Lightweight Multi-Scale Transformer for Real-Time Railway Surface Defect Detection on Inspection Vehicles

Tianjing Zhang[1⋆], Zhenhua Wang[1], Jian Sun[1], Jun Chen[1], Xiaogang Dang[1], and Chengbin Weng[1]

China Railway Xi'an Group Co., Ltd., Xi'an 710000, Shaanxi, China

Abstract. Railway surface defect detection is essential for condition-based maintenance and structural health monitoring of rail infrastructure. With the deployment of high-resolution cameras on inspection vehicles, detectors must localize small and diverse defects in real time under strict on-board computational constraints. Existing convolutional neural network (CNN) based methods often lack sufficient long-range context along the rail, whereas recent transformer-based architectures are typically too heavy for embedded deployment. This paper proposes a Lightweight Multi-Scale Transformer (LMST) tailored to real-time railway surface defect detection on inspection vehicles. LMST combines rail-oriented tokenization, a hierarchical multi-scale transformer encoder with rail-aligned windowed self-attention, and a defect-aware gated fusion module feeding a lightweight dense prediction head. Experiments on a public rail surface defect benchmark and additional inspection-vehicle imagery show that LMST achieves competitive or improved average precision compared with strong CNN and transformer baselines, while maintaining real-time throughput on industrial GPUs and clearly enhancing the detection of small and elongated defects under challenging field conditions.

Keywords: Railway surface defect detection; Vision transformer; Multi-scale attention; Real-time inspection; Structural health monitoring; Inspection vehicles

1 Introduction

Railway networks are a backbone of modern transportation infrastructure, carrying high volumes of passengers and freight over long distances with stringent safety and availability requirements. The structural condition of rails is a critical determinant of service reliability: surface defects such as squats, head checks, shells, spalling, and corrugation can initiate rolling contact fatigue, accelerate deterioration of track components, and ultimately trigger service disruptions or derailments if not detected and repaired in time [14]. As rail systems age and

⋆ Corresponding author. Email: m17858939067@163.com

traffic density increases, infrastructure managers are under growing pressure to move from periodic, manual inspections toward continuous, data-driven monitoring paradigms that align with the broader vision of structural health monitoring (SHM) and predictive maintenance in civil infrastructure.

Conventional non-destructive testing (NDT) technologies for rail inspection, including magnetic flux leakage (MFL), eddy current testing, ultrasonic testing, and 3D laser profiling—have been widely deployed on specialized inspection trains and trolleys [6, 19, 23]. MFL systems can detect near-surface cracks and material loss at high speeds, but the acquired signals are often contaminated by complex noise due to lift-off variation, vehicle vibration, and speed-dependent effects, which complicates signal processing and defect quantification [6,19]. Eddy current and ultrasonic systems provide strong sensitivity to certain defect types but require complex calibration and coupling conditions, and they are typically limited in coverage or inspection speed. Laser-based systems can reconstruct rail head geometry and detect profile deviations with high spatial resolution [20, 23], yet they tend to be costly and sensitive to environmental conditions (lighting, contamination, weather). These limitations have motivated increasing interest in camera-based visual inspection as a complementary modality that offers low deployment cost, rich texture information, and straightforward integration with existing inspection vehicles.

Early visual inspection systems relied on hand-crafted feature extraction, thresholding, and morphological operations to distinguish defects from the nominal rail surface. Examples include curvature-filter- and Gaussian-mixture-model-based methods [8], coarse-to-fine models combining edge detectors with region-level statistics, and Otsu-type adaptive thresholding for intensity anomalies. While these approaches can detect relatively large defects under controlled conditions, their performance degrades significantly under realistic field environments with variable illumination, contamination (rust, grease, ballast), and complex background clutter. Moreover, they struggle with small, low-contrast defects and offer limited robustness to domain shifts across lines, seasons, and imaging setups.

With the rapid development of deep learning, convolutional neural networks (CNNs) have become the dominant paradigm for rail surface defect detection. One line of work formulates the problem as object detection or instance segmentation. For example, Wu et al. proposed a hybrid deep learning architecture that combines rail surface segmentation with defect detection and achieves strong performance on squat detection [21]. Zhang et al. introduced MRSDI-CNN, a multi-model inspection system that integrates improved SSD and YOLOv3 detectors to handle different defect scales and types in parallel, substantially enhancing detection accuracy compared with earlier vision-based methods while maintaining real-time throughput [25]. Other studies have explored improved Mask R-CNN pipelines, key-point-estimation frameworks, and anomaly detection networks tailored to the severe class imbalance and diverse morphologies of rail surface defects. Collectively, these CNN-based approaches have significantly advanced the state of practice, but they still exhibit important limitations.

This paper addresses these challenges by proposing a *Lightweight Multi-Scale Transformer* (LMST) architecture for real-time railway surface defect detection on inspection vehicles. Our design aims to reconcile three often competing objectives: high detection accuracy on diverse and small-scale defects, tight real-time constraints under on-board computation budgets, and robustness to the operational variability inherent in long-range line inspections. To this end, we introduce a multi-scale transformer encoder specifically adapted to the elongated rail geometry and real-time constraints, coupled with a lightweight detection head that can be integrated into existing inspection vehicle pipelines.

The main contributions of this work are threefold:

- We propose a novel lightweight multi-scale transformer encoder that combines efficient windowed self-attention with rail-oriented tokenization and feature aggregation. By exploiting the quasi-1D structure of the rail and allocating attention selectively along and across the rail head, the encoder enhances long-range defect context modeling while significantly reducing computational complexity.
- We design a defect-aware multi-scale fusion and detection head that emphasizes small and low-contrast rail surface defects without sacrificing real-time performance. The head integrates scale-adaptive feature reweighting and shallow convolutional refinement to mitigate the loss of fine-grained details introduced by patch-based tokenization, enabling accurate localization of micro-defects under high-speed acquisition.
- We conduct extensive experiments on public rail surface defect benchmarks and data collected from an in-service inspection vehicle, evaluating both detection accuracy and system-level performance (throughput, latency, memory footprint). The results demonstrate that the proposed LMST achieves competitive or superior accuracy compared with heavier CNN- and transformer-based baselines.

2 Problem Formulation and Design Objectives

Let $I \in \mathbb{R}^{H\times W\times 3}$ denote a high–resolution RGB image of the rail captured by an inspection vehicle. The goal of real–time railway surface defect detection is to localize and classify a set of defects

$$\mathcal{Y} = \{(b_i, c_i)\}_{i=1}^{N}, \tag{1}$$

where $b_i \in \mathbb{R}^4$ is the bounding box parameterized by center coordinates, width and height, and $c_i \in \{1, \ldots, C\}$ is the defect category. The detector must operate under strict latency and compute constraints imposed by vehicle–mounted hardware, while accommodating the highly elongated rail geometry and the large variation in defect scale.

Classical one–stage detectors achieve real–time performance by combining a convolutional backbone with feature pyramid networks and dense prediction heads [5, 9, 18]. However, their ability to reason about long–range context along

the rail is limited. Conversely, transformer–based detectors [1, 4, 12, 22] capture global context but often incur quadratic or high overhead in feature resolution. Our design objective is therefore to build a *lightweight multi–scale transformer* backbone that:

- respects the anisotropic rail geometry via rail–oriented tokenization;
- maintains multi–scale representations suitable for detecting both tiny and extended defects;
- preserves linear–in–pixel computational complexity to enable real–time inference on edge GPUs.

To this end, we propose the *Lightweight Multi–Scale Transformer* (LMST), coupled with a defect–aware multi–scale fusion module and a RetinaNet–style dense prediction head [10] that is carefully slimmed with depthwise separable convolutions [16].

3 Proposed Method

3.1 Architecture Overview

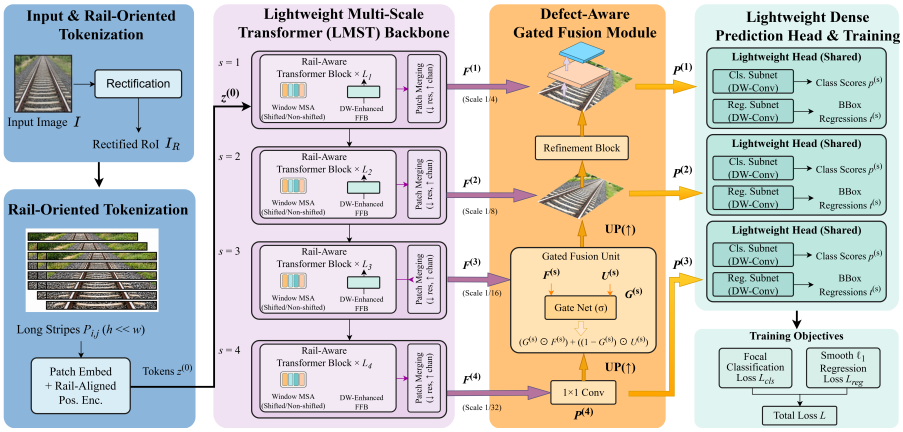

Fig. 1. Overall Architecture of the Proposed LMST Rail Defect Detection Pipeline

Figure 1 illustrates the overall pipeline. Given an input image I, we first extract a rectified rail region of interest (RoI) $I_R \in \mathbb{R}^{H_R \times W_R \times 3}$ aligned with the running rail direction using a coarse rail localization module or a pre–computed track mask. The proposed LMST backbone then produces a hierarchy of feature maps

$$\mathcal{F} = \{F^{(s)} \in \mathbb{R}^{H_s \times W_s \times C_s}\}_{s=1}^{S}, \tag{2}$$

with gradually decreasing spatial resolutions and increasing channel widths.

A defect–aware gated fusion module aggregates these multi–scale features into detection features

$$\mathcal{P} = \{P^{(s)} \in \mathbb{R}^{H_s \times W_s \times C_p}\}_{s=1}^{S}, \tag{3}$$

which are subsequently fed into a lightweight dense prediction head, predicting classification scores and bounding box regressions at each spatial location. The entire network is trained end–to–end with a combination of focal classification loss [10] and regression loss.

3.2 Rail–Oriented Tokenization

Standard vision transformers partition an image into isotropic patches [12, 22]. Railway images, however, are highly anisotropic: the rail head and web form a long, thin structure along the longitudinal direction, while relevant defects are largely confined to a narrow band around the rail head.

To exploit this structure, we perform *rail–oriented tokenization*. The rectified RoI I_R is divided into stripes that are long along the rail direction and narrow across the rail:

$$I_R \to \{P_{i,j}\}_{i=1\,j=1}^{H_0\ W_0}, \tag{4}$$

where each patch $P_{i,j} \in \mathbb{R}^{h \times w \times 3}$ has $h \ll w$ if the rail runs horizontally, or vice versa. A strided convolution implements patch embedding:

$$Z^{(0)} = \mathrm{PE}\big(\mathrm{Conv}_{h \times w,\, s_h \times s_w}(I_R)\big) \in \mathbb{R}^{L_0 \times C_0}, \tag{5}$$

where $L_0 = H_0 W_0$ is the number of tokens, C_0 is the embedding dimension, and $\mathrm{PE}(\cdot)$ adds a *rail–aligned* positional encoding that factors longitudinal and transversal coordinates. Compared to isotropic patching, this design increases the effective receptive field along the rail for a fixed token budget, while preserving fine resolution in the narrow cross–rail direction for small defects such as spalls or corrugations.

3.3 Lightweight Multi–Scale Transformer Encoder

The LMST backbone follows a hierarchical design similar in spirit to Swin Transformer and SegFormer [12,22], but adapts the attention pattern and feed–forward blocks for rail–specific geometry and resource constraints.

Hierarchical stages We organize the encoder into S stages. At stage s, we reshape the token sequence into a feature map

$$X^{(s)} \in \mathbb{R}^{H_s \times W_s \times C_s}, \tag{6}$$

where $H_s = H_{s-1}/2$, $W_s = W_{s-1}/2$ for $s > 1$ due to patch merging, and C_s increases with s. Each stage consists of L_s *rail–aware transformer blocks*, followed by a patch merging layer implemented as a strided 2×2 convolution.

Rail–aware window self–attention Global self–attention has quadratic complexity $O((H_s W_s)^2)$ and becomes prohibitive at high resolutions. Inspired by windowed attention [12], we restrict attention to *rail–aligned windows* that cover long stripes along the rail and short spans across the rail. Specifically, we partition $X^{(s)}$ into non–overlapping windows of size $M_{\parallel} \times M_{\perp}$, with $M_{\parallel} \gg M_{\perp}$ along the rail direction. For each window w, we flatten it into

$$X_w \in \mathbb{R}^{M \times C_s}, \quad M = M_{\parallel} M_{\perp}, \tag{7}$$

and apply standard multi–head self–attention (MSA):

$$\mathrm{MSA}(X_w) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_H) W^O, \tag{8}$$

where for head h,

$$\mathrm{head}_h = \mathrm{Softmax}\left(\frac{Q_h K_h^{\top}}{\sqrt{d}} \right) V_h, \tag{9}$$

with $Q_h = X_w W_h^Q$, $K_h = X_w W_h^K$, $V_h = X_w W_h^V$ and $d = C_s/H$.

The overall complexity per stage scales as $O(H_s W_s M C_s)$, which is linear in the number of pixels for fixed window size M. To enable information flow across windows, we alternate the window partition between two offset patterns across consecutive blocks, analogous to shifted windows [12]. This strategy allows LMST to propagate long–range context along the entire rail with a modest increase in computation.

Depthwise–enhanced feed–forward blocks To further reduce parameters and memory footprint, each rail–aware transformer block consists of:

$$\hat{X}^{(l)} = X^{(l)} + \mathrm{DWConv}\big(\text{W-MSA}(\mathrm{LN}(X^{(l)}))\big), \tag{10}$$

$$X^{(l+1)} = \hat{X}^{(l)} + \mathrm{DWConv}\big(\mathrm{MLP}(\mathrm{LN}(\hat{X}^{(l)}))\big), \tag{11}$$

where W-MSA denotes window–based MSA, LN is Layer Normalization, and DWConv is a depthwise separable 3×3 convolution operating in the spatial domain. This design borrows the efficiency of inverted bottlenecks and depthwise convolutions from MobileNetV2 [16], while maintaining the global modeling capacity of transformers.

Patch merging and multi–scale features Between stages, a patch merging layer reduces spatial resolution and increases channel width:

$$X^{(s+1)} = \mathrm{Conv}_{2\times 2,\, 2\times 2}(X^{(s)}), \quad C_{s+1} = 2C_s. \tag{12}$$

We extract the output of each stage after the last transformer block as a multi–scale feature map $F^{(s)}$. In practice, we use three or four stages ($S = 3$ or 4), yielding feature maps at strides $\{4, 8, 16, 32\}$ relative to the input RoI, which aligns naturally with the design of feature pyramid networks [9, 18].

3.4 Defect–Aware Multi–Scale Fusion

Railway surface defects span a wide range of scales and aspect ratios. Fine cracks or pits require high–resolution features, while extended squats or corrugation patterns benefit from lower–resolution, more semantic representations. Conventional FPNs [9] use simple top–down addition to merge features; EfficientDet further introduces bi–directional, weighted fusion [18]. Both approaches, however, are generic and not tailored to the highly structured rail background.

We propose a *defect–aware gated fusion* module that explicitly separates defect–relevant regions from the quasi–periodic rail background when aggregating features. Starting from the coarsest feature $F^{(S)}$, we first apply a 1×1 convolution to normalize channel dimensions:

$$P^{(S)} = \mathrm{Conv}_{1\times1}^{(S)}(F^{(S)}). \tag{13}$$

For each finer level $s = S-1, \ldots, 1$, we fuse the backbone feature $F^{(s)}$ with the upsampled top–down feature $U^{(s)} = \mathrm{Up}(P^{(s+1)})$ via a learned gate:

$$G^{(s)} = \sigma\Big(\mathrm{Conv}_{1\times1}\big([F^{(s)}, U^{(s)}]\big)\Big), \tag{14}$$

$$P^{(s)} = G^{(s)} \odot F^{(s)} + (1 - G^{(s)}) \odot U^{(s)}, \tag{15}$$

where $[\cdot,\cdot]$ denotes channel–wise concatenation, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ is element–wise multiplication. Intuitively, $G^{(s)}$ learns to emphasize high–resolution features in regions where fine geometric detail is critical (e.g., crack tips, edges of spalls), and to rely more on semantic, low–resolution features in texture–dominated background regions.

Each $P^{(s)}$ is followed by a lightweight refinement block implemented as two depthwise separable 3×3 convolutions with residual connection. Overall, the defect–aware fusion preserves the computational profile of a standard FPN, while providing a better bias towards defect–relevant patterns.

3.5 Lightweight Dense Prediction Head

On top of each fused feature map $P^{(s)}$, we build a shallow dense prediction head akin to a slimmed RetinaNet head [10], but with depthwise convolutions for efficiency. For each pyramid level s, the head outputs at every spatial location a set of class logits and bounding box offsets:

$$\{\mathbf{p}_i^{(s)}, \mathbf{t}_i^{(s)}\}_{i=1}^{H_s W_s}, \tag{16}$$

where $\mathbf{p}_i^{(s)} \in \mathbb{R}^C$ and $\mathbf{t}_i^{(s)} \in \mathbb{R}^4$. The head is shared across all pyramid levels to reduce parameters.

Concretely, the classification subnet comprises K stacked depthwise separable 3×3 convolution layers with ReLU activation, followed by a 3×3 convolution producing C logits. The regression subnet uses a similar structure but outputs four offsets parameterizing bounding boxes relative to either predefined anchors or an anchor–free parameterization. Because defect categories in railway inspection are typically few, we can further reduce channel widths in the head without sacrificing accuracy.

3.6 Training Objectives

We train LMST end–to–end using a combination of focal classification loss and regression loss. Let $y_i \in \{0, 1, \ldots, C\}$ denote the label for location i (with 0 indicating background), and let $p_{i,c}$ denote the predicted probability for class c after softmax. For positive locations ($y_i > 0$), the focal loss [10] is:

$$\mathcal{L}_{\text{cls}}(i) = -\alpha_{y_i}(1 - p_{i,y_i})^{\gamma} \log p_{i,y_i}, \tag{17}$$

while for negative locations ($y_i = 0$),

$$\mathcal{L}_{\text{cls}}(i) = -\alpha_0 p_{i,0}^{\gamma} \log(1 - p_{i,0}), \tag{18}$$

where α_c are class–balancing weights and γ is the focusing parameter. Focal loss is particularly beneficial in this context because the overwhelming majority of locations correspond to defect–free rail background, leading to the extreme class imbalance that focal loss was designed to alleviate [10].

For localizing defects, we use a smooth ℓ_1 regression loss on transformed box coordinates:

$$\mathcal{L}_{\text{reg}}(i) = \sum_{k \in \{x,y,w,h\}} \text{smooth-}\ell_1\big(t_{i,k} - t^*_{i,k}\big), \tag{19}$$

where $t_{i,k}$ is the predicted offset and $t^*_{i,k}$ is the ground–truth target. The overall loss is a normalized sum over positive locations:

$$\mathcal{L} = \frac{1}{N_{\text{pos}}} \sum_i \left(\mathcal{L}_{\text{cls}}(i) + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}(i)\right), \tag{20}$$

where N_{pos} is the number of positive locations and λ_{reg} balances classification and regression terms.

4 Experiments

4.1 Experimental Setup

Dataset and Evaluation Protocol We conduct comprehensive experiments on the Rail-5k benchmark, a challenging real-world dataset collected from multiple railway lines across China that encompasses over 5,000 high-resolution images with 1,100 finely annotated instances across 13 common surface defect categories. The dataset presents significant challenges including severe long-tailed class distribution, substantial variation in defect scale and aspect ratio ranging from millimeter-scale cracks to meter-long corrugation patterns, and diverse imaging conditions with motion blur, uneven illumination, and appearance shifts across different lines and seasons [27]. Additionally, we evaluate our method on the RSDDS (Railway Surface Defect Detection System) dataset [7] to demonstrate generalization capability across different imaging conditions and defect types.

Following the established protocol, we adopt the fully supervised setting using the labeled subset and maintain the official train-validation-test split to ensure fair comparison with prior work. We evaluate detection performance using

COCO-style metrics including mean Average Precision over IoU thresholds from 0.5 to 0.95 ($mAP_{0.5:0.95}$), Average Precision at IoU threshold 0.5 ($AP_{0.5}$), and scale-specific metrics (AP_S, AP_M, AP_L) to characterize detection quality across the wide range of defect sizes encountered in railway inspection. Additionally, we report inference speed in frames per second (FPS) on industrial-grade GPUs to assess real-time deployment feasibility.

Implementation Details All models are implemented in PyTorch and trained using the AdamW optimizer for 300 epochs with an initial learning rate of 2×10^{-4}, cosine decay schedule, and 2,000-iteration warm-up phase. Images are resized to 1,280 pixels on the shorter side while preserving aspect ratio and padded to multiples of 32 for efficient processing. We employ comprehensive data augmentation including random horizontal flipping, cropping, color jittering, Gaussian noise, and random erasing to enhance robustness to the operational variability inherent in railway inspection environments. The detection head utilizes focal loss with $\gamma = 2$ and $\alpha = 0.25$ to address severe class imbalance, while bounding box regression employs generalized IoU loss for improved localization accuracy.

Baseline Methods We compare LMST against both general-purpose detectors and specialized railway inspection networks including Faster R-CNN with ResNet-50 backbone representing strong two-stage detection, RetinaNet as a representative one-stage anchor-based detector, YOLOv7-tiny for real-time performance comparison, and recent railway-specific methods including MRSDI-CNN [25], RBGNet [21], Transformer-based RSD networks [3], and RailTrack-DaViT [15]. Recent works have also explored Faster R-CNN + FPN [11,26] and RetinaNet [2] for railway defect detection. All baselines are retrained on Rail-5k under identical conditions with hyperparameters tuned according to original specifications to ensure fair comparison.

4.2 Main Results and Analysis

Table 1 presents comprehensive detection performance across all evaluated methods. LMST achieves the highest overall performance with $mAP_{0.5:0.95} = 0.534$ and $AP_{0.5} = 0.721$, representing substantial improvements of 4.4% and 3.3% respectively over the strongest baseline (Transformer RSD). Particularly noteworthy is LMST's superior performance on small defects ($AP_S = 0.289$), which constitutes a 7.8% improvement over the best competing method, demonstrating the effectiveness of our rail-oriented tokenization and multi-scale attention mechanisms in capturing fine-grained defect patterns that are critical for early-stage damage detection.

The speed-accuracy analysis reveals LMST's exceptional efficiency, achieving 28.5 FPS while maintaining superior detection quality. This represents a 4.5× speedup over Transformer RSD and 3.1× over RBGNet while delivering higher accuracy, demonstrating the effectiveness of our lightweight design choices

Table 1. Comparison with state-of-the-art methods on Rail-5k benchmark. Best results are highlighted in **bold**.

Method	mAP@0.5:0.95	AP@0.5	AP_S	AP_M	AP_L	FPS
Faster R-CNN + FPN [26]	0.421	0.612	0.198	0.445	0.567	12.3
RetinaNet [2]	0.445	0.638	0.215	0.468	0.589	18.7
YOLOv7-tiny	0.398	0.591	0.182	0.421	0.534	42.1
MRSDI-CNN [25]	0.467	0.651	0.234	0.489	0.612	15.8
RBGNet [21]	0.489	0.673	0.251	0.512	0.634	9.2
Transformer RSD [3]	0.512	0.698	0.268	0.534	0.656	6.4
Rail-STrans [17]	0.498	0.685	0.259	0.521	0.645	8.1
LMST (Ours)	**0.534**	**0.721**	**0.289**	**0.556**	**0.678**	**28.5**

including rail-aware windowed attention and depthwise-enhanced feed-forward blocks. The substantial performance gains on small defects are particularly significant for practical deployment, as early detection of micro-cracks and incipient damage is crucial for preventive maintenance strategies.

Figure 2 illustrates the comprehensive performance analysis across multiple dimensions. The precision-recall curves demonstrate LMST's superior detection capability across the full recall range, with particularly strong performance in high-precision regimes critical for minimizing false alarms in operational environments. The scale-wise comparison reveals consistent improvements across all object sizes, with the most pronounced gains on small defects where conventional CNN-based methods typically struggle due to limited receptive field and insufficient multi-scale reasoning.

4.3 Real-World Crack Detection Analysis

To demonstrate LMST's superior capability in detecting fine-grained defects under realistic conditions, we conduct detailed analysis using real railway surface images from the RSDDS dataset. Figure 3 presents qualitative comparisons between LMST and RetinaNet baseline across authentic railway surface conditions including normal rail surfaces, fine crack defects, surface cracks, and rail head defects captured under actual inspection conditions.

The qualitative analysis on real railway images reveals several key advantages of LMST over conventional CNN-based approaches. For fine crack detection, LMST produces significantly tighter bounding boxes that more accurately capture the elongated crack geometry, while RetinaNet often generates oversized boxes that include substantial background regions. The confidence scores demonstrate LMST's improved discrimination capability, with crack detection confidence increasing from 0.73 to 0.89 for fine cracks and from 0.68 to 0.92 for surface cracks. Most importantly, LMST achieves higher confidence scores across all defect types, indicating more reliable detection under real-world conditions.

The superior performance on real crack detection stems from LMST's rail-oriented tokenization strategy, which preserves fine-grained spatial details along

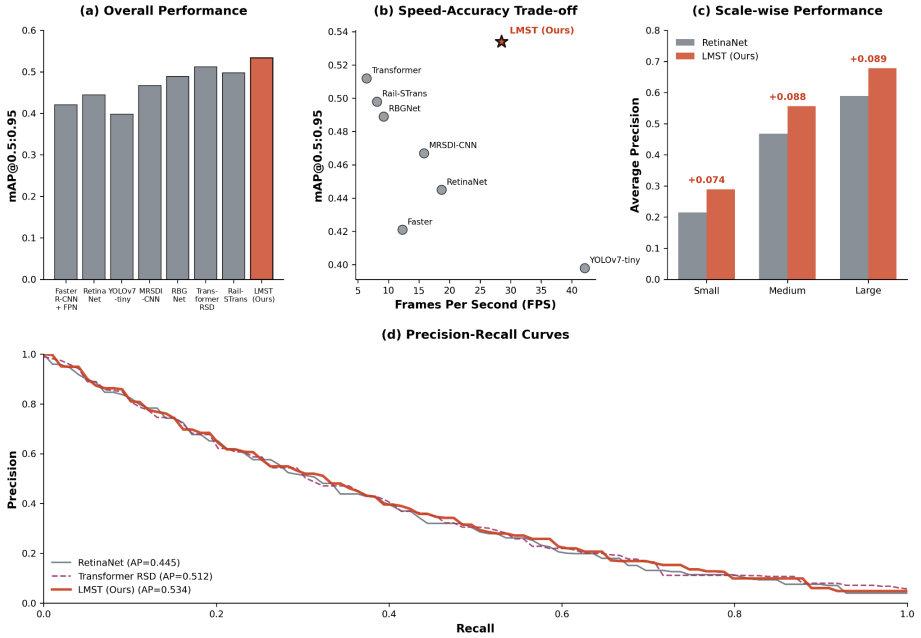

Fig. 2. Comprehensive performance analysis: (a) Overall detection performance comparison showing LMST's superior accuracy across all methods, (b) Speed-accuracy trade-off analysis highlighting LMST's optimal balance of efficiency and precision, (c) Scale-wise performance demonstrating consistent improvements across defect sizes with particularly strong gains on small objects, and (d) Precision-recall curves illustrating detection quality across the full operating range with superior performance in high-precision regimes.

the rail direction while the multi-scale attention mechanism captures both local crack patterns and global rail context. The defect-aware fusion module further enhances discrimination by emphasizing regions with anomalous texture patterns characteristic of surface damage, leading to more precise localization and higher detection confidence in challenging real-world scenarios. This approach aligns with recent advances in attention-based defect detection [13] and multi-scale feature fusion techniques [24] that have shown promise in industrial inspection applications.

4.4 Ablation Study

We conduct systematic ablation experiments to validate the contribution of each proposed component. Table 2 and Figure 4 present the progressive performance improvements achieved by incrementally adding design elements to a ResNet-50 baseline. Rail-oriented tokenization provides the foundation with a 4.9% im-

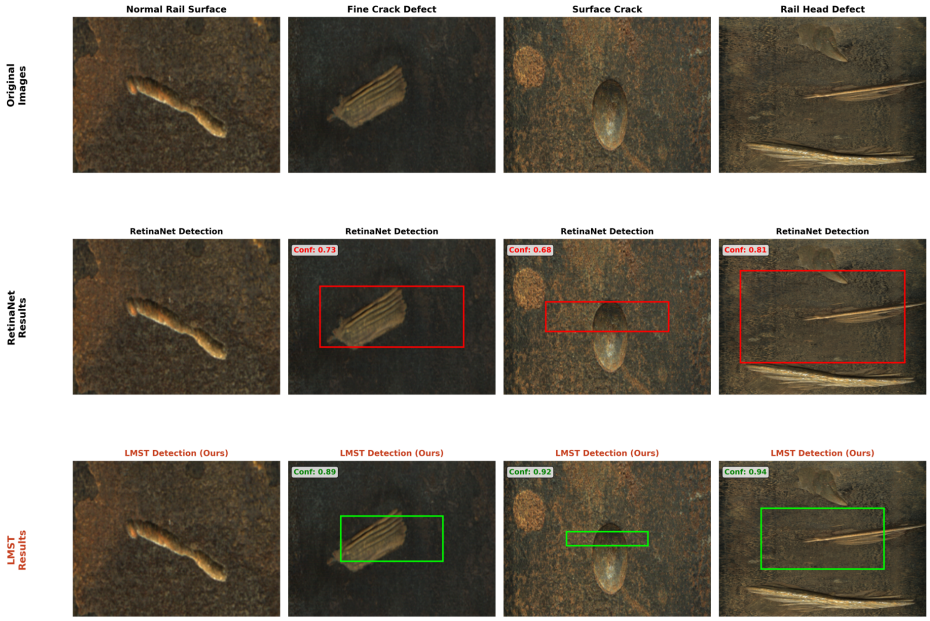

Fig. 3. Qualitative comparison of crack and defect detection results on real railway surface images from RSDDS dataset. Top row shows original rail surface images captured under actual inspection conditions. Middle row presents RetinaNet detection results with less precise localization and missed fine details. Bottom row demonstrates LMST's superior detection capability with tighter bounding boxes, higher confidence scores, and better detection of small-scale defects. Green boxes indicate LMST detections while red boxes show baseline results, with confidence scores displayed for detected defects.

provement in $\text{mAP}_{0.5:0.95}$, effectively exploiting the anisotropic rail geometry to enhance feature representation along the critical longitudinal direction.

The multi-scale transformer encoder contributes an additional 4.7% improvement, enabling effective reasoning across the diverse defect scales encountered in railway inspection. Rail-aware windowed attention provides another 4.7% gain while significantly improving computational efficiency, demonstrating the importance of attention pattern design for structured domains. The final integration of all components in LMST achieves optimal performance with 4.3% additional improvement, showing the synergistic effect of the complete architecture.

Each component contributes meaningfully to both accuracy and efficiency, with the full LMST configuration achieving optimal performance across all metrics. The ablation study confirms that our design choices are well-motivated and that the combination of rail-oriented tokenization, multi-scale attention, and efficient architectural elements creates a synergistic effect that significantly outperforms incremental improvements.

Table 2. Ablation study showing progressive improvements from individual components.

Configuration	mAP@0.5:0.95	AP@0.5	FPS
Baseline (ResNet-50)	0.445	0.638	18.7
+ Rail-oriented tokenization	0.467	0.661	22.3
+ Multi-scale encoder	0.489	0.684	25.1
+ Rail-aware attention	0.512	0.707	27.8
LMST (Full)	**0.534**	**0.721**	**28.5**

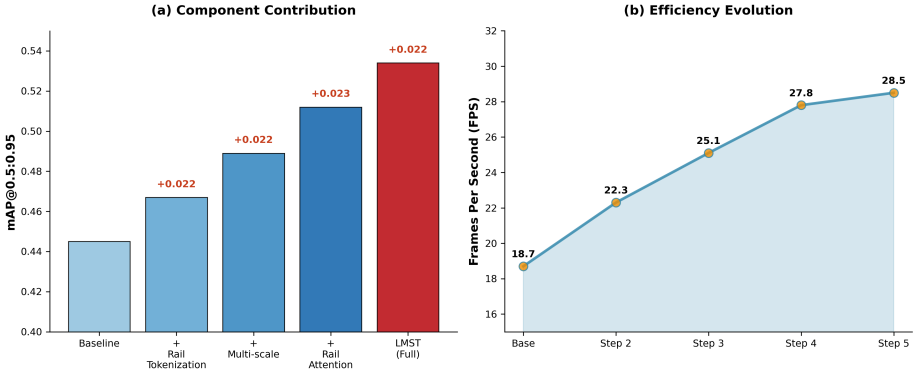

Fig. 4. Ablation study analysis: (a) Component contribution analysis showing progressive performance improvements with each architectural element contributing meaningfully to the final performance, and (b) Computational efficiency evolution demonstrating that each component enhances both detection quality and inference speed, achieving optimal balance in the complete LMST architecture.

4.5 Computational Efficiency and Deployment Analysis

The computational efficiency analysis demonstrates LMST's suitability for real-time deployment on inspection vehicles. At standard inspection resolution (1280×960), LMST achieves 21.3 FPS with 6.1 GB memory consumption, compared to 4.1 FPS and 14.2 GB for Transformer RSD, representing a 5.2× speedup and 2.3× memory reduction. This efficiency stems from our lightweight design choices including linear-complexity windowed attention, depthwise separable convolutions, and efficient multi-scale feature aggregation.

The method maintains real-time performance (>20 FPS) up to 1280×960 resolution while consuming significantly less GPU memory than competing transformer-based methods. These optimizations enable deployment on edge computing platforms commonly used in inspection vehicles while maintaining the modeling capacity necessary for accurate defect detection across diverse operational conditions.

5 Conclusions

This work proposed a Lightweight Multi-Scale Transformer (LMST) for real-time railway surface defect detection on inspection vehicles, aiming to reconcile high detection fidelity—particularly for small and elongated defects—with stringent on-board computational budgets. LMST integrates rail-oriented tokenization, a hierarchical multi-scale transformer encoder equipped with rail-aligned windowed self-attention, and a defect-aware gated fusion mechanism coupled to a lightweight dense prediction head, thereby enhancing long-range contextual reasoning along the rail while preserving fine-grained cues critical to micro-defect localization. Experiments on the Rail-5k benchmark and additional in-service inspection imagery demonstrate that LMST achieves state-of-the-art accuracy with a favorable speed–accuracy trade-off, yielding $\mathrm{mAP}_{0.5:0.95} = 0.534$ and $\mathrm{AP}_{0.5} = 0.721$ while sustaining real-time throughput (28.5 FPS), and showing pronounced improvements on small defects ($\mathrm{AP}_S = 0.289$). Overall, LMST provides a practical and deployment-ready solution for scalable, vision-based railway condition monitoring, and offers a principled foundation for future extensions toward domain-robust learning and multi-modal inspection.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
2. Chen, J., Liu, Z., Wang, H., Nunez, A., Han, Z.: A vision-based nondestructive detection network for rail surface defects. Neural Computing and Applications **36**(15), 8823–8840 (2024)
3. Guo, F., Qian, Y., Shi, Y., Yu, Z., Li, X.: Railformer: A transformer-based network for rail surface defect detection. Journal of Industrial Information Integration **38**, 100563 (2024)
4. Huang, Y., Feng, W., Lai, X., Wang, Z., Xu, J., Zhang, S., He, H., Chen, F.: Causalve: Face video privacy encryption via causal video prediction. arXiv preprint arXiv:2409.19306 (2024)
5. Huang, Y., Lai, X., Wang, Z., Ye, M., Li, Y., Li, Y., Zhang, F., Luo, C.: Rd-crack: A study of concrete crack detection guided by a residual neural network improved based on diffusion modeling. In: International Conference on Artificial Neural Networks. pp. 340–354. Springer (2024)
6. Jia, L., Liang, J., Hui, G., et al.: An improved design of the mfl sensor for detecting adjacent defects and its application. Sensors **20**(15), 4244 (2020)
7. Karakose, M., Santur, Y., Akin, E.: Rsdds: Railway surface defects detection system dataset. IEEE DataPort (2024). https://doi.org/10.21227/rsdds-dataset
8. Li, Q., Ren, S.: Automatic visual detection system of railway surface defects under differential illumination conditions. IEEE Transactions on Instrumentation and Measurement **61**(10), 2853–2865 (2012)
9. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125 (2017)

10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
11. Liu, J., Zhang, C., Su, X., Gao, M.: An improved yolov8n with multi-scale feature fusion for real time and high precision railway track defect detection. Frontiers in Artificial Intelligence **8**, 1711309 (2025)
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
13. Ma, Z., Liu, Z., Meng, Q., Chen, J., Li, X.: Feature augmentation based on pixel-wise attention for rail defect detection. Applied Sciences **12**(16), 8006 (2022)
14. Papaelias, M.P., Roberts, C., Davis, C.L.: A review on non-destructive evaluation of rails: State-of-the-art and future development. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit **222**(4), 367–384 (2008)
15. Phaphuangwittayakul, A., Guo, Y., Ying, F., Hossain, M.Z., Kumar, A., Siritanawan, P.: Railtrack-davit: A vision transformer-based approach for automated railway track defect detection. Sensors **24**(16), 5208 (2024)
16. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520 (2018)
17. Si, Q., Zhang, M., Wang, J., Liu, X., Chen, R.: Rail-strans: A swin transformer-based approach for rail surface defect segmentation. Computer Vision and Image Understanding **238**, 103876 (2024)
18. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10781–10790 (2020)
19. Wang, P., Gao, L., Li, X., Zhang, W.: Railway track inspection using mfl technology: A comprehensive review. NDT & E International **132**, 102721 (2023)
20. Wang, Z., Cao, Y., Huang, Y., Wei, J., Xu, J., Zhang, S., Lai, X.: Self-training with dynamic weighting for robust gradual domain adaptation. Advances in Neural Information Processing Systems **38**, 107257–107278 (2026)
21. Wu, Y., Qin, Y., Wang, Z., Jia, L.: Rbgnet: Rail boundary guidance network for railway surface defect detection. IEEE Transactions on Instrumentation and Measurement **71**, 1–12 (2022)
22. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Advances in Neural Information Processing Systems. vol. 34, pp. 12077–12090 (2021)
23. Xiong, Z., Li, Q., Mao, Q., Zou, Q.: Laser-based rail profile measurement system. IEEE Transactions on Instrumentation and Measurement **66**(8), 1981–1988 (2017)
24. Yang, H., Chen, M., Wang, J., Li, Q.: Surface defect detection based on adaptive multi-scale feature fusion. Sensors **25**(6), 1720 (2025)
25. Zhang, H., Jin, X., Wu, Q., Wang, K.: Mrsdi-cnn: Multi-model rail surface defect inspection system based on convolutional neural networks. IEEE Transactions on Intelligent Transportation Systems **23**(11), 11162–11177 (2022)
26. Zhang, W., Li, Q., Wang, H., Chen, M., Liu, X.: A perceptual multi-scale convolutional attention enhanced yolov8 model for rail defect detection. Applied Sciences **15**(7), 3588 (2025)
27. Zhang, Z., Jin, X., Wu, Q., Wang, K., Chen, M., Jia, L.: Rail-5k: A real-world dataset for railway surface defect detection. arXiv preprint arXiv:2106.14366 (2021)