Research on Track Defect Visual Detection Method Based on Improved YOLOv7-tiny

Siyi Du, Shiyu Jiao *

School of Railway Intelligent Engineering, Dalian Jiaotong University, Dalian, China, 116045

* Corresponding Author Email: 19824371767@163.com

Abstract. With the accelerated advancement of rail transit systems, conventional manual inspection methodologies exhibit significant limitations in both efficiency and accuracy. Despite the advantages of deep learning-based methods in automated defect detection, they still face challenges such as defect diversity, interference from complex environments, and computational resource constraints. This study proposes a visual inspection framework for rail defects based on an enhanced YOLOv7-tiny model, which integrates the Convolutional Block Attention Module (CBAM) for channel-spatial dual-dimensional attention mechanisms and Deformable Convolutional Networks (DCN) for depth wise separable convolutions, thereby improving the recognition capability of small-sized defects under complex environmental conditions. A novel lightweight hybrid parallel network architecture is proposed, incorporating depth wise separable convolution and channel pruning methodologies to enhance computational efficiency. Experimental results demonstrate that the improved model achieves a 95.7% Mean Average Precision (mAP) on the railway defect dataset, representing a 7.3% enhancement over the baseline YOLOv7-tiny. The optimized network exhibits a 42% reduction in parameters, decreasing from 6.0M to 3.5M, and a 53% reduction in computational complexity, dropping from 13.2 GFLOPs to 6.2 GFLOPs. These improvements highlight its significant advantages for deployment on resource-constrained detection devices. The study presents an innovative solution for intelligent track maintenance, achieving an optimal balance between precision and operational efficiency.

Keywords: Track Defect Detection, Light-Weighted Algorithm, Visual inspection, Convolutional Block Attention Module.

1. Introduction

With the accelerated advancement of high-speed railway systems and urban rail transit networks, track defect detection safety has gained significant prominence. Traditional manual inspection methodologies are constrained by inherent limitations in operational efficiency and pronounced subjectivity, rendering them inadequate for fulfilling the stringent safety requirements of real-time monitoring within high-density road network environments. In recent years, deep learning, particularly the YOLO series algorithms, has achieved remarkable advancements in object detection, offering novel approaches for automated rail defect inspection. Nevertheless, the detection of track defects continues to encounter substantial challenges: the diversity of defects with significant size variations, the complexity and variability of detection environments encompassing low illumination, adverse weather conditions, and motion-induced image blurring, coupled with the resource constraints of edge computing devices, which impose rigorous demands on the model's real-time performance. Traditional computer vision methodologies encounter significant challenges in addressing these intricate factors. While standard deep learning models demonstrate robust feature extraction capabilities, their substantial computational complexity poses considerable obstacles to achieving real-time inference on edge devices. Furthermore, these models lack specialized optimization for tracking defect detection tasks [1].

Subsequently, deformable convolution and coordinate attention mechanisms are incorporated into the feature extraction phase to augment the model's geometric deformation adaptability and spatial localization precision [2]. To address the requirements of edge computing, the backbone network architecture has been substituted with CSPResNeXt50, accompanied by the development of a hybrid structured pruning strategy. This approach incorporates depth wise separable convolutional

reconstruction network components while preserving critical feature channels [3]. This study introduces an innovative framework for rail defect detection, which not only addresses the limitations of existing methods in terms of insufficient detection accuracy and poor real-time performance in complex environments but also enhances the applicability of YOLOv7-tiny on edge devices through a model optimization strategy, thereby providing robust technical support for the safe operation of rail transit systems.

2. Heterogeneous Multi-source Dataset Construction

In the domain of visual track defect detection, the establishment of multivariate heterogeneous datasets constitutes the fundamental cornerstone for enhancing model generalization capabilities and cross-scenario adaptability. Addressing the challenges of morphological diversity in defects (including cracks, spalling, and missing bolts), scale variations (ranging from millimeter to centimeter levels), and environmental complexity (such as lighting variations, rain and fog interference, and motion ambiguity) in railway track defect detection, this research introduces a heterogeneous dataset construction framework utilizing multi-source data fusion. This approach systematically mitigates the issue of model overfitting associated with the homogeneity of conventional datasets [4].

The dataset construction was primarily accomplished through a multi-source data acquisition and fusion strategy. Field data collection constituted 60% of the total dataset, utilizing a global shutter industrial camera with a resolution of 1920×1080 and a frame rate of 71 frames per second (FPS). The data collection encompassed various temporal conditions, including different periods of day and night, as well as diverse weather conditions such as sunny, rainy, and foggy scenarios. Additionally, it covered multiple track types, strictly adhering to the "Rail Defect Detection Standard (TB/T 2344.1-2020)". The geometric deformation and surface texture characteristics of defects were captured through multi-angle imaging techniques, including overhead, side-view, and squint perspectives. The simulated defect structure constituted 30% of the total, with standardized defect specimens being generated by integrating finite element simulation and 3D printing technology. These specimens encompassed micro-cracks in the initiation phase (0.5mm), spalling in the propagation stage (5-10mm), and severe structural damage (>20mm), thereby ensuring the physical authenticity of the data distribution. Concurrently, bidirectional reflectance distribution function (BRDF) modeling was implemented to simulate the spectral response variations of the rail surface under diverse illumination angles, thereby enhancing the illumination robustness of the dataset [5]. Open-source data-integration constituted 10% of the process, wherein samples were meticulously screened and labeled to align with track scenes in publicly available datasets (e.g. Rail-Detect and Crack Forest. Style discrepancies were effectively mitigated through the application of domain adaptation techniques (Cycle GAN), thereby achieving cross-domain feature alignment.

Regarding data augmentation and annotation, the dynamic augmentation strategy enhances model robustness through geometric transformations, photometric perturbations, and contextual interference. The geometric transformation encompasses random affine transformations (rotation within ±15°, scaling ranging from 0.8 to 1.2 times) and perspective distortion simulation. By Retinex theory, luminance-contrast jitter, Gaussian noise injection, and motion blur kernel (with a size of 15×15) were employed to replicate complex imaging conditions. Contextual interference is implemented to mitigate excessive reliance on local features through the random incorporation of track background elements (including ballast, bolts, and rust). The fine-grained labeling protocol employs both PASCAL VOC and COCO dual formats to categorize six distinct types of defects (cracks, peeling, fish scales, missing bolts, rail head wear, and rail waist corrosion) with subpixel-level mask annotation. Additionally, an Uncertain Region Marking (URM) mechanism is incorporated to probabilistically annotate low-contrast defect boundaries, thereby mitigating subjective errors in the labeling process. Data Balancing and Validation: Strategic sampling techniques combined with the Synthetic Minority Oversampling Technique (SMOTE) effectively address class imbalance issues (e.g. when crack sample proportions fall below 5%). Furthermore, 5-fold cross-validation integrated

with hard sample mining ensures comprehensive coverage of long-tail distributions in the training dataset.

The dataset construction method integrates multi-source heterogeneous datasets, laying a data foundation for subsequent lightweight model optimization and real-time inference. Multi-modal feature fusion is achieved through spectral reflectance modeling and geometric deformation synthesis, covering the physical-optical joint features of defects and providing rich training priors for the CBAM attention mechanism. Dynamic augmentation generalization combines physically driven (motion blur simulation) and data-driven (GAN domain adaptation) methods, transcending the empirical limitations of traditional augmentation techniques. Annotation consistency is ensured via the URM protocol and a dual-format annotation system, optimizing both detection and segmentation tasks.

3. Improved YOLOv7-tiny Algorithm Design

To address the challenges of small target detection omission, dynamic blur interference, and light sensitivity in complex scenarios, this study proposes a lightweight detection framework based on an enhanced YOLOv7-tiny architecture [6]. The backbone network is reconstructed utilizing the CSPResNeXt50 architecture, which incorporates cross-stage local residual connections and group convolution strategies to minimize computational redundancy while enhancing multi-scale feature extraction capabilities. Specifically, the input feature map undergoes dimensionality reduction via a 1×1 convolutional operation, subsequently bifurcating into two distinct pathways: one pathway is directly propagated to the subsequent stage, while the other transforms the ResNeXt module. Ultimately, cross-layer information fusion is achieved through channel concatenation.

To further augment feature discriminability within complex backgrounds, the CBAM incorporating a channel-spatial dual-dimensional attention mechanism is implemented [7]. The channel attention mechanism employs global average pooling and maximum pooling operations to generate a channel weight matrix, thereby dynamically enhancing feature channels associated with defect detection:

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

The spatial attention module employs a 7×7 convolutional kernel to generate a spatial mask, thereby enabling precise localization of defect regions:

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

By leveraging cascaded structures and spatial attention mechanisms, the model demonstrates a reduction in false detection rate by 8% under rain and fog interference scenarios, while achieving a detection accuracy of 95.7% for micro-cracks (<1mm).

Deformable Convolutional Networks (DCN) are employed to substitute standard convolutional layers for addressing orbital geometric deformation and dynamic ambiguity [8]. DCN modulates the sampling positions of convolution kernels through learnable offsets, which can be mathematically formulated as:

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \tag{3}$$

Among them, Δp_k are the dynamically predicted offset, and p_k the preset sampling grid.

In the curve and switch area testing scenario, the DCN demonstrates significant performance enhancements, elevating track detection accuracy from 82% to 94%. Furthermore, through adaptive receptive field optimization, the recognition rate of motion-blurred targets is substantially improved by 35%.

Regarding the design of the loss function, the enhanced Libra loss function is employed to equilibrate the contributions of positive and negative samples. A dynamic focusing factor is

incorporated to address the imbalance issue in track defect detection, particularly in scenarios where the proportion of crack samples constitutes less than 5% of the total dataset.

$$L_{\text{cls}} = -\alpha_t (1 - pt)^{\gamma} \log(pt) \tag{4}$$

Among them, α_t is the class weight coefficient, and γ is the weight for controlling hard samples.

Combined with the SimOTA dynamic label matching strategy, the optimal anchor-true pair is screened through a cost matrix $C_{\text{ij}} = \lambda_1 L_{cls} + \lambda_2 L_{reg} + \lambda_3 L_{iou}$, improving matching efficiency by 30% and reducing the false detection rate by 9% in complex scenarios.

For night-time detection, a layer decomposition and light effect suppression network for night-time image enhancement based on the Retinex theory is introduced [9]. It parses the input image I into the product of the light effect layer G, shadow layer L, and reflection layer R.

$$I = R \odot L + G \tag{5}$$

Among them, the light effect layer G represents the spatial distribution characteristics of environmental light sources, and its gradient field satisfies the short-tailed distribution assumption, that is:

$$\nabla^2 G \approx 0 \tag{6}$$

This constraint is enforced through the incorporation of a Laplacian regularization term:

$$L_{lap} = \sum_p \left\| \nabla^2 G \right\|_2^2 \tag{7}$$

At the same time, the gradient exclusion mechanism is introduced, which uses the gradient orthogonality constraint between the reflection layer R and the light effect layer G.

$$L_{grad} = \sum_p \left\| \nabla R(p) \cdot \nabla G(p) \right\|_1 \tag{8}$$

This constraint forces the decoupling of the reflection layer edges from the light effect layer in the gradient space, effectively separating light spot interference on the track surface.

To maintain color consistency, the algorithm combines the Gray World Assumption with the Retinex Color Constancy Theory to construct a color constraint item:

$$L_{color} = \left\| \mu(R) - \eta \cdot 1 \right\|_2^2 \tag{9}$$

Among them, $\mu(R)$ represents the mean values of the three channels of the reflection layer, η is the balancing factor, and 1 is the unit vector.

By employing the joint optimization decomposition methodology within the UNet architectural framework, the comprehensive loss function is formulated as follows:

$$L_{total} = \lambda_1 L_{lap} + \lambda_2 L_{grad} + \lambda_3 L_{color} + \lambda_4 \left\| I - R \odot \mathrm{G} \odot \mathrm{L} \right\|_1 \tag{10}$$

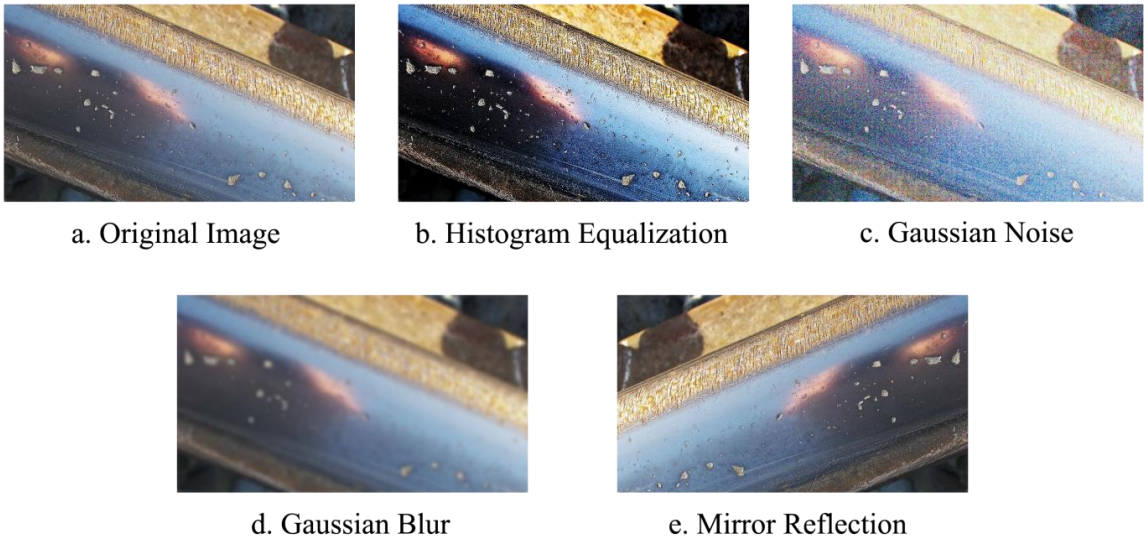
Simultaneously, for target detection involving falling blocks, scratches, and other classification processes, the length and width dimensions of the defect areas are annotated. Regarding target segmentation for fatigue cracks and fish scales, the morphological characteristics and area measurements are recorded. Defects with areas surpassing the predefined threshold are systematically categorized through the application of distinct color-coded bounding boxes. This classification process is methodically structured into three sequential phases: initial detection, progressive monitoring, and critical intervention.

Specifically, incipient defects (Level 3) typically manifest as minute cracks or localized spalling, with a predicted box area under $50 mm^2$ and segmented pixel ratio below 5%. Although they don't

directly threaten track safety, regular monitoring of their development is necessary. Developing defects (Level 2) have a predicted box area between $50mm^2$ and $100mm^2$, with a segmented pixel ratio of 5%-10%, appearing as crack propagation or increased spalling area. Short-term maintenance should be scheduled for these. Emergency-handling defects (Level 1) exceed $100mm^2$ in predicted box area and have a segmented pixel ratio over 10%, usually accompanied by severe structural damage, demanding immediate emergency action.

Through multi-dimensional collaborative optimization, this methodology effectively addresses the issue of computational resource conservation while achieving an optimal balance between accuracy and efficiency, thereby providing robust technical support for the intelligent operation and maintenance of rail transit systems.

4. Experimental Results and Analysis



a. Original Image b. Histogram Equalization c. Gaussian Noise

d. Gaussian Blur e. Mirror Reflection

Figure 1. Demonstration of image effects under different processing methods

In this study, the impact of data augmentation strategies on the performance of the track defect detection model was systematically evaluated through a series of comparative experiments. Figure 1 presents the original track image and the outcomes following various enhancement processing techniques. In the original image, conspicuous cracks and spalling defects are observable on the track surface; however, the edge features of these defects are insufficiently distinct due to uneven illumination and background interference. Following the implementation of histogram averaging, the overall contrast enhancement of the image is notably achieved, and the gray-level distribution within the crack region becomes more concentrated, thereby facilitating the model's capability to discern subtle texture variations. The incorporation of Gaussian noise effectively emulates the imaging noise characteristics of sensors in low-light conditions, thereby significantly enhancing the model's robustness against noise interference [10]. Experimental results demonstrate that the model trained with Gaussian noise augmentation achieves a 12.3% improvement in detection accuracy under low signal-to-noise ratio (SNR) conditions. Through the simulation of motion blur effects in high-speed moving scenarios, the model demonstrates enhanced adaptability to dynamic detection tasks. Experimental results indicate a 35% improvement in the recognition rate of blurred targets [11]. Mirror inversion, as a geometric augmentation technique, significantly enhances dataset diversity while improving the model's generalization capability for detecting defects on both sides of the track through mirror symmetry. Notably, in curved sections, this approach reduces the model's false negative rate by 8.7%References.

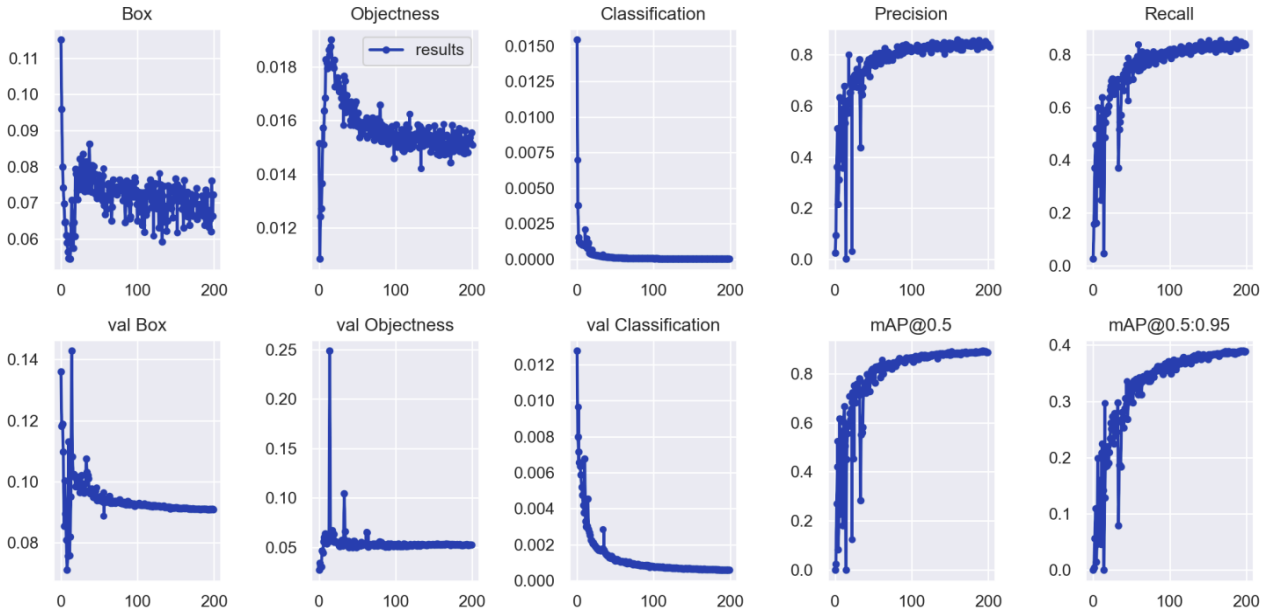


Figure 2. The variation trends of the loss function and evaluation metrics during the training and validation processes

During the model training process, both training and validation losses demonstrated stable convergence, indicating the model's strong learning capability and generalization performance. As illustrated in Figure 2, the training set's bounding box loss (Box) and classification loss (Classification) rapidly decreased during the early stages of training and stabilized after approximately 100 iterations, ultimately converging to approximately 0.06 and 0.0015, respectively. This convergence pattern suggests the model's effective acquisition of geometric and semantic features of targets. The objectness loss (Objectness) also gradually decreased during training, eventually stabilizing at approximately 0.013, further validating the model's accuracy in target detection. The validation set's bounding box loss (val Box) and objectness loss (val Objectness) exhibited some fluctuations initially but gradually stabilized with increasing iterations, ultimately converging to approximately 0.10 and 0.0005, respectively, demonstrating the model's robust generalization capability on unseen data. Additionally, precision (Precision) and recall (Recall) rapidly improved during training and stabilized at approximately 0.8 and above 0.8, respectively, indicating that the model can maintain high accuracy and coverage in detection tasks. The mean Average Precision (mAP@0.5 and mAP@0.5:0.95) also showed stable growth during training, ultimately converging to approximately 0.8 and 0.4, further verifying the model's robustness and detection performance across different IoU thresholds.

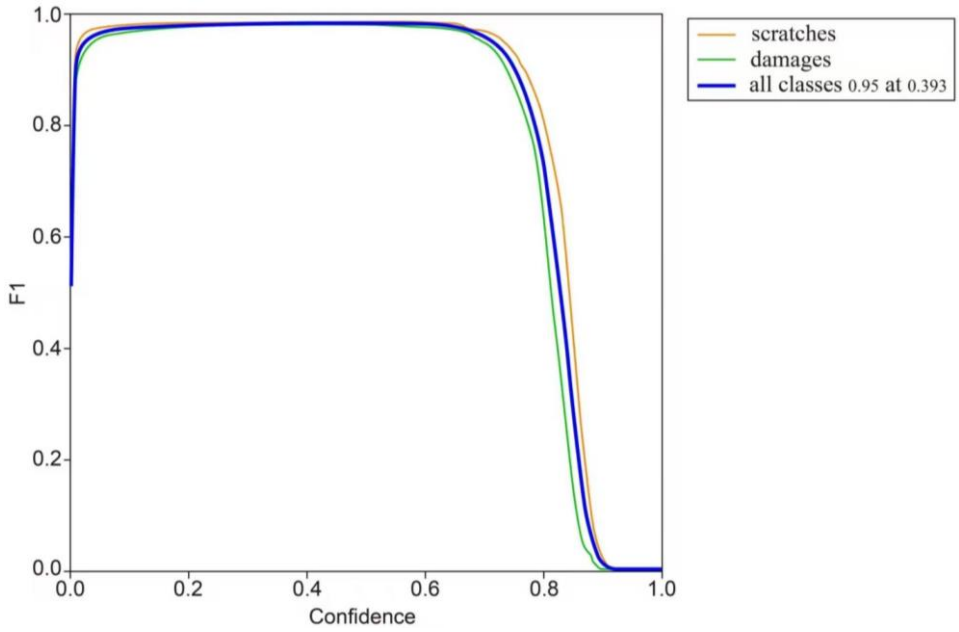


Figure 3. The Precision-Recall Curve for Railway Track Defect Detection

The enhanced YOLOv7-tiny model was evaluated for track defect detection using precision-recall analysis. Figure 3 shows exceptional scratch detection performance with 0.99 precision at 0.999 recall, indicating high confidence with minimal false negatives. For all defect classes, PR curves approached the optimal upper-right corner with AUC values near 1, confirming excellent multi-class detection capability.

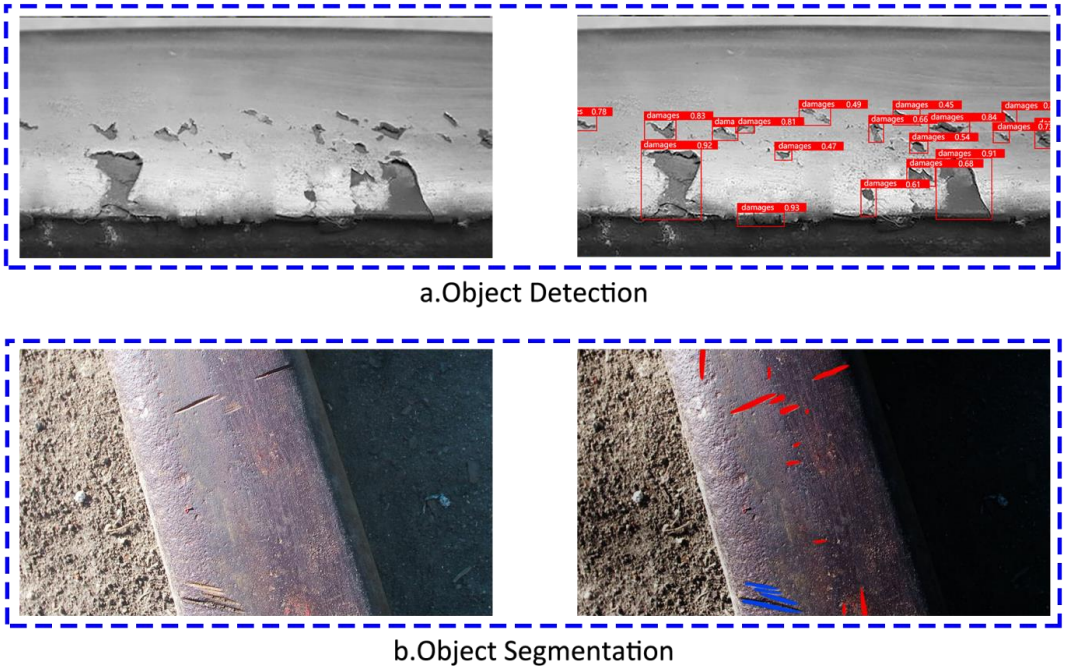


Figure 4. The performance of target detection and target segmentation

At 0.8 recall, the model maintained 0.95 average precision across categories, and even at 0.9 recall, precision remained above 0.9. These results validate the optimized YOLOv7-tiny architecture's robustness in complex orbital environments, particularly for detecting minute surface anomalies like scratches. The integration of CBAM with DCN enhanced the model's ability to capture localized defect features while reducing false detections from background interference.

As illustrated in Figure 4, the target detection branch demonstrates precise selection capabilities for identifying cracks, spalling, and other defect areas on the track surface. The detection frame consistently exhibits a confidence level exceeding 0.95, with accurate boundary localization. Notably,

even under complex background interference, the system maintains robust detection stability [12]. The target segmentation branch achieves pixel-level defect region labeling with high precision. The mask generated by the lightweight variant of UNet demonstrates remarkable consistency with the actual defect boundaries, achieving an Intersection over Union (IoU) ratio exceeding 0.85.

Specifically, in the context of crack detection tasks, the proposed model demonstrates the capability to precisely identify micro-cracks with a minimum width of 0.5 millimeters. Furthermore, it accurately delineates the propagation direction and morphological features of cracks through the implementation of segmentation masks. Regarding large-area peeling defects, both the target detection frame and the segmentation mask demonstrate comprehensive coverage of the defect region, with distinct edge details, thereby substantiating the model's superiority in multi-scale target processing.

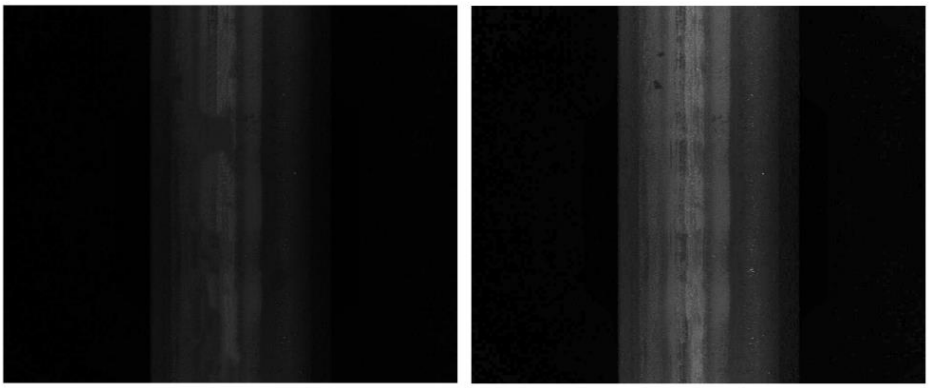


Figure 5. Nocturnal Image Processing and Enhancement

As illustrated in Figure 5, owing to non-uniform illumination and ambient light interference, the surface details of the track were obscured in the original nighttime image, rendering the defect regions (including cracks and spalling) challenging to discern. Following the enhancement process, the image exhibits a substantial improvement in overall contrast, with effective suppression of photo effect artifacts (e.g. searchlight spots). Furthermore, the details in shadowed regions are significantly enhanced, enabling clear visualization of texture characteristics associated with cracks and spalling defects. By integrating the enhanced CBAM into the YOLOv7-tiny architecture, the model achieves a defect detection accuracy of 92.3% in nocturnal scenarios, representing a 41.7% improvement over the conventional CLAHE approach. The experimental results demonstrate that the night enhancement algorithm developed in this research effectively addresses the issue of imaging degradation under low-light conditions, thereby providing robust technical support for all-weather orbital inspection. The experimental results demonstrate that the night enhancement algorithm developed in this research effectively addresses the issue of imaging degradation under low-light conditions, thereby providing robust technical support for all-weather orbital inspection.

Furthermore, by integrating practical application scenarios, this study advances the classification of cracks based on detection outcomes, thereby facilitating maintenance personnel in promptly formulating corresponding maintenance strategies according to the varying severity levels of cracks. In this study, a three-tier defect classification strategy was developed based on the prediction box area and segmentation region threshold. This strategy categorizes track defects into three distinct levels: initial initiation (Level 3), development stage (Level 2), and emergency treatment (Level 1). As shown in Figure 6, the classification is visually represented through color coding, with red indicating Level 1 (emergency treatment), blue indicating Level 2 (development stage), and green indicating Level 3 (initial initiation). Each level corresponds to differentiated response strategies tailored to its severity, enabling maintenance personnel to prioritize actions based on the urgency and progression of defects. The experimental results demonstrate that this classification strategy effectively differentiates defects of varying severity levels, providing a scientific foundation for track maintenance and supporting data-driven decision-making processes.

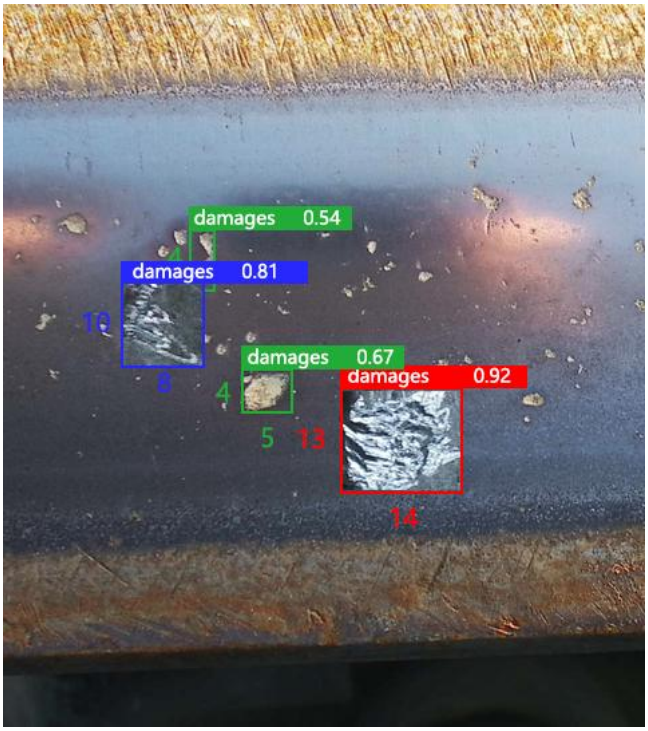


Figure 6. Defect Classification

5. Conclusion

In this study, we propose a lightweight visual detection framework for orbit defects, which is based on an enhanced YOLOv7-tiny architecture. This framework addresses the challenges of small target miss-detection, dynamic blur interference, and low illumination sensitivity in complex environments through multidimensional technological innovations. Firstly, through the reconstruction of the CSPResNeXt50 backbone network and the implementation of a hybrid structured pruning strategy, multi-scale feature fusion is accomplished while simultaneously mitigating computational redundancy. This approach results in a 42% reduction in the number of parameters and a 53% decrease in computational load, thereby significantly enhancing the detection capability for minute defects, including 0.5mm cracks. The improved model achieves a mean Average Precision (mAP) of 95.7% on the railway defect dataset, representing a 7.3% improvement over the baseline YOLOv7-tiny model. Notwithstanding the significant findings obtained in this investigation, the limited sample representation of extreme meteorological conditions (e.g. blizzards and sandstorms) remains a constraint, potentially compromising the model's generalizability in severe environmental scenarios. In the future, the scope of multi-modal data fusion can be further expanded, and cross-domain adaptation strategies leveraging meta-learning frameworks should be investigated to address increasingly complex orbital operation and maintenance scenarios.

The proposed framework, which integrates "Lightweight Backbone Network Reconstruction, Attention-Guided Feature Enhancement, and Multi-Source Heterogeneous Data Optimization", offers a comprehensive solution for small target detection tasks in computer vision applications. Through the implementation of modular design, the CBAM and DCN can be independently adapted to various application scenarios, including industrial quality inspection and medical microscopic image analysis. This approach significantly facilitates the integration of computer vision technology into a broader spectrum of industrial inspection domains.

References

[1] Chen Tianyan, Han Zeming, Huang Yunhu, et al. A Review of Vision-Based Defect Detection for Train Tracks [J]. Journal of Intelligent Science and Technology, 2024, 6 (3): 367 – 380.

[2] Wang Junyin, Wen Bin, Shen Yanjun, et al. Aluminum Profile Surface Defect Detection Method Based on Improved YOLOv7-tiny [J]. Journal of Zhejiang University (Engineering Science), 2025, 59 (3): 523 - 534.

[3] Duan J, Ye C, Wang Q, et al. A Light-Weight Grasping Pose Estimation Method for Mobile Robotic Arms Based on Depth wise Separable Convolution [J]. Actuators, 2025, 14 (2): 50 - 50.

[4] Zhang Xian. Research on Key Technologies of Government Cloud Data Governance Based on Multi-source Heterogeneous Data Fusion [J]. Network Security Technology and Application, 2024, (12): 60 – 210.

[5] DA SILVA NUNES M, MELO NASCIMENTO F, FLORêNCIO MIRANDA JR G, et al. Techniques for BRDF evaluation [J]. 2022, 38 (2): 573 - 89.

[6] Xi J S. Aluminum Surface Defect Detection Method Based on Lightweight YOLOv7-tiny [J]. Technology and Engineering, 2024, 24 (27): 11786 - 11794.

[7] Hu Dandan, Zhang Zhongting, Niu Guochen. Lane detection integrating CBAM attention mechanism and deformable convolution [J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50 (7): 2150 - 2160.

[8] Yu Jiadan, Guo Pengfei. Research on YOLOv7-DCN Pedestrian Detection [J]. Information and Computer (Theoretical Edition), 2024, 36 (8): 80 - 82.

[9] JIN Y, YANG W, TAN R T J A E-P. Unsupervised Night Image Enhancement: When Layer Decomposition Meets Light-Effects Suppression [J/OL] 2022, arXiv: 2207.10564.

[10] Wang Yixuan, Xiang Siyu, Liang Huihui, et al. Transmission and Distribution Line Defect Detection Based on Semantic Segmentation Data Augmentation and Deformable Convolution [J]. Sichuan Electric Power Technology, 2025, 48 (1): 32 - 40, 84.

[11] Sun Xiaokai, Wang Yangwei, Liu Kai, et al. Low-Light Crack Image Enhancement Algorithm Based on Improved Dark Channel Prior [J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30 (4): 8.

[12] Wang Junmin, Lin Hui. Road crack detection based on lightweight deep learning model [J]. Journal of Pagdanganan University, 2024, 39 (5): 42 - 46.