frontiers | Frontiers in Mechanical Engineering

TYPE Original Research
PUBLISHED 09 January 2026
DOI 10.3389/fmech.2025.1708579

OPEN ACCESS

EDITED BY
Georgios Mavropoulos,
School of Pedagogical and Technological Education, Greece

REVIEWED BY
Aleksandar Ašonja,
Business Academy University (Novi Sad), Serbia
Il Song Han,
Independent researcher, United Kingdom

*CORRESPONDENCE
Jun Ma,
Junma159689@hotmail.com

RECEIVED 23 September 2025
REVISED 19 December 2025
ACCEPTED 19 December 2025
PUBLISHED 09 January 2026

CITATION
Ma J, Xue X and Chen B (2026) Automatic identification of high-speed railway wheelset defects by integrating PointNet++ and Swin Transformer.
Front. Mech. Eng. 11:1708579.
doi: 10.3389/fmech.2025.1708579

Automatic identification of high-speed railway wheelset defects by integrating PointNet++ and Swin Transformer

Jun Ma[1]*, Xu Xue[2] and Bingzhi Chen[1]

[1]School of Mechanical Engineering, Dalian Jiaotong University, Dalian, Liaoning, China, [2]School of Electrical Engineering, Dalian Jiaotong University, Dalian, Liaoning, China

In order to address the technical challenges of detecting defects in high-speed railway wheelsets under complex conditions such as dynamic lighting, foreign object occlusion, and microscale anomalies, this paper proposes a dual-mode deep learning framework that integrates PointNet++ and Swin Transformer. This paper enhances defect recognition through cross modal feature collaboration, and combines cross modal attention (CMA) mechanism for dynamic feature alignment and geometric guidance suppression strategy for reducing occlusion noise. The experimental results showed an accuracy of 0.985, an F1 value of 0.982, and a recognition rate of 0.938 for defects smaller than 1 millimeter. Research has shown that the model maintains robust accuracy under different lighting conditions (strong/weak/reflective) and up to 40% occlusion, while optimized deployment on edge devices can achieve 23FPS with only 12M parameters. This work significantly improves the intelligence and reliability of the high-speed railway wheelset detection system.

KEYWORDS

automatic defect recognition, dynamic lighting, foreign object occlusion, high-speed rail wheels, PointNet++ model, Swin Transformer model

1 Introduction

High-speed rail wheelsets (Deng et al., 2021; Guo et al., 2020) are core components for train power transmission and contact with tracks. Their surface defects (Liao et al., 2023; Zhang et al., 2021) directly affect operational safety and the safety of passengers' lives and property. The structural integrity of high-speed train wheelsets directly determines their operational safety and stability. During long-term, high-load, and high-frequency service, wheelset surfaces are prone to defects such as cracks, abrasions, spalling, and pitting due to contact fatigue, wear, impact, or environmental corrosion. If these defects are not detected in time, they may lead to stress concentration, crack propagation, or even wheelset failure, seriously threatening the operational safety of high-speed trains. Therefore, achieving accurate and automated detection of wheelset defects is a crucial prerequisite for ensuring the reliable operation and intelligent maintenance upgrades of the high-speed train system.

Traditional manual inspection (Zhang et al., 2022a; Deng et al., 2024) relies on the experience and judgment of maintenance personnel and has bottlenecks such as low efficiency and high missed detection rate, making it difficult to meet the high-frequency and high-precision operation and maintenance requirements of high-speed railways. With the

development of deep learning technology, automated defect detection methods based on images or point clouds (Shaikh et al., 2025; Wang et al., 2025) are gradually being applied to industrial scenarios, but the complexity of high-speed rail wheels still leads to significant challenges: 1) Dynamic lighting interference, such as alternating strong/weak light inside and outside the tunnel, and reflection interference that invalidates image texture features; 2) Foreign matter occlusion, oil stains, and sand coverage lead to partial loss of defective areas; 3) It is difficult to identify tiny defects, and cracks <1 mm are easily masked by point cloud sparsity or image noise. The above problems urgently require an intelligent detection framework that integrates multi-modal data and has both geometric structure sensitivity and texture robustness to break through the performance bottleneck of traditional single-modal methods.

Existing high-speed rail wheel defect detection methods can be divided into two categories: single-mode and multi-mode. Among the single-mode methods, the point cloud driven model (Geng et al., 2023; Yu et al., 2022) extracts geometric features through hierarchical abstraction, which is robust to occlusion and illumination changes but has difficulty in capturing surface texture details; the image driven model (Song et al., 2024; Zhang et al., 2022b) is good at identifying texture anomalies but is easily disturbed by dynamic illumination. Under the trend of multi-modal fusion, early fusion strategies lead to defect localization deviation due to the loss of geometric information, while late fusion introduces redundant noise due to the cross-modal semantic gap. The continuous progress of railway technology system is inseparable from a large amount of in-depth research and the application of new maintenance methods. The fundamental goal of these efforts is to provide higher quality equipment, thereby achieving more reliable and efficient railway transportation services. Especially in the field of high-speed railway, wheelsets are the core of train operation safety, and the intelligent upgrading of their condition monitoring and defect identification technology is an important manifestation of this development trend. By integrating advanced sensing technology and artificial intelligence methods, the automated and accurate detection of wheelset defects is achieved, which is of key significance for improving the reliability and operation and maintenance efficiency of the entire railway system (Vuković and Kovalevskyy, 2024). Recent studies have attempted to introduce a cross-modal attention mechanism to align image and text/point cloud features through a Query-Key-Value architecture but face problems such as high computational complexity and insufficient alignment accuracy in industrial scenarios. In addition, existing methods generally lack targeted optimization for small defects, such as not designing geometric consistency constraints to enhance the curvature abnormal response of cracks <1 mm, resulting in a high missed detection rate. Therefore, a new fusion framework that takes into account both multi-modal feature alignment efficiency and defect sensitivity is urgently needed.

This paper applies a dual-modal deep learning classification framework that integrates PointNet++ and Swin Transformer to systematically solve the problem of missed detection of high-speed rail wheel defects under dynamic lighting, foreign object occlusion, and micro-scale. A two-stream feature extraction network is designed: PointNet++ extracts point cloud geometric features through a hierarchical Set Abstraction module, and Swin Transformer performs a multi-scale window attention mechanism on the image after the illumination invariance is enhanced to capture texture features such as edges. CMA is applied to map image features to the point cloud density of each level of PointNet++ through a multi-granularity alignment strategy, and the association weights between geometric queries and texture key-values are calculated to dynamically suppress noise interference in oil/silt occlusion areas. A geometric consistency regularization term is applied to constrain the second-order derivative of the point cloud curvature change. CMA guides texture weight allocation through geometric features and applies geometric consistency regularization terms to constrain the second-order derivative of curvature so that the recognition rate of defects <1 mm reaches 0.938; the accuracy rates of strong light/weak light/reflective scenes reach 0.968, 0.962, and 0.956, respectively, and it can accurately identify defects under different occlusion conditions. Through channel pruning and 8-bit quantization, the model is deployed to the Jetson AGX Xavier edge device, achieving an inference speed of 23 FPS, meeting the real-time detection requirements of high-speed rail (≥20 FPS) and promoting the upgrade of manual detection to intelligent detection.

2 Related work

As a direct representation of three-dimensional geometric information, point cloud has irreplaceable advantages in industrial defect detection. The early MLP (Multilayer Perceptron)-based PointNet achieved point cloud classification through global feature pooling, but its defect of ignoring local neighborhood relationships led to insufficient sensitivity to tiny defects. PointNet++ (Zeng et al., 2023) applied local structure perception capabilities through the hierarchical Set Abstraction module, making breakthroughs in tasks such as part recognition. Its sampling and Ball Query grouping strategies can abstract multi-scale geometric features layer by layer. However, its fixed-radius ball query mechanism is prone to local feature loss in sparse point cloud scenarios. PointTransformer (Jing and Wang, 2023; Wan et al., 2023) applies the Transformer architecture to point cloud processing and models the local geometric relationship of disordered point clouds through a position-aware self-attention mechanism. PCT (Point Cloud Transformer) (Kang et al., 2023) combines learnable kernel functions with spatially aware attention to significantly improve the recognition rate of small defects. PointNeXt (Huaiyu et al., 2024; Yang et al., 2025) surpasses PointNet++ in point cloud classification tasks through ResNet-style modules and enhanced data augmentation strategies. Although the above methods perform well in single-modal point cloud processing, their robustness to foreign object occlusion is still limited, and they need to be combined with image texture features for complementary enhancement. In addition, existing methods mostly focus on static scenes and it is difficult to deal with point cloud deformation caused by vibration in dynamic acquisition of high-speed rail wheelsets, which urgently needs to be combined with modal registration technology for optimization.

Visual transformer has shown its potential in industrial defect detection under complex backgrounds due to its global modeling capabilities. Swin Transformer (Li et al., 2022; Tang et al., 2023)

effectively balances computational complexity and multi-scale feature extraction capabilities through a sliding window attention mechanism and hierarchical feature map design, and outperforms traditional CNN (Convolutional Neural Network) models in tasks such as surface defect detection. Its multi-head self-attention mechanism can capture long-range dependencies and identify fracture features and patchy peeling patterns at crack edges in images. The further improved Swin Transformer (Zheng and Lin, 2023; Wang et al., 2024) applies an illumination invariant embedding layer and suppresses strong/weak light interference through a channel attention module, making it suitable for dynamic lighting scenes. Another type of model, such as FocalNet (Li B. et al., 2024), uses a dynamic sparse attention mechanism to reduce the amount of computation while maintaining high-resolution details, but its complexity is still limited by the global interaction requirement of $O(n^2)$. CrossViT (Xu et al., 2024) extracts multi-scale features of images through a dual-branch architecture and combines cross-attention to fuse global and local information, but it is easily disturbed by foreign textures in occluded scenes. Although the visual Transformer has significant advantages in image texture modeling, its robustness to dynamic lighting and occlusion still needs to be optimized in coordination with geometric features.

Multi-modal fusion, by combining point cloud geometric features with image texture features, has become a key direction to break through the limitations of single modality in the field of industrial quality inspection. Existing fusion strategies can be divided into three categories according to the interaction stage: early fusion (data level) renders the point cloud into a pseudo depth map or normal vector map and then inputs it into CNN processing (Guo et al., 2023), but the dimensionality reduction loss of geometric information leads to defect localization deviation; mid-term fusion (feature level) integrates modal information in the middle layer of the model. For example, simple feature concatenation introduces redundant noise due to the cross-modal semantic gap (Yang et al., 2023; He et al., 2022), while attention-based fusion (the multi-modal CrossViT framework proposed by Kang et al. (2025), which integrates image features with 3D spatial information and optimizes through contrastive learning) or point cloud-image spatial attention module (Sun et al., 2024) can screen key features; late fusion (decision level) (Xie et al., 2024) processes the post-modal fusion results independently, which is computationally efficient but difficult to resolve inter-modal conflicts. However, existing methods generally lack targeted optimization for minor defects and have insufficient cross-modal alignment efficiency. The CMA applied in this paper belongs to the mid-term fusion paradigm, and its innovation lies in: 1) Multi-granularity alignment strategy: The multi-scale image features (C1-C3) output by Swin Transformer are accurately mapped to the point cloud density of each level of PointNet++ through bilinear interpolation to achieve spatial adaptation of geometric structure and texture details; 2) Dynamic occlusion suppression mechanism: The geometric features are used as the query to dynamically calculate the association weight with the texture key-value, and the geometric guidance weight based on the curvature gradient is applied to suppress the noise response of the oil/silt occlusion area; 3) The joint geometric consistency regularization term is used to constrain the abnormal curvature changes of small defects, thereby significantly improving the robustness of defect recognition while maintaining computational efficiency.

3 Methods

This study employs a research paradigm combining experimental and computational methods, comprehensively utilizing core technologies such as 3D visual perception, deep learning, and multimodal information fusion. At the data acquisition level, a synchronous acquisition system consisting of a line laser scanner and a high-speed camera was used, with a hardware trigger controller ensuring spatiotemporal alignment. Data processing and algorithm development were primarily based on the PyTorch deep learning framework, with model training and optimization performed on an NVIDIA A100 GPU cluster. Furthermore, the model was compressed using channel pruning and 8-bit quantization techniques, and finally deployed on the Jetson AGX Xavier edge computing platform to verify its feasibility for real-time inference in industrial settings. This synergistic application of methods, technologies, and tools constitutes the complete methodological system of this study.

3.1 Data preprocessing and registration

To ensure strict alignment of point cloud and image data in time and space, a synchronous acquisition system of line laser scanner + high-speed camera is used. The core parameters are shown in Table 1.

When the laser scanner emits the laser beam, the trigger controller generates a pulse signal, which synchronously triggers the high-speed camera to expose, ensuring that the point cloud and image are strictly aligned in the space-time coordinate system.

To eliminate the coordinate deviation between LiDAR (Light Detection and Ranging) and the camera, a two-stage registration strategy of rigid transformation matrix calibration + ICP (Iterative Closest Point) algorithm optimization is adopted.

Calibration target: chessboard (10 × 7 corner points, grid spacing 20 mm) + spherical target (diameter 50 mm). Multi-view chessboard point cloud and image are collected, and the 3D coordinates and pixel coordinates of the corner points are extracted by minimizing the reprojection error of the sphere center.

$$\mathcal{L}_{\text{reproj}} = \sum_i \| \pi(R \cdot C_i + t) - c_i \|_2^2 \tag{1}$$

In Equation 1, C_i is the center of the sphere detected by LiDAR; c_i is the pixel coordinates of the center of the sphere detected by the image; and π is the camera projection function. R, t is the rotation and translation matrix.

Calibrated LiDAR point cloud $\mathcal{P} = \{p_i \in \mathbb{R}^3\}_{i=1}^N$ and camera depth map $\mathcal{D} = \{d_j \in \mathbb{R}^3\}_{j=1}^M$.

Minimizing the sum of squared distances between corresponding points in the point cloud and the depth map:

$$\min_{R,t} \sum_{i=1}^N \| (R \cdot_i + t) - \mathcal{N}(p_i) \|_2^2 \tag{2}$$

TABLE 1 Synchronous acquisition information.

Components	Model/Parameter	Function
Line laser scanner	Keyence LJ-V7000 (accuracy ±0.1 mm)	Acquire the 3D point cloud of the wheelset surface at a frequency of 200 Hz (density ≥10^4 points/m^2) and output point cloud data
High-speed camera	Basler ace acA2440-35uc (2,448 × 2048 pixels, 200 fps)	Synchronize the capture of wheelset images (resolution 224 × 224)
Trigger controller	National instruments PXIe-6660	Synchronize the acquisition timing of the laser scanner and camera through the hardware trigger signal, and the clock synchronization error is <1 μs
Data storage	Capacity ≥2 TB	Store point cloud and image data in real time to support subsequent offline processing

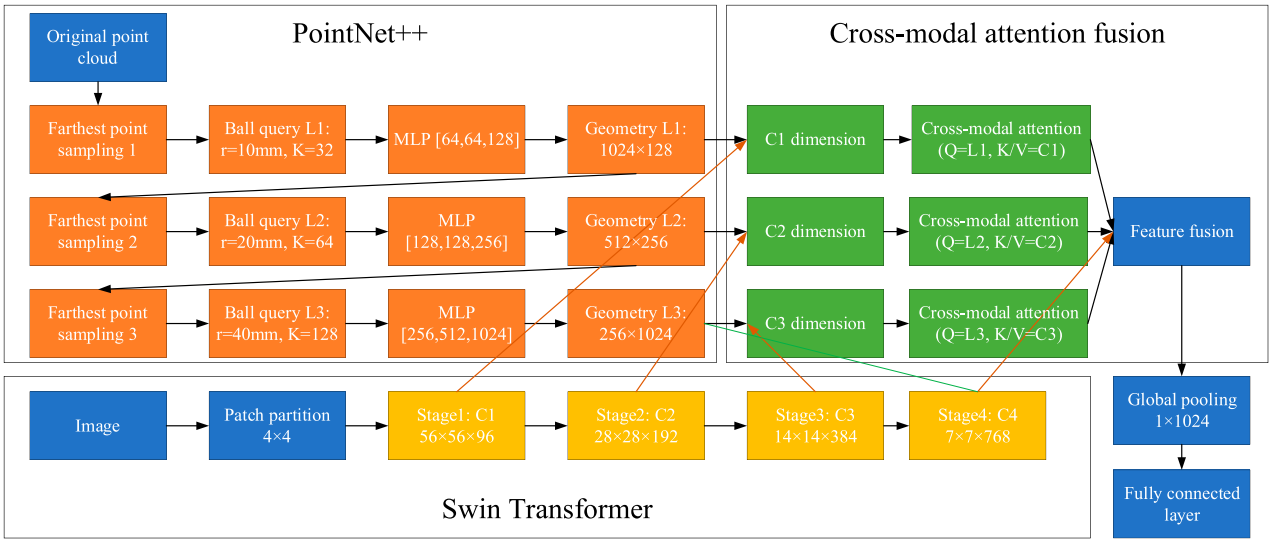

FIGURE 1
Two-stream feature extraction network.

In Equation 2, $\mathcal{N}(p_i)$ is the point closest to p_i in the depth map.

To improve the quality of feature extraction, targeted preprocessing is performed on the point cloud and image, respectively. The local density mean and standard deviation of the point cloud are calculated to remove outliers, as shown in the Equation 3:

$$\| p_i - \mu \| > 3\sigma \tag{3}$$

Photo-invariant enhancement: Based on the single-scale Retinex theory, logarithmic domain enhancement is performed on the image, as shown in the Equation 4:

$$I_{\text{enhanced}}(x, y) = \log(I(x, y)) - \log(G_\sigma * I(x, y)) \tag{4}$$

3.2 Two-stream feature extraction network

The two-stream feature extraction network is shown in Figure 1.

This paper proposes a dual-stream feature extraction network based on PointNet++ and Swin Transformer to realize multi-modal feature fusion analysis of high-speed rail wheel defects. The PointNet++ branch extracts point cloud geometric features step by step through the hierarchical Set Abstraction module: the first layer (L1) uses the farthest point sampling to obtain 1,024 points, combines the ball query to build a local neighborhood, and outputs 128-dimensional geometric features through MLP ([64, 64, 128]); the second layer (L2) is further downsampled to 512 points, and the MLP is upgraded to 256 dimensions; the third layer (L3) extracts 256 points (r = 40 mm, K = 128) of global features (1,024 dimensions). The Swin Transformer branch takes a 224 × 224 image as input and generates multi-scale texture features through a 4-stage sliding window attention mechanism: Stage 1 outputs C1 (56 × 56 × 96) to capture edge gradients, Stage 2's C2 (28 × 28 × 192) models patch textures, Stage 3's C3 (14 × 14 × 384) extracts directional textures, and Stage 4's C4 (7 × 7 × 768) fuses global context. The cross-modal fusion module adopts a multi-granularity alignment strategy: C1-C3 features are mapped to the point cloud density of each level of PointNet++ through bilinear interpolation, and occlusion noise is dynamically suppressed through the cross-modal attention mechanism. The fused features are globally pooled and input into the fully connected layer, combined with Focal Loss to alleviate the category imbalance problem, and a geometric consistency regularization term is applied to constrain the second-order derivative of the point cloud curvature change for defect identification.

PointNet++ (Xiang et al., 2025; Huang et al., 2023) abstracts the geometric features of point clouds layer by layer through a hierarchical Set Abstraction module. Its core process includes farthest point sampling → ball query grouping → multi-layer perceptron feature extraction, and finally outputs multi-scale geometric features {L1, L2, L3}, corresponding to the levels of point cloud density of 1,024, 512, and 256, respectively.

In farthest point sampling, for a given point cloud $\mathcal{P} = \{p_i \in \mathbb{R}^3\}_{i=1}^{N}$, the point farthest from the selected point set is iteratively selected as the sampling point to ensure the uniformity of the point cloud distribution:

$$\mathrm{FPS}(\mathcal{P}, K) = \arg \max_{p_j \in \mathcal{P}} \min_{q \in S} \| p_j - q \|_2 \tag{5}$$

In Equation 5, S is the set of selected points, and K is the number of target points.

For the sampling point $q_j \in S$, search for the neighborhood point $\mathcal{N}(q_j) = \{p_i \in \mathcal{P} \mid \|p_i - q_j\|_2 \le r\}$ within the radius r to build a local neighborhood relationship. The radius r is dynamically adjusted according to the defect scale.

For each set of neighborhood point $\mathcal{N}(q_j)$, features are extracted through MLP with shared weights:

$$f_j = \mathrm{MLP}\left(\left[q_j; \mathrm{MaxPool}\left(\{p_i - q_j\}_{p_i \in \mathcal{N}(q_j)_i}\right)\right]\right) \tag{6}$$

In Equation 6, the MaxPool operation aggregates the local geometric information of the neighborhood points relative to the coordinates $p_i - q_j$, and finally outputs the feature $f_j \in \mathbb{R}^d$.

Hierarchical feature output:

L1 (1,024 points): The original point cloud is sampled at 1,024 points at the farthest point, and MLP outputs 64-dimensional geometric features.

L2 (512 points): The L1 features are sampled at 512 points at the farthest point, and the MLP is upgraded to 128 dimensions.

L3 (256 points): The L2 features are sampled at 256 points at the farthest point, and the MLP is upgraded to 256 dimensions.

The geometric features output by PointNet++ include point cloud normal vector, curvature, and neighborhood point distance distribution. The normal vector estimates the surface direction through the difference in coordinates of the neighborhood points. The Equation 7 is as following:

$$n_j = \frac{\partial f}{\partial x} \times \frac{\partial f}{\partial y} \tag{7}$$

The curvature feature is usually simplified to the minimum eigenvalue of the neighborhood point covariance matrix.

The average Euclidean distance from the neighborhood point to the center point is calculated, and the neighborhood distance distribution is as the Equation 8:

$$\mu_j = \frac{1}{\left|\mathcal{N}(q_j)\right|} \sum_{p_i \in \mathcal{N}(q_j)} \| p_i - q_j \|_2 \tag{8}$$

Swin Transformer (Li Y. et al., 2024; Si et al., 2024) extracts global texture features of images through a multi-scale sliding window attention mechanism. Its architecture is based on the Swin-Tiny configuration, with an input image of resolution 224 × 224 and an output multi-scale feature map $\{C_1, C_2, C_3, C_4\}$.

The input image is divided into 4 × 4 non-overlapping blocks, each of which is flattened into a one-dimensional vector, and the embedded features are generated by linear projection:

$$\mathcal{E}_0 = \mathrm{Linear}(\mathrm{Patch}(x)) \in \mathbb{R}^{(H/4)(W/4) \times C_0} \tag{9}$$

In Equation 9, $H = W = 224$, $C_0 = 96$.

The core innovation of Swin Transformer (Irsal et al., 2024; Ma et al., 2025) lies in the sliding window mechanism, which divides the feature map into multiple windows. Through multi-head self-attention, for each point i in the window, the Query (Q), Key (K), and Value (V) are calculated, as following Equation 10 and 11:

$$\mathrm{MSA}(\mathcal{E}) = \mathrm{Concat}(\mathrm{Head}_1, \ldots, \mathrm{Head}_h) W^O \tag{10}$$

$$\mathrm{Head}_k = \mathrm{Softmax}\left(\frac{Q_k K_k^T}{\sqrt{d_k}}\right) V_k \tag{11}$$

The next layer shifts the window by half to achieve cross-window interaction and multi-scale feature map output:

C1 (56 × 56): shallow features, capturing edge gradients.

C2 (28 × 28): mid-level features, extracting patch textures.

C3 (14 × 14): high-level features, modeling directional textures.

C4 (7 × 7): global features, integrating multi-scale contexts.

Edge features capture crack boundaries through gradient amplitude, and the Equation 12 is as following:

$$G = \sqrt{G_x^2 + G_y^2} \tag{12}$$

Using local binary pattern statistics to calculate texture repeatability, the patch formula is as the Equation 13:

$$\mathrm{LBP}(p) = \sum_{i=0}^{7} s(g_i - g_c) 2^i \tag{13}$$

Directional texture: The scratch direction is detected by Gabor filter response, and the filter kernel is as the Equation 14:

$$g(x, y) = \exp\left(-\frac{x'^2 + y^2 y'^2}{2\sigma^2}\right) \cos(2\pi f x') \tag{14}$$

The parameter information of the PointNet++ + Swin Transformer model is shown in Table 2.

3.3 Cross-modal attention fusion

In order to achieve cross-modal interaction between point cloud geometric features and image texture features, a multi-granularity alignment strategy is designed to map the multi-scale texture features {C1, C2, C3, C4} output by Swin Transformer to the point cloud density of each level of PointNet++ through bilinear interpolation. Specifically:

C1→L1: The C1 feature map (56 × 56 × 96) is mapped to 1,024 points of the L1 level through bilinear interpolation to generate 1,024 × 96 texture features.

C2→L2: The C2 feature map (28 × 28 × 192) is mapped to 512 points of the L2 level, and 512 × 192 features are output.

C3→L3: The C3 feature map (14 × 14 × 384) is mapped to 256 points of the L3 level to generate 256 × 384 features.

TABLE 2 Model parameter information.

Components	Parameter name	Values/Configuration
PointNet++	Set abstraction layers	3
	Farthest point sampling	L1: 1,024 points, L2: 512 points, L3: 256 points
	Ball query	Radius r: L1 = 10 mm, L2 = 20 mm, L3 = 40 mm
	Number of neighboring points K	L1 = 32, L2 = 64, L3 = 128
	MLP structure	L1: [64, 64, 128], L2: [128, 128, 256], L3: [256, 512, 1,024]
	Activation function	ReLU
Swin transformer	Architecture type	Swin-tiny
	Input resolution	224 × 224
	Patch size	4 × 4
	Stages	4 stages
	Window size	7 × 7
	Number of self-attention heads	8
	MLP expansion ratio	4

C4 global fusion: The C4 feature map (7 × 7 × 768) is flattened into a 1-dimensional vector and concatenated with the global features of L3.

Mapping formula:

$$F_{\text{tex}}^{(l)} = \text{Bilinear}(C_l) \in \mathbb{R}^{N_l \times D_l} \tag{15}$$

In Equation 15, l represents the level; N_l is the number of points in the lth layer of PointNet++; and D_l is the texture feature dimension.

For the multi-scale feature maps {C1, C2, C3} output by Swin Transformer, this paper maps them to the spatial coordinates of PointNet++point clouds at various levels through bilinear interpolation. The interpolation process is based on the normalized image coordinates (u, v) of each sampling point in the point cloud, weighted and summed by the eigenvalues of the surrounding four pixels, where $F_{\text{tex}}^{(l)}(p) = \sum_{i=1}^{4} w_i \cdot C_l(x_i, y_i)$. (x_i, y_i) is the coordinate of adjacent pixels in the image, and w_i is the weight based on the relative position between point p and the pixel. This process achieves precise mapping from image feature space to point cloud geometric space, ensuring spatial consistency of texture features in geometric structure and providing a reliable feature alignment foundation for subsequent cross modal attention fusion.

The CMA module (Shi et al., 2024) dynamically calculates the association weights between geometric features and texture features through the Query-Key-Value architecture. Query generates the geometric features $F_{\text{geo}}^{(l)}$ of PointNet++, and generates the Query matrix through linear projection to obtain, as shown in the Equation 16:

$$Q = W_q F_{\text{geo}}^{(l)} \in \mathbb{R}^{N_l \times d_k} \tag{16}$$

The texture feature projection of Swin Transformer (Gao et al., 2022; Zhu et al., 2023) is as shown in Equation 17, 18:

$$K = W_k F_{\text{tex}}^{(l)} \in \mathbb{R}^{N_l \times d_k} \tag{17}$$

$$V = W_v F_{\text{tex}}^{(l)} \in \mathbb{R}^{N_l \times d_v} \tag{18}$$

The similarity matrix between Query and Key is calculated, and the attention weight is generated through Softmax normalization. The attention weight A is used to perform weighted summation on the Value feature V to generate the fusion feature, as shown in the Equation 19:

$$F_{\text{fusion}}^{(l)} = A \cdot V \in \mathbb{R}^{N_l \times d_v} \tag{19}$$

In order to solve the texture feature interference caused by oil/silt occlusion, a geometry-guided attention weight suppression strategy is designed. Dynamic weights are generated based on the geometric features (curvature κ_i) of PointNet++, as shown in the Equation 20:

$$\alpha_i = \sigma\left(\frac{\kappa_i}{\| \nabla \kappa_i \|_2 + \epsilon}\right) \tag{20}$$

Multiply the geometric weight α_i with the attention weight to suppress the response of the occluded area:

$$\tilde{A}_{ij} = \alpha_i \cdot A_{ij} \tag{21}$$

In Equation 21, $\tilde{A}_{ij}$ is the corrected attention weight, i is the Query point index, and j is the Key point index.

The corrected attention weight is used to weight the Value feature, as shown in the Equation 22:

$$\tilde{F}_{fusion}^{(l)} = \tilde{A} \cdot V \tag{22}$$

The final fusion feature is the concatenation of the weighted results layer by layer, as shown in the Equation 23:

$$F_{\text{final}} = \text{Concat}\left(\tilde{F}_{\text{fusion}}^{(1)}, \tilde{F}_{\text{fusion}}^{(2)}, \tilde{F}_{\text{fusion}}^{(3)}\right) \in \mathbb{R}^{(N_1+N_2+N_3) \times d_v} \tag{23}$$

3.4 Classification head and loss function

After the fused features are output by the cross-modal attention module, the feature dimensions need to be further compressed and mapped to the defect category space. The classification head in this paper adopts a two-stage fully connected network (Zhou et al., 2025; Zhang et al., 2023). Global maximum pooling is performed on the fused features to extract the maximum response value of each channel, as shown in the Equation 24:

$$f_{\text{global}} = \text{GMP}(F_{\text{fusion}}) \in \mathbb{R}^{1\times d} \tag{24}$$

This operation retains key characteristic responses (such as high curvature areas at the crack edge) and generates a 1 × 1024-dimensional global feature vector.

Dropout (dropout rate 0.5) is inserted between fully connected layers to prevent overfitting, and the Xavier initialization strategy is adopted to ensure training stability.

In order to solve the problems of category imbalance and micro-defect recognition, this paper designs a dual-objective loss function to jointly optimize classification accuracy and geometric constraints.

In order to solve the problem of unbalanced distribution of high-speed rail wheel defect samples, Focal Loss is used to alleviate the impact of category imbalance:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\alpha_i(1-p_i)^{y}\log(p_i) \tag{25}$$

In Equation 25, p_i is the predicted probability of the i-th sample, and α_i is the category weight.

In order to enhance the geometric sensitivity of small defects, the L2 regularization term of the point cloud curvature change is introduced:

$$\mathcal{L}_{\text{geo}} = \lambda\cdot\frac{1}{N}\sum_{i=1}^{N}\parallel\nabla^2\kappa_i\parallel_2^2 \tag{26}$$

In Equation 26, κ_i is the curvature of the i-th point.

The total loss is the weighted sum of the classification loss and the geometric constraint, as shown in the Equation 27:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{geo}} \tag{27}$$

The joint training uses the AdamW optimizer (learning rate 5×10^{-4}, weight decay 10^{-4}) to jointly optimize the classification and geometric constraints.

4 Experiment

4.1 Dataset construction

The dataset used in this study was constructed based on the publicly available FaultSeg dataset through large-scale expansion and enhancement. The FaultSeg dataset provides an image foundation for train wheel defects. To meet the requirements of the multimodal fusion method proposed in this article, 3D point cloud data strictly paired with each defect image was synchronously collected, and the sample size and diversity of defect scenes were significantly expanded. The final dataset consists of 10,918 strictly paired image point cloud samples, as shown in Table 3. Compared with the original FaultSeg, the main extensions of this dataset include: (1) the introduction of point cloud mode, which achieves the upgrade of data from 2D to 3D; (2) Added defect categories to make defect types more complete; (3) Systematically collected data under dynamic lighting conditions (strong light, weak light, reflection) and simulated foreign object occlusion conditions; (4) The total sample size has significantly increased from thousands of images in FaultSeg to over 10,000 paired samples. This enhanced dataset provides a solid foundation for validating cross modal defect detection algorithms under complex operating conditions.

The dataset in this paper is based on the FaultSeg dataset and synchronously collects point cloud data information. It contains 10,918 sets of samples, covering dynamic lighting, foreign object occlusion and small defect scenes. The distribution of the dataset is shown in Table 3.

The definition of tiny defects is a diameter of <1 mm, small-scale defects (1–3 mm), and medium-to-large defect sizes (>3 mm). Strong light environment refers to an extreme brightness scene with a light intensity exceeding 100,000 lux, simulating the working conditions of high-speed rail wheelsets under direct sunlight or strong reflective light interference outside the tunnel. Weak light environment refers to a low-light scene with a light intensity below 1,000 lux, simulating the complex working conditions of high-speed rail wheelsets during maintenance in tunnels or at night. Reflective environment refers to a scene where the metal surface of the wheelset is partially too bright due to specular reflection, simulating the working conditions of the wheelset in a humid environment.

The reflection scene is clearly defined and strictly controlled in data collection: it refers to the saturation highlight appearing in local areas due to the mirror reflection characteristics of the metal surface of the wheel, with pixel brightness values exceeding 200 (in the gray range of 0–255), and the highlight area accounting for 5%–30% of the area of interest of the wheel. This condition is accurately simulated in a laboratory environment by adjusting a point light source at a specific angle to reproduce the real working conditions of the wheel under wet or specific lighting conditions. All samples labeled as "reflection" in the dataset were screened using this standard, ensuring the uniformity and measurability of the testing conditions.

The dataset is divided into two parts: image and point cloud data. The wheelset image information is shown in Figure 2.

The five key surface conditions of high-speed rail wheels are categorized as follows: Crack, characterized by narrow fissures potentially propagating due to fatigue and stress concentration, posing a high hazard level risk of structural failure. Scuffing, appearing as linear wear marks typically caused by braking slippage, is assessed as a medium-level hazard. Spalling, where surface metal layers detach due to contact fatigue or thermal stress, represents a medium-high hazard by exacerbating vibration and noise. Pit, manifesting as small yet deep depressions from foreign matter intrusion or corrosion, is a medium-hazard defect that can initiate crack formation. Normal denotes a defect-free surface, indicating a safe operational state. This classification system comprehensively covers typical failure modes, from micro-damage to macro-defects, providing a clear basis for automated defect identification and safety assessment.

During long-term operation, high-speed rail wheelsets (i.e., the combination of wheels and axles) may suffer from various types of surface or structural defects due to the huge loads, friction, and impact. The dataset is divided into training set, test set, and validation set in the form of 8:1:1.

TABLE 3 Dataset distribution.

Category	Specific category	Number of samples	Defect size distribution	Lighting conditions	Occlusion rate
Defect	Crack	880	<1 mm (400)	Strong light (293), weak light (293), reflection (294)	0%–50%
	Scuffing	1,050	<1 mm (500)	Strong light (350), weak light (350), reflection (350)	
	Spalling	950	<1 mm (300)	Strong light (317), weak light (317), reflection (316)	
	Pit	1,038	<1 mm (400)	Strong light (346), weak light (346), reflection (346)	
Normal	Normal	7,000	-	Strong light (2,333), weak light (2,333), reflection (2,334)	0%

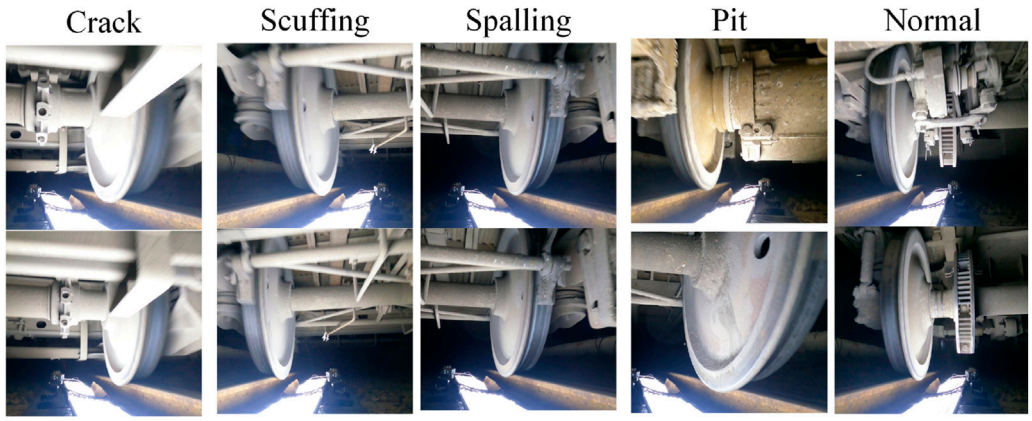

FIGURE 2
High-speed rail wheelset image.

TABLE 4 Comparison of high-speed rail wheel set type information.

Category	Description	Causes	Hazard level
Crack	Narrow cracks, which may extend	Fatigue, stress concentration	High
Scuffing	Linear wear marks	Braking skidding, slipping	Medium
Spalling	Surface metal layer has fallen off	Contact fatigue, thermal stress	Medium-high
Pit	Small but deep depressions	Foreign matter intrusion, corrosion	Medium
Normal	No defect	Normal use	Safe

The comparison of high-speed rail wheel set type information is shown in Table 4.

4.2 Experimental environment and index evaluation

To ensure the efficiency of model training and deployment, this experiment uses a graphics card cluster: NVIDIA A100 × 4, which supports mixed precision training. The processor uses Intel Xeon Gold 6,330, with 1 TB of memory, to accelerate data preprocessing and multi-threaded loading. PyTorch 1.13 is used to build a deep learning framework. And Jetson AGX Xavier is deployed.

The training hyperparameters in this paper are set as:

Optimizer: AdamW (learning rate 5×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$).

Learning rate schedule: Cosine annealing (initial learning rate = 5××10^{-4}, final 10^{-6}).

Regularization: weight decay 10^{-4}, dropout rate 0.5.

Batch size: batch size = 64 during training, batch size = 1 during deployment (single frame processing).

In order to comprehensively evaluate the performance of the model in high-speed rail wheel defect detection, this paper adopts an

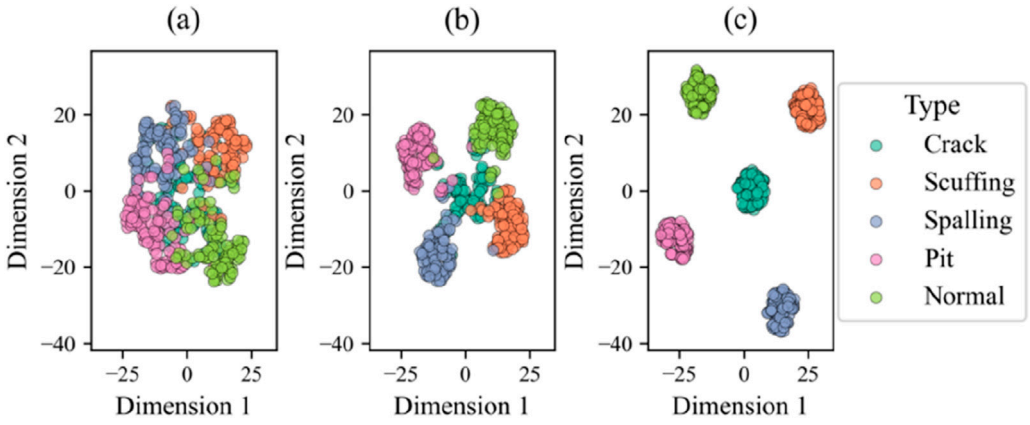

FIGURE 3
t-SNE visualization results. **(a)** Geometric features. **(b)** Global texture features. **(c)** Fusion features.

evaluation system that combines macro-level indicators with category-level indicators. The accuracy measures the overall classification performance of the model and is defined as the ratio of the number of samples predicted correctly to the total number of samples, as shown in the Equation 28:

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{28}$$

Macro precision measures the average precision of the model for all categories to avoid the impact of category imbalance, as shown in the Equation 29:

$$\mathrm{Precision}_{\mathrm{macro}} = \frac{1}{C}\sum_{i=1}^{C}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i} \tag{29}$$

Macro recall measures the average recall of the model for all categories, as shown in the Equation 30:

$$\mathrm{Recall}_{\mathrm{macro}} = \frac{1}{C}\sum_{i=1}^{C}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{30}$$

The macro F1 score is the harmonic mean of the macro precision and the macro recall, which comprehensively measures the model performance, as shown in the Equation 31:

$$\mathrm{F1}_{\mathrm{macro}} = \frac{2}{\sum_{i=1}^{C}\frac{1}{\mathrm{Precision}_i} + \frac{1}{\mathrm{Recall}_i}} \tag{31}$$

5 Results and analysis

5.1 t-SNE visualization of data features

t-SNE (t-Distributed Stochastic Neighbor Embedding) visualization is used to analyze the distribution separability of geometric features, global texture features, and fusion features, and verify that the inter-class separation is improved after cross-modal fusion. The t-SNE visualization results are shown in Figure 3.

Figure 3a shows that the geometric features of the five types of samples in the t-SNE space overlap significantly, and the boundaries between normal samples and defective samples are blurred, indicating that the local geometric features (such as curvature and normal vector) extracted by PointNet++ have limitations in class distinction. Although defects such as cracks and scratches have slight differences in local density distribution, the overall distribution is loosely mixed and no obvious independent clustering structure is formed, indicating that the geometric features have a weak ability to characterize small surface defects, which may be limited by the sparsity of point clouds or the sampling bias of neighborhood information.

Figure 3b shows that the distribution of normal samples and defective samples in the global texture features gradually separates, and defects such as cracks and pits tend to form weak clustering locally, reflecting the Swin Transformer's ability to model image edge gradients and patch repeatability. However, there is still overlap in categories such as scratches and peeling, indicating that texture features are easily affected by lighting changes or foreign body occlusion, leading to misjudgment of similar texture patterns.

In Figure 3c, the five types of samples show tight and separated clusters in the fusion feature space, with clear boundaries between classes and significantly enhanced compactness within classes, verifying the effective integration of geometric and texture features by the cross-modal attention mechanism. The weight distribution guided by geometric features strengthens the response of the defective area, suppresses the interference of oil/sand occlusion, and completely decouples the distribution of normal samples from defective samples, supporting more robust classification performance.

5.2 Changes in model loss values

The model in this paper is compared with the existing mainstream models to analyze the changes in loss values. The results are shown in Figure 4.

The initial training loss value of the model in this paper is the smallest and converges the fastest. The fundamental reason is the

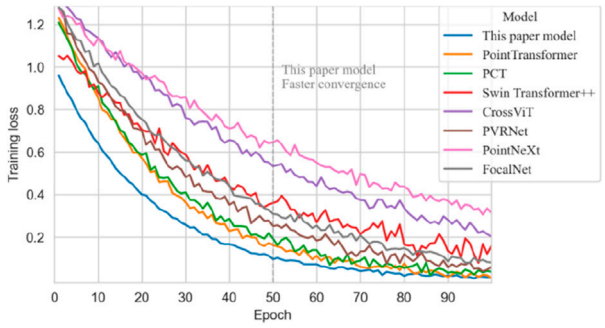

FIGURE 4
Changes in loss value.

synergistic optimization effect of the cross-modal attention mechanism and the geometric consistency regularization term. The multi-granularity alignment strategy maps the texture features of Swin Transformer to the point cloud density of each level of PointNet++ through bilinear interpolation, so that the geometric features and texture features can be aligned with high precision at the beginning of feature interaction, avoiding redundant calculations caused by dimensional mismatch in traditional early fusion. The cross-modal attention module uses geometric features to dominate the attention weight allocation for query, dynamically suppresses noise interference in oil/silt occlusion areas, and enables the loss function to accurately focus on key defect areas in the early stages of training, significantly reducing initial classification losses. In addition, the geometric consistency regularization term constrains the change of the second-order derivative of curvature, forcing the model to prioritize learning the local geometric anomalies in the defect area, reducing the oscillation in the gradient direction and accelerating parameter convergence. The model in this paper has both fast convergence and low loss characteristics under complex working conditions, verifying the advantages of the cross-modal fusion architecture in gradient direction optimization and feature space alignment.

5.3 Overall classification performance

The overall classification performance of the test set is shown in Figure 5.

From the overall classification performance of the test set, the proposed model significantly outperforms other comparison models with an accuracy of 0.985 and a macro F1 of 0.982, verifying its excellent performance in high-speed rail wheel defect classification. PointNeXt (accuracy 0.972, macro F1 0.974) and PCT (accuracy 0.969, macro F1 0.951) rank second and third, respectively, indicating that the improved version of PointNet++ (PointNeXt) is better than the traditional Transformer architecture (such as PointTransformer, accuracy 0.939, macro F1 0.942) in geometric feature modeling, but is still limited by occlusion interference. Swin Transformer++ (accuracy 0.968, macro F1 0.945) performs better than FocalNet (accuracy 0.965, macro F1 0.953), reflecting that its sliding window attention mechanism is more stable in strong light/reflective scenes. Among the multi-modal models, CrossViT (accuracy 0.948, macro F1 0.943) and PVRNet (Point-View Relation Neural Network) (accuracy 0.943, macro F1 0.936) perform worse than the model in this paper, mainly due to the redundant calculation and registration error of early fusion. It is worth noting that the macro F1 (0.942) of PointTransformer is higher than the accuracy (0.939), indicating that it has a strong recall ability for minority classes, but the overall classification performance is limited by the sparsity of point clouds and occlusion interference. The proposed model dynamically suppresses occlusion noise through a cross-modal attention mechanism and uses a geometric consistency regularization term to constrain curvature changes, making both accuracy and macro F1 better than other models.

The ROC-AUC index further reveals the global discrimination ability of the proposed model under complex working conditions: the proposed model has an AUC of 0.936, which is significantly higher

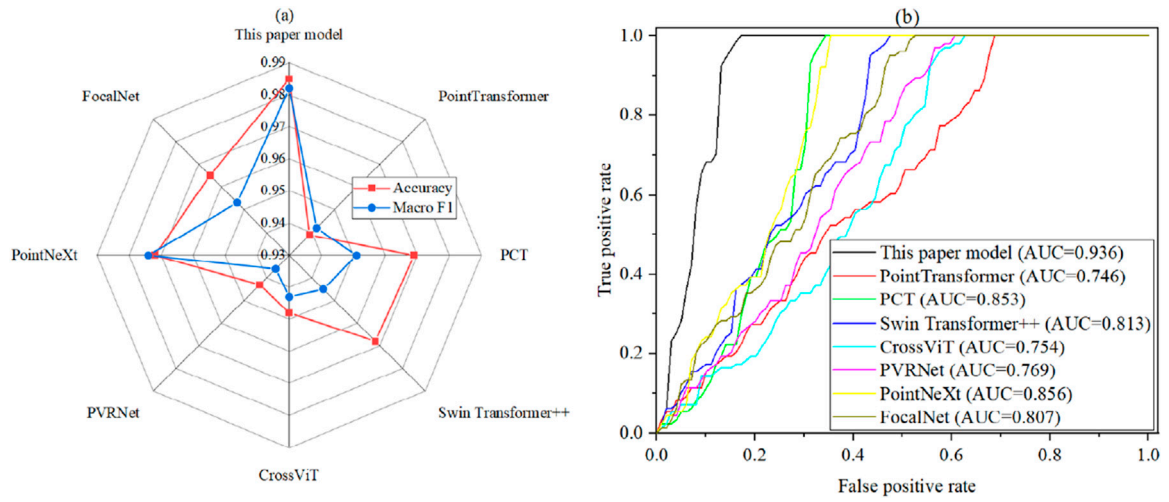

FIGURE 5
Overall classification performance. **(a)** Accuracy and macro F1 of different models. **(b)** ROC (Receiver Operating Characteristic) - AUC (Area Under the Curve) curve.

TABLE 5 Ablation experiment comparison information.

Abbreviation	Model	Data modality	Fusion strategy
M1	PointNet++	Point cloud (geometric features)	Unimodal processing
M2	Swin transformer	Image (texture features)	Unimodal processing
M3	PointNet++ + swin transformer (feature concatenation only)	Point cloud + image	Early fusion
M4	This paper model	Point cloud + image	Cross-modal attention fusion

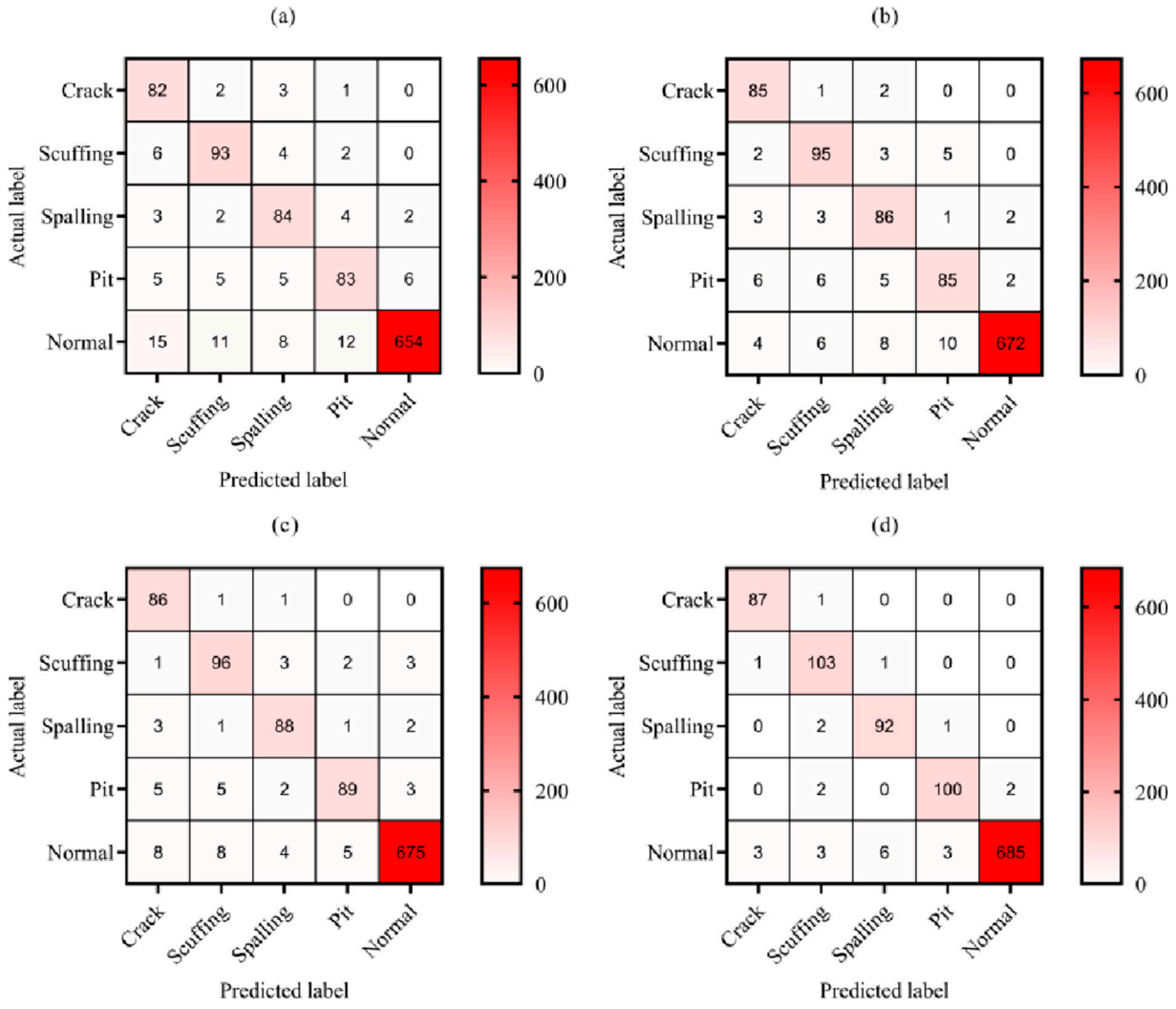

FIGURE 6
Confusion matrix. **(a)** M1 confusion matrix. **(b)** M2 confusion matrix. **(c)** M3 confusion matrix. **(d)** M4 confusion matrix.

than PointNeXt (0.856) and PCT (0.853), indicating that its robustness to category imbalance and occlusion interference is better than that of the point cloud single-modality method. The AUC values of Swin Transformer++ (AUC = 0.813) and FocalNet (AUC = 0.807) are lower than those of the proposed model, reflecting the limitations of image single-modality under illumination changes. Among the multi-modal models, the AUC values of CrossViT (AUC = 0.754) and PVRNet (AUC = 0.769) are significantly lower than that of the proposed model. PointTransformer (AUC = 0.746) and CrossViT have the lowest AUC values, indicating the shortcomings of the traditional Transformer architecture in local geometric modeling and the bottleneck of cross-modal alignment efficiency. The AUC advantage of the proposed model directly reflects the synergy of cross-modal fusion (geometric features guide attention weight allocation) and geometric consistency regularization (curvature second-order derivative constraint).

In order to systematically verify the performance of each component of the model in this paper, an ablation experiment is designed, and the comparison information is shown in Table 5.

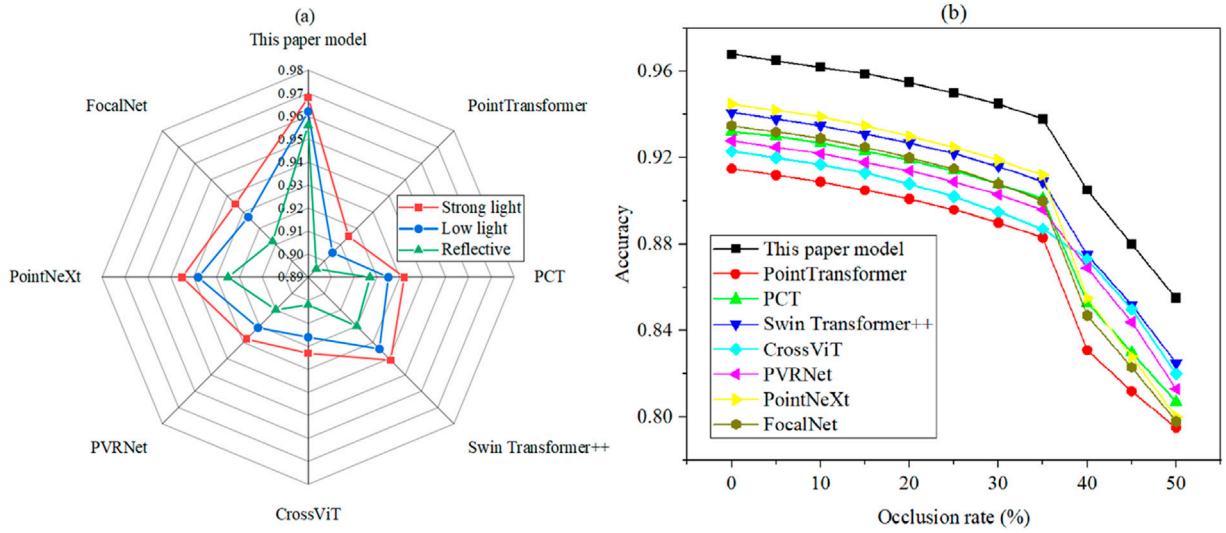

FIGURE 7
Robustness test results. **(a)** Lighting environment. **(b)** Occlusion.

The confusion matrix of the model involved in the ablation experiment is shown in Figure 6.

The total number of test samples is 1,092 and the number of samples correctly classified by each model shows that the model performance is ranked as follows: M4 (1,067) > M3 (1,034) > M2 (1,023) > M1 (996). M4 is significantly better than M3, with 33 more correct samples than M3, reflecting the substantial improvement of the classification performance by the dynamic alignment strategy (geometric feature-guided weight allocation) and the CMA module. Image texture features (M2) are better than point cloud geometric features (M1): M2 has 27 more correct samples than M1, indicating that texture features have a stronger ability to capture surface details in high-speed rail wheel defect detection. M3 has limited optimization of single-modal performance: M3 has only 38 more correct samples than M1 (M3-M1 = 38), indicating that simple feature splicing fails to fully exploit the potential of multi-modal collaboration.

5.4 Robustness test

To ensure the statistical validity of the evidence, all accuracy indicators of the reports are calculated based on independent test sets randomly divided from the complete dataset (accounting for 10% of the total, i.e. 1092 sample groups). For robustness testing of lighting and occlusion, specific subsets are extracted from the test set. Each subset contains a sufficient number of samples to avoid indicator fluctuations caused by too few samples.

The robustness of the model is analyzed through lighting environment and occlusion, and the change in accuracy is observed. The results are shown in Figure 7.

Figure 7a Classification accuracy across different lighting environments. The test subset comprises samples captured under strictly defined strong light (>100,000 lux), weak light (<1,000 lux) and reflected light.

Figure 7b Classification accuracy under increasing levels of occlusion. The occlusion rate is synthetically generated by applying random masks to the image and point cloud data, simulating foreign object coverage.

From the lighting robustness test, it can be seen that the model in this paper maintains the optimal accuracy in strong light (0.968), weak light (0.962) and reflection (0.956) scenes, and the difference between the three is small, indicating that the cross-modal attention mechanism guides the weight allocation through geometric features, effectively suppressing strong light overexposure, weak light low contrast and reflection saturation interference. In contrast, the accuracy of image unimodal models (such as Swin Transformer++, FocalNet) drops significantly in reflective scenes (Swin: 0.941→0.920; FocalNet: 0.935→0.912), reflecting their sensitivity to lighting fluctuations; point cloud unimodal models (such as PointNeXt, PCT) perform more stably (PointNeXt: 0.945→0.925; PCT: 0.932→0.917), but are weaker than the model in this paper. Early fusion methods (such as CrossViT and PVRNet) have limited illumination adaptability due to feature redundancy and semantic gap (CrossViT: 0.923→0.902; PVRNet: 0.928→0.910). The model in this paper dynamically compensates for the influence of illumination through a multi-granularity alignment strategy and a geometric consistency regularization term, which reduces the accuracy fluctuation in strong light/low light/reflective scenes.

The accuracy of our model changes from 0% to 50% in the range of occlusion rate (0.968→3), which is significantly better than other models. Its cross-modal attention mechanism guides texture weight allocation through geometric features and still maintains an accuracy of 0.905 under 40% occlusion, verifying the effectiveness of dynamic occlusion suppression. The multi-granularity alignment strategy (such as L1-L3 hierarchical mapping) of the model in this paper is coordinated with the geometric consistency regularization term to maintain the optimal inter-class separation in occlusion scenarios, supporting the stability of high-speed railway wheelset defect detection under foreign object coverage conditions.

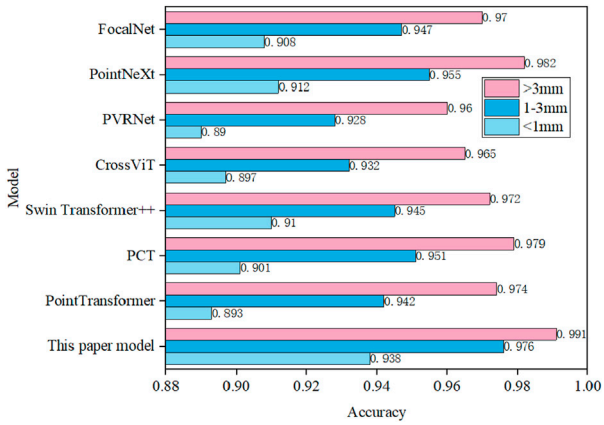

FIGURE 8
Defect size recognition effect.

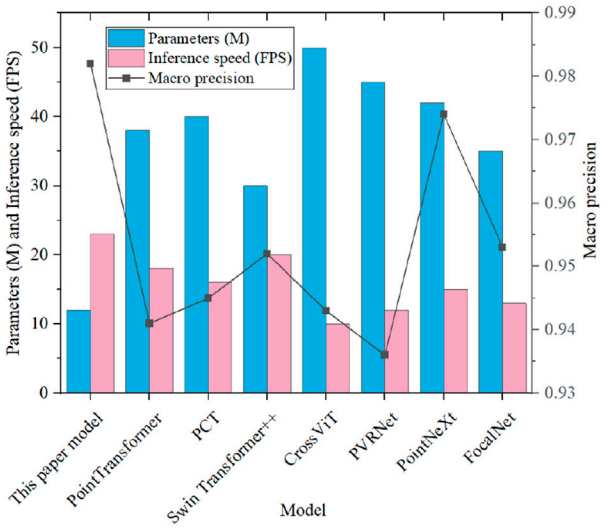

FIGURE 9
Deployment performance.

5.5 Defect size recognition effect

Defect sizes include: micro defects (<1 mm), small-scale defects (1–3 mm), and medium-to-large-scale defects (>3 mm). The defect recognition performance for different sizes is shown in Figure 8.

The proposed model maintains the best accuracy (0.938→0.976→0.991) in the classification of defects <1 mm, 1–3 mm and >3 mm, reflecting the targeted optimization of the CMA module and the geometric consistency regularization term for small defects. Among defects <1 mm, the accuracy of the proposed model is significantly higher than PointNeXt (0.912) and Swin Transformer++ (0.910), verifying that the texture weight allocation guided by geometric features effectively enhances the curvature abnormal response. The performance of each model is close to that of the medium- and large-scale defects (>3 mm) (0.960–0.991), but the proposed model still leads with an accuracy of 0.991, reflecting the accurate modeling of macroscopic structural damage by fusion features. In contrast, the single-modal model is limited by the lack of local information (such as PointTransformer is only 0.893 for defects <1 mm), and the early fusion models (such as CrossViT and PVRNet) have lower accuracy than the proposed model due to redundant calculations and registration errors. This result systematically proves that cross-modal attention fusion can significantly improve the ability to identify tiny defects by dynamically aligning geometric-texture features, while maintaining stable judgment of medium and large-scale defects, supporting the full-size coverage requirements of high-speed rail wheelset defect detection.

5.6 Parameter quantity and inference efficiency

The feasibility and effectiveness of deployment on industrial edge devices are verified by comprehensive analysis of parameter quantity, inference speed, and recognition accuracy. The results are shown in Figure 9.

The model in this paper significantly outperforms other models while maintaining a macro precision of 0.982 through channel pruning and 8-bit quantization (inference speed increased to 23 FPS). Although PointNeXt (42M, 15 FPS) and Swin Transformer++ (30M, 20 FPS) have moderate parameter counts, they do not solve the sensitivity of single-modal features to occlusion/illumination changes. CrossViT (50M, 10 FPS) is limited in edge device deployment due to redundant computation of cross-modal splicing. The cross-modal attention mechanism of this model maintains high macro accuracy through dynamic weight allocation (geometric feature dominance) and multi-granularity alignment strategy, and channel pruning and quantization reduce the number of parameters to 12M, which meets the computing power constraints of Jetson AGX Xavier and verifies its feasibility of industrial edge device deployment.

6 Conclusion

This study presents a dual-stream cross-modal defect classification framework that integrates PointNet++ for point cloud geometric feature extraction and Swin Transformer for image texture analysis. By employing a Cross-Modal Attention (CMA) mechanism, the framework achieves dynamic alignment of geometric and texture features, while a geometry consistency regularization term is introduced to enhance sensitivity to micro-defect curvature anomalies. Evaluated on a dedicated dataset of 10,918 multimodal samples, the model achieves an overall accuracy of 0.985 and a macro F1-score of 0.982. Notably, it attains a recognition rate of 0.938 for defects smaller than 1 mm and maintains 0.905 accuracy under 40% occlusion. Through channel pruning and 8-bit quantization, the model is compressed to 12M parameters and achieves real-time inference at 23 FPS on a Jetson AGX Xavier edge device, demonstrating its practical deployability. The proposed geometry-guided fusion strategy shows superior performance compared to state-of-the-art models such as PointNeXt and CrossViT, facilitating the transition from manual to intelligent defect inspection in high-speed rail wheelset maintenance.

This study holds significant scientific value by advancing the paradigm of multimodal fusion in industrial inspection, establishing a novel geometry-texture alignment mechanism that enhances both interpretability and robustness in defect detection under complex conditions. It contributes methodologically to the fields of 3D vision and deep learning by integrating structured point cloud processing with vision transformer-based feature learning. Socially, the work directly supports the safety, efficiency, and intelligence of high-speed railway operations. By enabling accurate and real-time defect identification even in challenging environments, the proposed system helps prevent potential failures, reduces maintenance downtime, and promotes the transition from labor-intensive manual checks toward automated, data-driven predictive maintenance, thereby enhancing rail transport reliability and public safety.

Despite these advances, the model exhibits performance degradation in extreme occlusion scenarios (>40%) and remains dependent on high-quality multimodal registration. Future work will focus on integrating unsupervised domain adaptation techniques to improve generalization across varying operational environments. Furthermore, extending the framework to support multi-task learning and broader multimodal collaborative prediction could enhance its applicability and robustness, thereby contributing to the further intelligence of industrial quality inspection systems.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JM: Writing – original draft, Investigation, Methodology, Data curation. XX: Writing – review and editing, Project administration, Methodology. BC: Software, Writing – review and editing, Visualization, Formal Analysis.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Deng, F., Ding, H., Lv, H., Hao, R., and Liu, Y. (2021). A method for fault diagnosis of high-speed railway wheelset bearings based on lightweight neural network. *J. Eng. Sci.* 43 (11), 1482–1490. doi:10.13374/j.issn2095-9389.2020.12.09.001

Deng, F., Cai, Y., Wang, R., and Zheng, S. (2024). Train wheelset bearing damage identification method based on convolution and transformer fusion framework. *J. Eng. Sci.* 46 (10), 1834–1844. doi:10.13374/j.issn2095-9389.2024.01.02.003

Gao, L., Zhang, J., Yang, C., and Zhou, Y. (2022). Cas-VSwin transformer: a variant swin transformer for surface-defect detection. *Comput. Industry* 140, 103689. doi:10.1016/j.compind.2022.103689

Geng, Y., Wang, Z., Jia, L., Qin, Y., Chai, Y., Liu, K., et al. (2023). 3DGraphSeg: a unified graph representation-based point cloud segmentation framework for full-range high-speed railway environments. *IEEE Trans. Industrial Inf.* 19 (12), 11430–11443. doi:10.1109/tii.2023.3246492

Guo, M., Junya, Y., Xuanbing, Q., Guo, A., Wang, J., Haiyan, X., et al. (2020). Embedded laser online detection of wheelset wear. *J. Henan Normal Univ. Nat. Sci. Ed.* 48 (5), 43–48. doi:10.16366/j.cnki.1000-2367.2020.05.007

Guo, Y., Wu, D., and Qingmin, W. (2023). A review of point cloud 3D object detection methods based on deep learning. *Appl. Res. Computers/Jisuanji Yingyong Yanjiu* 40 (1), 20. doi:10.19734/j.issn.1001-3695.2022.05.0251

He, J., Hou, N., Zhang, C., Hu, X., and Liu, J. (2022). Train wheel set tread damage diagnosis based on pyramid split attention. *J. China Saf. Sci.* 32 (5), 35. doi:10.16265/j.cnki.issn1003-3033.2022.05.2166

Huaiyu, C., Yang, C., Cui, Z., Yi, W., and Chen, X. (2024). Highway spill detection and positioning method based on image guidance and point cloud space constraints. *Optoelectron. Eng.* 51 (3), 230317. doi:10.12086/oee.2024.230317

Huang, Z., Gu, X., Wang, H., and Zhang, X. (2023). 3D point cloud data semantic segmentation and disordered grasping system based on improved PointNet++ transmission tower point cloud semantic segmentation model based on. *China Electr. Power* 56 (3), 77–85.

Irsal, R. B. P., Utaminingrum, F., and Ogata, K. (2024). Swin transformer adaptation into YOLOv7 for road damage detection. *Bull. Electr. Eng. Inf.* 13 (4), 2527–2536. doi:10.11591/eei.v13i4.7556

Jing, J., and Wang, H. (2023). Defect segmentation with local embedding in industrial 3D point clouds based on transformer. *Meas. Sci. Technol.* 35 (3), 035406. doi:10.1088/1361-6501/ad1289

Kang, L., Li, Z., Zhao, X., Zhao, Z., and Braun, T. (2023). Spatial–temporal point cloud transformer for sensing activity based on mmWave. *IEEE Internet Things J.* 11 (6), 10979–10991. doi:10.1109/jiot.2023.3329236

Kang, J., Mpabulungi, M., and Hong, H. (2025). Multi-modal CrossViT using 3D spatial information for visual localization. *Multimedia Tools Appl.* 84 (5), 2059–2083. doi:10.1007/s11042-024-20382-w

Li, Y., Xiang, Y., Guo, H., Liu, P., and Liu, C. (2022). Swin transformer combined with convolution neural network for surface defect detection. *Machines* 10 (11), 1083. doi:10.3390/machines10111083

Li, B., Meng, Q., Li, X., Wang, Z., Liu, X., Kong, S., et al. (2024). Enhancing YOLOv8's performance in complex traffic scenarios: optimization design for handling long-distance dependencies and complex feature relationships. *Electronics* 13 (22), 4411. doi:10.3390/electronics13224411

Li, Y., Tang, X., Liu, W., Huang, Y., and Li, Z. (2024). An improved method for detecting crane wheel–rail faults based on YOLOv8 and the swin transformer. *Sensors* 24 (13), 4086. doi:10.3390/s24134086

Liao, X., Jiang, H., Lin, P., Liu, H., Cai, Y., Lin, J., et al. (2023). Study on vibration characteristics of high-speed railway axle box system under bearing roller defect excitation. *J. Railw. Sci. and Eng.* 20 (9), 3262. doi:10.19713/j.cnki.43-1423/u.T20221893

Ma, Z., Zhou, S., and Lin, C. (2025). Defect detection in freight trains using a lightweight and effective multi-scale fusion framework with knowledge distillation. *Electronics* 14 (5), 925. doi:10.3390/electronics14050925

Shaikh, M. Z., Jatoi, S., Baro, E. N., Das, B., Hussain, S., Chowdhry, B. S., et al. (2025). FaultSeg: a dataset for train wheel defect detection. *Sci. Data* 12 (1), 309. doi:10.1038/s41597-025-04557-0

Shi, Q., Xu, W., and Miao, Z. (2024). Image-text multi-modal classification *via* cross-attention contextual transformer with modality-collaborative learning. *J. Electron. Imaging* 33 (4), 043042. doi:10.1117/1.jei.33.4.043042

Si, C., Luo, H., Han, Y., and Ma, Z. (2024). Rail-STrans: a rail surface defect segmentation method based on improved swin transformer. *Appl. Sci.* 14 (9), 3629. doi:10.3390/app14093629

Song, Y., Ji, Z., Guo, X., Hsu, Y., Feng, Q., Yin, S., et al. (2024). A comprehensive laser image dataset for real-time measurement of wheelset geometric parameters. *Sci. Data* 11 (1), 462. doi:10.1038/s41597-024-03288-y

Sun, F., Lin, G., Rui, X., Zhu, J., Zhou, Y., Wu, W., et al. (2024). Research on active obstacle recognition and distance perception based on fusion of visual image and 3D point cloud. *Mach. Tool and Hydraulics* 52 (16), 80. doi:10.3969/j.issn.1001-3881.2024.16.012

Tang, B., Song, Z. K., Sun, W., and Wang, X. D. (2023). An end-to-end steel surface defect detection approach *via* swin transformer. *IET Image Process.* 17 (5), 1334–1345. doi:10.1049/ipr2.12715

Vuković, V., and Kovalevskyy, S. (2024). Development of an innovative technical solution for the application of segmental managan inserts on the wear surface of the clamp of the tamping railway machines. *Adv. Eng. Lett.* 3, 42–51. doi:10.46793/adeletters.2024.3.2.1

Wan, R., Zhao, T., and Zhao, W. (2023). Pta-det: point transformer associating point cloud and image for 3d object detection. *Sensors* 23 (6), 3229. doi:10.3390/s23063229

Wang, D., Gao, K., Yuan, H., Yang, Y., Wang, Y., Kong, L., et al. (2024). Underwater image enhancement based on color correction and TransFormer detail sharpening. *J. Jilin Univ. Eng. Ed.* 54 (3), 785–796. doi:10.13229/j.cnki.jdxbgxb.20220483

Wang, Y., Miao, B., Zhang, Y., Huang, Z., and Xu, S. (2025). A novel rail damage fault detection method for high-speed railway. *Sensors* 25 (10), 3063. doi:10.3390/s25103063

Xiang, Y., Long, G., and Zhang, J. (2025). 3D point cloud data semantic segmentation and disordered grasping system based on PointNet++ network. *J. Mech. and Electr. Eng.* 42 (1), 146. doi:10.3969/j.issn.1001-4551.2025.01.016

Xie, Y., Zhang, L., Yu, X., and Xie, W. (2024). YOLOv5 target detection algorithm based on visible light-infrared feature interaction and fusion. *Control Theory and Applications/ Kongzhi Lilun Yu Yinyong* 41 (5), 914. doi:10.7641/CTA.2023.20475

Xu, H., Zheng, T., Liu, Y., Zhang, Z., Xue, C., Li, J., et al. (2024). A joint convolutional cross ViT network for hyperspectral and light detection and ranging fusion classification. *Remote Sens.* 16 (3), 489. doi:10.3390/rs16030489

Yang, N., Miao, Z., Wang, W., Zhang, J., and Sun, Y. (2023). Train wheel set tread defect recognition based on RP image attention fusion network. *J. Railw. Sci. and Eng.* 20 (12), 4811. doi:10.19713/j.cnki.43-1423/u.T20230152

Yang, L., Du, Y., and Dong, G. (2025). Lightweight point cloud classification model based on biased attention mechanism. *Laser and Optoelectron. Prog.* 62 (10), 1015006. doi:10.3788/LOP242058

Yu, X., He, W., Qian, X., Yang, Y., Zhang, T., Ou, L., et al. (2022). Real-time rail recognition based on 3D point clouds. *Meas. Sci. Technol.* 33 (10), 105207. doi:10.1088/1361-6501/ac750c

Zeng, N., Li, J., Zhang, Y., Gao, X., and Luo, L. (2023). Scattered train bolt point cloud segmentation based on hierarchical multi-scale feature learning. *Sensors* 23 (4), 2019. doi:10.3390/s23042019

Zhang, L., Huang, D., Liao, S., Yu, S., Ye, J., Wang, X., et al. (2021). Wheelset tread defect detection method based on target detection network. *Laser and Optoelectron. Prog.* 58 (4), 0410020. doi:10.3788/lop202158.0410020

Zhang, C., Hu, X., He, J., Liu, J., and Hou, N. (2022a). Train wheelset tread defect classification model based on SimAM and SpinalNet. *J. China Saf. Sci.* 32 (6), 38. doi:10.16265/j.cnki.issn1003-3033.2022.06.2563

Zhang, C., Hu, X., He, J., and Hou, N. (2022b). Yolov4 high-speed train wheelset tread defect detection system based on multi-scale feature fusion. *J. Adv. Transp.* 2022 (1), 1172654. doi:10.1155/2022/1172654

Zhang, Y., Wang, Y., Jiang, Z., Zheng, L., Chen, J., and Lu, J. (2023). Domain adaptation *via* transferable swin Transformer for tire defect detection. *Eng. Appl. Artif. Intell.* 122, 106109. doi:10.1016/j.engappai.2023.106109

Zheng, C., and Lin, H. (2023). YOLOv5 helmet wearing detection method based on Swin transformer. *Comput. Meas. and Control* 31 (3), 15. doi:10.16526/j.cnki.11-4762/tp.2023.03.003

Zhou, K., Lu, N., Jiang, B., and Ye, Z. (2025). FEV-Swin: multi-source heterogeneous information fusion under a variant swin transformer framework for intelligent cross-domain fault diagnosis. *Knowledge-Based Syst.* 310, 112982. doi:10.1016/j.knosys.2025.112982

Zhu, W., Zhang, H., Zhang, C., Zhu, X., Guan, Z., and Jia, J. (2023). Surface defect detection and classification of steel using an efficient swin transformer. *Adv. Eng. Inf.* 57, 102061. doi:10.1016/j.aei.2023.102061