Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detection Method of Ballasted Track Fasteners in Dark Environment Based on Deep Learning

Tangbo Bai

baitangbo@bucea.edu.cn

Beijing University of Civil Engineering and Architecture

Yufei Wang

Beijing University of Civil Engineering and Architecture

Houliang Xiang

Beijing University of Civil Engineering and Architecture

Xiaolan Wang

Beijing University of Civil Engineering and Architecture

Wangyi Li

Beijing University of Civil Engineering and Architecture

Research Article

Keywords: deep learning, D-fine, fault detection, fasteners

Posted Date: September 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10778294/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Detection Method of Ballasted Track Fasteners in Dark Environment Based on Deep Learning

Tangbo Bai[1,2], Yufei Wang[1,2], Houliang Xiang[1,2], Xiaolan Wang[1,2], Wangyi Li[1,2]

[1] School of Mechanical-Electronic and Vehicle Engineering, Beijing University of Civil Engineering and Architecture, Beijin, China, 100044

[2] Beijing Key Laboratory of Intelligent Detection, Operation and Maintenance for Urban Rail Transit Vehicles, Beijing University of Civil Engineering and Architecture, Beijing, China, 100044

E-mail: baitangbo@bucea.edu.cn

Keywords: deep learning; D-fine; fault detection; fasteners

Abstract

Railway infrastructure facilities have long been subjected to vehicle loads and environmental factors. Track fasteners along the line are prone to spring clip loss, displacement, and damage, seriously threatening safe railway operation. To address insufficient illumination in dark environments, complex background near rails, and severe interference from ballast stones on ballasted tracks, a rail fastener state detection model based on D-fine is proposed to improve detection accuracy. First, the ASFF attention mechanism is introduced to mitigate noise and redundant information caused by poor lighting. Second, SCConv convolution is adopted to enhance feature extraction for complex fastener backgrounds. Finally, CoTAttention is introduced to reduce the impact of ballast stones. Experimental results show that the improved model achieves increases of 2.7%, 4.8%, and 2.0% in mAP, AP50, and AR50, respectively.

1. Introduction

Rail fasteners play a crucial role in reducing the interaction forces between the rails and sleepers, thereby ensuring the safety of track vehicles[1]. Damage to fasteners can cause more pronounced vibrations and jolts when trains pass[2][3], which not only affects passenger ride comfort[4] but may also lead to further damage to the track structure and the trains themselves[5]. Currently, common maintenance practices primarily rely on inspectors conducting regular patrols along the track to visually determine whether fasteners are missing or damaged. Although this method does not depend on complex equipment, it has limitations such as long inspection cycles, low efficiency, and susceptibility to subjective human judgment[6].

In recent years, with the rapid development of computer vision and artificial intelligence technologies, neural networks have demonstrated significant potential in applications such as object recognition and fault detection[7]. Current deep learning-based detection models are mainly divided into two categories: one is the two-stage model based on region proposals, represented by Faster R-CNN and Mask R-CNN. These models offer high detection accuracy but suffer from complex structures, high computational costs, and slow detection speeds, making it difficult to meet real-time detection requirements and challenging to deploy on terminal devices. For example, literature [8] proposed a two-stage classification model based on improved Faster R-CNN and Support Vector Data Description (SVDD) for the rapid and accurate detection of fastener conditions. The other category is the one-stage regression-based model, represented by SSD and YOLO. These models significantly improve detection speed while maintaining a certain level of accuracy, offering better real-time performance. For instance, literature [9] built a lightweight model based on YOLOv5n, replacing the C3 modules in the original

backbone and neck with C3Fast modules to reduce the number of parameters and computational load, and introduced the SIoU loss to enhance convergence speed and recognition accuracy, while adding the SA attention mechanism in the neck to improve feature fusion. Literature [10] proposed the LFGB-YOLO model, which includes a Light-Fast part and a GB-Neck part. The former reduces parameters, computation, and memory access frequency while maintaining high accuracy, and the latter reduces the complexity of the feature fusion network while preserving feature extraction capability, advancing the research on fastener screw detection in complex environments. Literature [11] used an improved SSD model to locate fastener regions from railway images; literature [12] proposed the YOLO-Fastener algorithm, combining channel and spatial attention mechanisms to enhance the extraction capability of key features of ballastless track fastener defects; literature [13] integrated the SRConv module into the RT-DETR backbone and introduced channel attention to improve computational efficiency; literature [14] proposed the GMS-YOLO method, replacing the YOLOv8 backbone network with GhostHGNetv2, and introduced multi-scale convolutional attention (MSCA) and a shared convolutional detection head (SCDH) to enhance the model's detection robustness in narrow and complex environments. Literature [15] proposed the AMEA-YOLO lightweight object detection algorithm, which integrates FasterNet, enhanced deep second-order channel attention, lightweight modules, and the HardSwish activation function, thereby effectively reducing the number of model parameters while ensuring high-resolution remote sensing vehicle detection accuracy.

In terms of detection in low-light environments, literature [16] proposed a method for object detection in complex dark environments based on YOLO-AS. By introducing the Zero-DCES image enhancement module, image quality in dark environments is improved through adaptive contrast enhancement. Subsequently, the YOLO-AS detection model was constructed, integrating ECA_ASPP and the SK attention mechanism to effectively enhance multi-scale feature expression capabilities. Literature [17] proposed the HDR-YOLO framework, which integrates an environment-agnostic dynamic feature extraction (DEAF) module and is optimized using a hybrid loss combining mean squared error and Log-Cosh to improve object detection performance in harsh environments. Literature [18] proposed an edge detection-based multi-scale feature enhancement network (EDMFEN) for low-light image enhancement tasks. It preserves edge details using the Sobel operator and integrates multi-scale feature extraction blocks with a spatial attention mechanism to fully restore image content and enhance enhancement performance. Literature [19] proposed the POP-YOLOv8 model to improve the detection performance of partially occluded pedestrians in nighttime traffic scenes. By integrating key improvements such as a feature enhancement module, the POPAM attention mechanism, the Ghost lightweight structure, and mixed-precision inference, it achieves higher detection accuracy and real-time performance under low-light and occlusion conditions. Literature [20] proposed a lightweight model called SFS-YOLO based on YOLOv11n. By integrating the SPD-Conv module, a fine-grained channel attention mechanism, and the SIoU loss function, it achieves accurate detection of swimming crab behavior under different lighting conditions. Literature [21] proposed a cascaded imaging method combining low-light enhancement and visible-light-long-wave infrared image fusion. Through joint training of LLENet and IXNet with customized loss functions, it effectively preserves salient targets and texture details on multiple datasets, generating high-quality fused images. Literature [22] proposed a cross-guided fusion attention network for RGB-T semantic segmentation, where attention mechanisms mutually guide modal weights, and global information and a dual decoder are introduced to ultimately generate three prediction maps.

Although the above methods have achieved intelligent detection through deep learning, numerous challenges remain in detecting rail fasteners in low-light environments. Firstly, insufficient lighting

conditions significantly impact the scene, severely interfering with the model's detection of fasteners and turnouts. Secondly, ballasted tracks are affected by ballast interference, making it difficult to extract fastener regions from the ballast[23]. Finally, the composition of the on-site rail environment is relatively complex[24], often including various interfering factors such as bolts, wires, and oil stains, which further increases the difficulty of detection.

The D-FINE[25] model is a real-time object detection model based on the Transformer architecture, with its core innovation lying in the redefinition of the bounding box regression task. It achieves high precision while maintaining an extremely high inference speed. For fastener detection tasks that require real-time processing of video streams, it ensures latency-free real-time monitoring. Its Fine-grained Distribution Refinement (FDR) transforms the localization task into a probability distribution optimization problem, rather than forcibly predicting imprecise fixed values. Additionally, its adaptive weighting strategy enhances the weight of the target's central region, compensating for the loss of edge information caused by insufficient illumination. However, interference from noise and redundant information generated by low-light conditions still has a significant impact on detection performance.

To address these issues, this paper proposes an object detection method based on D-FINE. The main contributions are as follows: 1) A high-definition camera was used to conduct targeted multiple experiments along a total railway length of 20 km, acquiring 10,000 fastener image data covering fastener states and defect types under insufficient lighting conditions. The collected data were organized and analyzed, and annotation was completed. A total of seven annotation labels were defined: normal fastener, wire occlusion, bolt occlusion, foreign object occlusion, missing fastener, fastener deflection, and fastener fracture; 2) Subsequently, research was carried out based on the characteristics of the data and the specific features of rail fasteners. According to the data analysis results, targeted optimization and improvement of the deep learning algorithm were performed. First, to reduce the interference of noise and redundant information caused by insufficient lighting, the ASFF feature fusion module was introduced. Through an adaptive spatial feature fusion mechanism, it can effectively filter out conflicting information; 3) Second, to address the problem of complex backgrounds, the SCConv convolutional module was introduced. By reconstructing the feature maps in both spatial and channel dimensions, it effectively improves the model's ability to extract detailed features and its detection accuracy; 4) Finally, to mitigate the interference of ballast stones on ballasted tracks, the CoTAttention mechanism was introduced. It simultaneously models the interdependencies among features and effectively integrates contextual information, thereby significantly enhancing the feature representation capability.

In addition, the main structure of this paper is as follows: First, the model improvement methods adopted to address the difficulties of recognition under insufficient lighting conditions and the severe interference of ballast stones on ballasted tracks are described; second, the collection and processing procedure of the rail fastener dataset, as well as the algorithm evaluation metrics used, are introduced; finally, the effectiveness of the proposed improvement methods is verified through ablation experiments and comparative experiments with different algorithms, and the research conclusions are summarized accordingly.

2.Theoretical Background

2.1. Method framework

This paper addresses the current difficulties in detecting ballasted track fasteners under dark environments, including noise and redundant information interference caused by insufficient lighting, complex backgrounds, and interference from ballast stones. It aims to propose an intelligent detection

method based on D-FINE to achieve real-time monitoring of fastener conditions. This method mainly consists of two parts: dataset construction and deep learning algorithm improvement. Its specific implementation framework is described below and illustrated in Figure 1.

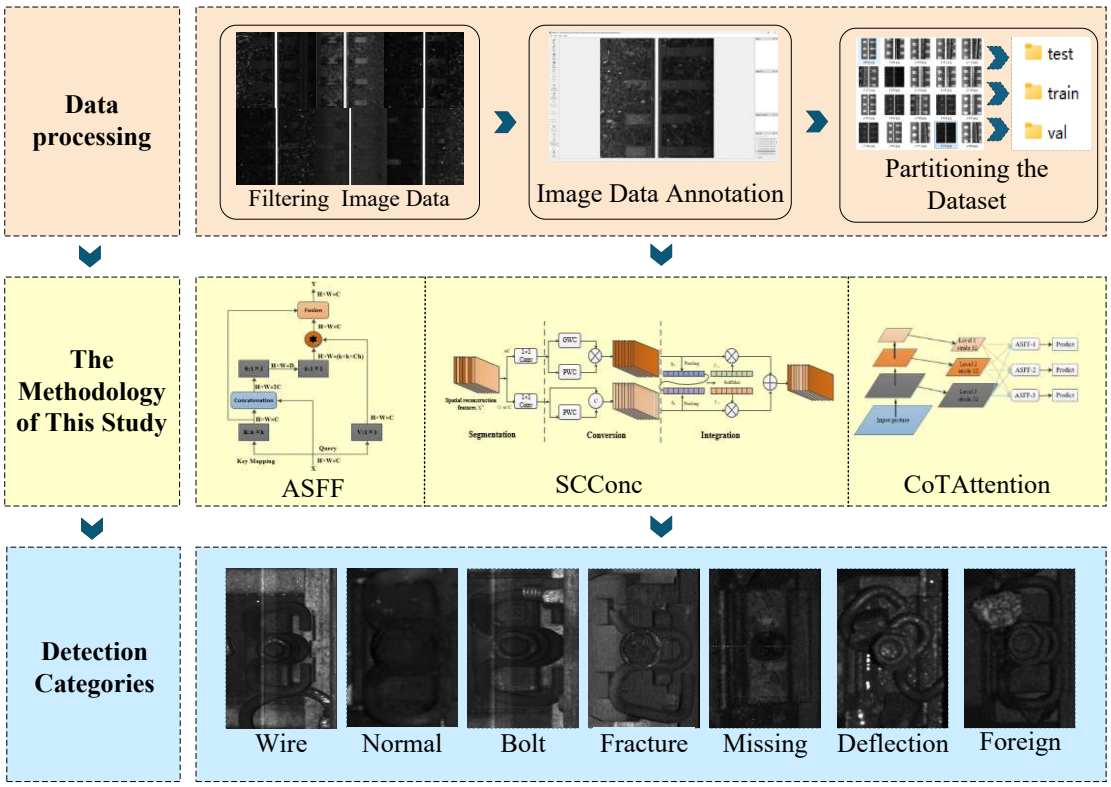


Figure 1. Method Framework

1) A high-definition camera was used to conduct multiple experiments along a ballasted track railway line for image acquisition. Effective samples containing rail fasteners in dark environments were selected from the collected images. The screened images were analyzed and organized. According to the actual fastener conditions and common interference factors, the recognition targets were divided into seven categories: normal fastener, wire occlusion, bolt occlusion, foreign object occlusion, missing fastener, fastener deflection, and fastener fracture. Subsequently, data annotation was completed, and the dataset was divided as needed into training set, validation set, and test set. To address the problem of missing edge information caused by insufficient lighting, improvements were made based on the D-FINE series algorithm framework. The main network structure of the D-FINE series algorithm is shown in Figure 2. Its adaptive weighting strategy enhances the weight of the target center region, thereby compensating for missing edge information. At the same time, it has an extremely high inference speed, which is beneficial for real-time fastener monitoring requirements.

2) To address the problems of shallow features containing a large amount of noise and deep features having blurred spatial positions in dark environments, the ASFF feature fusion module[26] was introduced. By learning weights, it can automatically suppress the fuzzy interference of deep features in dark areas while enhancing the weak but effective signals in shallow features, thereby restoring clearer target features. ASFF learns adaptive weights to spatially filter out conflicting information, retaining only useful features, thus eliminating inconsistencies between features of different scales and improving feature extraction capability.

3) To address the complex fastener detection environment, the SCConv convolutional module[27] was introduced into the backbone network, enabling finer processing of image features. SCConv decomposes the spatial convolution kernel in traditional convolution into a spatial reconstruction unit

(SRU) and a channel reconstruction unit (CRU), which respectively reconstruct the spatial information and channel information of the image, thereby improving feature extraction capability.

4) To address the interference of ballast stones in ballasted tracks, the CoTAttention mechanism[28] was introduced into the backbone network to increase the model's attention to fastener conditions. The CoTAttention mechanism explicitly mines and encodes local contextual information between neighboring pixels to form static features, then dynamically learns attention weights to capture long-range dependencies, and fuses the static local features with dynamic global attention features for output. In summary, the optimization methods proposed in this paper effectively improve the target detection performance of the model for rail fasteners.

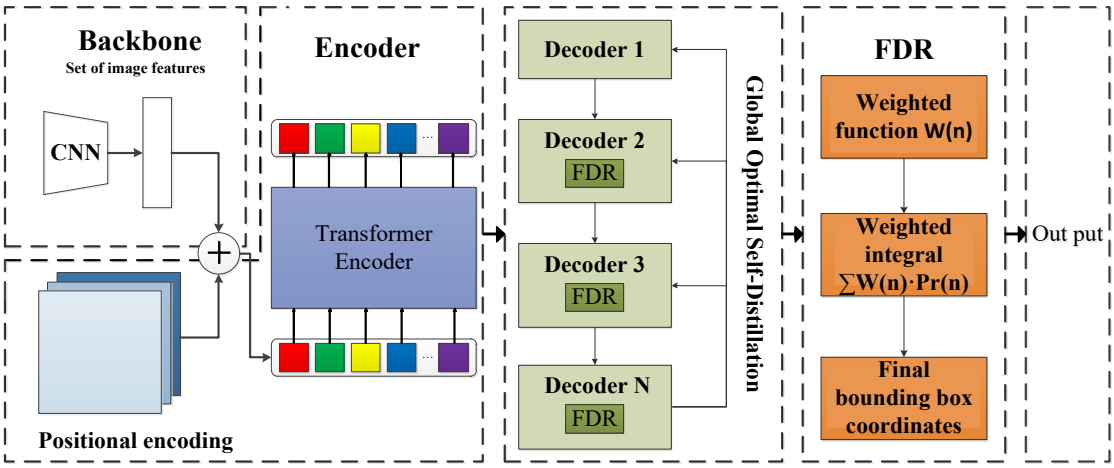


Figure 2. Network Structure

2.2 ASFF Structure

In the task of detecting ballasted track fasteners in low-light environments, insufficient lighting conditions cause shallow features to contain significant noise and deep features to suffer from blurred spatial locations, thereby interfering with the extraction of effective features. Methods such as Feature Pyramid Networks and multi-level feature pyramids are widely adopted; however, in single-stage detectors, the inconsistency between features at different scales limits further improvement in model performance. To address this issue, this paper introduces the Adaptively Spatial Feature Fusion (ASFF) module, the structure of which is shown in Figure 3.

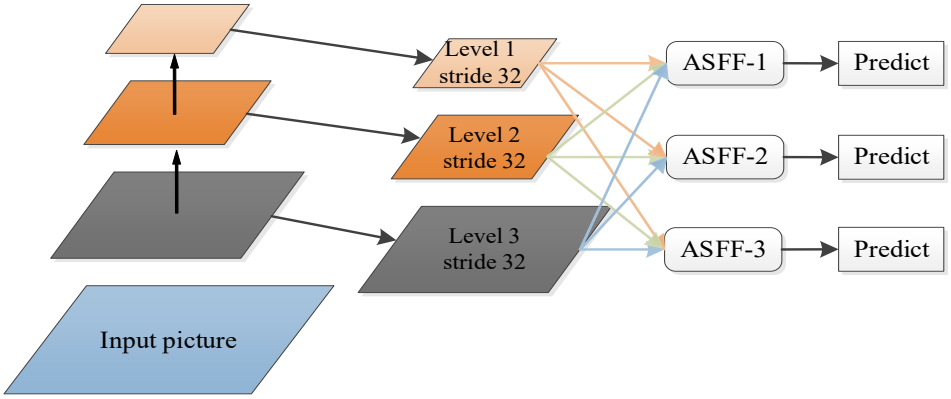


Figure 3. ASFF Network Structure

The ASFF method was proposed for single-stage object detection tasks, aiming to address the inconsistency issue between features at different scales. Its main innovation lies in introducing an adaptive spatial feature fusion mechanism that can effectively filter out conflicting information, thereby enhancing the model's robustness to scale variations. The core idea of this method is to achieve

adaptive fusion, i.e., dynamically adjusting the fusion weights of features from different layers based on the content of feature maps at each scale and their importance in the current task, rather than simply concatenating features. During the fusion process, ASFF evaluates the importance of features from each layer and dynamically weights the fusion accordingly to avoid contradictions between multi-scale features.

The working principle of the ASFF detection head is as follows: First, features from different levels are upsampled or downsampled to have the same spatial dimensions. In the figure, Level 1, Level 2, and Level 3 represent feature layers with different resolutions in the feature pyramid; ASFF-1, ASFF-2, and ASFF-3 denote the fused features after applying the ASFF mechanism to the corresponding levels. Taking ASFF-3 as an example (illustrated in the enlarged portion of the figure), features from other levels are adjusted to the same size as the third layer and then fused through learnable weight maps to generate the final fused features used for prediction.

In this way, ASFF can adaptively select the most effective features at each spatial position, thereby improving detection accuracy. This mechanism enables the model to flexibly determine which feature levels are most critical for the final prediction based on the specific location and scale context. In nighttime environments, where lighting conditions are complex and variable, the fusion of multi-level features helps enhance feature representation capabilities, capturing more details and contextual information, thus improving the detection performance of target fasteners in low-light conditions.

2.3 Backbone Network Optimization

As the core structure responsible for feature extraction in deep learning models, the performance of the backbone network is directly related to the accuracy of rail fastener detection. In real-world scenarios, the track environment faces numerous challenges, being significantly affected by insufficient lighting conditions while also being subject to complex environmental noise interference. Additionally, rail fasteners themselves come in varied forms and are often occluded by objects such as wires and bolts, which further increases the difficulty of detection. This paper enhances the feature extraction capability of the backbone network in low-light environments by introducing the CoTAttention mechanism and the SCConv convolution module.

2.3.1 SCConv Convolution Module

In the task of detecting ballasted track fasteners in low-light environments, targets are often affected by factors such as oil stains, bolts, and wires, resulting in a complex environment that makes it difficult to extract effective features. SCConv is an innovative convolutional structure that enhances the model's ability to extract detailed features and improves detection accuracy by reconstructing the spatial and channel dimensions of features, respectively. Traditional convolution kernels have limited performance in such scenarios, whereas SCConv refines the input features through its Spatial Reconstruction Unit (SRU) and Channel Reconstruction Unit (CRU): the SRU is used to separate and reconstruct spatial features, reducing spatial redundancy; the CRU adopts a split-transform-fusion strategy to reduce channel redundancy. This mechanism enables SCConv to extract more discriminative key features from complex backgrounds, the structure of which is shown in Figure 4.

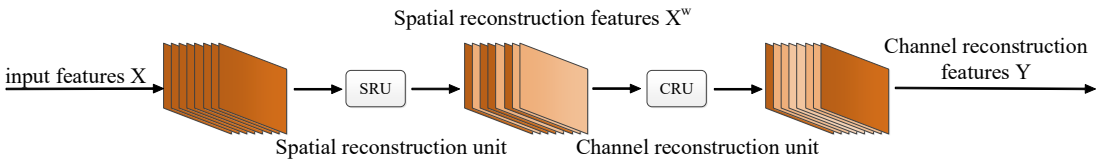


Figure 4. SCConv Architecture

The structure of traditional convolutions is typically fixed in both spatial and channel dimensions, which often limits their expressive capability when dealing with features of varying sizes and shapes. In complex scenarios such as detecting small targets or weak features, convolutions with a fixed structure struggle to fully extract critical information, easily leading to feature omission. SCConv introduces a dual reconstruction mechanism for both spatial and channel dimensions, enabling the convolution kernel to adaptively adjust its form and scale based on the input features, thereby better adapting to data environments where target features are not prominent. By utilizing spatial transformation matrices and channel transformation matrices, this method optimizes the spatial structure of feature maps and fuses channel information, effectively enhancing the model's expressive capability. SCConv is primarily composed of two core modules: the Spatial Reconstruction Unit (SRU) and the Channel Reconstruction Unit (CRU).

1. Spatial Reconstruction Unit (SRU): The operation of SRU can be summarized into two steps: separation and reconstruction. Its structure is shown in Figure 5.

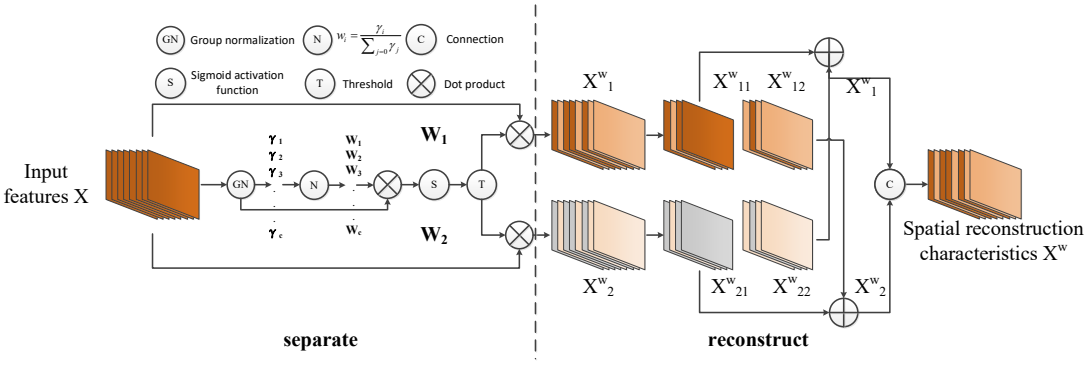


Figure 5. Spatial Reconstruction Unit Structure

The separation operation involves a given intermediate feature map $X \in R^{N \times C \times H \times W}$, where N represents the batch axis, C represents the channel axis, and H and W represent the spatial height and width axes. The GN layer denotes the group normalization layer, with μ and σ being the mean and standard deviation of X, ε being a very small positive constant, and γ and β being trainable parameters. First, the input feature X is normalized by subtracting the mean μ and dividing by the standard deviation σ, as shown in Equation (1).

$$X_{out}=GN(X)=\gamma\frac{X-\mu}{\sqrt{\sigma^2+\varepsilon}}+\beta \tag{1}$$

In the formula, the trainable parameter $\gamma \in R^C$ serves as a measure of the spatial pixel variance for each batch and channel. The normalized correlation weights $W_\gamma \in R^C$ are obtained from Equation (2), representing the importance of different feature maps.

$$W_\gamma=\{w_i\}=\frac{\gamma_i}{\sum_{j=1}^{C}\gamma_j},i,j=1,2,L,C \tag{2}$$

After reweighting the feature map of W_γ, the weights are obtained. These weights are then mapped to the range (0, 1) using the Sigmoid function and gated through a threshold. In the module, weights above the threshold are set to 1 to obtain the information weights W_1, while weights below the threshold are set to 0 to obtain the non-information weights W_2. The entire process of obtaining W is shown in Equation (3).

$$W = Gate(Sigmoid(W\gamma(GN(X)))) \tag{3}$$

The high-information feature map X^{w_1} and the low-information feature map X^{w_2} are obtained. X^{w_1} contains informative and expressive spatial content, while X^{w_2} contains little or no information and is considered redundant. To reduce spatial redundancy, a reconstruction operation is further proposed, which combines the information-rich features with the less informative features to generate more informative features, thereby saving space. The SRU structure employs a cross-reconstruction operation to fully integrate the two weighted feature maps with different information levels, strengthening the information flow between them. The cross-reconstructed features X^{w_1} and X^{w_2} are then concatenated to obtain the spatially refined feature map X^{w}. The entire process is represented by Equation (4), where $\otimes$ denotes element-wise multiplication, $\oplus$ denotes element-wise addition, and $\cup$ denotes the concatenation operation.

$$\begin{cases} X_1^w = W_1 \otimes X \\ X_2^w = W_2 \otimes X \\ X_{11}^w \oplus X_{22}^w = X^{w1} \\ X_{21}^w \oplus X_{12}^w = X^{w2} \\ X^{w1} \cup X^{w2} = X^w \end{cases} \tag{4}$$

2. Channel Reconstruction Unit (CRU): The goal of CRU is to reduce redundancy in the channel dimension by employing a split-transform-fusion strategy to explore channel redundancy in features and extract high-level representative information. Its structure is shown in Figure 6.

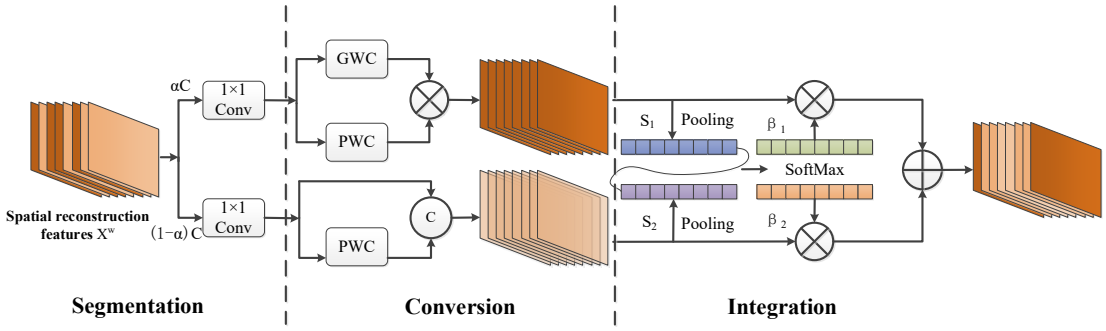


Figure 6. Channel Reconstruction Unit Structure

In the split phase, for a given spatially refined feature $X_w \in R^{c \times h \times w}$, the channels of X^w are first divided into two parts: one part with αC channels and the other with (1-w)C channels. In this study, α is set to 0.5. Subsequently, 1×1 convolutions are used to compress the channels of the feature maps to improve computational efficiency, and a compression ratio r is introduced to control the feature channels of the CRU to balance computational cost. After the split and compression operations, the spatially refined feature X^w is divided into an upper part X_{up} and a lower part X_{low}. In the transformation phase, X^w is fed into the upper transformation stage to act as a "feature extractor." Efficient, low-computation-cost grouped convolutions (GWC) and pointwise convolutions (PWC) are used to extract features. In this setup, the group size in GWC is set to 16. The aforementioned transformation stage can be expressed as Equation (5).

$$Y_1 = M^G X_{up} + M^{P_1} X_{up} \tag{5}$$

In the formula, M^G and M^{P1} are the learnable weights of GWC and PWC, respectively. The input and output feature maps of the upper stage are X_{up} and Y_1, respectively. X_{low} is fed into the lower transformation stage, as expressed in Equation (6).

$$Y_2 = M^{P_2} X_{low} \cup X_{low} \tag{6}$$

In the formula, M^{P2} represents the learnable weights of PWC, and $\cup$ denotes the concatenation module. X_{low} and Y_2 are the input and output feature maps of the lower stage, respectively. In the fusion stage, the SKNet method is used to adaptively merge the output features Y_1 and Y_2 from the upper and lower transformation stages. Global average pooling is applied to gather global spatial information, and the global channel descriptor S_m is calculated as shown in Equation (7).

$$S_m = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} Y_c(i,j), m=1,2 \tag{7}$$

Subsequently, the global channel descriptors S_1 and S_2 from the upper and lower parts are stacked together, and a channel attention operation is applied to generate the importance vectors β_1 and β_2, as shown in Equation (8).

$$\beta_1 = \frac{e^{S_1}}{e^{S_1} + e^{S_2}}, \beta_2 = \frac{e^{S_2}}{e^{S_1} + e^{S_2}} \tag{8}$$

Finally, the channel-reconstructed features can be represented as $\beta_1 Y_1 + \beta_2 Y_2$. In the SCConv module, all parameters are concentrated in the transformation stage, which can significantly reduce the number of parameters compared to traditional convolution operations.

2.3.2 CoTAttention Mechanism

In the task of detecting ballasted track fasteners in low-light environments, the detection targets are often affected by ballast interference, making it difficult to accurately identify fastener regions within complex, wide-field images containing ballast. In traditional attention mechanisms, each position in the input feature map independently generates a query and a key, representing the features of that position and its relationships with other positions, respectively. However, in such mechanisms, the query-key associations are learned independently, focusing only on local positional features and failing to adequately model global contextual information. In visual tasks, the lack of contextual information can limit feature representation capabilities, thereby affecting model performance improvements. To mitigate these issues, the CoTAttention module integrates context mining with self-attention learning into a unified architecture, the specific structure of which is shown in Figure 7.

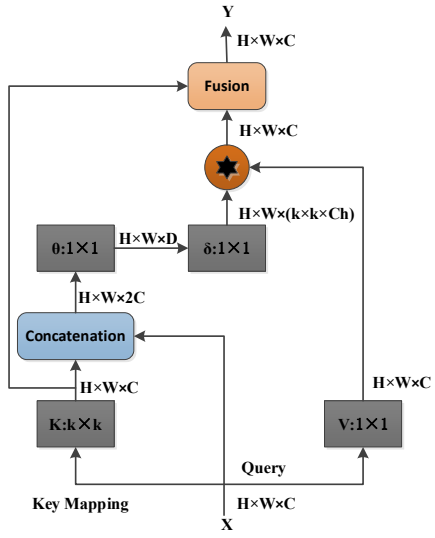


Figure 7. CoTAttention Network Structure

Through this design, CoTAttention can simultaneously model the interdependencies between

features and effectively integrate contextual information, thereby significantly enhancing feature representation capabilities. Assuming the input feature map X has dimensions H × W × C, its specific workflow is as follows:

(1) Contextual Information Encoding

In the CoT Block, the Key, Query, and Value are still generated from the input feature map X:

$$Q=XW，\ K=XW_K，\ V=XW_V \tag{9}$$

Unlike traditional self-attention mechanisms that use 1×1 convolutions to generate the Key, CoTAttention first applies k×k grouped convolutions at each spatial position of the Key, with a receptive field covering a local neighborhood. This operation explicitly incorporates neighborhood information, so that the generated contextual Key K encodes not only its own features but also captures contextual relationships with surrounding regions. At this point, the dimensions of K remain H×W×C, but its feature representation is significantly enhanced.

(2) Local Attention Computation

The core design of the CoTAttention lies in that it does not rely on traditional pointwise Query-Key calculations. Instead, it constructs a local attention matrix based on the Query feature Q and the Key feature K, which contains contextual information. This mechanism introduces contextual guidance into the Query-Key relationship at each position, thereby enabling the attention to have stronger global perception capabilities.

(3) Context-Enhanced Feature Expression

Guided by contextual information, CoTAttention can model broader pixel dependencies while strengthening the feature expression of target regions. By integrating the contextual information K extracted by convolutions into the Value, the update of the feature map V includes not only the Query-Key attention weights but also explicitly incorporates contextual cues from local neighborhoods.

3. Experimental Results and Analysis

3.1. Experimental parameters

In the experiment, the operating system used is Windows 11, the GPU and CPU are RTX 5090 (8GB) and Intel Core Ultra 9 285K respectively, the deep learning framework is PyTorch 2.4.1, the programming language is Python 2.12.3, and CUDA version is Cuda 12.4.

3.2. Data acquisition and processing

The rail fastener states include normal, deflection, missing, fracture, foreign object occlusion, bolt occlusion, and wire occlusion. Using a high-definition acquisition device (resolution 2048×2048, global shutter CMOS sensor, supporting GigE interface and PoE power supply), multiple experiments were conducted on a railway section with a total length of 20 km, and 10,000 fastener image data (2048×2048) were selected. After post-processing, a dataset of 5,000 fastener images was finally obtained, as shown in Figure 8.

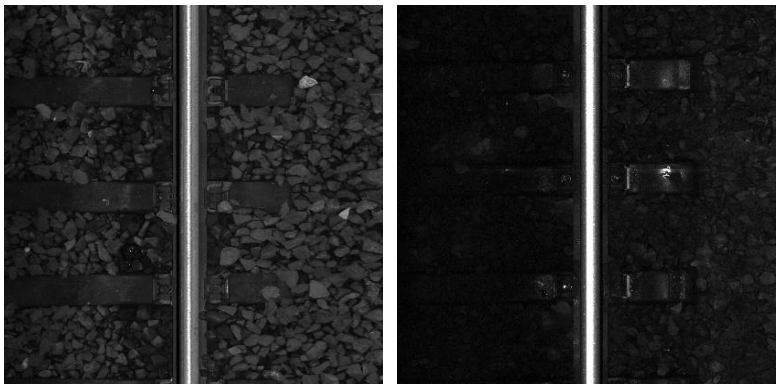


Figure 8. Images of the fastener dataset

3.3. Evaluation indicators

To evaluate the detection performance of the improved algorithm on rail fastener conditions, a series of ablation experiments were conducted to optimize the algorithm's performance. The mean Average Precision (mAP), Average Precision at an IoU threshold of 0.5 (AP50), Average Recall at an IoU threshold of 0.5 (AR50), Floating Point Operations (FLOPs), and Multiply-Accumulate Operations (MACs) were used as performance indicators. The calculation formulas are as follows.

$$AP = \int_0^1 P(R)dR \tag{10}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{11}$$

$$AR = \frac{1}{N}\sum_{i=1}^{N} R_i \tag{12}$$

3.4 Analysis of Experimental Results

3.4.1 Ablation Experiments

Table 1. Ablation Experiment Results

ASFF	SCConv	CoTAttention	mAP	AP50	AR50	FLOPs(G)	MACs(G)
-	-	-	0.580	0.832	0.766	7.1	3.5
+	-	-	0.602	0.841	0.770	7.2	3.6
-	+	-	0.581	0.844	0.769	7.1	3.5
-	-	+	0.586	0.855	0.772	7.1	3.5
+	+	-	0.591	0.820	0.781	7.2	3.6
+	-	+	0.592	0.835	0.799	7.2	3.6
-	+	+	0.585	0.841	0.777	7.1	3.5
+	+	+	0.607	0.880	0.786	7.3	3.7

From Table 1, it can be observed: ① Experiment 1 introduces the ASFF module based on the original model. Compared to the baseline model, mAP, AP50, and AR50 increased by 2.2%, 0.9%, and 0.4%, respectively, while FLOPs and MACs only increased by 0.1 each. The ASFF module improves model accuracy without significantly increasing computational load. ② Experiment 2 embeds the SCConv module into the original model, resulting in increases of 0.1%, 1.2%, and 0.3% in mAP, AP50, and AR50, respectively, with no changes in FLOPs and MACs. The SCConv module does not add to the

computational load but enhances the model's feature extraction capability. ③ Experiment 3 introduces the CoTAttention mechanism into the original model, leading to improvements of 0.6%, 2.3%, and 0.6% in mAP, AP50, and AR50, respectively, with no changes in FLOPs and MACs. The CoTAttention mechanism mitigates the interference of ballast in ballasted tracks, increasing model accuracy. ④ Experiment 4 shows that when ASFF and SCConv are introduced simultaneously, AP decreases by 1.3%, while mAP and AR increase by 1.1% and 1.5%, respectively. ⑤ Experiment 5 introduces ASFF and CoTAttention into the original model, resulting in increases of 1.2%, 0.3%, and 3.3% in mAP, AP50, and AR50, respectively. ⑥ Experiment 6 introduces SCConv and CoTAttention into the original model, leading to increases of 0.5%, 0.9%, and 1.1% in mAP, AP50, and AR50, respectively. ⑦ Experiment 7 introduces ASFF based on Experiment 6, achieving increases of 2.7%, 4.8%, and 2.0% in mAP, AP50, and AR50, respectively, compared to the baseline model. This configuration minimizes missed detections of damaged fasteners while reducing the probability of misclassifying normal fasteners as damaged.

To comprehensively evaluate the performance of each improved module across multiple key indicators, the quantitative results of all experimental configurations were summarized and compared using a histogram. This provides a more intuitive visualization of the performance differences among different combination strategies.

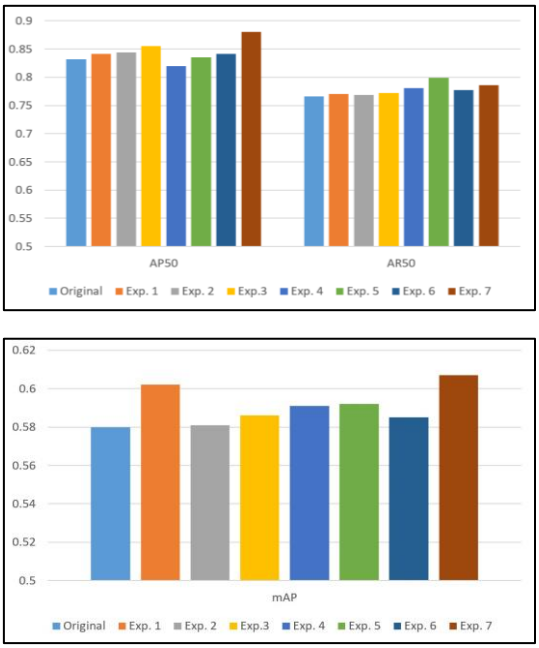
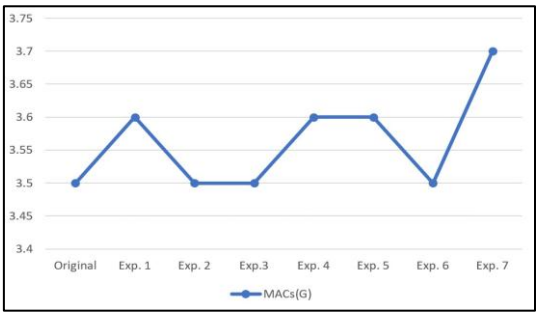


Figure 9. Performance comparison of ablation experiments



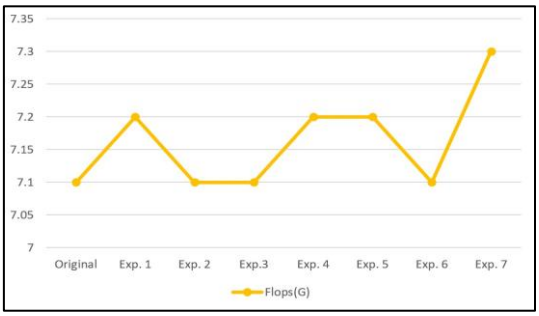


Figure 10. Computational speed comparison of ablation experiments

The ablation experiment results demonstrate that the proposed improved method effectively enhances the model's capability for multi-scale feature extraction and fine detail capture in low-light environments, significantly improving both the average precision and average recall of detection. The experimental results prove that the proposed solution achieves significant performance improvements while maintaining a low parameter count and computational complexity, demonstrating strong engineering applicability and practical application potential.

3.4.2 Comparative experiments

We selected mainstream detection models, including D-fine, YOLOv8, YOLO11n, YOLOv12n, and RT-DETR50, to conduct comparative experiments under the same training parameters, in order to verify the superiority of the proposed model. The results are shown in Table 2.

Table 2. Comparative Experiment Results

Model	mAP	ap50	AR50	Flops	MACs
D-fine	0.580	0.832	0.766	7.1	3.5
YOLOv8	0.603	0.744	0.616	6.9	3.4
YOLO11n	0.592	0.740	0.626	6.4	3.2
YOLOv12n	0.597	0.742	0.625	6.5	3.3
Rt-detr50	0.557	0.559	0.544	130.5	65.2
My model	0.607	0.880	0.786	7.3	3.7

From Table 2, it can be seen that compared to the baseline D-fine model, the proposed model achieves increases of 2.7%, 4.8%, and 2.0% in mAP, AP50, and AR50, respectively, with only a slight increase of 0.2 in both FLOPs and MACs. Compared to the YOLOv8 network model, mAP, AP50, and AR50 are improved by 0.4%, 13.6%, and 17.0%, respectively, with FLOPs and MACs increasing by only 0.4 and 0.3. Compared to the YOLO11n network model, mAP, AP50, and AR50 are enhanced by 1.5%, 14.0%, and 6.0%, respectively. Compared to the YOLOv12n network model, mAP, AP50, and AR50 are increased by 1.0%, 13.8%, and 16.1%, respectively. Compared to the RT-DETR50 network model, mAP, AP50, and AR50 are improved by 5.0%, 32.1%, and 24.2%, respectively, while FLOPs and MACs are significantly lower than those of the RT-DETR50 model. In summary, compared to other mainstream models, the proposed model achieves improvements in mAP, AP50, and AR50 with only minimal increases in FLOPs and MACs. Therefore, this model is reasonable and effective for the task of rail fastener detection.

Based on the detailed comparative data presented in Table 4, Figure 10 further provides a visual comparison between the proposed model and several mainstream object detection models from the perspective of key performance indicators. Specifically, a comprehensive analysis is conducted from three dimensions: model complexity, detection accuracy, and computational efficiency, aiming to demonstrate that the proposed model achieves an effective balance between high performance and

model efficiency.

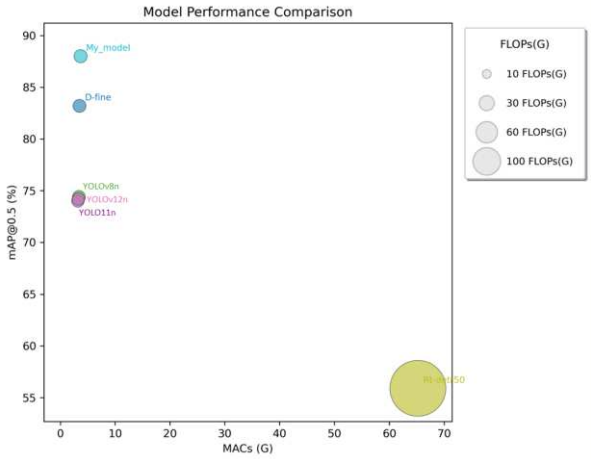


Figure 11. Performance Comparison of Improved Models

In summary, the improved D-FINE model proposed in this paper achieves excellent detection performance while maintaining a smaller model scale, making it suitable for fastener detection in low-light environments and demonstrating significant application potential and engineering value.

To visualize the model's performance, a comparison of the detection results between the original model and the improved experimental model is presented in the figure below.

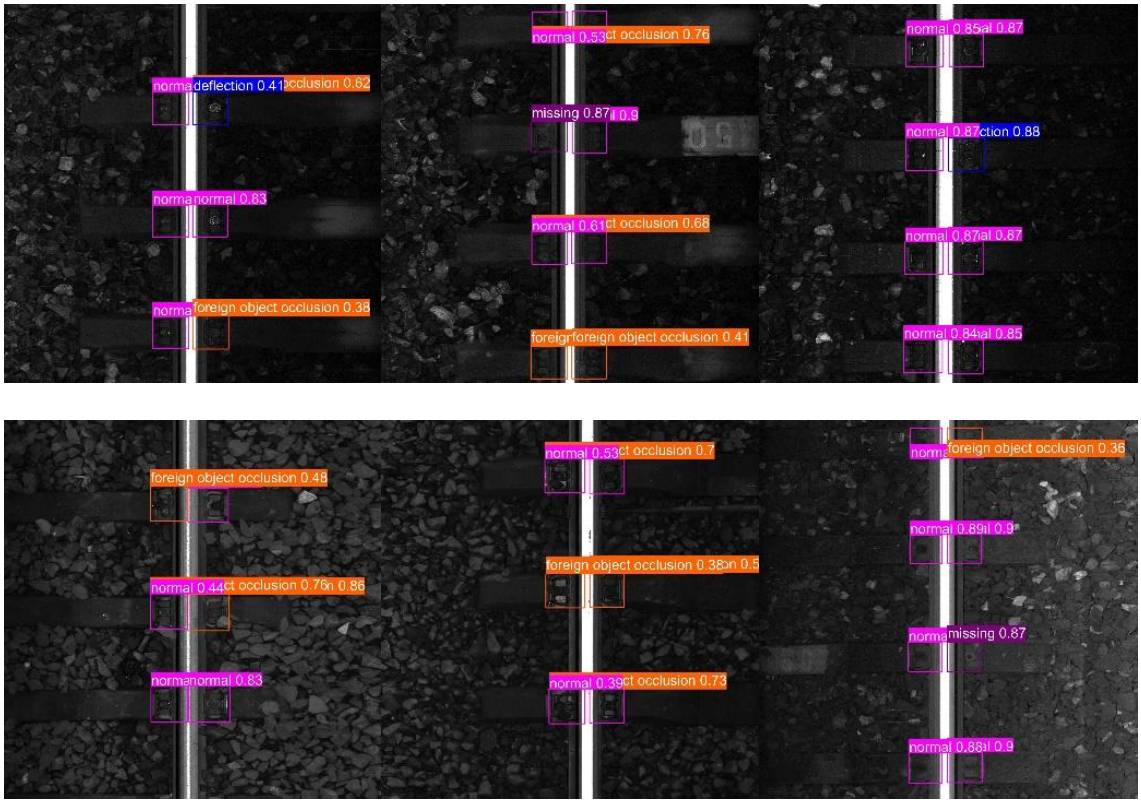


Figure 12. D-fine Detection Effect Diagram

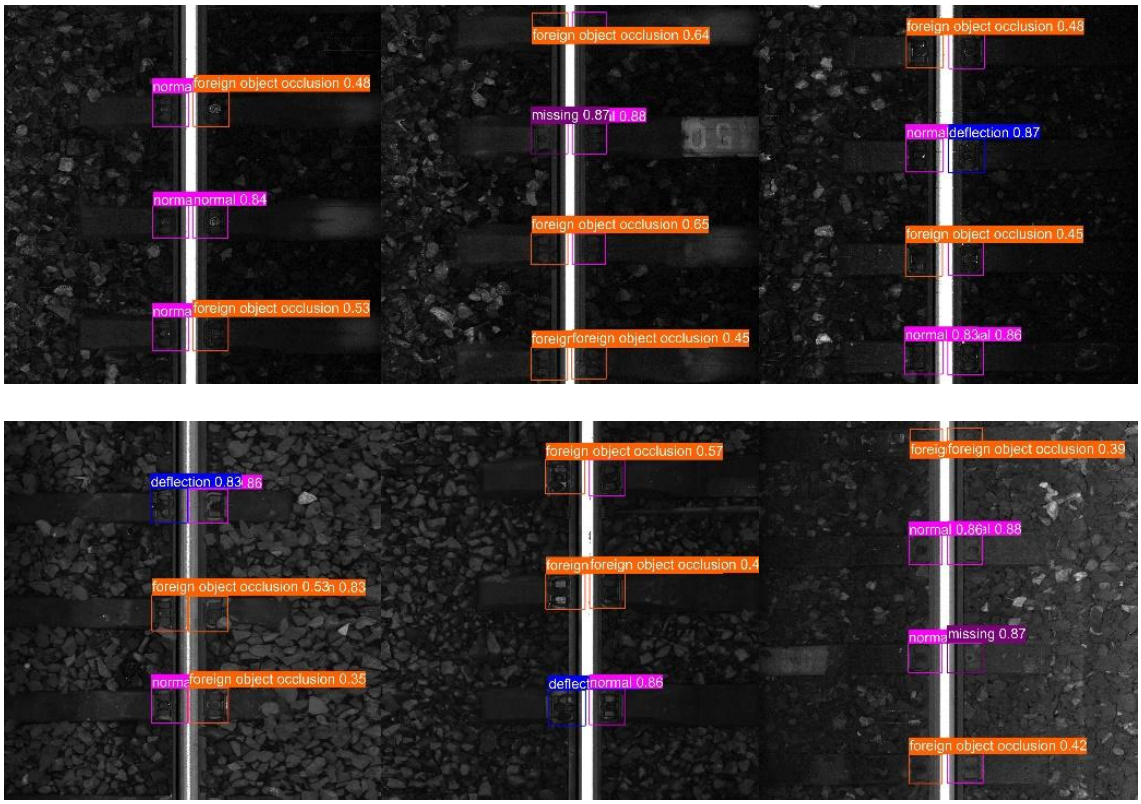


Figure 13. Improved Detection Effect Diagram

As shown in the figure: In the detection results of the unimproved model, due to insufficient illumination, the image features are weak, leading to uncertainty in the model's detection results and the problem of duplicate labels. Meanwhile, because the original model has weak anti-interference ability against complex backgrounds, it becomes confused in distinguishing between normal fasteners with oil stains and fasteners with foreign object intrusion. Moreover, under low-light conditions, the model's detection confidence is low. In contrast, the proposed model alleviates the effect of insufficient illumination through the ASFF feature fusion module, enhances the model's feature extraction capability via the SCConv convolutional module, and addresses the issue of weak features of small fastener targets with the CoTAttention mechanism. These three improvements enable the proposed model to correctly identify various types of fasteners and successfully distinguish normal fasteners from those with foreign object intrusion while avoiding interference from oil stains. At the same time, the detection confidence is also improved compared to the original model. This further verifies the superiority of the proposed model.

To further investigate the model's attention mechanism on fastener regions in low-light environments, this study employs Grad-CAM technology[29] to generate heatmaps for analysis, as shown in Figures 13 and 14. The original D-Fine model exhibits a relatively dispersed attention distribution in fastener regions and often incorrectly focuses on non-defect areas. In contrast, the improved model demonstrates a more concentrated attention distribution, with its generated feature heatmaps more accurately focusing on fastener regions. This indicates that the improved model can more precisely identify fastener defect features, effectively enhancing recognition accuracy while reducing false detections in non-defect areas. Therefore, the model's feature extraction capability for fasteners in low-light environments is significantly strengthened, enabling more accurate localization of defect regions.

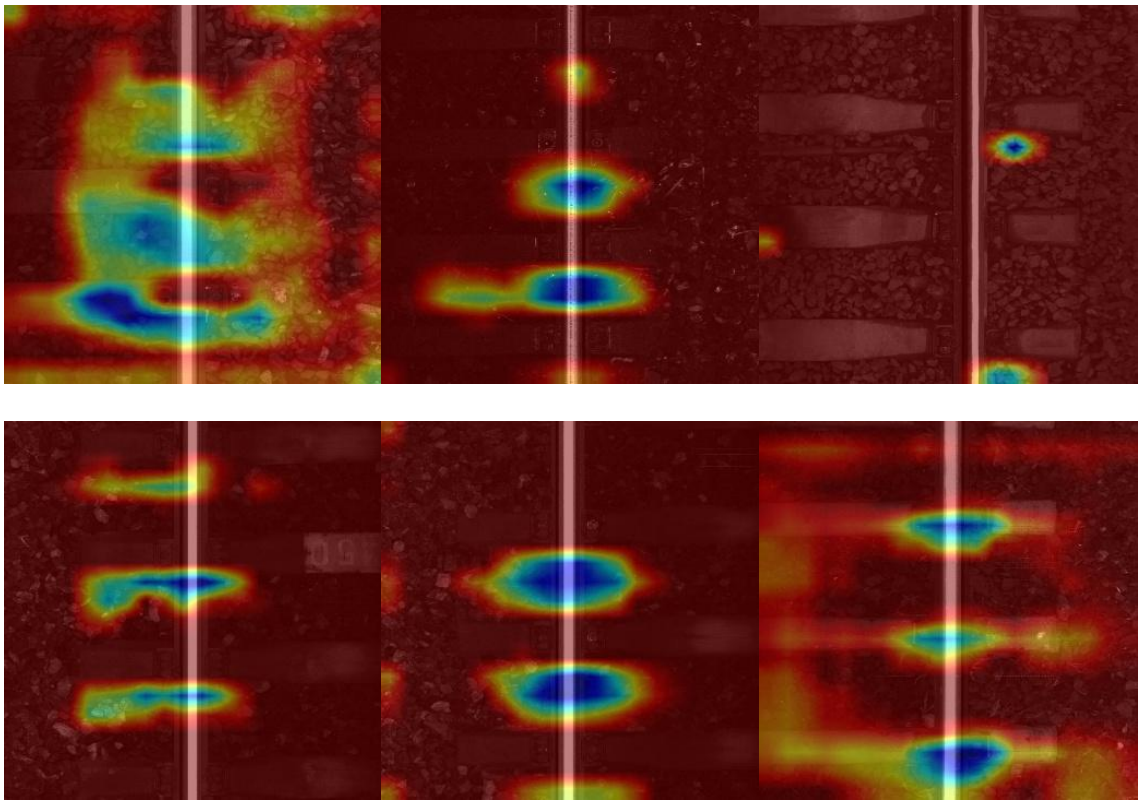


Figure 14. D-fine Model Heatmap

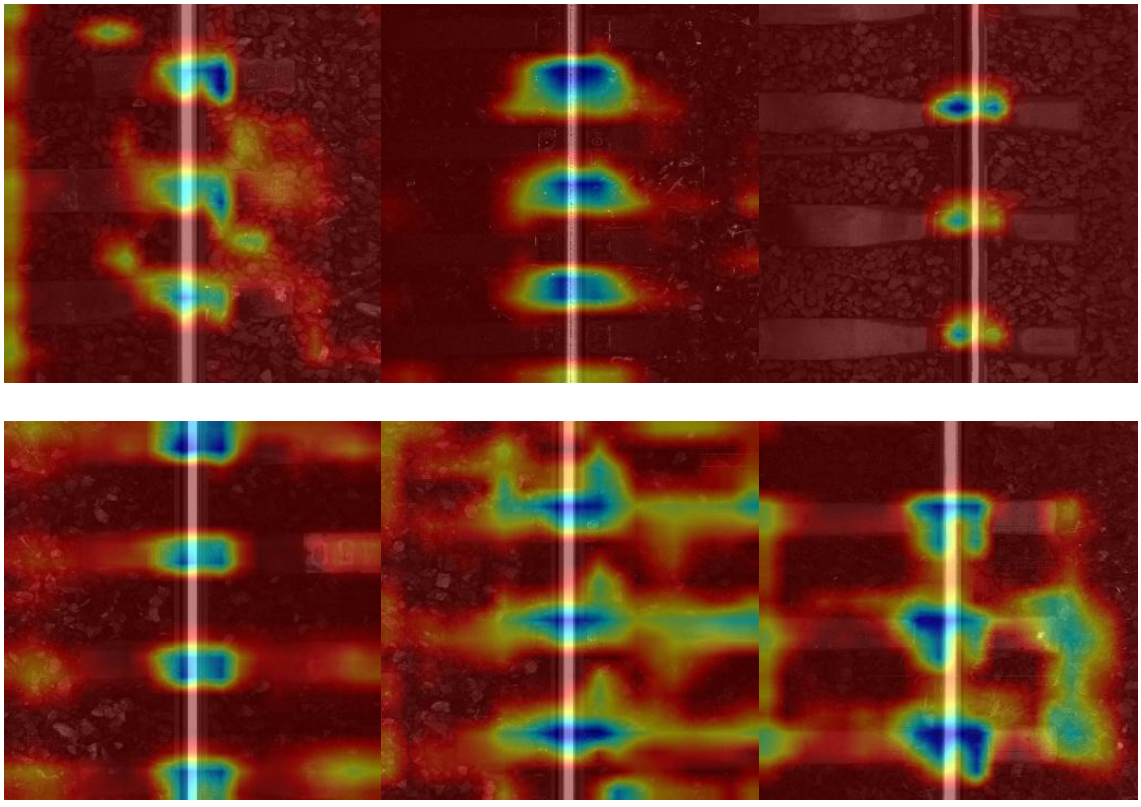


Figure 15. Improved Model Heatmap

In summary, the experimental results fully validate the significant performance improvement of the improved model in the task of fastener detection in low-light environments. Both the quantitative evaluation of detection accuracy for various defect types and the attention distribution characteristics

revealed by heatmaps demonstrate that the improved model possesses superior defect detection capabilities, as well as more precise feature recognition and localization accuracy. These findings confirm that the proposed improvement strategies effectively address detection challenges in complex environments, achieving a favorable balance between model performance and efficiency, and providing solid technical support for the accurate identification of fastener defects in low-light conditions.

4. Summarizes

To address the problems of insufficient lighting conditions, severe environmental interference such as oil stains, and weak features due to the small area proportion of fasteners in fastener detection under dark environments, an improved D-fine model-based dark-environment fastener image detection model is proposed. Experiments are conducted on a self-built fastener dataset to verify the improvement in model accuracy and performance. The conclusions are as follows:

1) By introducing the ASFF feature fusion module, the interference of noise and redundant information caused by insufficient lighting is alleviated. The ASFF feature fusion module introduces an adaptive spatial feature fusion mechanism that can effectively filter out conflicting information, thereby improving the model's robustness to scale variations.

2) The SCConv convolutional structure is introduced into the backbone network to enhance the model's ability to recognize effective information. SCConv is an innovative convolutional structure that reconstructs features in both spatial and channel dimensions. Compared with the original D-fine model, the improved model achieves increases of 2.7%, 4.8%, and 2.0% in mAP, AP50, and AR50, respectively, while FLOPs and MACs increase by only 0.2 each.

3) The CoTAttention mechanism is introduced, which simultaneously models the interdependencies among features and effectively integrates contextual information. As revealed by the attention distribution characteristics shown in the heatmap, the model's ability to extract fastener features in ballasted track environments is improved.

In addition, compared with mainstream detection models, the detection accuracy of the proposed model is also improved, further verifying the effectiveness of the improvements. Experimental results show that although the improved model has a slight increase in parameter count and computational complexity, its detection accuracy is significantly enhanced, providing a reliable technical means for ballasted track fastener detection under dark environments.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China 52272385, National Natural Science Foundation of China under Grant 52272385.

Declarations

Conflict of Interest

The authors declare no competing interests.

References

[1] Yin, X., Wei, X., & Zheng, H. (2022). Railway track vibration analysis and intelligent recognition of fastener defects. *Advanced Theory and Simulations*, *5*(10), 2200027.

[2] Huang, Y., Wang, J., Le, W., Zhang, L., & Su, J. (2021). Study on mechanical behaviours of rail fasteners and effects on seismic performance of urban rail viaduct. *Structures*, *33*, 3822–3834.

[3] Zhang, Y., Yang, X., & Liu, S. (2024). Design and parameters influence analysis of dynamic vibration absorber for fastener clips in high-speed railway. *Journal of Vibration and Control*, *30*(3–4), 472–486.

[4] Wei, J., Liu, C., Ren, T., Liu, H., & Zhou, W. (2017). Online condition monitoring of a rail fastening system on high-speed railways based on wavelet packet analysis. *Sensors*, *17*(2), 318.

[5] Chen, M., Zhai, W., Zhu, S., Xu, L., & Sun, Y. (2022). Vibration-based damage detection of rail fastener using fully convolutional networks. *Vehicle System Dynamics*, *60*(7), 2191–2210.

[6] Germoglio Barbosa, I., Lima, A. D. O., Edwards, J. R., & Dersch, M. S. (2024). Development of track component health indices using image-based railway track inspection data. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, *238*(6), 706–716.

[7] Wang, N., Jia, L., Qin, Y., Yao, D., Yang, J., & Wang, Z. (2025). Dual-stage manifold preserving mixed supervised learning for bogie fault diagnosis under variable conditions. *Engineering Applications of Artificial Intelligence*, *149*, 110512.

[8] Bai, T., Yang, J., Xu, G., & Yao, D. (2021). An optimized railway fastener detection method based on modified Faster R-CNN. *Measurement*, *182*, 109742.

[9] Cai, Y., He, M., Tao, Q., Xia, J., Zhong, F., & Zhou, H. (2024). Fast rail fastener screw detection for vision-based fastener screw maintenance robot using deep learning. *Applied Sciences*, *14*(9), 3716.

[10] Cai, Y., He, M., & Chen, B. (2024). Efficient deep learning based rail fastener screw detection method for fastener screw maintenance robot under complex lighting conditions. *Scientific Reports*, *14*(1), 28957.

[11] Liu, J., Liu, H., Chakraborty, C., Yu, K., Shao, X., & Ma, Z. (2021). Cascade learning embedded vision inspection of rail fastener by using a fault detection IoT vehicle. *IEEE Internet of Things Journal*, *10*(4), 3006–3017.

[12] Ye, W., Ren, J., Lu, C., Zhang, A. A., Zhan, Y., & Liu, J. (2024). Intelligent detection of fastener defects in ballastless tracks based on deep learning. *Automation in Construction*, *159*, 105280.

[13] Song, W., Liao, B., Ning, K., & Yan, X. (2024). Improved real-time detection transformer-based rail fastener defect detection algorithm. *Mathematics*, *12*(21), 2227-7390.

[14] Ding, Q., Li, W., Xu, C., Zhang, M., Sheng, C., He, M., & Shan, N. (2024). GMS-YOLO: An algorithm for multi-scale object detection in complex environments in confined compartments. *Sensors*, *24*(17), 5789.

[15] Wang, S. B., Gao, Z. M., Jin, D. H., Gong, S. M., Peng, G. L., & Yang, Z. J. (2024). AMEA-YOLO: a lightweight remote sensing vehicle detection algorithm based on attention mechanism and efficient architecture: S.-B. Wang et al. *The Journal of supercomputing*, *80*(8), 11241-11260.

[16] Ren, B., Xu, Z., Zhao, J., Hao, R., & Zhang, J. (2025). Complex dark environment-oriented object detection method based on YOLO-AS. *Scientific Reports*, *15*(1), 21873.

[17] Lyu, Z., & An, W. (2024). HDR-YOLO: Adaptive object detection in haze, dark, and rain scenes based on YOLO. *International Journal of Pattern Recognition and Artificial Intelligence*, *38*(5), 2450006.

[18] Li, C., Song, S., Gao, P., Huang, W., & Bi, L. (2024). EDMFEN: Edge detection-based multi-scale feature enhancement network for low-light image enhancement. *KSII Transactions on Internet and Information Systems (TIIS)*, *18*(4), 980–997.

[19] Liu, H., Zhang, Z., & Feng, B. (2026). POP-YOLOv8: an object detection framework for partially occluded pedestrians in nighttime traffic environments. Scientific Reports, 16(1), 4841.

[20] Li, L., Ren, Z., Tang, C., Lu, S., & Liang, Y. (2025). A deep learning-based detection model and illumination-adaptive behavioral analysis for soldier crabs in the intertidal zone. *Ecological Informatics*, 103338.

[21] Huang, H., Liu, H., Wang, H., Yang, Y., Guo, C., Chen, M., & Han, K. (2025, December). A Cascaded Enhancement-Fusion Network for Visible-Infrared Imaging in Darkness. In *Photonics* (Vol. 12, No. 12, p. 1231). MDPI.

[22] Fu, Y., Chen, Q., & Zhao, H. (2022). CGFNet: cross-guided fusion network for RGB-thermal semantic segmentation. The Visual Computer, 38(9), 3243-3252.

[23] Fu, J., Chen, X., & Lv, Z. (2022). Rail fastener status detection based on MobileNet-YOLOv4. *Electronics*, *11*(22), 3677.

[24] Li, L., Lv, Z., Chen, X., Qiu, Y., Li, L., Ma, A., Zheng, S., & Chai, X. (2021). Research on track fastener positioning method based on local unidirectional template matching. *Science progress*, *104*(2), 368504211026131.

[25] Peng, Y., Li, H., Wu, P., Zhang, Y., Sun, X., & Wu, F. (2024). D-FINE: Redefine regression task in DETRs as fine-grained distribution refinement. *arXiv preprint arXiv:2410.13842*.

[26] Liu, S., Huang, D., & Wang, Y. (2019). Learning spatial fusion for single-shot object detection. *arXiv preprint arXiv:1911.09516*.

[27] Li, J., Wen, Y., & He, L. (2023). Scconv: Spatial and channel reconstruction convolution for feature redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6153–6162).

[28] Li, Y., Yao, T., Pan, Y., & Mei, T. (2022). Contextual transformer networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(2), 1489–1500.

[29] Draelos, R. L., & Carin, L. (2022). Explainable multiple abnormality classification of chest CT volumes. *Artificial Intelligence in Medicine*, *132*, 102372.