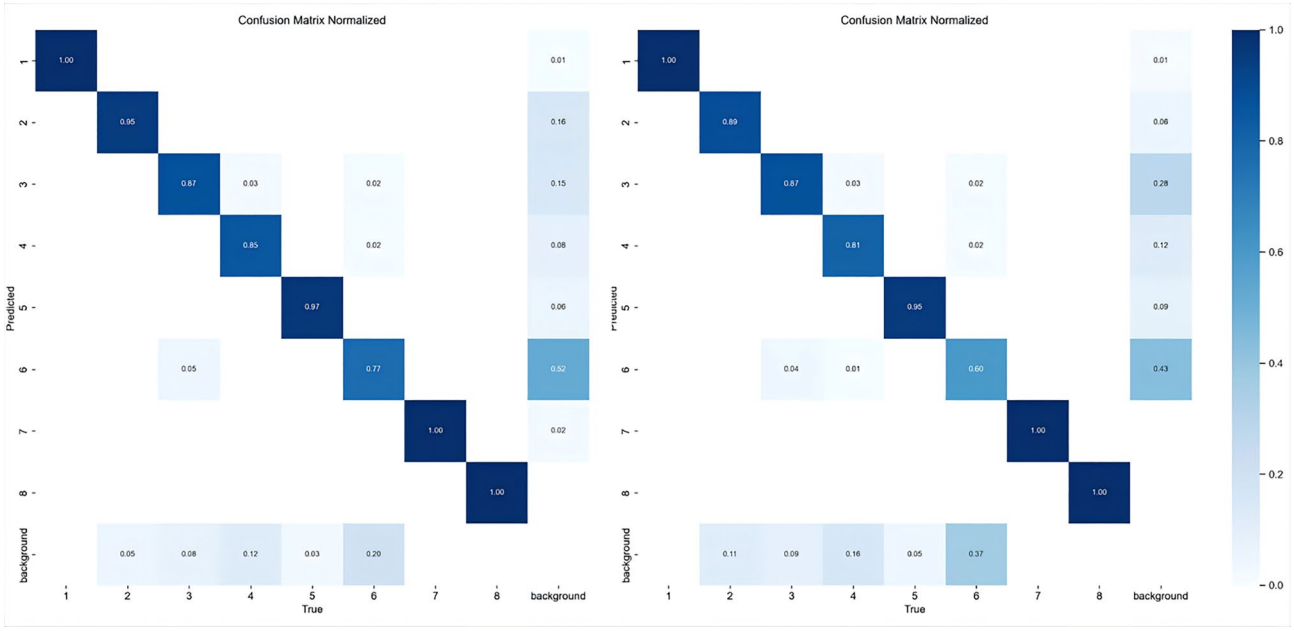

Figure 12. Confusion matrix (the right figure represents the baseline model RT-DETR, while the left figure represents our improved model).

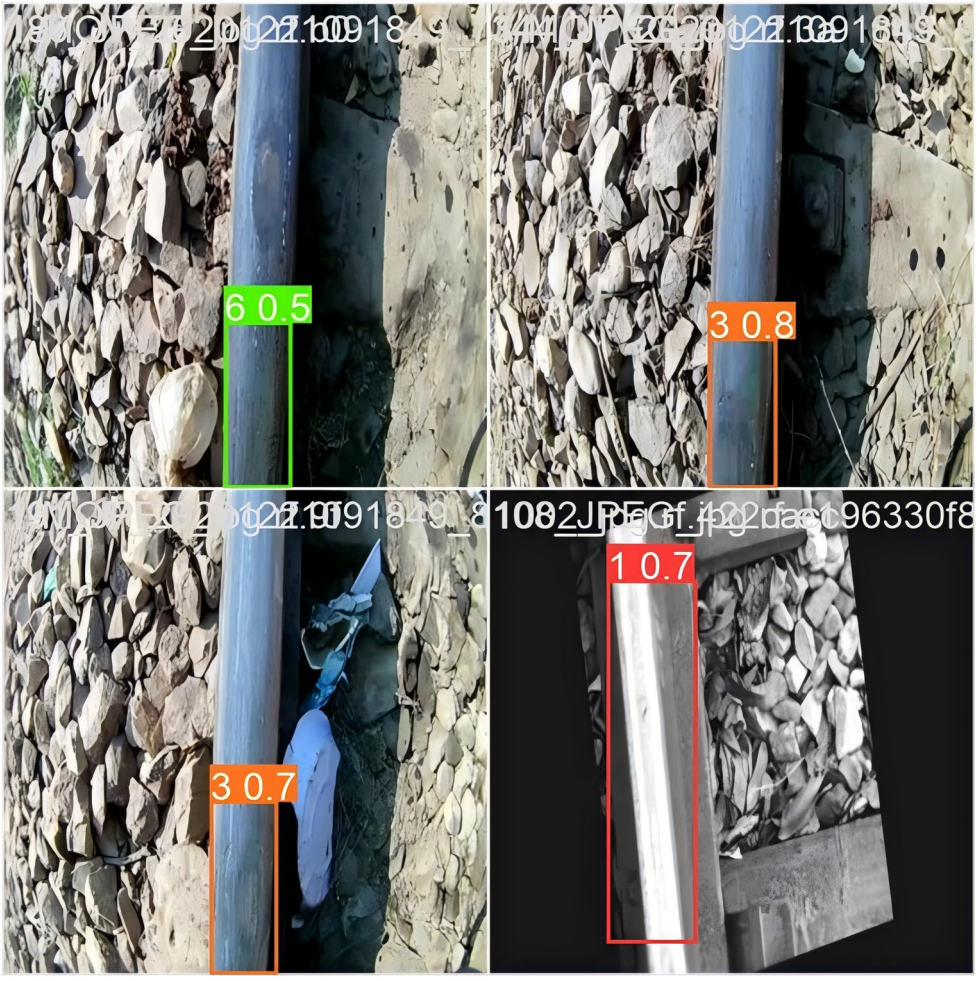

Figure 13. Model's detection results.

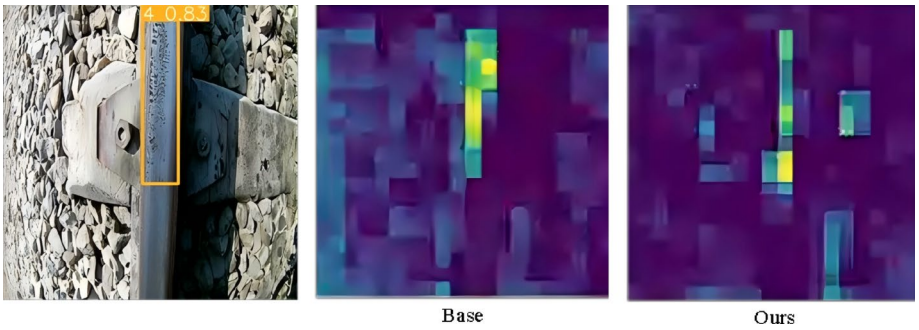

Figure 14. Visualization of feature maps from the backbone network.

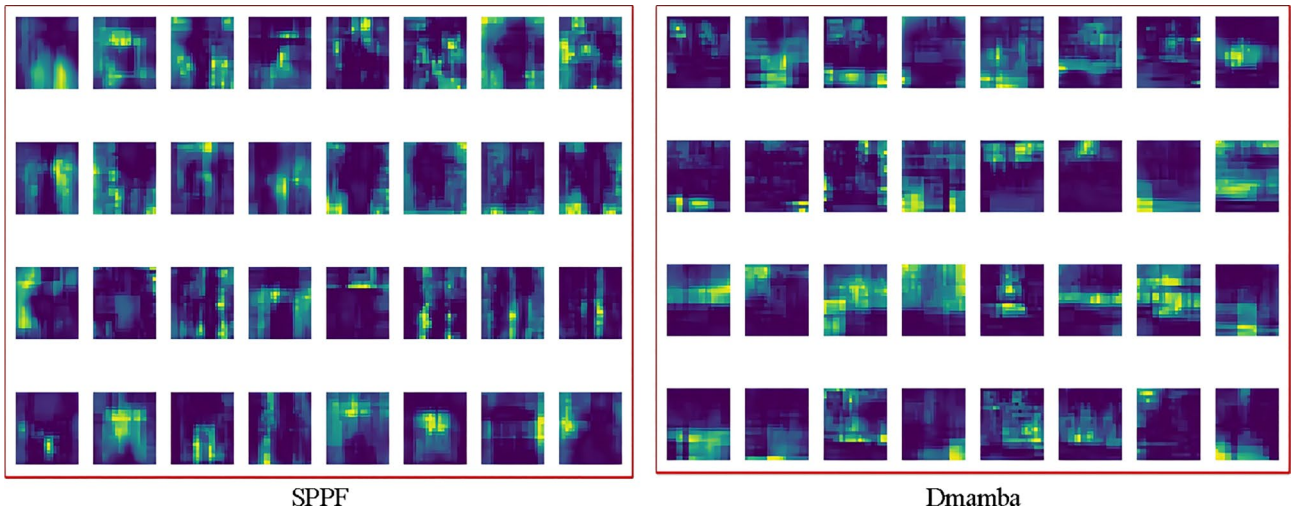

Figure 15. Visualization of feature maps from the DMamba module.

(a) Ablation Study on the Backbone Network The backbone network of our model is an improved version of CSPDarknet, which is derived from the ResNet architecture. In this design, the SAGFE module is used to replace the original C2f module, while the DMamba module is integrated into the final layer of the backbone network.

As shown in Table 2, an ablation study was conducted on the two key modules proposed in this work—SAGFE and DMamba. When only the SAGFE module was incorporated into the baseline model, the number of parameters significantly decreased (from 16.7M to 13.2M), while both recall and precision improved. This demonstrates that SAGFE enhances feature extraction capability while simultaneously reducing model complexity. After integrating the DMamba module, the model achieved a further improvement in mAP (from 0.894 to 0.897), validating its effectiveness in global contextual modeling. When both modules were applied together, the model achieved the best overall performance, with P = 0.870, R = 0.892, and mAP = 0.906. These results indicate that local spatial-semantic modeling and global context modeling are highly complementary, jointly contributing to superior detection performance of railway defects under complex environmental conditions.

(b) Comparative Study on the Backbone Network To further validate the advantages of the proposed model, we conducted comparative experiments with several mainstream backbone networks. The comparison includes convolution-based architectures such as DenseNet and EfficientNet, as well as attention-based architectures such as Swin Transformer and CoT Transformer. In addition, the proposed method was also compared with the latest vision state-space model, Vision Mamba, to comprehensively demonstrate the superiority of our approach.

As shown in Table 3, in the comparative experiments on backbone networks, both DenseNet and EfficientNet achieved lower performance than Transformer-based and state-space-based models, indicating that convolutional architectures alone are insufficient for effectively modeling complex global semantic relationships. Among the Transformer-based networks, Swin Transformer and CoT Transformer demonstrated strong global modeling capabilities, with CoT Transformer achieving an mAP of 0.900, highlighting the advantage of its collaborative attention mechanism. Vision Mamba exhibited stable performance in both precision and recall, confirming the potential of state-space models in visual tasks.

In contrast, the proposed method achieved the best results across all three metrics: Precision (P) = 0.870, Recall (R) = 0.892, and mAP = 0.906, further improving detection accuracy. These results demonstrate that the

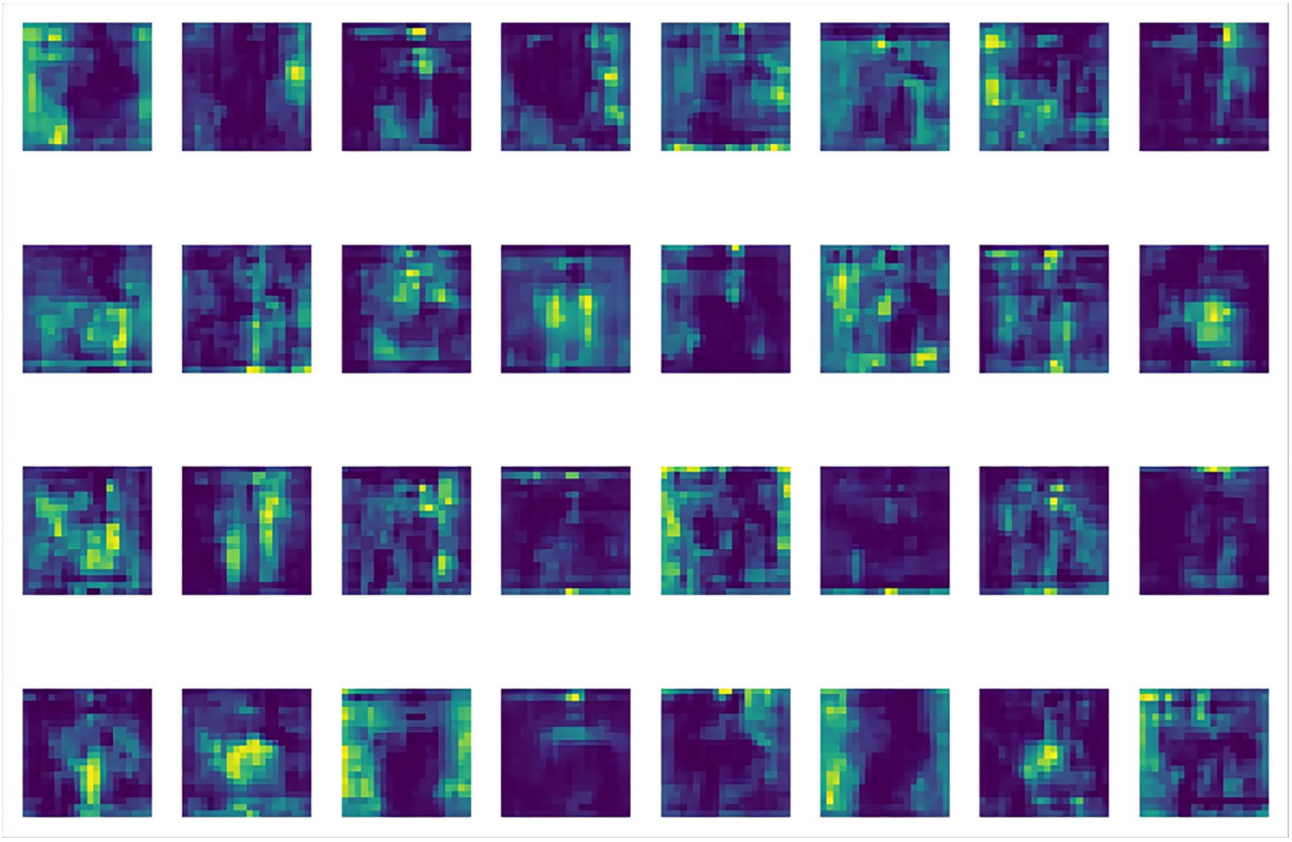

Figure 16. Relational pattern features extracted by the HyperGCN module.

Figure 17. Relational pattern features integrated by the SCConv module.

SAGFE	DMamba	P	R	mAP	Para (M)
		0.864	0.884	0.894	16.7
✓		0.868	0.889	0.890	13.2
	✓	0.867	0.888	0.897	18.1
✓	✓	0.870	0.892	0.906	15.6

Table 2. Ablation study on the backbone network.

BackBone	P	R	mAP
DenseNet	0.851	0.862	0.871
EfficientNet	0.867	0.877	0.893
Swin transformer	0.875	0.881	0.895
CoT transformer	0.857	0.887	0.900
Vision mamba	0.865	0.871	0.899
Ours	**0.870**	**0.892**	**0.906**

Table 3. Comparative study on the backbone network.

HyperGCN	SCConv	P	R	mAP	Para (M)
		0.864	0.884	0.894	16.7
✓		0.867	0.892	0.898	18.1
	✓	0.871	0.889	0.894	17.9
✓	✓	0.874	0.892	0.910	19.3

Table 4. Ablation study on the neck network.

Neck	P	R	mAP
FPN	0.848	0.863	0.877
PANet	0.864	0.884	0.894
BiFPN	0.866	0.880	0.895
GraphFPN	0.845	0.831	0.847
Ours	**0.872**	**0.892**	**0.910**

Table 5. Comparative study on the neck network.

proposed spatial-semantic relational modeling strategy effectively integrates local and global features, leading to superior detection performance in complex railway defect scenarios.

(c) Ablation Study on the Neck Network

The neck network of our model is built upon the PANet architecture. It primarily employs the HyperGCN module to extract guided relational features and utilizes the SCConv module to achieve adaptive calibration and feature fusion.

As shown in Table 4, the ablation results of the neck network demonstrate that the proposed modules contribute significant improvements to model performance. The SAGFE module in the backbone enhances local spatial-semantic representation while reducing parameters, yielding more discriminative features. The DMamba module effectively captures global contextual dependencies, improving detection robustness under complex backgrounds. In the neck structure, the HG-FPN integrates hypergraph reasoning and self-calibrated convolution to strengthen cross-scale semantic interactions and optimize feature fusion efficiency. When combined, these modules deliver consistent gains in precision, recall, and mAP, confirming their complementarity in local modeling, global context learning, and multi-scale feature integration.

(d) Comparative Study on the Neck Network To further validate the superiority of the proposed model, we conducted comparative experiments with several mainstream neck network architectures, including the conventional FPN, PANet, the bidirectional fusion network BiFPN, the high-order interaction-based HorNeck, and the graph reasoning-based GraphFPN.

As shown in Table 5, the comparative experiments on the neck network indicate that the proposed method outperforms mainstream multi-scale feature fusion structures such as FPN, PANet and BiFPN, achieving up to a 1.5% improvement in mAP. It also shows clear superiority over GraphFPN, further demonstrating the advantages of hypergraph reasoning and self-calibrated convolution in complex visual scenarios.

(e) Overall Ablation Study As shown in Table 6, the overall ablation results further validate the effectiveness of the proposed method. Independently improving either the backbone or the neck network leads to a significant performance gain, among which the HG-FPN demonstrates particularly notable improvements in multi-scale feature fusion, achieving an mAP of 0.910. When both enhanced components are applied jointly, the model attains the best overall performance, with Precision (P), Recall (R), and mAP all reaching their highest values, and the mAP increases to 0.916, representing a 2.2% improvement over the baseline.

(f) Robustness Evaluation Under Complex Interference Conditions To evaluate the robustness of the proposed method, we introduce various interference conditions into the test set through image processing techniques, constructing an augmented dataset that better simulates real-world railway inspection scenarios, including extreme illumination and severe surface contamination.

BackBone	Neck	P	R	mAP	Para (M)
		0.864	0.884	0.894	16.7
✓		0.870	0.892	0.906	15.6
	✓	0.874	0.892	0.910	19.3
✓	✓	0.878	0.898	0.916	18.2

Table 6. Overall ablation results.

Method	P	R	mAP
Standard test set	0.878	0.898	0.916
Extreme illumination	0.862	0.841	0.872
Noise	0.824	0.871	0.826

Table 7. Robustness evaluation results under different interference conditions.

BackBone	P	R	mAP	Flops	FPS
YOLO-V5	0.872	0.857	0.864	109G	190
YOLO-V8	0.876	0.894	0.910	165.2G	170
YOLO-V11	0.861	0.876	0.888	86.9G	230
DETR	0.855	0.807	0.823	187G	32
Dino-DETR	0.872	0.892	0.910	279G	13
RT-DETR	0.864	0.884	0.894	136G	185
Ours	**0.878**	**0.898**	**0.916**	**152G**	**56**

Table 8. Overall comparative results.

As shown in Table 7. Under extreme illumination conditions, the overall performance shows a noticeable decline, with mAP decreasing from 0.916 to 0.872. This degradation can be attributed to the reduced local contrast and the distortion of fine structural details, which makes it more difficult for the model to detect small-scale or low-contrast defects, leading to missed detections.

In contrast, the interference simulated by local noise, representing oil stains or environmental contamination, has a more pronounced impact on performance, reducing mAP from 0.916 to 0.826. This is because such noise may exhibit similar texture and appearance to certain defect patterns, thereby introducing false positives. In addition, the occlusion caused by noise can weaken the feature representation of defects, further affecting detection performance.

Although the proposed method demonstrates strong performance on the standard test set, these results indicate that vision-based defect detection methods still face challenges in complex railway environments, particularly in distinguishing true defects from visually similar interference. On the other hand, considering that practical railway inspection is typically based on continuous image sequences, such information loss can be partially mitigated through temporal consistency.

In future work, we will explore illumination-invariant feature learning and the integration of multi-modal information to further improve robustness under complex conditions.

(g) Overall Comparative Study

As shown in Table 8, compared with these representative methods, the proposed model achieves the best performance across all three evaluation metrics—P = 0.878, R = 0.898, and mAP = 0.916—exceeding the best-performing YOLOv8 and DINO-DETR by 0.6 percentage points in mAP. This demonstrates that, benefiting from the synergistic integration of local semantic modeling, global contextual reasoning, and multi-scale feature fusion, our method exhibits superior detection capability and robustness in complex scenarios.

Overall, YOLO series achieve better speed—efficiency trade-offs, while DETR-based methods provide stronger global modeling but suffer from high computational cost and low FPS. RT-DETR improves this limitation with balanced performance. Compared with all baselines, our method achieves the best accuracy (mAP=0.916) with moderate FLOPs (152G), but at the cost of reduced speed (56FPS), indicating a trade-off favoring detection performance over real-time efficiency.

(h) Analysis of Limitations Despite the encouraging performance, the proposed method still has several limitations. First, the introduction of relational modeling and state-space components increases the architectural complexity to some extent. Although the additional computational overhead is controlled through sparse and region-level designs, the overall inference efficiency remains slightly lower than that of the baseline model. This may limit its applicability in scenarios with strict real-time constraints.

Second, the proposed method relies primarily on visual appearance for defect detection. As observed in the experimental analysis, its performance degrades under challenging conditions such as strong illumination and

heavy surface contamination. In particular, visually similar interference patterns may still lead to false positives, indicating that the current feature representation is not fully robust to complex environmental variations.

In future work, we will focus on improving computational efficiency, enhancing robustness under complex conditions, and exploring more adaptive mechanisms for parameter selection.

Conclusion

This paper proposes a rail defect detection method based on topology-enhanced relational modeling, which performs feature representation and fusion around local, global, and high-order semantic associations. First, the dynamic multi-order feature extraction module based on a semantic association graph effectively constructs a regional relational topology and employs sparse attention and graph convolution to extract features, thereby accurately capturing the semantic characteristics of defects. Second, the multi-directional random cross-scanning state-space modeling enables the extraction of global information under various spatial orderings, effectively enhancing the discriminability between defect and background features. Finally, the hypergraph-guided feature fusion pyramid effectively captures high-order dependency relationships among targets, strengthening semantic associations across multiple objects. Experimental results on a public railway defect dataset demonstrate that the proposed model achieves strong performance and robustness, verifying its effectiveness and generalization capability.

It should be noted that the model still has room for improvement in terms of lightweight design and deployment efficiency. Future work will focus on further optimizing computational efficiency across different application scenarios and developing more powerful inference mechanisms to handle severe occlusion and highly dynamic environments.

Data availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Received: 12 November 2025; Accepted: 14 April 2026
Published online: 28 April 2026

References

1. Zhao, Y. et al. A review on rail defect detection systems based on wireless sensors. *Sensors* **22**(17), 6409. https://doi.org/10.3390/s22176409 (2022).
2. Taştimur, C., Karaköse, M., Akın, E., & Aydın, İ. Rail defect detection with real time image processing technique. In *2016 IEEE 14th International Conference on Industrial Informatics (INDIN)*, pp. 411–415 (2016). https://doi.org/10.1109/INDIN.2016.7819208 . IEEE
3. Sresakoolchai, J. & Kaewunruen, S. Railway defect detection based on track geometry using supervised and unsupervised machine learning. *Struct. Health Monit.* **21**(4), 1757–1767. https://doi.org/10.1177/14759217211027085 (2022).
4. Kumar, A. & Harsha, S. P. A systematic literature review of defect detection in railways using machine vision-based inspection methods. *Int. J. Transp. Sci. Technol.* **18**, 207–226. https://doi.org/10.1016/j.ijtst.2024.06.003 (2025).
5. Deutschl, E., Gasser, C., & Niel, A., et al. Defect detection on rail surfaces by a vision based system. In *IEEE Intelligent Vehicles Symposium*, 2004, pp. 507–511. IEEE, Parma, Italy. https://doi.org/10.1109/IVS.2004.1336399 (2004).
6. Xu, Q., Zhao, Q., & Yu, G., et al. Rail defect detection method based on recurrent neural network. In: 2020 39th Chinese Control Conference (CCC), pp. 6486–6490. IEEE, Shenyang, China. https://doi.org/10.23919/CCC50068.2020.9189377 (2020).
7. Mi, Z., Chen, R. & Zhao, S. Research on steel rail surface defects detection based on improved yolov4 network. *Front. Neurorobot.* **17**, 1119896. https://doi.org/10.3389/fnbot.2023.1119896 (2023).
8. Wei, X. et al. Railway track fastener defect detection based on image processing and deep learning techniques: a comparative study. *Eng. Appl. Artif. Intell.* **80**, 66–81. https://doi.org/10.1016/j.engappai.2019.01.001 (2019).
9. Ma, H., Sun, Z., Su, Y., Wang, H., Li, S., Yu, Z., Kang, Y., & Xu, H. Cross-modal two-stream target focused network for video anomaly detection. In *International Conference in Communications, Signal Processing, and Systems*, pp. 69–78. Springer (2022).
10. Girshick, R. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169 (2015).
11. Jiang, P. et al. A review of yolo algorithm developments. *Proc. Comput. Sci.* **199**, 1066–1073. https://doi.org/10.1016/j.procs.2022.01.135 (2022).
12. Liu, W., Anguelov, D., & Erhan, D., et al. Ssd: single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 21–37. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-46448-0_2 (2016).
13. Li, Z. et al. Development and challenges of object detection: A survey. *Neurocomputing* **598**, 128102 (2024).
14. Wang, A. et al. Yolov10: Real-time end-to-end object detection. *Adv. Neural. Inf. Process. Syst.* **37**, 107984–108011 (2024).
15. Wu, Y., Qiang, F., & Zhou, W., et al. Pfcnet: enhancing rail surface defect detection with pixel-aware frequency conversion networks. *IEEE Signal Process. Lett.* (2025).
16. Gibert, X., Patel, V. M. & Chellappa, R. Deep multitask learning for railway track inspection. *IEEE Trans. Intell. Transp. Syst.* **18**(1), 153–164. https://doi.org/10.1109/TITS.2016.2568756 (2016).
17. Zhang, D. et al. Two deep learning networks for rail surface defect inspection of limited samples with line-level label. *IEEE Trans. Industr. Inf.* **17**(10), 6731–6741. https://doi.org/10.1109/TII.2020.2984314 (2020).
18. Zhou, X. et al. Dense attention-guided cascaded network for salient object detection of strip steel surface defects. *IEEE Trans. Instrum. Meas.* **71**, 1–14. https://doi.org/10.1109/TIM.2021.3104051 (2021).
19. Niu, M. et al. Unsupervised saliency detection of rail surface defects using stereoscopic images. *IEEE Trans. Industr. Inf.* **17**(3), 2271–2281. https://doi.org/10.1109/TII.2020.2991634 (2020).
20. Gao, Y., Cao, Z., Qin, Y., Lian, L., Kou, L., & Wang, Y. Segmented-aggregated framework with internal-external constraint mechanism for unsupervised track anomaly detection. *IEEE Trans. Instrum. Meas.* (2025).
21. Ge, X. et al. An anomaly detection method for railway track using semisupervised learning and vision-lidar decision fusion. *IEEE Trans. Instrum. Meas.* **73**, 1–15 (2024).
22. Yu, J., Huang, J., Zhong, W., Jiang, Q., Yan, X., & Yang, X. Multipeeling of homogeneous stationarity and heterogeneous nonstationarity with differentiated learning for process monitoring. *IEEE Trans. Cybern.* (2025).
23. Yu, J. & Yan, X. Whole process monitoring based on unstable neuron output information in hidden layers of deep belief network. *IEEE Trans. Cybern.* **50**(9), 3998–4007 (2019).

24. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974 (2024).
25. Qu, Z. et al. A photovoltaic cell defect detection model capable of topological knowledge extraction. *Sci. Rep.* **14**(1), 21904. https://doi.org/10.1038/s41598-024-61333-2 (2024).
26. Li, Y. et al. Contextual transformer networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(2), 1489–1500. https://doi.org/10.1109/TPAMI.2022.3145672 (2022).
27. Zhu, L., Wang, X., & Ke, Z., et al. Biformer: vision transformer with bi-level routing attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10323–10333 (2023). https://doi.org/10.1109/CVPR52729.2023.00996.
28. Nguyen, E. et al. S4nd: modeling images and videos as multidimensional signals with state spaces. *Adv. Neural. Inf. Process. Syst.* **35**, 2846–2861 (2022).
29. Gu, A., & Dao, T. Mamba: linear-time sequence modeling with selective state spaces. In: Proceedings of the First Conference on Language Modeling (2024). arXiv preprint arXiv:2312.00752.
30. Yang, Y., Su, Y., & An, S. Vdssa: Ventral & dorsal sequential self-attention autoencoder for cognitive-consistency disentanglement. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 693–705. Springer (2022).
31. Liu, Y. et al. Vmamba: visual state space model. *Adv. Neural. Inf. Process. Syst.* **37**, 103031–103063 (2024) arXiv:2401.10166.
32. Chen, K. et al. Rsmamba: remote sensing image classification with state space model. *IEEE Geosci. Remote Sens. Lett.* **21**, 1–5. https://doi.org/10.1109/LGRS.2024.3376113 (2024).
33. Lei, M., Li, S., & Wu, Y., et al. Yolov13: real-time object detection with hypergraph-enhanced adaptive visual perception. arXiv preprint arXiv:2506.17733 (2025).
34. Li, J., Wen, Y., & He, L. Scconv: spatial and channel reconstruction convolution for feature redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6153–6162. https://doi.org/10.1109/CVPR52729.2023.00602 (2023).
35. Xing, M., Feng, Z., Su, Y., & Oh, C. Learning by erasing: Conditional entropy based transferable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 6261–6269 (2024).

Acknowledgments

The authors sincerely acknowledge Universiti Putra Malaysia (UPM) for providing the academic platform and institutional support for this study. They also thank the authors of the publicly available code used in this research, whose contributions have greatly supported academic progress in this field. The authors further express their sincere appreciation to the proofreaders and editors for their valuable efforts in enhancing the quality and clarity of the manuscript.

Author contributions

Conceptualization, Q.W. and S.A.R.; methodology, Q.W. J.W. and S.A.R.; software, Q.W.; validation, Q.W. and S.A.R.; data curation, Q.W.; writing—original draft preparation, Q.W.; writing—review and editing, Q.W. S.A.R. S.H.T. and M.A.A.; supervision, Q.W. and S.A.R.; All authors have read and agreed to the published version of the manuscript.

Funding

The authors declare that this research received no funding from any organization or institution.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.b.A.R.