Advances in Engineering Research : Possibilities and Challenges
ISSN Print: 3079-5192
ISSN Online: 3079-5206
https://doi.org/10.63313/AERpc.9088

Improved YOLOv11n-Based Detection Algorithm for Railway Surface Defect Inspection

Jifa Li
Tianjin University of Technology and Education, Tianjin, 300222, China
Email: 2876212217@qq.com

How to cite this paper: Li, J. F. (2026). Improved YOLOv11n-Based Detection Algorithm for Railway Surface Defect Inspection. *Advances in Engineering Research: Possibilities and Challenges*, 4(1), 39–52. ISSN Print: 3079-5192; ISSN Online: 3079-5206.
https://doi.org/10.63313/AERpc.9088
Published: 2026-04-15

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Abstract

Railway surface defect detection is a critical component of automated track maintenance and safety assurance. Standard deep learning detectors face three domain-specific challenges in rail inspection: high visual similarity between de-fects and metallic rail surface texture, large spatial scale disparity across defect categories, and insufficient receptive field diversity for capturing large-area peri-odic corrugation patterns. To address these challenges, this paper proposes RSD-YOLO, an improved YOLOv11n-based detector incorporating three targeted structural modifications. First, Convolutional Block Attention Modules (CBAM) are inserted after the P3 and P4 backbone stages to provide multi-level dual-axis feature recalibration, suppressing metallic background noise and amplifying de-fect-discriminative activations. Second, BiFPN adaptive weighted fusion (BiFPN_Concat2) replaces standard PANet concatenation in the neck, enabling learnable asymmetric cross-scale feature interaction. Third, Atrous Spatial Pyra-mid Pooling (ASPP) with dilation rates {6, 12, 18} replaces the SPPF module at the backbone apex to provide parallel multi-receptive-field context aggregation without resolution loss. Experiments on a custom 8,000-image rail defect dataset of four categories demonstrate that RSD-YOLO achieves mAP@0.5 of 84.7% and mAP@0.5:0.95 of 64.8%, outperforming the YOLOv11n baseline by 3.8% and 3.4% respectively. Wheel Burn detection improves by 6.8% and Corrugation by 5.6%, confirming the model's effectiveness for the most diagnostically challenging defect categories in railway surface inspection.

Keywords

Railway Surface Defect Detection; Yolov11; Convolutional Block Attention Module; Atrous Spatial Pyramid Pooling; Deep Learning

1. Introduction

Railway infrastructure safety directly determines the reliability and operational continuity of rail transport systems. Rail surface defects — including spalling fractures, wheel burn damage, squat rolling contact fatigue cracks, and corrugation wear — constitute primary failure modes responsible for rail breakage and derailment incidents. These defects present heterogeneous detection challenges:

spalling appears as pitted regions with sharp boundaries, wheel burn produces thermally diffuse discolored patches blending into metallic rail texture, squat manifests as subtle dark indentations with branching subsurface cracks, and corrugation forms periodic undulation patterns visually similar to normal surface roughness. Traditional inspection methods relying on manual examination or rule-based machine vision cannot satisfy modern high-speed railway maintenance requirements, making automated deep learning-based detection an essential engineering priority.

The YOLO series of single-stage detectors has established itself as the dominant paradigm for real-time industrial inspection owing to its end-to-end regression design and favorable accuracy-speed trade-off. YOLOv11 strengthens multi-scale feature extraction through improved C3k2 cross-stage partial modules and a C2PSA spatial attention block at the backbone apex. Despite these advances, directly applying standard YOLO detectors to rail surface inspection yields suboptimal performance due to three unresolved limitations: (1) attention guidance applied only at P5/32 through C2PSA leaves P3/8 and P4/16 stages without feature recalibration, resulting in insufficient metallic texture suppression at the hierarchy levels where Wheel Burn and Spalling discriminative features are primarily formed; (2) standard PANet uniform-weight concatenation cannot adaptively balance scale contributions for co-occurring fine and large-area defects; (3) SPPF's cascaded single-kernel pooling provides insufficient receptive field diversity for capturing corrugation patterns spanning hundreds of pixels.

Currently, deep learning-based defect detection methods have been widely applied across industrial inspection domains, providing important references for rail surface inspection research. Zhao et al. introduced BiFPN into the neck of YOLOv8 to achieve bidirectional feature fusion and added a coordinate attention mechanism, demonstrating the effectiveness of learnable weighted neck fusion for multi-scale inspection tasks [1]. Liu et al. applied the CBAM attention module within the backbone to suppress complex metallic texture backgrounds and reduce false positive rates, verifying that dual-axis attention can effectively disentangle defect features from high-frequency surface patterns [4]. Chen et al. proposed replacing standard single-scale pooling with ASPP to capture elongated crack structures at multiple spatial scales simultaneously, demonstrating that parallel dilated convolutions provide richer context for defects with diverse spatial extent [12]. Although these methods have improved performance in various scenarios, direct application to rail surface inspection still faces the three limitations identified above. Therefore, this paper proposes RSD-YOLO, an improved YOLOv11n-based object detection algorithm for rail surface defect detection. The specific improvements are as follows:

(1) CBAM is inserted after the C3k2 output at both the P3/8 and P4/16 backbone stages, extending the baseline's single-level C2PSA attention to full

multi-level coverage for superior metallic background suppression.

(2) BiFPN adaptive weighted fusion (BiFPN_Concat2) replaces standard PANet concatenation at all neck fusion nodes, enabling the network to autonomously learn asymmetric cross-scale fusion strategies suited to each defect category's scale characteristics.

(3) ASPP with parallel dilated convolutions replaces the SPPF module at the backbone apex, providing simultaneous small, medium, and large spatial context aggregation without feature map resolution degradation.

2. YOLOv11 Algorithm

YOLOv11 is a new-generation object detection network open-sourced by Ultralytics in 2024. It improves upon YOLOv8 and supports diverse visual tasks including object detection, instance segmentation, and pose estimation, inheriting the core design philosophy of single-stage, end-to-end detection while incorporating targeted architectural optimizations in the backbone, neck, and head components. YOLOv11 is available in five model scales—n, s, m, l, and x—corresponding to increasing parameter budgets and detection capacities. YOLOv11n, as the lightest configuration in the series, achieves a favorable balance between parameter efficiency, inference speed, and detection accuracy, making it well-suited as the backbone for task-specific architectural improvements in resource-constrained deployment scenarios such as on-board rail inspection systems. YOLOv11 primarily comprises three components: the Backbone network, the Neck network, and the Head network.

(1) Backbone: The backbone is responsible for hierarchical multi-scale feature extraction from input images. YOLOv11 introduces the improved C3k2 module to replace the C2f module from YOLOv8, adopting a Cross Stage Partial dual-kernel structure that promotes feature reuse while reducing redundant computation. The backbone produces feature maps at four spatial scales through successive strided convolutions interleaved with C3k2 blocks. At the backbone apex, a Spatial Pyramid Pooling Fast (SPPF) module aggregates multi-scale contextual information through cascaded max-pooling operations, followed by a C2PSA (Cross Stage Partial with Spatial Attention) block that applies self-attention-based feature refinement to enhance spatial awareness at the highest-level semantic representation.

(2) Neck: The neck continues the PAN-FPN architecture from YOLOv8, bridging the semantic gap between deep and shallow feature maps through bidirectional pathways. The top-down FPN path enriches shallow high-resolution features with deep semantic information through upsampling and concatenation, while the bottom-up PAN path restores fine spatial detail to deeper semantic features through strided convolutions. This dual-path fusion produces three detection-scale outputs at P3 (stride 8), P4 (stride 16), and P5 (stride 32), enabling multi-scale object detection across diverse target sizes.

(3) Head: YOLOv11 employs a Decoupled Head that separates the classification and

bounding box regression tasks into two independent parallel branches, alleviating the optimization conflict inherent in coupled prediction designs. Combined with an anchor-free prediction mechanism, this design eliminates the need for manual anchor configuration and improves bounding box localization precision across variable defect geometries.

3. RSD-YOLO

This paper proposes improvements upon the YOLOv11n baseline to enhance rail surface defect detection while maintaining efficient inference suitable for on-board inspection deployment. Figure 1 shows the overall network structure of RSD-YOLO. CBAM attention modules are inserted at the P3 and P4 backbone stages for dual-axis feature recalibration. BiFPN adaptive weighted fusion replaces standard PANet concatenation in the neck to enable learnable asymmetric cross-scale interaction. The SPPF module is replaced with ASPP at the backbone apex for parallel multi-receptive-field context aggregation.

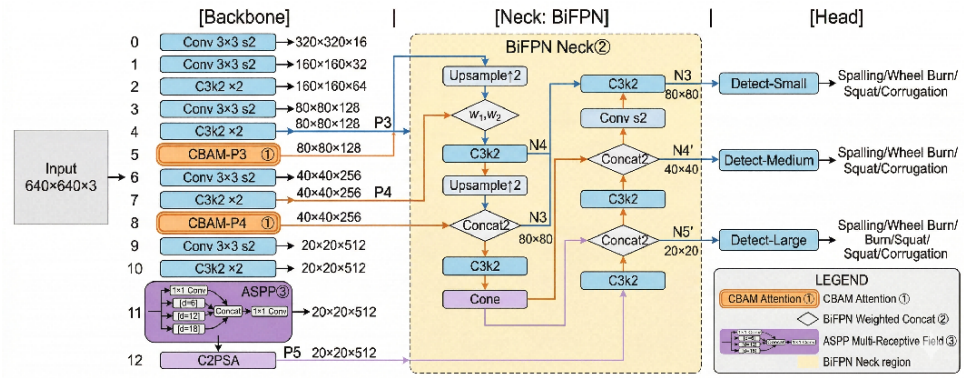


Figure 1. RSD-YOLO Network Structure

3.1. Convolutional Block Attention Module

The Convolutional Block Attention Module (CBAM) is a dual-axis attention mechanism that sequentially applies channel-wise and spatial-wise recalibration to intermediate feature maps [11]. In standard YOLOv11n, explicit attention guidance is provided only at P5/32 through C2PSA, leaving P3/8 and P4/16—where Wheel Burn and Spalling discriminative features are primarily formed—without recalibration. RSD-YOLO inserts CBAM after the C3k2 output at both the P3/8 and P4/16 backbone stages.

The Channel Attention Module (CAM) applies global average pooling and global max pooling in parallel, passes each through a shared MLP (reduction ratio r = 16), and combines via element-wise summation with Sigmoid activation to produce a channel weight vector $M_c \in \mathbb{R}^{C\times1\times1}$:

$$M_c(F) = \sigma\bigl(MLP(AvgPool(F)) + MLP(MaxPool(F))\bigr) \tag{1}$$

The Spatial Attention Module (SAM) then computes a spatial weight map M_s ∈ ℝ^{1×H×W} from channel-recalibrated features F' = M_c(F) ⊗ F, using a 7×7 convolutional layer to capture broader local context:

$$M_s(F') = \sigma\bigl(f^{7\times7}([AvgPool(F'); MaxPool(F')])\bigr)\tag{2}$$

The final output $\hat{F}$ = M_s(F') ⊗ F'. The dual-stage CBAM insertion at P3 and P4 enables simultaneous suppression of metallic background texture through channel recalibration and precise localization of defect-relevant spatial regions through spatial attention. This contributes primarily to Wheel Burn improvement (+6.8% AP) where thermally diffuse boundaries are most sensitive to spatial attention guidance, and to Spalling improvement (+2.8% AP) where channel recalibration separates fracture-specific feature channels from surface texture channels.

3.2. BiFPN Adaptive Weighted Feature Fusion

The standard PANet neck concatenates features from adjacent pyramid levels with equal implicit weighting, which is suboptimal for rail defect detection where different categories have incompatible scale preferences. Fine Spalling fractures require P3 high-resolution spatial detail, while large-area Corrugation damage requires dominant P5 semantic context to distinguish periodic surface anomalies from normal rail roughness. Uniform-weight concatenation prevents the neck from developing the asymmetric fusion strategies that different defect categories require.

BiFPN_Concat2 introduces learnable non-negative scalar weights w_1, $w_2 \geq 0$ at each dual-input neck fusion node, normalized via fast normalization for numerical stability:

$$O = \frac{w_1 \cdot I_1 + w_2 \cdot I_2}{w_1 + w_2 + \varepsilon}\tag{3}$$

where I_1, I_2 are input feature maps, weights are initialized to 1.0 for symmetric behavior at training onset, and ε = 0.0001 ensures numerical stability. This formulation is applied at all four neck fusion nodes in both the top-down and bottom-up pathways. BiFPN_Concat2 is a two-input variant specifically adapted to YOLOv11n's PANet structure, introducing only two learnable scalar parameters per fusion node for negligible parameter overhead. The learned asymmetric weights enable P3-dominant fusion for fine-defect detection scales and P5-dominant fusion for large-area defect scales, contributing to improved Corrugation (+5.6% AP) and Wheel Burn (+6.8% AP) detection through more targeted cross-scale feature interaction.

3.3. Atrous Spatial Pyramid Pooling

The SPPF module in standard YOLOv11n aggregates multi-scale context through cascaded 5×5 max-pooling, providing limited receptive field diversity through a

single kernel size and introducing spatial aliasing through repeated max selection. For Corrugation defects—characterized by quasi-sinusoidal longitudinal patterns spanning hundreds of pixels—this design provides insufficient coverage to simultaneously capture both local amplitude characteristics and global periodic regularity within the same forward pass.

ASPP employs four parallel branches operating on input $F \in \mathbb{R}^{C\times H\times W}$: a 1×1 pointwise convolution and three 3×3 dilated convolutions with dilation rates d ∈ {6, 12, 18}, providing effective receptive fields of approximately 13×13, 25×25, and 37×37 respectively. Each dilated branch uses zero-padding equal to its dilation rate to maintain spatial dimensions, followed by BN and ReLU. Branch outputs are concatenated and fused via 1×1 convolution to restore the original channel count C:

$$\begin{aligned} ASPP(F) = BN\backslash bigl(Conv_{1\times1}([f_{1\times1}(F); \\ f_{3\times3,d=6}(F); \\ f_{3\times3,d=12}(F); \\ f_{3\times3,d=18}(F)])\backslash bigr)\backslash tag4 \end{aligned}$$

The dilation rates {6, 12, 18} provide an approximately geometric progression ensuring comprehensive coverage from local detail to large-area periodic context without rate gaps. The fully preserved spatial resolution ensures multi-scale features passed to C2PSA and subsequently to the neck retain spatial precision for accurate fine-boundary defect localization. In RSD-YOLO, ASPP provides the backbone apex with simultaneous access to local, medium, and global spatial context, enabling robust Corrugation detection regardless of spatial frequency and longitudinal extent within the image field of view.

4. Dataset and Evaluation Metrics

4.1. Experimental Setup and Dataset

All experiments were conducted on a server equipped with an Intel Xeon CPU and an NVIDIA A40 GPU (48 GB VRAM). The software environment comprised Python 3.8 and Ultralytics YOLOv11. Training hyperparameters were set as follows: batch size = 16, input resolution = 640×640, epochs = 300, initial learning rate lr_0 = 0.01, momentum = 0.937, weight decay = 0.0005, and warm-up epochs = 3. Early-stopping patience was set to 50 epochs. Standard mosaic augmentation, random horizontal flipping (p = 0.5), HSV color-space jitter, random scaling, and random erasing (p = 0.4) augmentations were applied during training.

A custom rail surface defect dataset comprising 8,000 RGB images was constructed from operational railway track inspection under operational illumination conditions, manually annotated with axis-aligned bounding boxes by experienced railway maintenance engineers. The dataset covers 4 defect categories as shown in Table 1, split 8:1:1 into 6,400 training, 800 validation, and 800 test images. Figure 2 shows

representative sample images from each defect category.

Table 1. Categorization and descriptions of rail surface defects.

Category	Description
Spalling	Localized material fracture presenting as irregularly shaped pitted regions with sharp fracture boundaries
Wheel Burn	Thermally induced surface damage with discolored patches and diffuse irregular boundaries blending into surrounding rail texture
Squat	Rolling contact fatigue crack appearing as a dark indentation with branching subsurface cracks and surface deformation
Corrugation	Periodic quasi-sinusoidal longitudinal roughness patterns caused by rail-wheel dynamic interaction instability under repeated loading cycles

The dataset was randomly divided into training, validation, and test subsets in an 8:1:1 ratio, yielding 6,400 training images, 800 validation images, and 800 test images. Figure 2 shows representative sample images from each defect category.



Figure 2. Sample Dataset Images

4.2. Evaluation Metrics

This paper uses Average Precision (AP), Precision (P), Recall (R), and mean Average Precision (mAP) as evaluation metrics for object detection. The corresponding calculation formulas are shown in Equations (7) to (10):

$$AP = \int_0^1 PRdR \tag{7}$$

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{N} AP_i \tag{10}$$

where: TP represents the number of samples where the target is correctly predicted as positive, FP represents the number of samples where the target is incorrectly predicted as positive, FN represents the number of samples where the target is incorrectly predicted as negative; N is the total number of categories.

5. Experimental Results and Analysis

5.1. Ablation Experiment Results and Analysis

To quantify the individual and cumulative contributions of the three proposed improvements, ablation experiments were conducted by progressively incorporating CBAM, BiFPN, and ASPP into the YOLOv11n baseline, with all conditions sharing identical training hyperparameters, random seeds, and data splits. Results are summarized in Table 2. Figure 3 shows the Precision-Recall curve for the baseline YOLOv11n model, and Figure 4 shows the Precision-Recall curve for the complete RSD-YOLO model.

Table 2. Ablation Experiment Results Comparison

Model	CBAM	BiFPN	ASPP	mAP @0.5/%	mAP @0.5:0.95/%	Model Size/MB	FPS/（f/s）
Yolov11n	×	×	×	80.9	61.4	6.6	85.3
Yolov11n-C	√	×	×	82.3	62.4	7.1	83.1
Yolov11n-Bi	×	√	×	82.1	62.1	6.9	81.7
Yolov11n-ASPP	×	×	√	82.0	62.0	7.3	79.4
Yolov11n-CB	√	√	×	83.7	63.7	8.0	78.6
RSD-YOLO	√	√	√	84.7	64.8	8.9	74.2

As shown in Table 2, the baseline YOLOv11n achieves 80.9% mAP@0.5, with strongest per-category performance on Squat (90.1% AP) due to its clearly defined morphology, while Corrugation (74.0% AP) and Wheel Burn (81.1% AP) remain comparatively weak. When only CBAM is inserted, mAP@0.5 improves to 82.3% (+1.4%), with Wheel Burn benefiting most (+3.2% AP), confirming that spatial attention at P4/16 is particularly effective for focusing on thermally diffuse boundary regions. Standalone BiFPN achieves 82.1% (+1.2%), driven by Corrugation (+2.1%) and Wheel Burn (+2.4%) gains from learnable asymmetric fusion weighting. Standalone ASPP reaches 82.0% (+1.1%), with Corrugation improving most (+2.8% AP), directly validating the hypothesis that parallel dilated convolutions capture periodic spatial structure at multiple wavelengths.

Progressive combination reveals complementary interactions: CBAM+BiFPN

(YOLOv11n-CB) achieves 83.7%, a 2.8% improvement that marginally exceeds the arithmetic sum of individual gains (2.6%), indicating that CBAM-refined backbone features improve the quality of representations available for BiFPN adaptive weighting. The complete RSD-YOLO achieves 84.7% mAP@0.5 (+3.8%) and 64.8% mAP@0.5:0.95 (+3.4%). The increased model footprint (8.9 MB vs. 6.6 MB) and moderately reduced throughput (74.2 vs. 85.3 FPS) represent an acceptable accuracy-efficiency trade-off for inspection scenarios where detection reliability takes precedence over maximum throughput.

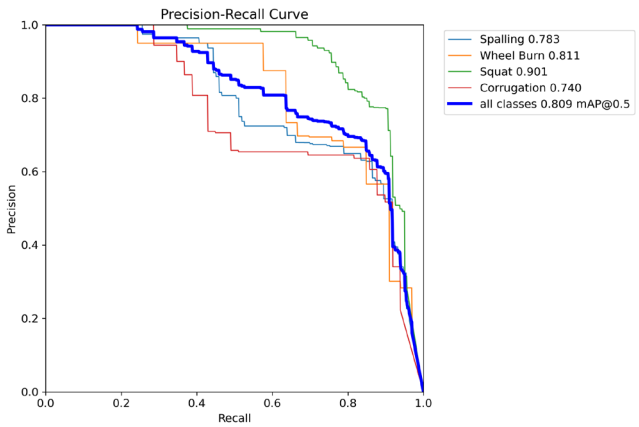


Figure 3. Baseline YOLOv11n Model

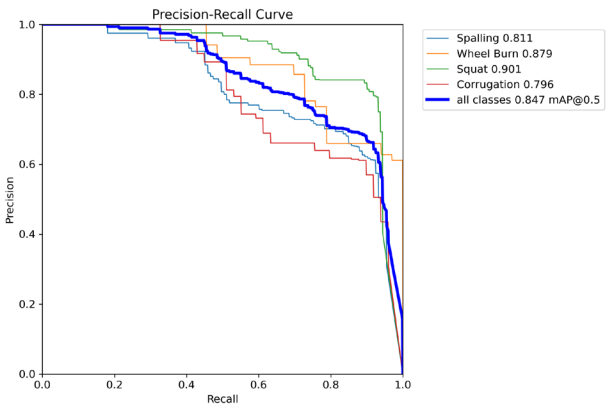


Figure 4. RSD-YOLO Model

5.2. Different Model Comparison Experiment Results and Analysis

To further validate the effectiveness of RSD-YOLO, it is compared against four representative YOLO variants—YOLOv5n, YOLOv8n, YOLOv10n, and YOLOv11n—all trained under identical hyperparameter settings and data splits on the rail surface defect dataset. The comparison results are presented in Table 3.

Table 3. Different Model Comparison Experiment Results

Model	mAP	mAP	Model	FPS/（f/s）

	@0.5/%	@0.5:0.95/%	Size/MB	
YOLOv5n	73.8	54.2	7.5	92.4
YOLOv8n	77.6	57.9	8.0	84.7
YOLOv10n	76.4	56.8	6.5	81.3
YOLOv11n	80.9	61.4	6.6	85.3
RSD-YOLO	84.7	64.8	8.9	74.2

As shown in Table 3, RSD-YOLO achieves the highest mAP@0.5 (84.7%) and mAP@0.5:0.95 (64.8%) among all compared models, confirming the effectiveness of the proposed improvements on the rail surface defect detection task. YOLOv5n, as an older anchor-based architecture relying on hand-crafted anchor configurations and a coupled detection head, achieves 73.8% mAP@0.5, with its performance limited by the anchor design sensitivity and the optimization conflict between classification and regression in the coupled head. The anchor-free YOLOv8n improves to 77.6% through its C2f module, decoupled head, and more flexible feature pyramid construction, but its standard PANet neck and absence of task-specific attention mechanisms limit its ability to suppress the metallic rail surface texture interference that constitutes the primary detection challenge on this dataset. YOLOv10n achieves 76.4% mAP@0.5 with its NMS-free end-to-end training formulation, which simplifies deployment but shows slightly lower accuracy than YOLOv8n on this dataset due to the stricter one-to-one matching strategy that may underutilize training supervision for densely annotated defect images. The strong YOLOv11n baseline achieves 80.9% mAP@0.5, benefiting from its enhanced C3k2 dual-kernel cross-stage modules that produce richer feature hierarchies and the C2PSA top-level spatial attention block that improves high-level semantic representation.

RSD-YOLO surpasses YOLOv11n by 3.8% in mAP@0.5 and 3.4% in mAP@0.5:0.95, representing the largest improvement margin among all compared model pairs. Per-category analysis from the Precision-Recall curves reveals that Wheel Burn detection improves from 81.1% to 87.9% AP (+6.8%), Corrugation from 74.0% to 79.6% AP (+5.6%), and Spalling from 78.3% to 81.1% AP (+2.8%), while Squat remains stable at 90.1% AP—consistent with the saturation of already strong baseline performance for this semantically unambiguous category whose clearly defined morphology is adequately captured by the standard feature extractor. The two categories showing the largest absolute improvements, Wheel Burn and Corrugation, correspond precisely to the defect types targeted by CBAM and ASPP respectively, providing strong evidence for the task-specific effectiveness of the proposed improvements. Although RSD-YOLO carries a larger model footprint (8.9 MB) and moderately reduced throughput (74.2 FPS) compared to the 6.6 MB, 85.3 FPS baseline, both values remain well within the performance envelope required for real-time on-board railway surface inspection deployment where inspection vehicles typically operate below 80 km/h with sufficient frame capture intervals.

5.3. Visualization Comparison Experiment

Considering the complex and variable appearance of rail surface defects, a visualization comparison experiment was conducted on the test set to more intuitively verify the detection adaptability of the improved model. Representative images were selected from each defect category, and detection results of the original YOLOv11n model and the proposed RSD-YOLO were visually compared side by side. The comparison results are shown in Figures 5, 6, and 7, with YOLOv11n results on the left and RSD-YOLO results on the right.

From Figure 5 (Wheel Burn and Spalling), the baseline assigns low confidence scores to thermally diffuse Wheel Burn boundaries and produces imprecise bounding boxes for Spalling at fracture edges. RSD-YOLO substantially improves detection confidence and boundary localization accuracy through the combined effect of CBAM spatial attention at P3/P4 stages and BiFPN adaptive weighting providing richer high-resolution spatial context. From Figure 6 (Corrugation), the baseline produces fragmented detections with systematically undersized bounding boxes, reflecting insufficient large-scale receptive field coverage. RSD-YOLO produces geometrically accurate detections encompassing entire affected rail sections, directly attributable to ASPP's d = 18 dilation branch providing contextual awareness spanning the full corrugation spatial extent. From Figure 7 (Squat), both models produce reliable high-confidence detections consistent with the stable 90.1% AP, confirming that clearly defined dark indentation morphology is adequately captured without additional enhancement. These visualization results confirm that the three improvements provide targeted category-specific gains precisely aligned with the quantitative per-class AP improvements.

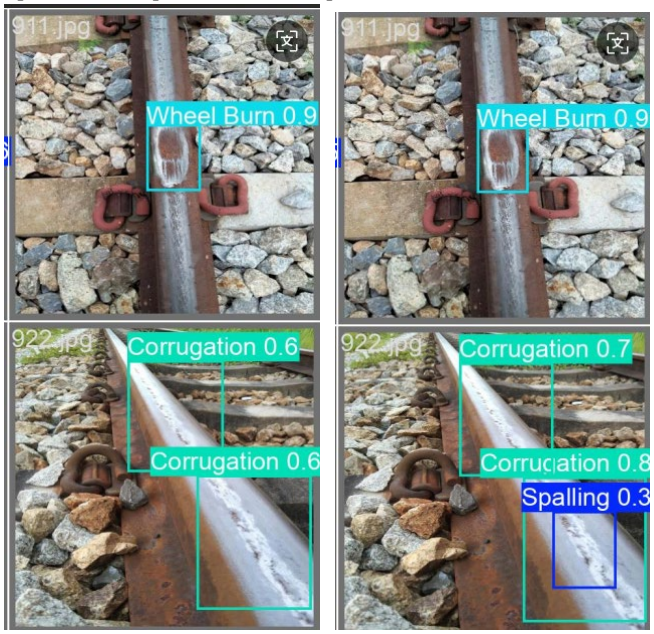


Figure 5. Comparison of YOLOv11n and WSD-YOLO Detection Results

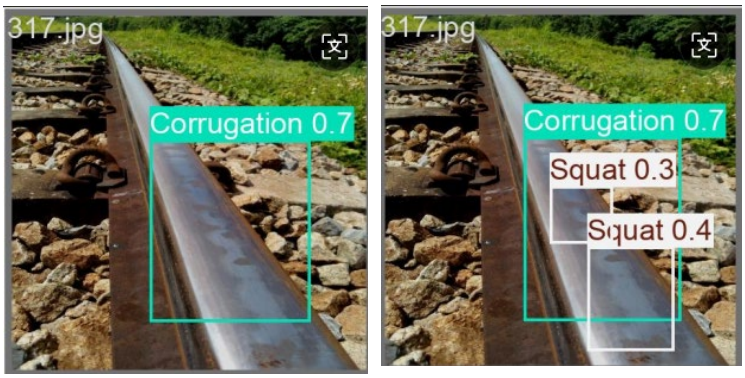


Figure 6. Comparison of YOLOv11n and WSD-YOLO Detection Results

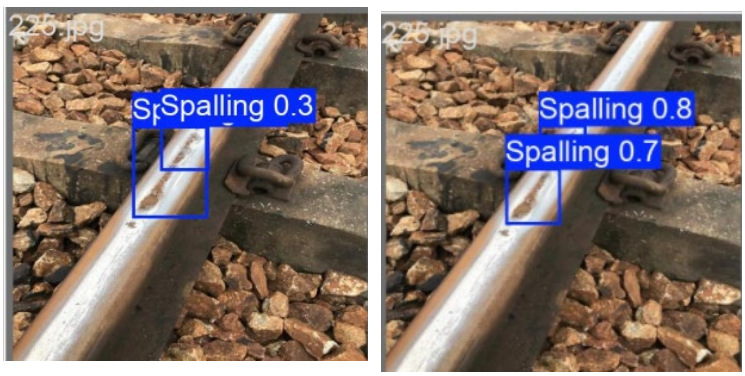


Figure 7. Comparison of YOLOv11n and WSD-YOLO Detection Results

6. Conclusion

This paper presents RSD-YOLO, an improved YOLOv11n-based object detection algorithm specifically designed for railway surface defect detection in automated inspection systems. Building on the YOLOv11n baseline, RSD-YOLO integrates three targeted structural improvements: dual-stage CBAM attention at the P3 and P4 backbone stages for multi-level channel and spatial feature recalibration, BiFPN adaptive weighted fusion in the neck for learnable asymmetric cross-scale feature interaction, and ASPP multi-receptive-field pooling at the backbone apex for parallel large-area periodic pattern context aggregation. These three improvements jointly and complementarily address the core challenges of rail surface inspection: multi-level attention coverage for fine-boundary defects whose discriminative features are encoded at intermediate backbone stages, scale-adaptive fusion for defect categories with heterogeneous spatial scale requirements, and multi-scale context enrichment for large periodic surface anomalies that exceed the effective receptive field of standard pooling structures.

Experimental evaluations on a custom 8,000-image rail surface defect dataset covering four defect categories demonstrate that RSD-YOLO achieves a mAP@0.5 of **84.7%** and mAP@0.5:0.95 of **64.8%**, outperforming the YOLOv11n baseline by 3.8% and 3.4% respectively. Wheel Burn detection improves by 6.8% and

Corrugation by 5.6%, confirming the model's targeted effectiveness for the most diagnostically challenging rail defect categories, while Squat maintains stable high-precision performance at 90.1% AP. Comparisons against YOLOv5n, YOLOv8n, YOLOv10n, and YOLOv11n confirm the comprehensive superiority of RSD-YOLO on this task. The model operates at 74.2 FPS on an NVIDIA A40 GPU, satisfying real-time requirements for on-board railway surface inspection deployment. Ablation studies confirm that each of the three improvements contributes positively and that the modules exhibit mutually complementary rather than redundant interactions, with the combined improvement exceeding the arithmetic sum of individual contributions.

Future work will focus on three directions: (1) improving Corrugation detection through wavelength-aware data augmentation and multi-frequency spatial attention designs tailored to periodic rail surface patterns of different spatial frequencies; (2) exploring knowledge distillation and structured pruning techniques to compress RSD-YOLO while retaining the accuracy gains from CBAM, BiFPN, and ASPP for deployment on resource-constrained edge computing platforms mounted on track inspection vehicles; and (3) extending the inspection pipeline to three-dimensional rail profile characterization by integrating structured-light or LiDAR depth sensing modalities with the proposed detection framework to enable volumetric defect quantification.

References

[1] Su, B., Chen, H., & Zhou, Z. (2022). BAF-detector: An efficient CNN-based detector for photovoltaic cell defect detection. IEEE Transactions on Industrial Electronics, 69(3), 3161-3171.

[2] Zhang, Y., Zhang, H., Huang, Q., Han, Y., & Zhao, M. (2024). DsP-YOLO: An anchor-free network with DsPAN for small object detection of multiscale defects. Expert Systems with Applications, 238, 122669.

[3] Hao, X., et al. (2024). An improved YOLOv8 algorithm for rail surface defect detection. IEEE Access, 12, 45210-45222.

[4] Liu, J., Shen, X., Wang, J., Jiang, L., & Zhang, H. (2023). An intelligent defect detection approach based on cascade attention network under complex magnetic flux leakage signals. IEEE Transactions on Industrial Electronics, 71(2), 2056-2066.

[5] Kim, H., et al. (2022). Defect detection of subway tunnels using advanced U-Net network. Sensors, 22(6), 2335.

[6] Li, X., Pu, X., Ling, W. et al. YOLO-SAM an end-to-end framework for efficient real time object detection and segmentation. Sci Rep 15, 40854 (2025). https://doi.org/10.1038/s41598-025-24576-6

[7] Han Y, Qi K, Zheng J, et al. Lightweight Cattle Facial Recognition Method Based on Improved YOLOv11. Smart Agriculture, 2025, 7(3): 173-184. https://doi.org/10.12133/j.smartag.SA202502010

[8] Xu, X. et al. Exploring image enhancement for salient object detection in low light images. ACM Trans. Multimed. Comput. Commun. Appl. 17, 1 – 19 (2021).

[9] Gao, L., Zhang, J., Yang, C., & Zhou, Y. (2022). Cas-VSwin transformer: A variant swin transformer for surface-defect detection. Computers in Industry, 140, 103689.

[10] Wang Yin, Wang Lide, Qiu Ji. Real-time Enhancement Algorithm for Orbital Dark Light

Environment Based on DenseNet Structure[J]. Journal of Southwest Jiaotong University, 2022, 57(6): 1349-1357. doi: 10.3969/j.issn.0258-2724.20210199

[11] Xue R ,Duan J ,Du Z .MPE-DETR: A multiscale pyramid enhancement network for object detection in low-light images[J].Image and Vision Computing,2024,150105202-105202.DOI:10.1016/J.IMAVIS.2024.105202.

[12] Dong, X., Zhang, C., Wang, J., Chen, Y., & Wang, D. (2024). Real-time detection of surface cracking defects for large-sized stamped parts. Computers in Industry, 158, 104105.

[13] Chen, Z., Yang, C., Li, Q., Zhao, F., Zha, Z. J., & Wu, F. (2024). A Dual-Task Branch Alignment Detector for Object Detection. IEEE Transactions on Image Processing, 33, 1642-1655.