Academic Journal of Emerging Technologies
ISSN Print: 3104-4417
ISSN Online: 3104-4425
CD ISSN: 3104-4433
https://doi.org/10.63313/AJET.9081

Rail Surface Defect Detection with Multi-Scale Detection Heads and Receptive Field Enhancement

Yuhao Ma*. Jincheng Li. Zhichao Hu. Taoyu Wang and Jie Xu
Tianjin University of Technology and Education, Tianjin 300222, China
Email: myh18335373024@163.com

How to cite this paper: Ma, Y., Li, J., Hu, Z., Wang, T., & Xu, J. (2026). Rail surface defect detection with multi-scale detection heads and receptive field enhancement. Academic Journal of Emerging Technologies, 4(1), 76–89. ISSN Print: 3104-4417, ISSN Online: 3104-4425.
https://doi.org/10.63313/AJET.9081
Published: 2026-11-18

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

Rail surface defects have diverse types and large scale variations, and standard detectors often miss fine-grained texture defects such as corrugation. This paper proposes a lightweight rail defect detection model based on an improved YOLOv11n. Three progressive improvements are introduced. First, a high-resolution P2 detection head is added to the neck to build a four-scale detection architecture, enhancing the perception of fine-grained textures. Second, C3k2 modules in the deep backbone are replaced with C2f modules, which refine gradient flow and deep semantic feature extraction while alleviating false detections caused by the shallow head. Third, an SPPF module is introduced in the P2 branch to enlarge the receptive field and compensate for weak semantics in shallow features. On a self-built Rail Surface Defect Dataset, the final model achieves 0.797 mAP@0.5 and 0.562 mAP@0.5-0.95 on the independent test set, improving the baseline by 11.78% and 14.93%, respectively. For corrugation, which is severely missed by the baseline, accuracy increases from 0.292 to 0.581. The model size increases only from 5.51 MB to 6.07 MB, achieving a good balance between accuracy and deployment cost for high-frequency, large-scale automated railway inspection.

Keywords

Rail Defect Detection; Yolov11; Multi-Scale Detection; Receptive Field Enhancement; Feature Fusion

1. Introduction

Railway transportation is a critical infrastructure of the national economy, and its operational safety is of paramount importance. As train speeds and axle loads continue to increase, rails inevitably develop various surface defects during long-term high-intensity service due to metal fatigue, friction, and environmental corrosion, such as spalling, wheel burn, squat, and corrugation. If these defects are not detected and treated in a timely manner, they accelerate the deterioration of the

track structure and can even lead to catastrophic accidents such as derailment. Traditional rail defect detection relies heavily on manual inspection and ultrasonic testing. However, manual inspection is inefficient, subjective, and prone to missed detections under fatigue, whereas ultrasonic equipment, although mature, is expensive and mainly targets internal cracks, with limited capability for surface and subsurface defects. Given the rapid growth of railway mileage, traditional methods can no longer meet the demands of large-scale, high-frequency modern inspection. In recent years, automated defect detection based on computer vision has become an important direction for addressing these problems owing to its non-contact, high-efficiency, and low-cost advantages.

Among computer vision algorithms, the YOLO (You Only Look Once) family, as a representative of one-stage object detectors, has been widely used in industrial defect detection because of its speed and accuracy. Deep learning-based rail surface defect detection has made considerable progress. For example, Wang and Zhao [1] introduced adaptive downsampling and a multi-scale feature fusion neck into YOLO11n, improving the preservation of fine-grained features for small rail defects. Ma et al. [2] proposed EADI-YOLO, a lightweight rail surface defect detector that integrates C2f-EMAN attention, AKConv, and DySample modules to balance accuracy and inference speed in complex railway environments. Wang and Zhou [3] introduced multidirectional rectangular convolution and cross-scale feature extraction into a one-stage detector, improving the identification of multiple rail surface defects.

Researchers have also explored dedicated optimizations for small-object detection. Mao et al. [4] combined SKNet attention with a C2f structure in an improved CenterNet, enabling the model to focus on key image regions and improving the detection of minor rail defects. Mi et al. [5] introduced Res2Net and CBAM attention to enlarge the receptive field for blurred, extremely small rail defect edges, strengthening the representation of subtle targets while reducing parameter redundancy. In addition, Wang [6] improved YOLO11 for strip steel surface defect detection in complex scenes, providing a cross-domain reference for handling strong background texture interference.

Despite these advances, and although the YOLO family has evolved from YOLOv5 to the latest YOLOv11 with C3k2, C2PSA, and SPPF, standard YOLOv11 still has limitations in rail defect detection. First, its multi-scale detection capability is insufficient: the standard model uses three detection scales (P3, P4, and P5, corresponding to 80×80, 40×40, and 20×20 feature maps), and performs poorly on defects with fine-grained repetitive textures such as corrugation; experiments show that the baseline achieves only 0.292 mAP@0.5 on corrugation. Second, shallow features have weak semantic information: even if a higher-resolution detection head (e.g., the 160×160 P2 layer) is added directly, the model struggles to distinguish true defects from complex normal rail textures because shallow layers lack sufficient

semantic context. Third, performance varies greatly across defect categories: reliable identification of every category is required for railway safety, yet the baseline's accuracy is highly imbalanced (mAP@0.5 from 0.292 to 0.633), severely limiting its practical deployment value.

To address these challenges, this paper proposes an improved YOLOv11n model for rail defect detection. The main contributions are as follows:

A four-scale detection architecture is constructed: a 160×160 P2 branch is added to the Neck to enhance the perception of fine-grained texture defects through higher-resolution feature maps, significantly reducing the baseline's severe missed detections of corrugation.

Deep feature extraction is optimized: C2f modules replace C3k2 modules in part of the backbone, enhancing deep semantic representation through finer-grained gradient flow and feature reuse, and mitigating the accuracy degradation of some categories after introducing the P2 head.

The shallow-feature receptive field is enhanced: an SPPF module replaces the conventional processing module in the P2 detection branch, enlarging the shallow receptive field without significantly increasing parameters and further improving overall detection accuracy.

2. Structure of the YOLOv11n Baseline Model

YOLOv11 is a state-of-the-art one-stage object detection architecture released by Ultralytics in 2024. Compared with previous versions, it deeply restructures feature extraction efficiency, cross-scale fusion, and parameter resource scheduling. Considering that rail surface defect inspection equipment usually faces limited edge computing resources and stringent high-frequency real-time inference requirements, YOLOv11n, the variant with the fewest parameters and the lowest computational cost, is selected as the baseline. Its overall architecture follows the classic 'backbone extraction–neck fusion–multi-scale prediction' paradigm and consists of three core components: Backbone, Neck, and Head.

(1) Backbone: It performs progressive downsampling and multi-level feature extraction. YOLOv11n uses deeply optimized C3k2 modules as its core skeleton, together with standard convolutions (Conv) for spatial downsampling and channel expansion. At the deep end (P5 stage), the baseline connects SPPF (fast spatial pyramid pooling) and a C2PSA attention module. The former broadens the global receptive field by concatenating multi-scale max-pooling results, while the latter dynamically guides the network to focus on key semantic channels, providing high-quality representations for subsequent feature interaction.

(2) Neck: It performs cross-level semantic alignment and multi-scale information aggregation. YOLOv11n follows the classic bidirectional FPN (feature pyramid network) and PAN (path aggregation network) architecture. The top-down FPN path propagates abstract semantic information from deep layers to shallow layers to

enhance classification discrimination, while the bottom-up PAN path transfers precise spatial cues from shallow layers to deep layers, complementing semantic representation and spatial localization.

(3) Head: It performs final category classification and bounding-box regression. Standard YOLOv11n adopts a decoupled structure with three detection branches of different receptive fields. These branches receive the P3 (80×80), P4 (40×40), and P5 (20×20) feature maps from the Neck (for a 640×640 input) and detect and localize objects of different sizes through a multi-scale divide-and-conquer strategy.

3. PCS-YOLO Rail Surface Defect Detection Model

Based on YOLOv11n, this paper proposes PCS-YOLO, an enhanced network architecture for rail surface defect detection. In railway safety inspection, the model must reliably detect defects of various types and scales, since severe missed detection of even one category can constitute a safety hazard. The core objective is therefore to significantly improve overall detection accuracy while maintaining a lightweight model. The name reflects three core optimizations: first, a new P2 multi-scale detection head (P2 Head) captures shallow high-resolution features to enhance fine-grained texture defect detection; second, C2f modules (C2f Module) replace some C3k2 modules in the deep backbone to enhance deep semantic feature extraction; third, an SPPF module is introduced in the P2 layer to enlarge the shallow receptive field through parameter-free pooling. The overall architecture of PCS-YOLO after these three stages is shown in Figure 1.

3.1. P2 Multi-Scale Detection Head

In practical rail surface defect detection, defect types differ greatly in texture complexity and spatial scale. Corrugation, in particular, appears as fine periodic ripples on the rail surface with highly fine-grained repetitive texture. Standard YOLOv11n provides only three detection levels: P3 (80×80), P4 (40×40), and P5 (20×20). The highest-resolution P3 level already has a downsampling factor of 8, so such fine textures suffer severe information loss during progressive downsampling. Experiments show that the baseline achieves only 0.292 mAP@0.5 on corrugation, far below the other three categories (0.502–0.633), making it the key bottleneck limiting overall performance.

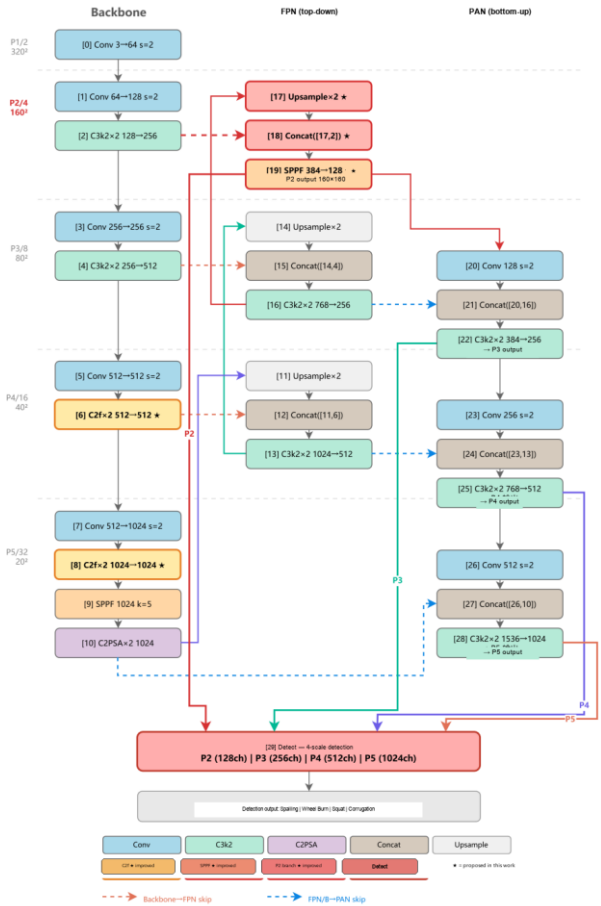


Figure 1. PCS-YOLO network architecture

To address the loss of fine-grained texture features, a P2 high-resolution detection branch is added to the top-down path of the Neck FPN. Feature map size is closely related to the downsampling stride. Let the input image resolution be R_{in}; the output resolution after network layers, R_{out}, is computed as Eq. (1):

$$R_{out} = \lfloor \frac{R_{in}}{S} \rfloor \tag{1}$$

where S is the cumulative downsampling stride. The standard P3 layer has $S = 8$, whereas the new P2 layer is taken directly from the C3k2 module at layer 2 of the backbone, for which $S = 4$. The P2 feature map therefore retains a resolution of 160×160 and preserves finer spatial geometry.

Deep features are then upsampled and fused with the P2 features from the backbone. The concatenation process is expressed as Eq. (2):

$$\mathrm{F_{P2}} = \mathrm{Concat}(\mathrm{Upsample}(\mathrm{F_{P3}}), \mathrm{F_{B2}}) \tag{2}$$

where F_{P3} is the deep feature map propagated top-down, F_{B2} is the high-resolution feature output by layer 2 of the backbone, Upsample denotes 2× upsampling, and Concat denotes channel-wise concatenation.

In the network restructuring, layers 17–19 are added (layer 17 performs upsampling and layer 18 performs concatenation), extending the original three-scale architecture (P3/P4/P5) to four scales (P2/P3/P4/P5). Experiments on the independent test set show that the P2 branch raises mAP@0.5 from 0.713 to 0.780, an improvement of 9.40%. The largest gain is on corrugation, whose mAP@0.5 jumps from 0.292 to 0.603 (+106.51%), indicating that the 160×160 feature map effectively preserves the spatial information of fine textures. The P2 detection head forms the foundational module (P) of PCS-YOLO.

3.2. C2f Module Replacement

Although the P2 head significantly improves overall mAP@0.5 (0.713→0.780), per-category analysis reveals that spalling detection drops from 0.502 to 0.420 (–16.33%). The underlying mechanism is that although the P2 layer has a 160×160 resolution and can extract clear low-level visual features such as edges, it is very shallow and lacks high-level semantic guidance. The C3k2 modules in the original YOLOv11n backbone have relatively coarse feature reuse granularity, so the deep network cannot extract sufficiently robust semantic features to feed back to shallow layers. Consequently, the P2 layer, lacking sufficient semantic discrimination, easily mistakes normal rail texture for defects or misses spalling defects with indistinct semantic features.

To enhance deep feature extraction and representation in the backbone, C2f modules are introduced to replace some C3k2 modules. C2f achieves finer-grained feature reuse through a distinctive split operation and multi-branch nonlinear transformation. Let the input feature be X; its channel split is expressed as Eq. (3):

$$X_1, X_2 = Split(X) \tag{3}$$

The split features undergo deep nonlinear transformation in multiple parallel Bottleneck branches and are then aggregated across channels. The overall feature evolution is computed as Eq. (4):

$$Y = Concat(X_1, X_2, B_1(X_2), B_2(B_1(X_2)), \ldots, B_n(\ldots)) \tag{4}$$

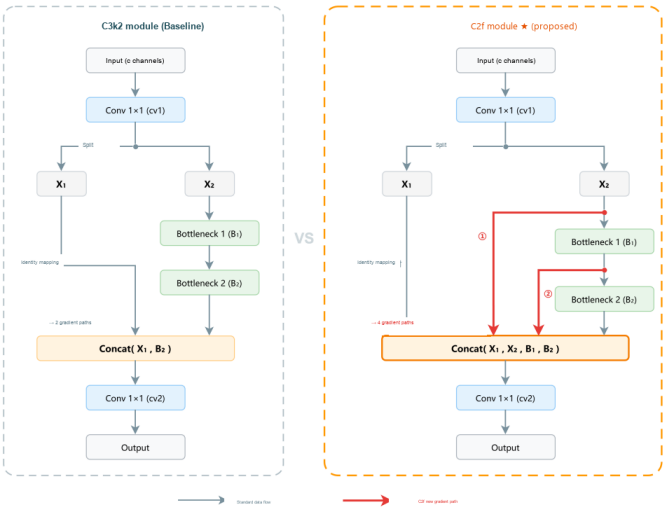


Figure 2. Comparison of C3k2 and C2f module structures and gradient paths

where X_1 serves as the identity mapping branch that preserves the original information, B_i denotes the deep nonlinear transformation of the i-th Bottleneck module, and Concat denotes channel-wise tensor concatenation. To illustrate the advantage of this structural difference more intuitively, Figure 2 compares the internal data flow of the baseline C3k2 and the improved C2f. As shown in Figure 2, C3k2 concatenates only the identity branch X_1 and the output of the last Bottleneck B_2, providing only two main gradient backpropagation paths. In contrast, C2f feeds the split X_2 and the intermediate output of every Bottleneck (e.g., B_1, B_2) directly into the final fusion layer (the new gradient-flow paths marked by red arrows in Figure 2). Compared with the limited gradient transfer of C3k2, the derivation in Eq. (4) and the design in Figure 2 jointly show that C2f constructs four richer gradient-flow paths, generating deeper semantic features with higher representational dimensionality and finer granularity.

To balance representational capability and model compactness, a partial replacement strategy is adopted: only the C3k2 modules in the P4 (layer 6) and P5 (layer 8) stages, which extract high-level semantics, are rebuilt as C2f modules with 512 and 1024 channels, respectively. On the independent test set, mAP@0.5 rises to 0.788, and mAP@0.5-0.95 increases from 0.558 to 0.570, indicating that the enhanced gradient flow of C2f improves bounding-box localization. More importantly, spalling accuracy recovers from 0.420 to 0.473, partially offsetting the degradation caused by the P2 head and confirming that deep semantic enhancement supports the shallow detection head. The C2f module provides the core semantic support (C) of PCS-YOLO.

3.3. SPPF Receptive Field Enhancement in the P2 Layer

After the first two improvements, mAP@0.5 increases from the baseline's 0.713 to

0.788, but further optimization is still possible. The high resolution of the P2 feature map (160×160) makes its effective receptive field extremely small under convolution. When predicting rail defects at P2, the model lacks sufficient global context such as the direction of surrounding rails and background illumination. For squat and similar defects that require surrounding rail context for accurate discrimination, accuracy can still improve (mAP@0.5 = 0.632). It is therefore necessary to enlarge the P2 receptive field without significantly increasing computational parameters.

An SPPF (spatial pyramid pooling – fast) mechanism is introduced to replace the original module. SPPF aggregates the feature space by cascading three 5×5 max-pooling operations. In a convolutional neural network, the equivalent receptive field of cascaded operations is computed as Eq. (5):

$$RF_n = RF_{n-1} + (k-1) \times \prod_{i=1}^{n-1} s_i \tag{5}$$

where RF_n is the equivalent receptive field after the n-th pooling, k is the current kernel size (5), and s_i is the stride of the i-th operation (to preserve the feature map size, all strides s are set to 1 with padding). With an initial receptive field of $RF_0 = 1$, Eq. (5) shows that after one, two, and three cascaded pooling operations the receptive field expands to 5, 9, and 13, respectively. Cascaded pooling is therefore mathematically equivalent to performing 5×5, 9×9, and 13×13 pooling in parallel.

The final feature aggregation of SPPF is expressed as Eq. (6):

$$F_{out} = Concat(X, P_{5\times5}(X), P_{5\times5}(P_{5\times5}(X)), P_{5\times5}(P_{5\times5}(P_{5\times5}(X)))) \tag{6}$$

where X is the input feature and $P_{5\times5}$ denotes 5×5 max pooling. This operation fuses multi-scale receptive fields with high computational efficiency and zero additional parameters.

In the actual network configuration, the C3k2 module at layer 19, originally used for P2 fused features, is replaced by SPPF. On the independent test set, the greatly expanded receptive field enables the P2 layer to exploit global context for defect prediction. mAP@0.5 rises further to 0.797, a cumulative improvement of 11.78% over the baseline. Squat accuracy increases from 0.632 to 0.671 (+6.17%), confirming the effectiveness of receptive field enhancement for categories that require contextual discrimination. The final PCS-YOLO model improves accuracy on all four rail defect categories while remaining lightweight, and the SPPF module constitutes its key receptive field enhancement design (S).

4. Dataset and Evaluation Metrics

4.1. Experimental Configuration and Dataset

The experiments are conducted on Linux (Ubuntu 5.4.0-65-generic) with Python 3.8.10 and PyTorch 1.11.0 (CUDA 11.3), using Ultralytics 8.3.168 for training and evaluation on an NVIDIA A40 GPU. Input images are resized to 640×640, the batch

size is 16, and the maximum number of epochs is 300. Early stopping (patience = 50) is enabled to prevent overfitting.

Experiments use a self-built Rail Surface Defect Dataset (RSDD) collected by rail inspection equipment at high resolution. Professional annotators labeled the images according to railway industry defect classification standards, covering four typical rail surface defects: spalling, wheel burn, squat, and corrugation. The dataset contains 2,396 images and 3,950 annotated instances, divided into training, validation, and test sets at approximately 80%:10%:10%. The split is summarized in Table 1.

Table 1. Dataset split statistics

Split	Number of images	Number of annotated instances
Training set	1916	3182
Validation set	240	375
Test set	240	393
Total	2396	3950

The class distribution is somewhat imbalanced. In the training set, squat is the most frequent (1,265 instances, 39.75%), followed by spalling (983, 30.89%) and corrugation (552, 17.35%), while wheel burn is the least frequent (382, 12.01%). The detailed distribution is given in Table 2. This imbalance is broadly consistent with the frequency of defects in real railway maintenance and better reflects detection capability in practical scenarios.

Table 2. Distribution of annotated instances by category in the training set

Category	Instances	Proportion /%
Spalling	983	30.89
Wheel Burn	382	12.01
Squat	1265	39.75
Corrugation	552	17.35
Total	3182	100.00

4.2. Evaluation Metrics

Precision (P), recall (R), average precision (AP), and mean average precision (mAP) are used to evaluate detection performance. The corresponding formulas are given in Eqs. (7)–(10):

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$AP = \int_0^1 P(R)\, dR \tag{9}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} A P_i \tag{10}$$

where TP is the number of correctly predicted positives, FP is the number of incorrectly predicted positives, and FN is the number of missed positives; N is the total number of categories. AP is obtained by integrating the precision–recall curve and measures per-category accuracy; mAP is the average of AP over all categories and reflects overall performance. This paper reports mAP@0.5 (mAP at an IoU threshold of 0.5) and mAP@0.5-0.95 (the average mAP over IoU thresholds from 0.5 to 0.95 in steps of 0.05). The former reflects basic detection capability, while the latter imposes stricter localization requirements. All metrics are evaluated on the independent test set using the final trained model to ensure objectivity and reliability.

5. Experimental Results and Analysis

5.1. Ablation Experiment Results and Analysis

To verify each improvement in PCS-YOLO, a progressive ablation study is designed in which modules are added successively under identical training settings and their contributions are evaluated. All models are initialized with COCO pre-trained weights, trained for 300 epochs on RSDD, and evaluated on the independent test set. The overall results are shown in Table 3, and per-category accuracies are compared in Table 4.

Table 3. Ablation experiment results

Model	P2	C2f	SPPF	mAP@0.5	mAP@0.5-0.95
YOLOv11n	×	×	×	0.713	0.489
+P2	√	×	×	0.780	0.558
+P2+C2f	√	√	×	0.788	0.570
PCS-YOLO	√	√	√	0.797	0.562

Table 4. Per-category mAP@0.5 comparison

Model	Spalling	Wheel Burn	Squat	Corrugation
YOLOv11n	0.502	0.530	0.633	0.292
+P2	0.420	0.582	0.627	0.603
+P2+C2f	0.473	0.560	0.632	0.612
PCS-YOLO	0.469	0.526	0.671	0.581

Table 3 shows that the YOLOv11n baseline achieves 0.713 mAP@0.5 and 0.489 mAP@0.5-0.95 on the RSDD test set. With the standard three-scale architecture (P3/P4/P5), it detects spalling, wheel burn, and squat to some extent, but reaches

only 0.292 on corrugation, indicating severe missed detections.

With the P2 high-resolution head, mAP@0.5 rises from 0.713 to 0.780 (+9.40%) and mAP@0.5-0.95 from 0.489 to 0.558 (+14.11%). The per-category results in Table 4 show that the largest gain is on corrugation, whose mAP@0.5 increases from 0.292 to 0.603 (+106.51%), demonstrating that the 160×160 feature map captures the fine texture of corrugation and substantially alleviates missed detections. However, spalling accuracy drops from 0.502 to 0.420 (–16.33%) because shallow features carry weak semantics. While the high-resolution map enhances fine-grained perception, it also introduces more background noise that interferes with large, boundary-blurred spalling defects.

Replacing deep C3k2 modules in the backbone with C2f further raises mAP@0.5 to 0.788 and mAP@0.5-0.95 to 0.570. The split–concat structure of C2f preserves richer gradient-flow information and improves deep semantic extraction. Per-category results show that spalling recovers from 0.420 to 0.473, partially offsetting the degradation caused by the P2 head, while corrugation improves slightly to 0.612. The enhanced deep representation therefore mitigates noise from the shallow head and yields a better multi-scale balance.

Finally, replacing the C3k2 module in the P2 branch with SPPF brings the final PCS-YOLO model to 0.797 mAP@0.5, a cumulative improvement of 11.78% over the baseline. Cascaded max pooling expands the P2 receptive field at zero parameter cost, allowing shallow features to incorporate richer context while retaining high resolution. Squat accuracy rises from 0.632 to 0.671 (+6.17%), confirming that receptive field enhancement benefits periodic defects that require contextual discrimination.

Overall, PCS-YOLO improves mAP@0.5 by 11.78% and mAP@0.5-0.95 by 14.93% over the baseline. The three improvements are complementary: the P2 head addresses severe missed detections of corrugation (+98.97%), C2f optimizes deep feature extraction and mitigates the side effects of P2, and SPPF enlarges the shallow receptive field and improves squat accuracy. The model file increases only from 5.51 MB to 6.07 MB, confirming the effectiveness of each strategy and the lightweight design goal.

5.2. Comparison with Different Models

To further verify the superiority of PCS-YOLO, three representative lightweight YOLO models—YOLOv5n, YOLOv8n, and YOLOv11n—are used as comparison baselines and trained and tested on the same RSDD dataset. The results are shown in Table 5.

Table 5. Comparison with different models

Model	mAP@0.5 (%)	mAP@0.5-0.95 (%)	Model Size /MB	FPS / (f/s)
YOLOv5n	75.77	46.49	7.3	102.17
YOLOv8n	77.93	51.70	8.0	69.32
YOLOv11n	71.23	48.95	6.5	96.44
PCS-YOLO	79.66	56.16	6.07	91.93

Table 5 shows that PCS-YOLO achieves the best accuracy among all compared models on RSDD, with 79.66% mAP@0.5 and 56.16% mAP@0.5-0.95. Compared with the second-best YOLOv8n, mAP@0.5 is 1.73 percentage points higher and mAP@0.5-0.95 is 4.46 points higher, while the model size is only 6.07 MB versus 8.0 MB, indicating clear advantages in both accuracy and lightness. Compared with the YOLOv11n baseline, PCS-YOLO improves mAP@0.5 by 8.43 points and mAP@0.5-0.95 by 7.21 points while being smaller (6.07 MB vs. 6.5 MB), validating the proposed strategies.

Notably, YOLOv11n achieves lower mAP@0.5 (71.23%) than YOLOv5n (75.77%) and YOLOv8n (77.93%) on this dataset, which may be related to the prevalence of fine-grained texture defects such as corrugation; the default three-scale architecture of YOLOv11n has insufficient perception for such defects. By introducing the P2 high-resolution head and SPPF receptive field enhancement, PCS-YOLO compensates for this weakness, surpassing both the YOLOv11n baseline and all compared models in mAP@0.5. In terms of inference speed, PCS-YOLO reaches 91.93 f/s, slightly below YOLOv5n (102.17 f/s) and YOLOv11n (96.44 f/s) but well above YOLOv8n (69.32 f/s), meeting the real-time requirements of railway inspection.

5.3. Training Process Analysis

Figure 3 shows validation mAP@0.5 curves during training for the four models. In terms of convergence, the baseline converges fastest in the early stage because of its simpler structure, reaching 0.735 mAP@0.5 at about epoch 80. The three improved models converge slightly more slowly in the first 50 epochs due to the more complex P2 detection head, but catch up and surpass the baseline after epoch 100. PCS-YOLO (AB03) reaches its best validation performance (mAP@0.5 = 0.828) at around epoch 230 and remains stable thereafter. Notably, PCS-YOLO reaches a maximum recall of 0.912 at epoch 286, the highest among all models, indicating that the improved multi-scale architecture substantially strengthens coverage of all defect categories.

In terms of training efficiency, the baseline takes about 13.3 s per epoch. Adding the P2 head (+P2 and +P2+C2f) increases this to about 41.5 s per epoch, mainly because the 160×160 P2 feature map introduces considerable extra computation. Replacing C3k2 with SPPF in the P2 branch reduces PCS-YOLO training to 17.1 s per epoch, only 29.1% more than the baseline, greatly lowering computational cost while maintaining detection performance.

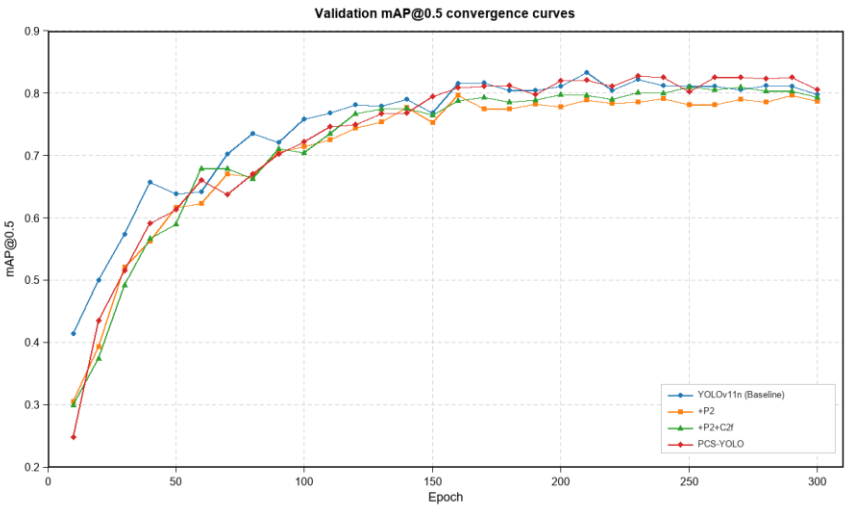
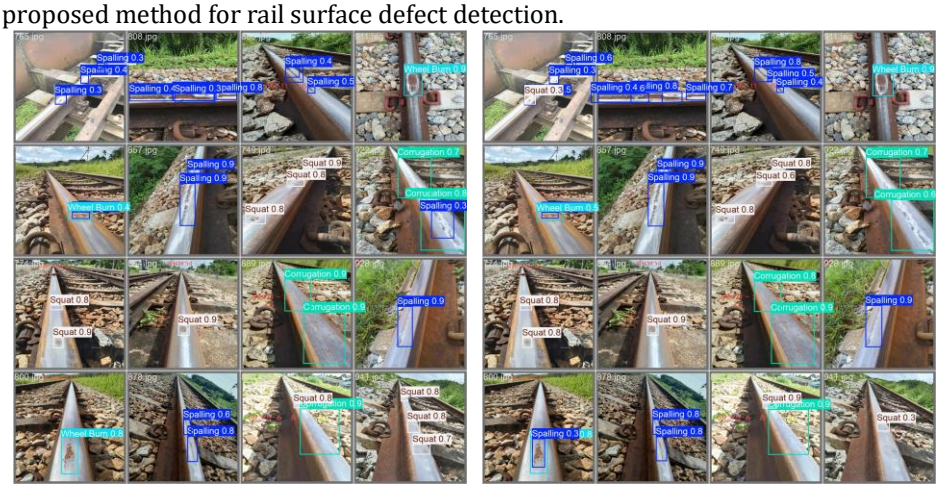


Figure 3. Variation of mAP@0.5 during training

5.4. Visual Comparison Analysis

To verify the detection performance of PCS-YOLO more intuitively, representative rail surface images are selected from the test set for a visual comparison between the YOLOv11n baseline and PCS-YOLO. The results are shown in Figure 4, with the baseline on the left and PCS-YOLO on the right.

The visualization shows that the baseline clearly misses corrugation defects because their fine periodic ripples are difficult to capture in the feature maps at P3 and above. With the P2 high-resolution head and SPPF receptive field enhancement, PCS-YOLO accurately identifies such fine-grained texture defects. For squat and similar defects that require combining local texture with global context, PCS-YOLO also achieves higher confidence and more precise localization, demonstrating the practicality of the proposed method for rail surface defect detection.



YOLOv11n PCS-YOLO

Figure 4. Visual comparison

References

[1] Wang, H. and Zhao, J. (2026) Research on Defect Detection on Steel Rails Based on Improved YOLO11n Algorithm. Applied Sciences, 16(2), 842. https://doi.org/10.3390/app16020842

[2] Ma, S., Li, R. and Hu, H. (2025) EADI-YOLO: A Lightweight and Efficient Model for Rail Surface Defect Detection. Research Square Preprint. https://doi.org/10.21203/rs.3.rs-7183947/v1

[3] Wang, M. and Zhou, Y. (2024) Autonomous Rail Surface Defect Identification Based on an Improved One-Stage Object Detection Algorithm. Journal of Performance of Constructed Facilities, 38(5), 04024041. https://doi.org/10.1061/JPCFEV.CFENG-4816

[4] Mao, Y., Zheng, S., Li, L., Shi, R. and An, X. (2024) Research on Rail Surface Defect Detection Based on Improved CenterNet. Electronics, 13(17), 3580. https://doi.org/10.3390/electronics13173580

[5] Mi, Z., Chen, R. and Zhao, S. (2023) Research on Steel Rail Surface Defects Detection Based on Improved YOLOv4 Network. Frontiers in Neurorobotics, 17, 1119896. https://doi.org/10.3389/fnbot.2023.1119896

[6] Wang, P. (2026) Complex Scene Strip Steel Surface Defect Detection Algorithm Based on Improved YOLO11. Computer Science and Application, 16(2), 240-250. https://doi.org/10.12677/CSA.2026.162055