Structural Durability & Health Monitoring

Tech Science Press

https://doi.org/10.32604/sdhm.2026.077736

ARTICLE

A Multimodal Defect Detection Method for Key Components of Rail Transit Systems

Haoyu Li[1], Jiayi Wang[1], Zhaoyu Wu[1], Shuo Yan[1], Ziqi Zhang[1], Yang Gao[2,3], Genwang Peng[2,3] and Zhiwei Cao[2,*]

[1]The Third Operation Branch Company Affiliated with Beijing Mass Transit Railway Operation Corp., Ltd., Beijing, China
[2]State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University, Beijing, China
[3]School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China

*Corresponding Author: Zhiwei Cao. Email: zhiwei@bjtu.edu.cn

Received: 16 December 2025; Accepted: 06 February 2026; Published: 18 May 2026

ABSTRACT: Key components of rail transit systems, such as tracks and vehicle bodies, are prone to developing various types and manifestations of defects during long-term operation. These defects not only accelerate component aging and failure but also pose serious threats to train operational safety. Among existing intelligent detection methods, they mostly rely solely on visible light images demonstrate limited robustness in complex scenarios. This limitation stems from their high dependence on ambient lighting conditions, rendering them insufficient to meet practical railway inspection requirements. While mainstream multimodal detection methods incorporate the complementary strengths of heterogeneous data sources, they fail to fully leverage the intrinsic associative value of data across different modalities. Furthermore, the occurrence frequency and sample size of various rail transit defects exhibit significant disparities, resulting in severe sample imbalance across detection datasets. This substantially reduces the detection accuracy for rare defect categories. To address these critical technical challenges, this paper proposes a multimodal defect detection method for key components of rail transit systems. The method introduces a multimodal input architecture by integrating Red, Green, Blue (RGB) visual data with depth geometric data. It incorporates a self-learning deep feature fusion module that enables complementary enhancement and deep coupling of heterogeneous modal information. This is achieved through targeted feature extraction and multi-round interactive fusion across different modalities. Additionally, we propose a weighted composite balanced loss function that employs dynamic adaptive weighting factors to adjust the model optimization direction in real time. This method effectively mitigates training bias caused by sample imbalance and improves detection performance for minority defect classes. Finally, experimental results on the track fastener defect RGBD dataset and the vehicle body fastener defect RGBD dataset demonstrate that the proposed method achieves optimal defect detection accuracy and meets real-time inspection requirements.

KEYWORDS: Track safety; track inspection; defect detection; multimodal data; object detection

1 Introduction

The rail transit system, as the core framework of modern urban transportation, comprises interconnected subsystems, including the track structure, vehicle equipment, signaling control system, and power supply network. Serving as the physical support and guiding foundation for train operation, the track is subjected to periodic vibrations and impacts from dynamic train loads over extended periods. Its structural integrity is easily compromised by the combined effects of environmental corrosion and fatigue, which can lead to typical defects such as loose fasteners and foreign objects on the track, posing risks to the stability

and safety of train operations. As the primary moving component of the rail transit system, the train car body features a highly integrated structural design, incorporating key elements such as the bogie, coupler, and car body frame. Over years of service, key vehicle components gradually develop typical failure modes, including fatigue cracks and connection failures, which directly threaten operational safety. The failure modes of key components in above-ground rail transit systems vary and include, but are not limited to, geometric deformation, spatial displacement, loss of structural integrity, and surface deterioration. If these defects are not identified and addressed promptly, they may be progressively amplified through dynamic coupling effects, significantly increasing the risk of failure. Therefore, rapid and accurate defect detection is becoming increasingly critical.

The diverse forms and various types of defects undoubtedly pose significant challenges to current manual inspection methods. With advancements in artificial intelligence, vision-based intelligent inspection techniques are increasingly being applied on-site to improve inspection efficiency and reduce labor costs. A commonly used approach is supervised deep learning detection based on single data sources. This method automatically extracts and learns key feature information from visible light images through neural networks, enabling the identification and localization of defects, thereby greatly enhancing efficiency. He et al. [1] proposed a fast and accurate fastener defect detection method combining multi-scale Convolutional Neural Networks (CNNs) with Transformers, enabling the localization and identification of fastener defects under complex conditions such as occlusion. Wu et al. [2] proposed an all-in-one YOLO network to realize rail and track fastener defects detection. An et al. [3] proposed a fast defect detection algorithm for fasteners based on T-YOLO and an overlapping reconstruction strategy. Chen et al. [4] proposed a two-stage framework for detecting defects in critical components of moving trains, termed CDDF. This approach first locates large and small train components separately, followed by defect detection for each component. Shaikh et al. [5] employed the Anchor-Free YOLOv8 model with a decoupled head module to identify bogie defects in moving trains. Wang et al. [6] proposed a high-speed train body paint film defect detection framework consisting of three components: reflection interference elimination, defect feature enhancement, and defect detection and classification. However, in complex rail transit environments such as entry/exit tunnels and nighttime operations, lighting conditions vary significantly. The quality and stability of visible light data collected by a single visible light camera are suboptimal. The key defect characteristics it captures are susceptible to interference and masking, or may even be lost, thereby compromising the reliability of detection algorithms.

Structured light cameras, due to their high adaptability to varying illumination conditions, are increasingly employed in industrial and railway applications for high-precision measurement and high-resolution imaging. They are particularly well-suited for measuring complex surfaces and can sensitively detect deformations in an object's surface shape. Structured light cameras enable the simultaneous capture of RGB and depth images with pixel-to-pixel alignment. These images provide surface texture information and spatial relative position data, respectively. Consequently, deep learning-based intelligent defect detection methods utilizing RGBD data have emerged. Wang et al. [7] proposed CLANet, a rail surface defect detection network that leverages fused RGB-D images to identify fastener defects. Gao et al. [8] introduced a fast-fitment anomaly detection method based on data-layer fusion, which combines grayscale images with their corresponding depth maps before inputting them into the network for detection. Ge et al. [9] proposed a railway track anomaly detection method based on decision-level fusion, comprising three components: supervised pre-detection, semi-supervised re-detection, and decision-level fusion, where results from the first two stages are integrated at the decision level. Wang et al. [10] fused RGBD cross-modal information at the feature level for rail surface defect detection. Based on these fusion approaches, intelligent defect detection methods utilizing RGBD data can be categorized into three layers: data layer, feature layer, and

decision layer. Existing fusion methods primarily focus on the data and decision layers, while the relatively few feature-layer fusion approaches often fail to achieve sufficient integration, limiting the enhancement of effective information. Moreover, current research on multimodal intelligent track inspection predominantly objects surface defects in rails, with limited investigation into defects in other critical track components. Research on vehicle body defects is even more scarce. Furthermore, these systems often lack versatility across different scenarios, hindering their ability to meet the demands of cross-scenario detection.

Additionally, the frequency and probability of defects occurring in key components of rail transit systems vary, resulting in differences in the volume of defect data collected across various categories. Consequently, datasets are prone to the long-tail effect, and improvements in detection accuracy for classes with few samples are limited by the available data volume. To address this issue, researchers have proposed various approaches. Tu et al. [11] achieved defect detection under category imbalance by analyzing the geometric morphology of track fasteners and the length of connected domains. A more promising approach involves refining the loss function, as it directly influences the network's learning direction. Effective adjustments can encourage the network to prioritize learning from categories with fewer samples, thereby enhancing detection accuracy. Qiu et al. [12] proposed a Center-Triplet loss function for fastener detection in track defects. An et al. [3] and Li et al. [13] introduced Quality Focal Loss to address accuracy shortcomings caused by sample imbalance. Existing research has primarily focused on optimizing a single loss function. While this approach can mitigate class imbalance, it has not fully leveraged the complementary strengths of different loss functions.

To address the immature fusion mechanisms in existing multimodal feature fusion networks and the common issue of sample imbalance in datasets, we propose a MultiModal Defect Detection (MMDD) method specifically designed for critical components in rail transit systems. This approach enables precise, real-time detection of defects in track fasteners and vehicle body fasteners across diverse scenarios. In this paper, we introduce an innovative deep fusion module for multimodal features that effectively achieves adaptive integration of semantic and spatial information from heterogeneous modalities. Additionally, we enhance the classification loss function to improve detection accuracy for sparsely represented categories affected by sample imbalance. The specific contributions of this paper are as follows.

(1) We propose a multimodal architecture for detecting defects in critical components of rail transit systems. By utilizing visible light images and depth maps as inputs, this approach enables precise, real-time detection and localization of defects in track fasteners and vehicle body fasteners.
(2) We propose a self-learning deep feature fusion module that effectively achieves multimodal feature complementarity and information gain by establishing a modality-specific feature extraction paradigm combined with a dynamic and adaptive fusion mechanism.
(3) We propose a weighted composite equilibrium loss function that effectively balances model attention by using weighting factors to emphasize underrepresented categories and explicitly adjusts the optimal decision boundary, thereby mitigating accuracy degradation caused by sample imbalance.

2 Method

2.1 Overview

This paper proposes a framework for detecting defects in key components of rail transit systems based on RGBD feature-level fusion. The overall architecture is illustrated in Fig. 1. This framework adopts YOLOv9s [14] as its baseline architecture and first constructs a dual-stream backbone network to enable parallel extraction of multimodal features. By capturing rich semantic representations from RGB images and geometric structural information from depth images through independent encoding branches. Subsequently,

we designed the Self-learning Deep Feature Fusion Module. This module establishes a cross-modal interaction mechanism specifically designed for multimodal heterogeneous characteristics, enabling the efficient deep fusion of complementary features across different modalities. To address the long-tail distribution issue in rail transit scenarios—where normal samples and abnormal samples exhibit extreme numerical imbalance—we propose the Weighted Composite Equilibrium Loss Function. By dynamically reallocating weights to counteract the effects of skewed data distribution, this approach significantly improves detection accuracy for rare-category defects. Finally, YOLOv9's Neck and Head networks perform multi-scale feature integration and decoding, outputting category labels, bounding box coordinates, and confidence scores for track fasteners and vehicle body fasteners. This enables precise localization and classification of defects.

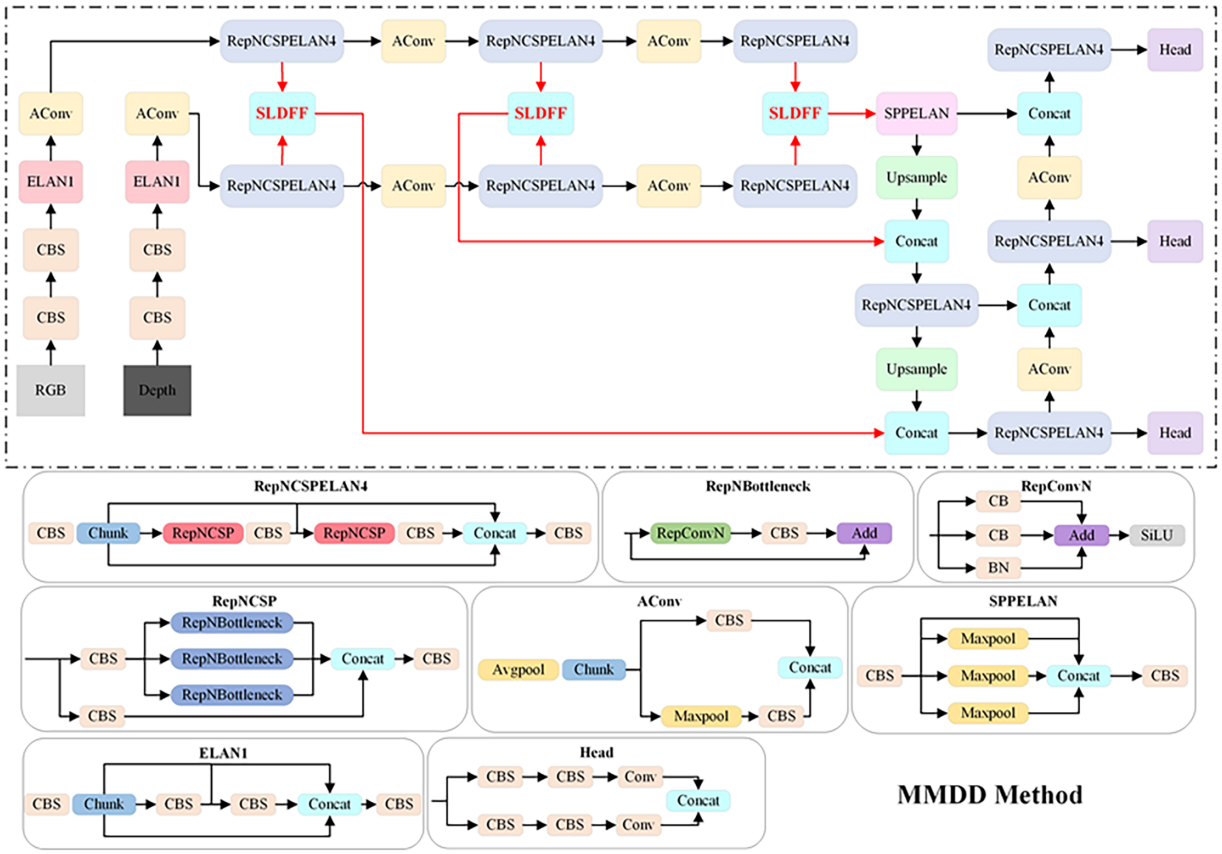

Figure 1: Overview of the MMDD method.

2.2 Self-Learning Deep Feature Fusion Module (SLDFF)

To address the heterogeneity of multimodal data, particularly the complex semantic information in RGB features and the precise spatial structural information in depth features, this paper proposes a self-learning depth feature fusion module. The schematic diagram of this module is shown in Fig. 2. This module designs specialized feature extraction paradigms tailored to each modality and achieves complementary integration of multimodal information through self-learning and adaptive fusion strategies.

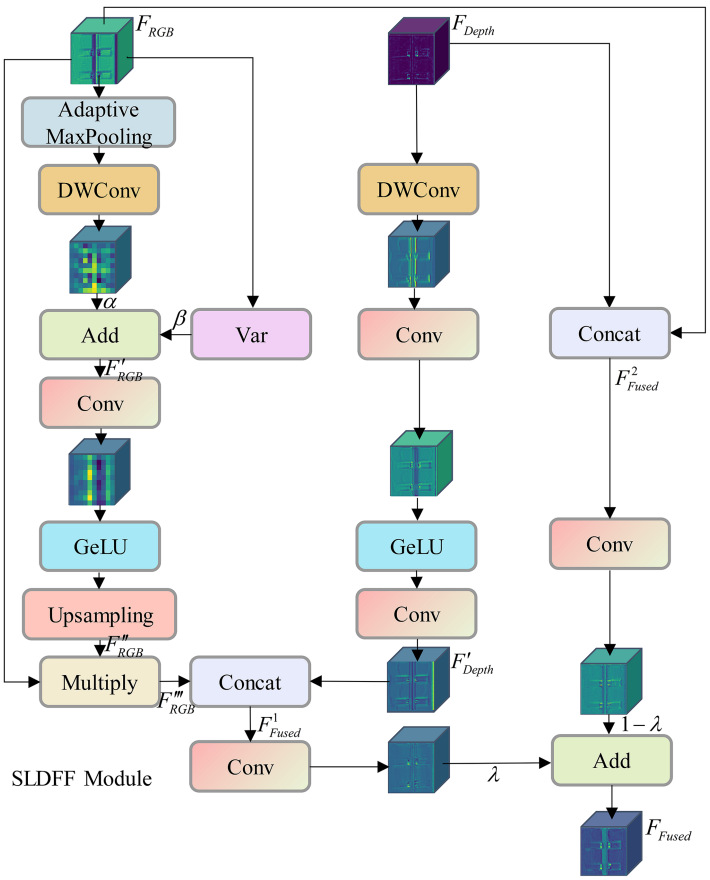

Figure 2: Self-learning deep feature fusion module.

For RGB features rich in semantic information, a dual-branch feature extraction architecture—comprising a variance branch and a main branch—is designed to enhance global context modeling capabilities. Specifically, the variance branch computes statistics that quantify the spatial variance across different channels in the RGB features, measuring the dispersion of pixel values within each channel. High-variance channels typically encode high-frequency salient information, while low-variance channels correspond to semantically similar regions. This variance-based channel importance modeling mechanism guides the network to focus on discriminative features. The main branch integrates adaptive max pooling and dilated convolutions to expand the receptive field while reducing spatial resolution, thereby achieving global feature modeling. To dynamically balance contributions from both branches, learnable scaling weights α and β are introduced. These parameters are continuously updated and optimized through end-to-end training, automatically adjusting the relative contributions of the variance and main branches. Subsequently, further modulation is performed using convolutions and Gaussian Error Linear Unit (GeLU) activation functions, which are then multiplied pixel-by-pixel with the original RGB features to yield representative RGB features. This residual structure mitigates information loss during the feature transformation process. The calculation is shown in Eqs. (1)–(3).

$$F'_{RGB} = \alpha * DWConv_{3\times3}\left(AMP\left(F_{RGB}\right)\right) + \beta * Var\left(F_{RGB}\right) \quad (1)$$

where, $DWConv3\times3(\cdot)$ is a 3 × 3 depth-wise convolutional layer. $AMP(\cdot)$ denotes adaptive max pooling operation. $Var(\cdot)$ represents the variance calculation of the channel. α and β are learnable scaling weights.

$$F''_{RGB} = UP\left(GeLU\left(Conv_{1\times1}\left(F'_{RGB}\right)\right)\right) \tag{2}$$

where, UP refers to upsampling, where nearest-neighbor interpolation is employed in this paper. $Conv1\times1(\cdot)$ denotes a 1 × 1 convolutional layer.

$$F'''_{RGB} = F_{RGB} \cdot F''_{RGB} \tag{3}$$

For deep features incorporating explicit geometric priors, a three-layer Multilayer Perceptron (MLP) architecture was designed that combines separable convolutions with standard convolutions to efficiently model spatial hierarchical information, resulting in representative deep features. Deep convolutions are employed to capture local spatial patterns while minimizing computational overhead. Cross-channel information integration is then achieved through pointwise convolution, enhancing the representation of depth-induced structural constraints. The calculation is shown in Eq. (4).

$$F'_{Depth} = Conv_{1\times1}\left(GeLU\left(Conv_{1\times1}\left(DWConv_{3\times3}\left(F_{Depth}\right)\right)\right)\right) \tag{4}$$

To fuse the modulated RGB features with the depth features, channel concatenation is employed for aggregation, followed by convolutional refinement. Additionally, to extract multi-level contextual information, feature reuse is achieved by concatenating the raw convolutional RGB features with the depth features. By introducing the modulation factor λ, we control the contribution ratio between the modulated features and the original features in the final fusion output. The optimal parameters are determined using a grid search method. The calculation is shown in Eqs. (5)–(7).

$$F^1_{Fused} = Concat\left(F'''_{RGB}, F'_{Depth}\right) \tag{5}$$

$$F^2_{Fused} = Concat\left(F_{RGB}, F_{Depth}\right) \tag{6}$$

$$F_{Fused} = \lambda * F^1_{Fused} + (1-\lambda) * F^2_{Fused} \tag{7}$$

This module achieves adaptive fusion of semantic and spatial information through its unique modeling architecture and self-learning parameters, thereby effectively enhancing the accuracy of defect detection.

2.3 Weighted Composite Equilibrium Loss Function

To address the limitations of single-loss functions in improving sample imbalance and the insufficient exploration of complementary properties in multi-loss functions, this paper proposes a weighted composite balanced loss function. By integrating the strengths of Quality Focal Loss [15] and LDAM loss [16], it aims to simultaneously tackle sample imbalance and the issue of blurred classification boundaries for sparse categories. This approach enhances the accuracy and robustness of defect detection for critical components in rail transit systems.

A typical loss function for class imbalance is Focal Loss [17], which is an extension of BCE Loss that introduces a scaling factor to adjust for sample quantity and difficulty. However, this loss function is only applicable to discrete labels and cannot directly handle continuous labels in joint classification-quality representations. Quality Focal Loss extends traditional Focal Loss from discrete labels to continuous labels. By dynamically modulating the loss contribution of easily classifiable samples, it enables the model to focus on samples from underrepresented categories. It expands the cross-entropy component to its full form and

introduces the absolute difference between predicted and actual values as a scaling factor. The calculation is shown in Eq. (8).

$$L_{QFL}(p, y) = -|y - p|^{\beta} (y \log(p) + (1 - y) \log(1 - p)) \tag{8}$$

where, y denotes the continuous label, p represents the predicted probability, and β serves as the modulation factor.

In imbalanced datasets, models often struggle to adequately learn the features of the minority class, leading to decision boundaries that favor the majority class. The core idea of LDAM Loss is to introduce stronger regularization for the minority class than for the majority class during training. By allocating a larger classification margin to the sparsely represented category, it forces the model to learn a more robust classification boundary. This enables the model to capture finer distinctions between categories, addressing the issue of sparsely represented categories being overwhelmed in long-tail distributions. The calculation is shown in Eq. (9).

$$L_{LDAM}(p, y) = -\log\left(e^{p_y - \gamma_y} / \left(e^{p_y + \gamma_y} + \sum_{j \neq y} e^{p_j}\right)\right) \tag{9}$$

where, for category y, the classification boundary γ_y is defined as the minimum distance from all samples in that category to the decision boundary. Through derivation, the classification boundary for category y is inversely proportional to the sample size n_y. Therefore, categories with fewer samples should be assigned a larger classification boundary. The calculation is shown in Eq. (10).

$$\gamma_y = C / n_y^{1/4} \tag{10}$$

where, C is a hyperparameter.

The proposed weighted composite loss function combines the continuous label optimization of Quality Focal Loss with the optimal classification boundary mechanism of LDAM Loss. By leveraging dynamic weight allocation, we fully exploit the complementary nature of loss functions with distinct emphases, simultaneously adjusting model attention while explicitly optimizing the decision boundary. This approach provides an efficient and versatile solution for dense detection tasks under long-tail distributions. The calculation is shown in Eq. (11).

$$L_{cls} = \varepsilon \cdot L_{QFL} + (1 - \varepsilon) \cdot L_{LDAM} \tag{11}$$

where, ε is a hyperparameter that balances the weights of both components and is determined through a grid search method.

In the regression task, to optimize bounding box localization accuracy and accelerate model convergence, this paper employs the CIOU loss function [18] as the computational criterion for regression loss. This function comprehensively evaluates geometric constraints, including the normalized center-to-center distance between predicted and ground-truth bounding boxes, the consistency of aspect ratios, and the overlap area. It effectively addresses the issue of traditional IoU metrics being insensitive to positional shifts. Consequently, it enables more efficient gradient propagation and more precise spatial localization during training. The calculation is shown in Eq. (12).

$$L_{box} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha \cdot v \tag{12}$$

where, IoU refers to the intersection-over-union ratio between the predicted bounding box and the ground truth bounding box. b and b^{gt} represent the center coordinates of the predicted box and the ground truth box, respectively, and denote the squared Euclidean distance. c is the diagonal length of the minimum bounding rectangle encompassing both boxes, v measures the consistency of aspect ratios, and α serves as the balance coefficient.

This study introduces the Distribution Focal Loss mechanism. This loss function enhances the model's ability to refine the probability distribution of bounding boxes by dynamically adjusting the weight distribution across different discrete position intervals. This approach effectively improves the accuracy and robustness of coordinate regression for occluded targets and ambiguously located objects in complex scenes. The calculation is shown in Eq. (13).

$$L_{dfl} = -\sum_{i=y_l}^{y_r} \left[\left(1 - \frac{y - y_l}{y_r - y_l} \right) \log\left(p_i\right) + \frac{y - y_l}{y_r - y_l} \log\left(p_{i+1}\right) \right] \tag{13}$$

where, y represents the true label, y_l and y_r denote the two integer coordinates closest to y, while p_i and p_{i+1} correspond to the probabilities predicted by the model at y_l and y_r.

The overall loss function of the network is composed of a weighted sum of three components. The calculation is shown in Eq. (14).

$$L_{total} = \lambda_{cls} L_{cls} + \lambda_{box} L_{box} + \lambda_{dfl} L_{dfl} \tag{14}$$

where, λ_{cls}, λ_{box}, and λ_{dfl} are 0.5, 7.5, and 1.5, respectively. These values represent the default configuration for the YOLOv9 model. They prioritize localization accuracy over classification because precise bounding boxes are more critical for detecting small defects, while classification is generally easier to learn. Using the same settings ensures fair and consistent comparisons with baseline models while maintaining training stability and detection performance.

3 Experiment

3.1 Dataset

This paper constructs a track fastener defect RGBD dataset and a vehicle body fastener defect RGBD dataset. The rail fastener defect RGBD dataset is collected by a rail inspection cart equipped with a 3D structured light camera, operating on a 60-m track in the laboratory. The laboratory illuminance is controlled at the level of a normal sunny day, with various defect types simulated on the track. The camera is positioned 80 cm above the track surface. The inspection vehicle moves at a speed of 20 km/h, with the camera continuously capturing images of track defects as it travels alongside the vehicle. The vehicle body fastener defect RGBD dataset is collected by maintenance personnel using 3D structured light cameras, capturing fasteners on the roofs of active trains operated by a subway company. The camera is mounted on a gantry spanning the tracks. Images are captured synchronously as trains pass through the gantry at conventional low speeds (5–15 km/h) under outdoor natural daylight conditions. The dataset annotation is conducted by railway and locomotive maintenance personnel with field experience, following detailed annotation guidelines. These guidelines define clear standards for each defect type, establish protocols for handling ambiguous or occluded cases. To maintain quality, each image is independently annotated by two annotators, with discrepancies resolved by a third staff member. Additionally, 15% of the final annotations are randomly selected for review by external inspectors, confirming high annotation reliability.

The track fastener defect RGBD dataset contains 1411 pairs of matched grayscale images and depth images, each with an original resolution of 2048 × 2048 pixels. The dataset is randomly split into training,

validation, and test sets in an 8:1:1 ratio. This dataset contains 13 types of track defects. Fig. 3 shows a set of images from the dataset and the 13 defect categories. The number of samples for each type of defect in the dataset is shown in Fig. 4. The number of normal fastener samples is much higher than that of defective samples, exhibiting a long-tail distribution, and the distribution of defective samples across different categories is relatively uneven.

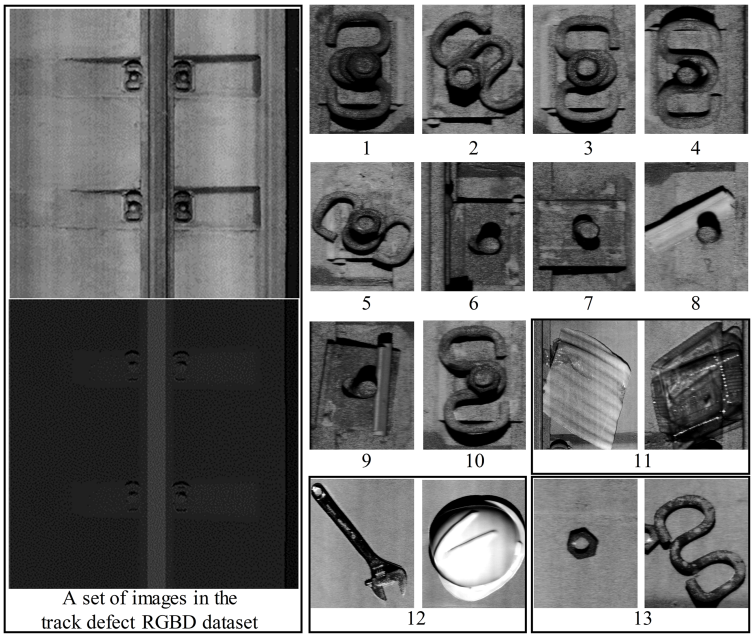

Figure 3: A set of images from the track fastener defect RGBD dataset and the defect categories in the dataset. From 1 to 13, they are: fastener, nut loose, nut miss, plain washer miss, elastic strip loose, elastic strip miss, plate loose, plate miss, baffle loose, baffle miss, foreign matter, tool, component.

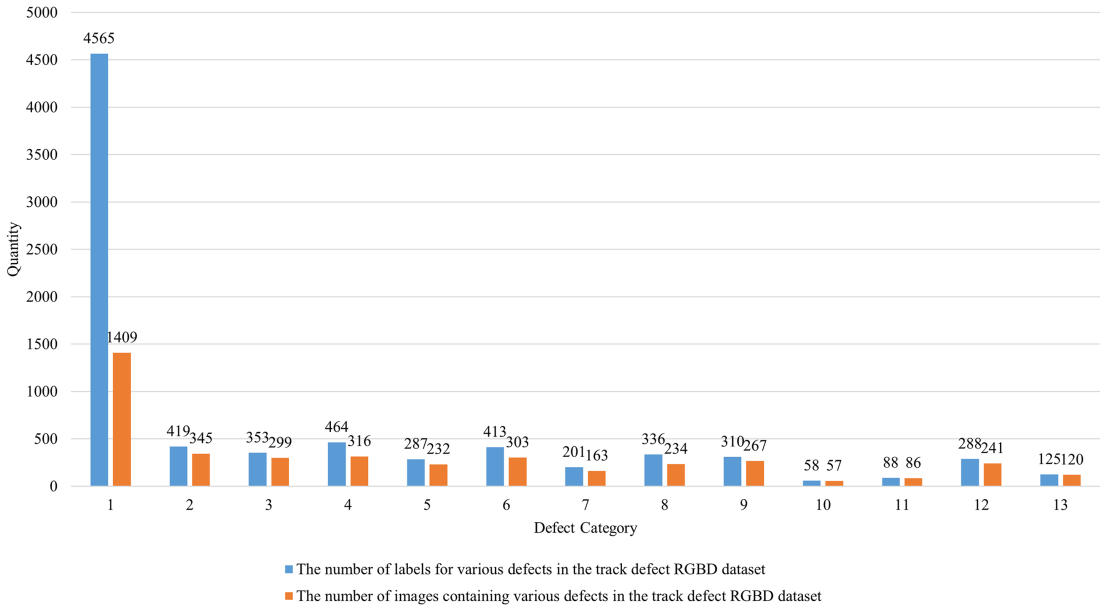

Figure 4: Statistical chart of the number of various defect samples in the track fastener defect RGBD dataset.

The vehicle body fastener defect RGBD dataset contains 860 pairs of matched grayscale and depth images, each with a resolution of 512 × 1300 pixels. The dataset is randomly split into training, validation, and test sets in an 8:1:1 ratio. The defect category in the dataset is loose fasteners on the roof panel, as shown in Fig. 5. The dataset contains a total of 869 samples.

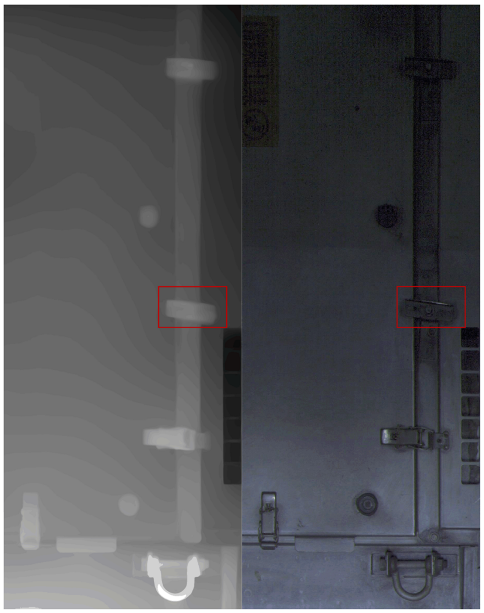

Figure 5: A set of images from the vehicle body fastener defect RGBD dataset.

3.2 Implementation Details

This algorithm is based on the PyTorch framework, with CUDA version 11.3. For the track fastener defect RGBD dataset, all algorithms in this paper are trained for 250 epochs with a batch size of 32. During training, input images are uniformly resized to 640 × 640 pixels. The momentum coefficient is set to 0.937, and the weight decay coefficient is set to 0.0005. The initial learning rate is set to 0.01 and decayed to 0.0001 via cosine annealing over the training epochs. A 3 epoch linear warm-up is applied at the beginning to gradually ramp up the learning rate from zero. Using the SGD optimizer. For the vehicle body fastener defect RGBD dataset, all algorithms in this paper are trained for 200 epochs. All other parameter settings remain consistent with those used during training on the track fastener defect RGBD dataset. All algorithms are trained and tested using the NVIDIA GeForce RTX 4090 GPU.

3.3 Evaluation Metrics

This paper evaluates the model's detection accuracy using five metrics: Precision, Recall, F1-Score, mAP0.5, and mAP0.5:0.95. F1-score is the harmonic mean of precision and recall, serving as a comprehensive measure of a model's accuracy and completeness in identifying positive samples during detection tasks. The mAP0.5 denotes the mAP (mean Average Precision) at an IoU (Intersection over Union) threshold of 0.5, while mAP0.5:0.95 is the average of mAP values computed over IoU thresholds from 0.5 to 0.95 in steps of 0.05, providing a more stringent assessment of localization accuracy. The calculation methods are shown in Eqs. (15)–(19).

$$P = \frac{TP}{TP + FP} \tag{15}$$

$$R = \frac{TP}{TP + FN} \tag{16}$$

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{17}$$

$$AP = \int_1^0 P(R)\, d(R) \tag{18}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{19}$$

where, TP represents true positives, FP represents false positives, TN represents true negatives, FN represents false negatives, and n represents the number of sample categories tested.

This article uses inference speed (FPS) to quantify detection speed, with higher FPS values indicating better real-time performance.

3.4 Experimental Results

3.4.1 Vehicle Body Fastener Defect Detection Results

To evaluate the effectiveness of the proposed method, we selected 12 mainstream object detection methods with similar parameters for qualitative and quantitative comparative evaluation. Specifically, these include YOLOv7 [19], YOLOv8 [20], YOLOv9 [14], YOLOv10 [21], YOLOv11 [22], and YOLOv12 [23], encompassing both single-modal visible light and RGBD-based modalities. The quantitative results are shown in Table 1. MMDD achieved the best performance across all four accuracy metrics, with a precision rate 2.7% higher than the next best model and an mAP0.5 0.9% superior to the runner-up. Inference speed is 64.3 FPS, meeting the requirements for real-time detection. As shown in Table 2, the detection accuracy of MMDD is less affected by training samples, and its detection performance remains relatively stable. Additionally, as shown in Table 3, we conducted five rounds of paired t-tests comparing our method with the baseline for the mAP0.5:0.95 metric. The results demonstrate statistically significant improvements. To visually demonstrate the detection performance of each algorithm, we have compared the visualizations of the algorithms with superior performance, as shown in Fig. 6. It can be seen that most methods produce false negatives and false positives. In contrast, the method proposed in this paper achieves high accuracy and recall while maintaining a high detection confidence level, which further validates the reliability of the algorithm.

Table 1: Quantitative comparative evaluation of methods performance on vehicle body fastener defect RGBD dataset.

Model	Size	Precision (%)	Recall (%)	mAP0.5 (%)	mAP0.5:0.95 (%)	F1-Score (%)	Params (M)	FPS
YOLOV7t	640	86.6	87.9	93.3	68.5	87.3	<u>8.39</u>	<u>225.7</u>
YOLOV8s	640	83.1	93.6	93.0	69.2	88.0	11.1	**263.7**
YOLOV9s	640	88.4	87.3	93.8	68.7	87.9	7.17	97.9
YOLOV10s	640	86.2	90.5	94.5	70.1	88.3	**8.04**	165.3
YOLO11s	640	<u>89.0</u>	88.5	94.5	<u>71.5</u>	<u>88.8</u>	9.42	203.5
YOLOV12s	640	83.0	92.7	92.9	69.5	87.6	9.24	123.5
YOLOV7t-RGBD	640	82.8	87.7	93.1	68.8	85.2	11.4	167.2
YOLOV8s-RGBD	640	82.2	92.7	93.9	68.3	87.1	16.5	188.4

(Continued)

Table 1 (continued)

Model	Size	Precision (%)	Recall (%)	mAP0.5 (%)	mAP0.5:0.95 (%)	F1-Score (%)	Params (M)	FPS
YOLOV9s-RGBD	640	88.9	87.2	94.2	70.5	88.0	9.86	71.7
YOLOV10s-RGBD	640	80.1	<u>95.3</u>	<u>94.7</u>	70.9	87.0	11.3	133.1
YOLO11s-RGBD	640	82.5	92.7	93.9	70.0	87.3	14.2	125.7
YOLO12s-RGBD	640	82.2	90.0	93.7	69.8	85.9	15.0	87.9
MMDD	640	**91.7**	**95.7**	**95.6**	**71.7**	**93.7**	11.8	64.3
Improvement(%)	–	+3.0	+0.4	+0.9	+0.2	+5.5	−31.8	−75.6

Note: The bold values indicate the best results, and the underlined values denote the second-best results.

Table 2: Statistical analysis of detection accuracy for the MMDD Method on vehicle body fastener defect RGBD dataset (%).

Sample	Precision	Recall	mAP0.5	mAP0.5:0.95	F1-score
1	91.7	95.7	95.6	71.7	93.7
2	91.6	95.5	95.7	71.6	93.5
3	91.4	95.6	95.3	71.5	93.5
4	91.5	95.5	95.7	71.4	93.5
5	91.7	95.6	95.5	71.5	93.6
Mean	91.58	95.58	95.56	71.54	93.56
Var	0.017	0.007	0.028	0.130	0.008
Std	0.130	0.084	0.167	0.114	0.089
95% CI	[91.42, 91.74]	[95.48, 95.68]	[95.35, 95.77]	[71.40, 71.68]	[93.45, 93.67]

Table 3: Experimental results and *p*-values for MMDD vs. Baseline Across Five Runs on vehicle body fastener defect RGBD dataset.

Method	Sample 1 (%)	Sample 2 (%)	Sample 3 (%)	Sample 4 (%)	Sample 5 (%)	Mean (%)	*p*-value
Baseline	70.5	70.1	70.8	70.5	70.3	70.44	–
MMDD	**71.7**	**71.6**	**71.5**	**71.4**	**71.5**	**71.54**	$\mathbf{6 \times 10^{-4}}$

3.4.2 Track Fastener Defect Detection Results

Similarly, we also present the quantitative comparison results for track fastener defect detection, as shown in Table 4. MMDD achieved the highest recall rate, which is crucial in railway scenarios. This indicates that the proposed algorithm achieves the lowest false-negative rate, thereby minimizing safety risks associated with missed detections to the greatest extent. Meanwhile, MMDD achieved the optimal mAP0.5:0.95, surpassing the next-best performer by 2.4%. Although it did not achieve optimal performance on precision and mAP0.5 metrics, the results still reached state-of-the-art levels. This also demonstrates that the proposed algorithm achieves superior detection performance across multiple datasets of critical components in urban rail transit systems under diverse scenarios. As shown in Table 5, the detection performance of MMDD remains stable in track defect detection. As shown in Table 6, we conducted five

rounds of paired *t*-tests for the mAP0.5:0.95 metric. The *p*-value was significantly less than 0.01, indicating that the performance improvement of this method over the baseline method is statistically significant.

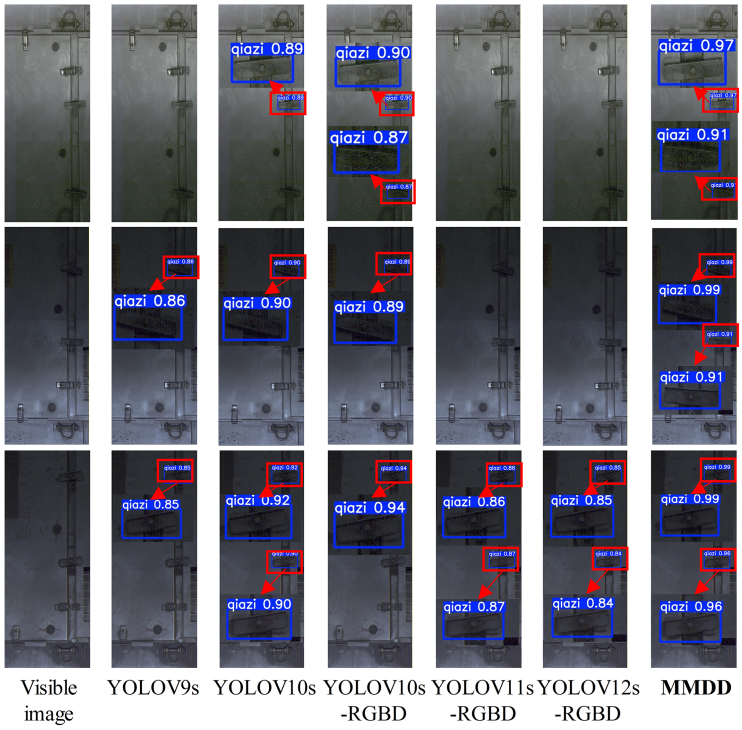

Figure 6: Visualization results of the state-of-the-art methods and MMDD method on vehicle body fastener defect RGBD dataset.

Table 4: Quantitative comparative evaluation of methods performance on track fastener defect RGBD dataset.

Model	Size	Precision (%)	Recall (%)	mAP0.5 (%)	mAP0.5:0.95 (%)	F1-Score(%)	FPS
YOLOV7t	640	84.8	84.2	87.6	62.4	84.5	**175.2**
YOLOV8s	640	<u>88.5</u>	87.2	**92.4**	66.8	87.8	47.37
YOLOV9s	640	85.3	85.6	87.8	65.1	85.4	78.5
YOLOV10s	640	78.0	78.0	82.8	59.6	78.0	124.8
YOLO11s	640	81.1	77.1	79.6	57.9	79.0	<u>162.5</u>
YOLOV12s	640	85.5	85.5	88.3	62.1	85.5	101.5
YOLOV7t-RGBD	640	85.1	79.5	81.7	60.4	82.2	114.2
YOLOV8s-RGBD	640	84.6	<u>89.4</u>	91.2	66.9	86.9	137.4
YOLOV9s-RGBD	640	**89.3**	86.9	91.1	66.6	<u>88.1</u>	65.3
YOLOV10s-RGBD	640	75.6	78.0	83.1	60.0	76.8	124.1
YOLO11s-RGBD	640	87.9	85.5	90.4	<u>67.2</u>	86.7	104.9
YOLO12s-RGBD	640	86.5	86.3	90.6	65.3	86.4	55.6
MMDD	640	87.5	**89.8**	<u>91.9</u>	**69.6**	**88.6**	52.8
Improvement(%)	–	−2.0	+0.4	−0.5	+3.5	+0.6	−69.9

Note: The bold values indicate the best results, and the underlined values denote the second-best results.

Table 5: Statistical analysis of detection accuracy for the MMDD Method on track fastener defect RGBD dataset (%).

Sample	Precision	Recall	mAP0.5	mAP0.5:0.95	F1-Score
1	87.5	89.8	91.9	69.6	88.6
2	87.6	89.5	91.8	69.4	88.5
3	87.6	89.9	92.0	69.4	88.7
4	87.5	89.6	91.7	69.6	88.5
5	87.2	89.6	91.8	69.5	88.4
Mean	87.48	89.68	91.84	69.50	88.54
Var	0.027	0.027	0.013	0.010	0.013
Std	0.164	0.164	0.114	0.100	0.114
95% CI	[87.28, 87.68]	[89.48, 89.88]	[91.70, 91.98]	[69.38, 69.62]	[88.40, 88.68]

Table 6: Experimental results and *p*-values for MMDD vs. Baseline Across Five Runs on track fastener defect RGBD dataset.

Method	Sample 1(%)	Sample 2(%)	Sample 3(%)	Sample 4(%)	Sample 5(%)	Mean(%)	*p*-value
Baseline	66.6	66.6	66.3	66.8	66.5	66.56	–
MMDD	**69.6**	**69.4**	**69.4**	**69.6**	**69.5**	**69.50**	$\mathbf{5 \times 10^{-5}}$

The following visualization results are presented to verify the superiority of the MMDD method, as shown in Fig. 7. In the first row of images, only MMDD and YOLOv8s-RGBD successfully detected defects in category 3, with MMDD providing higher confidence scores. In the second row of images, MMDD accurately identified all defects present. The third row of images contains five categories of rare and difficult-to-detect defects: categories 4, 6, 8, 9, and 10. The results indicate that, compared to other methods, only MMDD achieved complete detection of all five defect categories, while other methods exhibited missed detections. These results fully demonstrate that MMDD exhibits superior detection capabilities and robustness when confronted with the challenges of sparse and difficult samples.

3.5 Melting Experiment

3.5.1 Effect of Different Adjustment Factor in SLDFF Module

The core of the feature fusion module designed in this paper lies in the synergistic interaction between modulated features and original features, with their contribution levels regulated by a modulation factor. Therefore, we designed hyperparameter configurations with different adjustment factors based on the grid search method to investigate their quantitative impact on the final detection performance. As shown in Table 7 for the vehicle body fastener defect RGBD dataset, performance is generally significantly superior when the proportion of modulated features exceeds that of the original features. This may be due to the defect category in the vehicle body fastener defect data being classified as positional tilt, with no significant changes observed in the edge profile. Therefore, modulation characteristics carrying deep semantic information provide crucial insights for detecting this defect. When the modulation factor is 0.9, the performance achieves overall optimality. On the track fastener defect RGBD dataset, as shown in Table 8, performance is superior when the contribution of modulated features is relatively close to that of raw features, i.e., when the difference falls within [0–0.4], corresponding to a modulation factor ranging from 0.7 to 0.3. This is because the raw features of track fastener defect data contain relatively distinct edge contour information,

which is enhanced information common to both RGB and depth modalities. Modulation features contain rich semantic information, and both play an irreplaceable role in the final detection process.

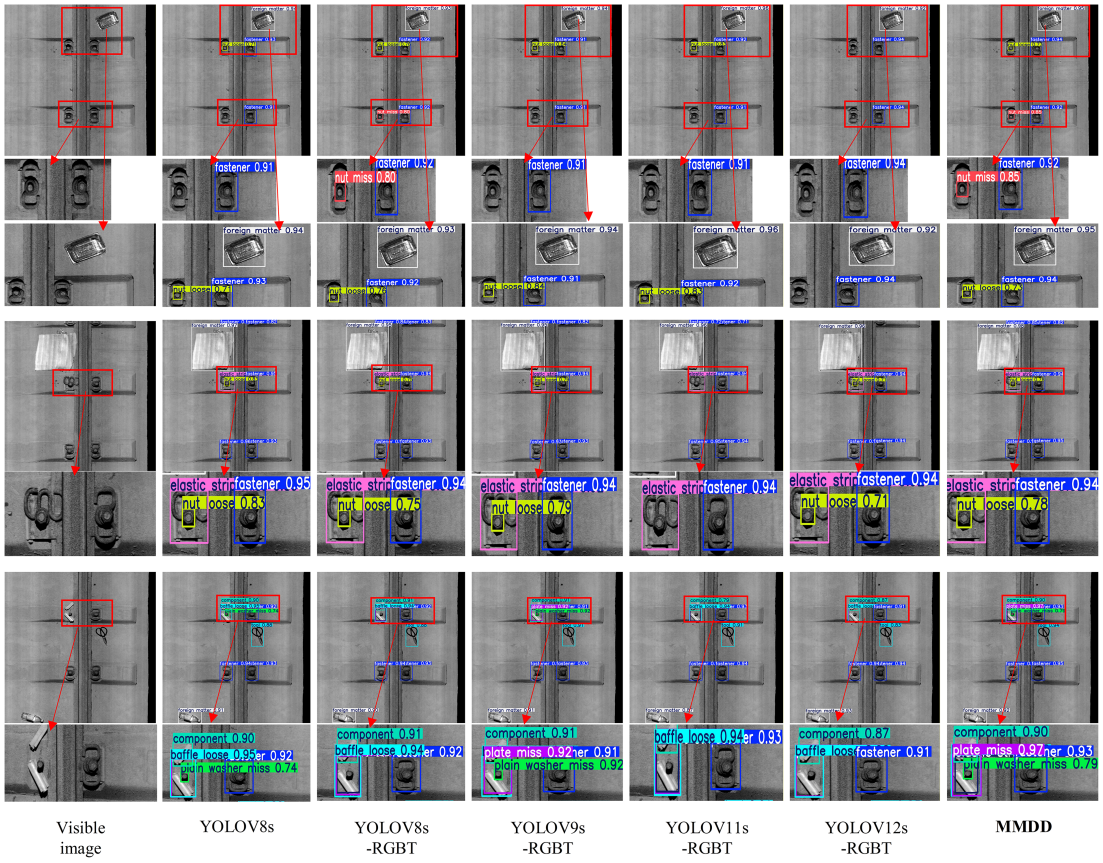

Figure 7: Visualization results of the state-of-the-art methods and MMDD method on track fastener defects RGBD dataset.

Table 7: Quantitative evaluation results of different adjustment factor in SLDFF module on vehicle body fastener defect RGBD dataset (%).

Adjustment Factor	Precision	Recall	mAP0.5	mAP0.5:0.95	F1-Score
(1, 0)	<u>88.3</u>	**97.3**	**96.0**	<u>71.3</u>	<u>92.6</u>
(0.9, 0.1)	**91.7**	95.7	<u>95.6</u>	**71.7**	**93.7**
(0.8, 0.2)	83.3	<u>96.4</u>	94.3	70.8	89.4
(0.7, 0.3)	86.8	90.9	94.0	69.7	88.8
(0.6, 0.4)	82.2	90.9	93.5	69.2	86.3
(0.5, 0.5)	84.0	92.7	93.4	69.1	88.1
(0.4, 0.6)	86.4	87.3	93.3	70.3	86.8
(0.3, 0.7)	83.0	90.0	93.3	70.0	86.4
(0.2, 0.8)	84.5	90.0	92.9	69.8	87.2
(0.1, 0.9)	82.5	90.9	93.3	68.3	86.5
(0, 1)	82.8	89.1	93.2	68.2	85.8

Note: The bold values indicate the best results, and the underlined values denote the second-best results.

Table 8: Quantitative evaluation results of different adjustment factor in SLDFF module on track fastener defect RGBD dataset (%).

Adjustment Factor	Precision	Recall	mAP0.5	mAP0.5:0.95	F1-Score
(1, 0)	61.0	85.0	74.9	57.3	71.0
(0.9, 0.1)	85.4	86.9	89.9	64.7	86.1
(0.8, 0.2)	78.3	**93.0**	86.2	63.0	85.0
(0.7, 0.3)	<u>86.9</u>	88.5	90.4	**67.6**	87.7
(0.6, 0.4)	**88.2**	86.5	**91.3**	67.0	87.3
(0.5, 0.5)	85.3	90.4	90.9	66.5	<u>87.8</u>
(0.4, 0.6)	86.7	89.1	90.6	<u>67.1</u>	**87.9**
(0.3, 0.7)	76.0	<u>91.0</u>	85.6	61.5	82.8
(0.2, 0.8)	85.1	89.3	<u>91.2</u>	67.0	87.1
(0.1, 0.9)	85.5	89.3	90.7	66.9	87.4
(0, 1)	85.0	88.3	90.5	66.6	86.6

Note: The bold values indicate the best results, and the underlined values denote the second-best results.

3.5.2 Effect of Different of Loss Function

This paper proposes a loss function to improve the accuracy of detecting rare defect categories in scenarios where data samples are imbalanced. Specifically, these include BCE Loss, Quality Focal Loss, and Varifocal Loss [24], which do not account for sample imbalance factors. The results for the four precision metrics are shown in Table 9. It can be seen that the proposed loss function achieved the best performance across all four metrics. It can be observed that an inappropriate loss function not only fails to improve detection performance but also leads to performance degradation. To specifically analyze the impact of loss functions on accuracy for classes with few samples, we report Tables 10 and 11. For categories 4, 6, 8, 9, and 10, as shown in Fig. 4, both the number of images and the number of examples are relatively small, making them low-sample categories. Additionally, their features are difficult to extract, classifying them as hard samples. The proposed algorithm achieves significant improvements in both mAP0.5 and mAP0.5:0.95 across these five categories. Taking category 10 as an example, it has the smallest number of samples and exhibits characteristics highly similar to category 1. The proposed loss function not only adjusts model attention based on sample quantity but, more crucially, enables the model to capture finer distinctions between categories, thereby optimizing decision boundaries. Finally, we analyze the contribution of each component to the overall performance improvement through ablation studies. According to the results presented in Table 12, using only the SLDFF module improves mAP0.5:0.95 by 1.0% and recall by 1.6%, indicating that multimodal fusion enhances the model's detection performance. The weighted composite balanced loss contributes most significantly to the overall improvement by effectively boosting the extremely low accuracy of minority categories, thereby substantially increasing the average metrics. Quantitatively, adding this loss function alone improves mAP0.5:0.95 by 1.8% and the F1-score by 0.4% compared to the baseline. When both components are integrated, the complete model achieves a 3.0% increase in mAP0.5:0.95 and a 0.5% improvement in the F1-score over the baseline.

Table 9: Quantitative evaluation results of different loss function on track fastener defect RGBD dataset (%).

Loss Function	Precision	Recall	mAP0.5	mAP0.5:0.95	F1-Score
BCE	86.9	88.5	90.4	67.6	87.7
QFL	80.2	87.8	88.9	65.2	83.8
VFL	82.3	89.5	90.1	66.8	85.7
Ours	**87.5**	**89.8**	**91.9**	**69.6**	**88.6**

Table 10: mAP0.5 of different loss function on track fastener defect RGBD dataset (%).

Loss	all	1	2	3	4	5	6	7	8	9	10	11	12	13
BCE	90.4	**99.4**	92.4	76.4	93.0	**99.0**	95.0	95.5	90.2	72.4	74.5	**96.5**	**95.0**	**95.9**
QFL	88.9	**99.4**	93.1	70.6	94.5	98.3	95.5	92.7	91.5	68.7	69.3	95.5	92.5	93.5
VFL	90.1	**99.4**	**94.9**	**79.9**	92.7	98.7	95.1	**95.9**	86.3	72.3	71.4	95.7	94.8	94.4
Ours	**91.9**	99.3	94.2	78.2	**97.3**	98.4	**96.0**	95.7	**94.3**	**73.7**	**82.5**	96.0	94.8	94.3

Note: The bold values indicate the best results, and the underlined values denote the second-best results.

Table 11: mAP0.5:0.95 of different loss function on track fastener defect RGBD dataset (%).

Loss	all	1	2	3	4	5	6	7	8	9	10	11	12	13
BCE	67.6	94.4	61.0	54.9	46.2	**71.9**	84.5	69.1	53.0	50.4	57.4	**81.7**	81.3	72.2
QFL	65.2	93.0	63.3	47.9	44.5	69.0	85.5	67.9	50.1	45.8	55.2	78.0	78.4	69.6
VFL	66.8	94.3	**65.7**	**57.9**	45.3	71.0	81.2	70.4	50.5	47.5	58.7	60.7	**88.4**	**76.3**
Ours	**69.6**	**94.8**	65.1	57.0	**47.1**	71.6	**88.1**	**70.8**	**55.7**	**50.7**	**65.7**	80.1	84.0	74.5

Note: The bold values indicate the best results, and the underlined values denote the second-best results.

Table 12: Quantitative evaluation results of fusion experiments on track fastener defect RGBD dataset.

Baseline	SLDFF Module	Weighted Composite Equilibrium Loss	Precision (%)	Recall (%)	mAP0.5 (%)	mAP0.5:0.95 (%)	F1-Score (%)	FPS
√			**89.3**	86.9	91.1	66.6	88.1	**65.5**
√	√		86.9	88.5	90.4	67.6	87.7	52.1
√		√	87.9	89.1	90.8	68.4	88.5	64.9
√	√	√	87.5	**89.8**	**91.9**	**69.6**	**88.6**	52.8

Note: The bold values indicate the best results, and the underlined values denote the second-best results.

4 Conclusion

This paper proposes a multimodal defect detection method for key components of rail transit systems. This method innovatively incorporates RGB visual data and depth geometric data to design a dual-stream feature extraction backbone network. A specially designed self-learning deep feature fusion module fully extracts and deeply integrates heterogeneous modal information by executing dedicated feature extraction paradigms on different modal data. Simultaneously, a weighted composite balanced loss function is proposed to address the issue of low detection accuracy for underrepresented categories caused by imbalanced calibration samples, by leveraging dynamically adaptive weighting factors. This loss-function-based approach to addressing sample imbalance belongs to the paradigm of active learning. It is not

constrained by specific dataset-style characteristics, and relies heavily on the model's ability to learn salient target features, thereby exhibiting high robustness. Comparative and ablation experiments conducted on the track fastener defect RGBD dataset and vehicle body fastener defect RGBD dataset demonstrate that the proposed algorithm achieves optimal performance across all detection metrics while meeting the real-time engineering requirements for inspection. This provides a reliable and precise technical solution for intelligent rail transit inspection, offering significant engineering application value and academic reference significance.

This method assumes relatively stable image acquisition quality. Under adverse conditions such as intense light or low illumination, sensor imaging performance may deteriorate, thereby affecting the model's detection accuracy. Future improvements may incorporate a preprocessing module to restore images captured in challenging environments.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by Beijing Natural Science Foundation, grant number L241078, the Postdoctoral Fellowship Program of CPSF, grant number GZC20251118 and Beijing Subway Operation Co., Ltd.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Haoyu Li, Yang Gao and Zhiwei Cao; methodology, Haoyu Li and Zhiwei Cao; software, Haoyu Li; validation, Haoyu Li; formal analysis, Jiayi Wang; investigation, Zhaoyu Wu; resources, Shuo Yan; data curation, Ziqi Zhang; writing—original draft preparation, Haoyu Li; writing—review and editing, Yang Gao and Genwang Peng; visualization, Haoyu Li; supervision, Yang Gao and Genwang Peng; project administration, Jiayi Wang; funding acquisition, Zhiwei Cao. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. He J, Wang W, Lv F, Luo H, Zhang G, Chen Z. Multi-scale CNN-transformer hybrid network for rail fastener defect detection. IEEE Trans Intell Transport Syst. 2025;26(6):8894–906. doi:10.1109/tits.2025.3540846.
2. Wu Y, Chen P, Qin Y, Qian Y, Xu F, Jia L. Automatic railroad track components inspection using hybrid deep learning framework. IEEE Trans Instrum Meas. 2023;72:1–15. doi:10.1109/tim.2023.3265636.
3. An Y, Li X, Cao Y, Su S, Sun Y, Wang F, et al. A T-YOLO and overlapping reconstruction-based method for rail fastener defect detection in heavy-haul railway. IEEE Trans Instrum Meas. 2025;74:1–12. doi:10.1109/tim.2025.3602605.
4. Chen C, Li K, Cheng Z, Piccialli F, Hoi SCH, Zeng Z. A hybrid deep learning based framework for component defect detection of moving trains. IEEE Trans Intell Transp Syst. 2022;23(4):3268–80. doi:10.1109/tits.2020.3034239.
5. Shaikh MZ, Ahmed Z, Baro EN, Hussain S, Milanova M. Deep learning based identification and tracking of railway bogie parts. Alex Eng J. 2024;107(1):533–46. doi:10.1016/j.aej.2024.07.064.
6. Wang S, Xu Z, Wang Y, Tan Z, Zhu D. A three-stage framework for accurate detection of high-speed train body paint film defects. Adv Eng Inform. 2024;62(7):102838. doi:10.1016/j.aei.2024.102838.
7. Wang J, Song K, Zhang D, Niu M, Yan Y. Collaborative learning attention network based on RGB image and depth image for surface defect inspection of No-service rail. IEEE/ASME Trans Mechatron. 2022;27(6):4874–84. doi:10.1109/tmech.2022.3167412.
8. Gao Y, Cao Z, Qin Y, Ge X, Lian L, Bai J, et al. Railway fastener anomaly detection via multisensor fusion and self-driven loss reweighting. IEEE Sens J. 2024;24(2):1812–25. doi:10.1109/jsen.2023.3336962.
9. Ge X, Cao Z, Qin Y, Gao Y, Lian L, Bai J, et al. An anomaly detection method for railway track using semisupervised learning and vision-lidar decision fusion. IEEE Trans Instrum Meas. 2024;73:1–15. doi:10.1109/tim.2024.3417537.

10. Wang J, Li G, Qiu G, Ma G, Xi J, Yu N. Depth-assisted semi-supervised RGB-D rail surface defect inspection. IEEE Trans Intell Transport Syst. 2024;25(7):8042–52. doi:10.1109/tits.2024.3387949.
11. Tu Z, Wu S, Kang G, Lin J. Real-time defect detection of track components: considering class imbalance and subtle difference between classes. IEEE Trans Instrum Meas. 2021;70:1–12. doi:10.1109/tim.2021.3117357.
12. Qiu Y, Liu H, Liu J, Shi B. Center-triplet loss for railway defective fastener detection. IEEE Sens J. 2024;24(3):3180–90. doi:10.1109/jsen.2023.3339883.
13. Li X, Cao Y, Wang F, Sun Y, Su S, Yang W, et al. Rail fastener defect detection of heavy haul railway based on improved YOLOv8. In: Proceedings of the 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC); 2024 Sep 24–27; Edmonton, AB, Canada. p. 825–30. doi:10.1109/itsc58415.2024.10920117.
14. Wang CY, Yeh IH, Liao HYM. YOLOv9: learning what you want to learn using programmable gradient information. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Proceedings of the Computer Vision—ECCV 2024. 2024 Sep 29–Oct 4; Milan, Italy. Vol. 15089, Cham, Switzerland: Springer; 2024. p. 3–19. doi:10.1007/978-3-031-72751-1_1.
15. Li X, Wang W, Wu L, Chen S, Hu X, Li J, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. arXiv:2006.04388. 2020. doi:10.48550/arXiv.2006.04388.
16. Cao K, Wei C, Gaidon A, Arechiga N, Ma T. Learning imbalanced datasets with label-distribution-aware margin loss. Adv Neural Inf Process Syst. 2019;32:107–17. doi:10.1109/icip42928.2021.9506389.
17. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2999–3007. doi:10.1109/iccv.2017.324.
18. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: faster and better learning for bounding box regression. Proc AAAI Conf Artif Intell. 2020;34(7):12993–3000. doi:10.1609/aaai.v34i07.6999.
19. Wang CY, Bochkovskiy A, Liao HM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 7464–75. doi:10.1109/cvpr52729.2023.00721.
20. Ultralytics. YOLOv8 [Internet]. [cited 2025 Jan 1]. Available from: https://github.com/ultralytics/ultralytics/tree/main/ultralytics/cfg/models/v8.
21. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. YOLOv10: real-time end-to-end object detection. arXiv:2405.14458. 2024. doi:10.48550/arXiv.2405.14458.
22. Ultralytics. YOLOv11 [Internet]. [cited 2025 Apr 1]. Available from: https://github.com/ultralytics/ultralytics/tree/main/ultralytics/cfg/models/11.
23. Tian Y, Ye Q, Doermann D. YOLOv12: attention-centric real-time object detectors. arXiv:2502.12524. 2025. doi:10.48550/arXiv.2502.12524.
24. Zhang H, Wang Y, Dayoub F, Sunderhauf N. VarifocalNet: an IoU-aware dense object detector. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 8510–9. doi:10.1109/cvpr46437.2021.00841.