A Lightweight YOLO26 Pavement Defect Detection Method for Complex Pavement Environments

Borong Zhu[1], Kaiying Tian[1], Qihao Xu[1], Yixuan Wang[1], Dingyu Wu[2], Tianyuan Liu[1,*]
[1]*School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, Henan, China*
[2]*School of Mechanical and Electrical Engineering, Henan University of Technology, Zhengzhou, Henan, China*
**Corresponding Author*

Abstract: Timely identification of pavement deterioration plays a vital role in upholding traffic safety standards and optimizing the efficiency of road maintenance operations. Current approaches often struggle with extracting discriminative features for small-scale defects embedded in cluttered backgrounds, while surface textures and cast shadows frequently trigger both missed detections and spurious alarms. To overcome these limitations, a lightweight road damage detection framework built upon YOLO26 is presented in this work, specifically tailored for challenging pavement conditions. Taking the recently released YOLO26n model from Ultralytics as the foundation, and while strictly constraining model parameters and underlying computational expenditure, an Efficient Multi-Scale Attention (EMA) mechanism is appended and integrated at the terminal stage of the backbone network to further enhance detection accuracy. Through adaptive orthogonal and cross-spatial calibration of channel-wise feature responses, this mechanism effectively suppresses irrelevant background clutter and reinforces the representational strength of critical damage signatures. Evaluation on the China partition of the publicly available RDD2022 dataset reveals that the proposed E-YOLO26 model attains a Precision of 78.0% and an mAP@0.5 of 74.0%, accompanied by an exceptionally low computational burden of merely 5.9 GFLOPs. These figures correspond to gains of 2.6 and 0.1 percentage points relative to the baseline YOLO26 model, respectively. The findings of this study furnish an optimized computational trade-off solution along with practical engineering support for automated detection on embedded devices deployed in real-world road inspection scenarios.

Keywords: Road Damage; YOLO26; Attention Mechanism; Deep Learning; Feature Enhancement; RDD2022

1. Introduction

Roads constitute a fundamental pillar of national transportation infrastructure, and their structural condition bears a direct influence on economic growth and the operational efficiency of society at large. Various forms of visible pavement deterioration not only serve as early indicators of declining structural load-bearing capacity but also act as primary catalysts for moisture infiltration and the accelerated degradation of deeper pavement layers. Should these early-stage damages fail to be identified and addressed in a timely and effective manner, the combined effects of heavy vehicular traffic and environmental exposure will cause them to rapidly progress into severe defects such as potholes and surface collapse [1], thereby driving full-lifecycle road maintenance costs upward at an exponential rate and posing serious threats to vehicular safety. Consequently, the ability to detect road damage promptly and accurately carries substantial engineering significance and economic value for the implementation of preventive maintenance strategies, the prolongation of road service life, and the assurance of fundamental traffic safety.

Historically, road inspection has predominantly depended on manual visual assessment, an approach characterized by low efficiency, considerable subjectivity, occupational hazards, and difficulties in producing digitized records suitable for systematic analysis. Such a method proves inadequate for addressing the pressing demands of modern, large-scale road network management. In recent years, computer vision

technologies anchored in deep learning have undergone transformative advancement, and object detection algorithms grounded in deep neural networks have been broadly adopted for damage identification across domains including roads and bridges [2].

The landscape of object detection algorithms can generally be delineated along two trajectories. The first encompasses single-stage detectors, which accomplish target localization and classification within a unified neural network architecture without the need for prior candidate region generation. This paradigm offers a streamlined structure and high computational efficiency, conferring notable advantages in terms of inference speed and resource expenditure, with YOLO and SSD standing as prominent exemplars. The second trajectory comprises two-stage detectors, which employ a sequential strategy by first proposing candidate regions and subsequently performing refined classification and spatial adjustment upon these regions; representative architectures include R-CNN and Faster R-CNN [2-4]. Two-stage detectors generally deliver superior detection accuracy, yet they incur greater structural complexity and elevated computational overhead. In application scenarios demanding real-time responsiveness, such as road defect detection, single-stage architectures prove more suitable owing to their compact design and rapid inference capabilities, thereby achieving wider practical adoption. Among these, YOLO (You Only Look Once) [5] stands as a seminal algorithm that formulates the detection task as an end-to-end regression problem, exerting a profound and lasting impact on the evolution of the entire object detection field through its remarkable speed and efficiency.

At present, research efforts dedicated to road damage detection leveraging YOLO-based frameworks have produced a wealth of theoretical contributions. Su Weiguo et al. [6] developed a road crack recognition system grounded in the YOLOv3 algorithm, fulfilling real-time operational requirements while sustaining high detection accuracy. Yuan et al. [7] introduced an enhanced variant, YOLO-CD, built upon YOLOv8n, which markedly elevated multi-scale crack detection performance through the incorporation of a Scale Sequence Feature Fusion (SSFF) module and attention mechanisms. Li [8] advanced the YOLOv8 framework by incorporating the Spatial Pyramid Pooling Network together with the Convolutional Block Attention Module (CBAM) to enable adaptive feature extraction for road defects spanning multiple scales. Liu Yuanyuan et al. [9] put forward a lightweight crack detection approach with adaptive characteristics by integrating multiple attention mechanisms including CEA, effectively boosting detection accuracy under complex environmental conditions. Ning et al. [10] devised a lightweight pavement crack detection method founded on YOLOv7, employing SimAM attention, and systematically evaluated the effects of attention placement at various positions within the network. Xuan Yiguo et al. [11] tackled the issues of missed and false detections by incorporating Deformable Convolution (DCN) alongside a small-target detection evaluation framework based on the Wasserstein distance to further refine model performance. Nevertheless, cracks typically manifest as highly elongated structures within images, and during the progressive downsampling inherent to deep convolutional networks, their faint gradient signals are readily overwhelmed by background noise, resulting in persistently elevated miss rates for subtle cracks when processed by general-purpose detection models.

A thorough examination of the existing literature indicates that current YOLO-based improvement strategies still exhibit notable shortcomings when confronted with realistic and complex inspection environments. First, the miss rate of prevailing models concerning the distinctive geometric and visual attributes of cracks, namely their extremely elongated morphology and weak contrast, remains unacceptably high. Second, most existing enhancements target older YOLO generations, leaving the potential of the most recent algorithmic architectures largely unexplored. Moreover, the excessive accumulation of complex modules and the resulting computational burden frequently stand in direct conflict with the stringent real-time and low-power requirements imposed by edge devices deployed in road inspection operations.

In response to these challenges, an ultra-lightweight pavement damage detection model termed E-YOLO26 (EMA-enhanced YOLO26), founded on cross-spatial feature encoding, is proposed in this study. The principal contributions are summarized as follows:

(1) Integration of an orthogonal cross-spatial

attention enhancement mechanism. To address the issue that complex damage morphologies such as pavement alligator cracking exhibit considerable variability and are easily confounded by environmental factors, the Efficient Multi-Scale Attention (EMA) mechanism is specifically embedded within the feature fusion network. This mechanism carries out orthogonal spatial encoding through 1D adaptive pooling applied along horizontal and vertical orientations, thereby strengthening the model's capacity to capture the continuous linear texture patterns characteristic of damage. As a result, it effectively suppresses interference from complex background artifacts including pavement shadows and tire marks at the fundamental feature level.

(2) Thorough performance validation across complex multi-classification scenarios. Multi-classification experiments were performed on five fine-grained defect categories, longitudinal cracks, transverse cracks, alligator cracks, potholes, and repair patches, within the China partition of RDD2022. The outcomes demonstrate that E-YOLO26 exhibits outstanding detection accuracy and generalization capability under an extremely constrained computational budget, offering an efficient and lightweight technical pathway for the intelligent and high-precision operation and maintenance of industrial-grade road inspection systems.

2. Methodology

2.1 YOLO26 Model

The fundamental architecture of YOLO26 comprises three principal components: the Backbone feature extraction network, the Neck feature fusion network, and the decoupled detection Head. The backbone is tasked with extracting hierarchical features from the input image and is primarily composed of Conv modules, C3K2 modules, SPPF modules, and C2PSA modules. The neck network aggregates features at different abstraction levels produced by the backbone, achieving effective integration and enhancement of multi-scale information. Subsequently, the detection head carries out parallel prediction of target bounding boxes and category labels based on the fused feature representations [5]. The complete framework of YOLO26 is illustrated in Figure 1.

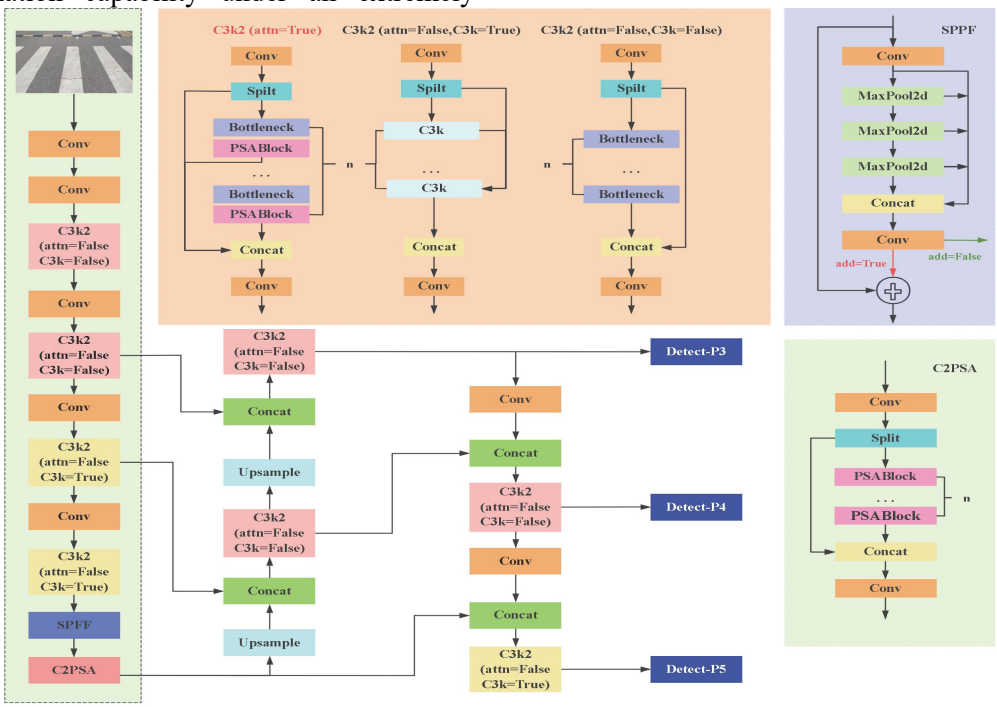


Figure 1. Overall Framework of YOLO26

Conventional YOLO architectures have traditionally sought accuracy gains through deeper networks and wider channel configurations, yet such strategies often introduce computational redundancy, rendering them ill-suited for lightweight deployment. YOLO26 departs from this evolutionary paradigm of parameter accumulation and fundamentally restructures the architectural design philosophy. By incorporating a

Non-Maximum Suppression Free (NMS-Free) inference mechanism, Progressive Loss balance (ProgLoss), and Small Target Label Assignment (STAL) strategies, it preserves a lightweight footprint while substantially improving both the detection rate and localization precision for targets exhibiting complex morphological characteristics.

2.2 Improved YOLO26 Model

Despite the strong performance demonstrated by the native YOLO26 algorithm across general natural image object detection tasks, pavement imagery captured in real open-world scenarios exhibits intrinsically unstructured properties, including rough background textures, highly variable defect scales, and divergent micro-topological structures. Under the constraint of maintaining an extremely lightweight baseline, a cross-spatial attention mechanism is incorporated to construct the E-YOLO26 (EMA-enhanced YOLO26) detection model. The complete network architecture of the enhanced E-YOLO26 is depicted in Figure 2, where the red-colored EMA module positioned at the lower left signifies the introduced cross-spatial attention component.

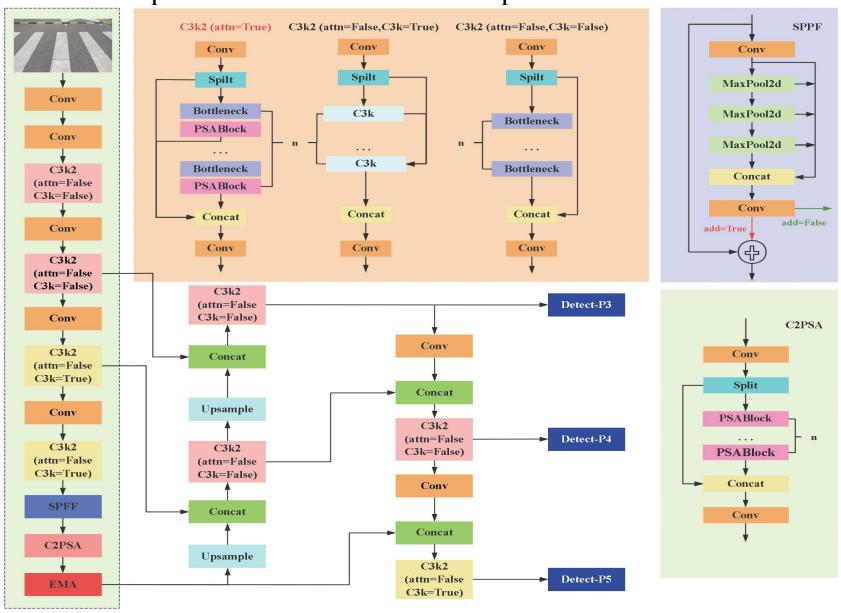


Figure 2. Overall Framework of the Improved E-YOLO26

The native YOLO26 architecture incorporates the C2PSA (Cross-Stage Partial Spatial Attention) self-attention module by default at the terminus of the feature fusion pipeline. Such modules are inclined to capture global long-range dependencies; however, when confronted with fine-grained, intricately interwoven texture patterns, they tend to induce excessive smoothing of local features, causing faint cross-crack gradient signals to be masked by expansive pavement shadow or tire mark noise. To bolster the network's capability in capturing divergent spatial features, the EMA module is appended deep within the backbone network in this work.

The EMA module introduces a cross-spatial learning strategy wherein input feature maps are partitioned into channel groups. Within each subgroup, parallel 1×1 and 3×3 convolutional branches are employed to extract multi-scale spatial responses. Following this, EMA performs concurrent Adaptive Average Pooling along the horizontal and vertical axes, encoding 2D spatial information into 1D directional vectors. A Softmax function is then applied to generate cross-dimensional attention weights. The core pooling computations can be expressed as:

$$z_c^h = \frac{1}{W}\sum_{i=1}^{W} x_c(h,i) \tag{1}$$

$$z_c^w = \frac{1}{H}\sum_{j=1}^{H} x_c(j,w) \tag{2}$$

Here, c denotes the c-th channel, while i and j represent spatial positional indices along the spatial dimensions of x. The terms z_c^h and z_c^w correspond to the pooling outputs of the feature map along the height H and width W directions, respectively.

This orthogonal 1D feature encoding approach aligns naturally with the horizontal and vertical interwoven texture orientations characteristic of mesh-like alligator cracks. The EMA module is

capable of adaptively attending to the continuous topological structure of cracks without substantially inflating the network's computational cost (GFLOPs), effectively mitigating interference from complex background artifacts and markedly enhancing the network's ability to extract fine-grained local features. The internal network topology of the EMA module is presented in Figure 3.

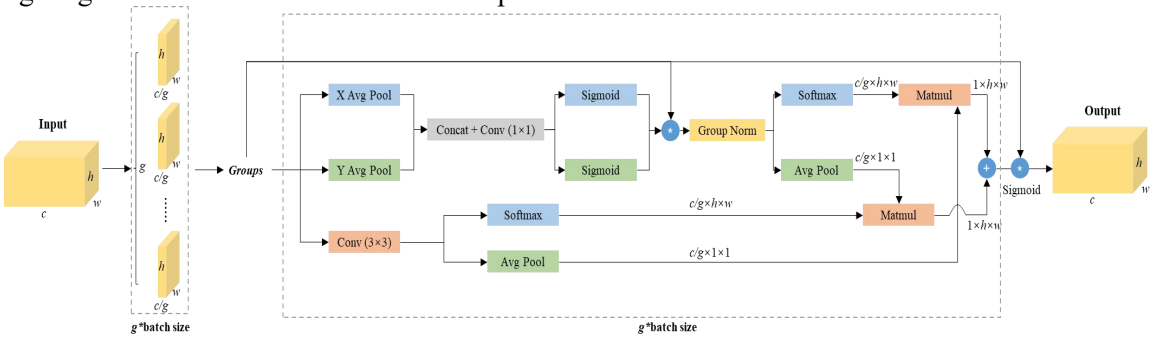


Figure 3. Internal Network Topology of the EMA Module

3. Experimental Results and Analysis

3.1 Dataset

The experiments conducted in this study employ the open-source RDD2022 (Road Damage Detection 2022) dataset [12]. This dataset has been predominantly collected via drones and vehicle-mounted imaging systems operating on real open roads and contains exceptionally complex background noise. The China partition was selected for experimentation, designated as RDD2022_China, encompassing a total of 4,378 images. To assess the effectiveness of the proposed algorithm within complex multi-classification tasks, five representative damage categories were filtered and retained: longitudinal cracks (D00), transverse cracks (D10), alligator cracks (D20), potholes (D40), and repair areas (Repair). Prior to experimentation, the dataset was randomly partitioned into training, validation, and testing subsets in an 8:1:1 ratio to ensure the objectivity of model assessment.

3.2 Experimental Environment

All experiments were carried out on a Windows 10 operating system equipped with an NVIDIA GeForce RTX 4050 discrete GPU featuring 6GB of VRAM. The implementation was developed in Python 3.9 using the PyTorch 2.3.0 deep learning framework, with CUDA enabled for accelerated computation. The training hyperparameters were configured as follows: the number of training epochs was set to 150, the batch size to 16, the initial learning rate to 0.01, the SGD optimizer was employed, and the input image resolution was uniformly fixed at 640×640. A complete summary of the parameter settings used during model training is provided in Table 1.

Table 1. Parameter Settings during the Training

Parameter	Setting
Input Image Size	640 × 640 pixels
Batch Size	16
Training Epochs	150
Optimizer	SGD
Initial Learning Rate	0.01
Momentum	0.937
Weight Decay	0.0005

3.3 Evaluation Metrics

Precision (P), Recall (R), Average Precision (AP), and mean Average Precision (mAP) were selected as the quantitative metrics for evaluating model performance. Precision quantifies the fraction of genuine positive samples among all instances classified as positive by the model. Recall measures the proportion of correctly identified positive samples relative to the total number of actual positive instances. Average Precision (AP) is defined as the area under the Precision-Recall curve; through interpolation of this curve, the integral of precision at varying recall levels is computed to provide a quantitative assessment of model performance. Accounting for the influence of the IoU threshold, both mAP@0.5 and the more stringent mAP@0.5:0.95 metrics are employed. The computational formulas for these metrics are as follows:

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{5}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{6}$$

In these formulations, TP denotes the count of correctly identified cracks; FP represents the count of negative samples erroneously classified as cracks (i.e., false positives); FN signifies the count of undetected true cracks (i.e., false negatives); N denotes the total number of ground-truth cracks evaluated; and i indexes the i-th crack.

3.4 Ablation Study

To systematically and quantitatively demonstrate the tangible contribution of each innovative component within the proposed E-YOLO26 model to overall detection performance, this section adopts the standard YOLO26n architecture as the baseline and designs ablation experiments following the principle of controlled variables. The investigation closely tracked the evolution of model parameters, computational overhead (GFLOPs), and mean Average Precision (mAP@0.5). The quantitative findings are summarized in Table 2. The corresponding mAP@0.5-0.95 values for both models across the five distress category labels are reported in Table 3.

Table 2. Results of Ablation Experiments before and After Improvement

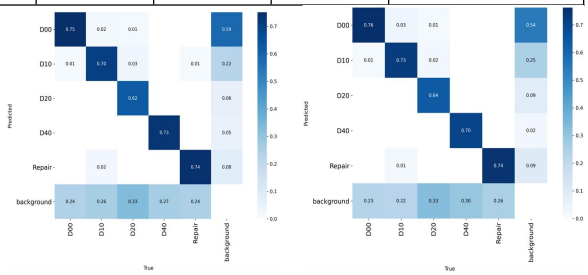
Group	Model Architecture	Precision (P, %)	Recall (R, %)	Parameters (M)	Computation (GFLOPs)	mAP@0.5(%)	mAP@0.5-0.95(%)
1	YOLO26n	75.4	65.7	2.50	5.8	73.9	45.4
2	E-YOLO26	78.0	64.7	2.51	5.9	74.0	47.8

Table 3. mAP@0.5-0.95 of Different Category Data Before and After Improvement

Model Architecture	mAP@0.5-0.95(%)				
	D00	D10	D20	D40	Repair
YOLO26n	40.4	39.1	29.8	41.4	60.3
E-YOLO26	47.8	41.9	34.4	44.5	64.7

As evident from the results presented in Tables 3 and 4, the incorporation of the EMA module elevated the model's overall Precision from the baseline value of 75.4% to 78.0%, while Recall experienced a marginal decline but remained at a comparable level. The comprehensive IoU performance (mAP@0.5-0.95) under strict regression conditions improved by 2-4 percentage points across all damage category labels, corroborating that the cross-spatial orthogonal encoding mechanism of EMA substantially enhanced the model's capacity to discriminate between structurally distinct damages. Concurrently, the introduction of this module did not precipitate any significant increase in computational complexity (GFLOPs), confirming that the model achieves a favorable balance between computational expenditure and detection performance.

Building upon the above quantitative indicators, and to conduct a deeper analysis of the EMA module's class-specific discrimination tendencies and resilience against background interference, Figure 4 presents a comparison of normalized confusion matrices before and after module integration, while Figure 5 displays the P-R curves of the improved model.



(a) YOLO26n (b) E-YOLO26

Figure 4. Comparison of Confusion Matrices before and After Introducing the EMA Module

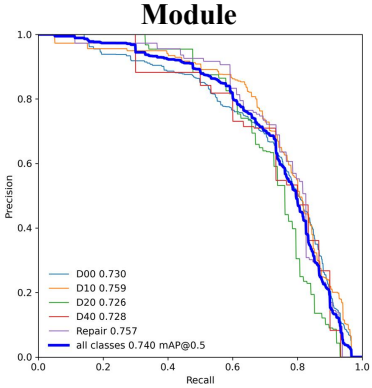


Figure 5. P-R Curves of the E-YOLO26 Model on the Test Set

From the confusion matrix shown in Figure 4, it is apparent that prior to EMA module incorporation, the baseline model exhibited a non-trivial proportion of false-positive misclassifications within the Background

category. Following the integration of the enhancement module, the rate at which the model erroneously classified background regions as various crack types was markedly suppressed. This provides direct evidence, grounded in the underlying data distribution characteristics, that the EMA orthogonal spatial filtering mechanism is capable of effectively isolating high-frequency interference sources such as pavement shadows and tire marks, thereby elevating overall Precision.

Examination of the P-R curve in Figure 5 reveals that the area enclosed by each curve and the coordinate axes, corresponding to the AP values per category, performs exceptionally well for core defects including longitudinal cracks (D00) and transverse cracks (D10), exhibiting fully convex envelopes. This further validates, across multiple category dimensions, the efficiency and robustness of the proposed algorithm when processing road damages characterized by extreme aspect ratios and weak feature signatures.

3.5 Comparative Experiments

Under identical experimental conditions, comparative evaluations were conducted between the E-YOLO26 model and several other mainstream YOLO variants to verify the effectiveness of E-YOLO26's performance. The experimental outcomes are presented in Table 4.

Table 4. Comparison of Evaluation Metrics across Different YOLO Models

Group	Model Architecture	Precision (P, %)	Recall (R, %)	Parameters (M)	Computation (GFLOPs)	mAP@0.5 (%)	mAP@0.5-0.95 (%)
1	YOLOv8n	81.5	74.1	3.01	8.1	81.4	53.7
2	YOLOv10n	77.6	67.1	2.71	6.5	74.9	48.2
3	YOLOv11n	82.5	73.5	2.59	6.3	80.9	51.8
4	YOLOv12n	72.5	70.0	2.56	6.3	72.6	44.6
5	YOLO26n	75.4	65.7	2.50	5.8	73.9	45.4
6	E-YOLO26	78.0	64.7	2.51	5.9	74.0	47.8

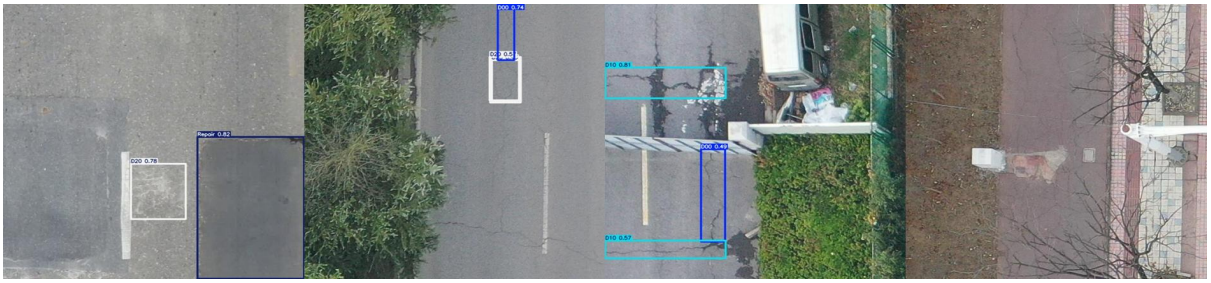
As indicated in Table 4, although YOLOv8n and YOLOv11n achieved mAP@0.5 scores of 81.4% and 80.9%, respectively, these performance advantages were attained at the cost of substantially higher computational demands, measured at 8.1 GFLOPs and 6.3 GFLOPs, respectively. Notably, the computational complexity of YOLOv8n exceeds that of the E-YOLO26 model proposed herein by approximately 37.3%.

Given the edge deployment constraints of road inspection equipment, the introduction of cross-spatial feature encoding enables E-YOLO26 to sustain a detection Precision of 78.0% under an exceptionally low computational budget. The model presented in this study achieves a superior equilibrium between low-power hardware environments and highly functional detection performance, rendering it particularly well-suited for integration into drone-mounted or portable vehicle-based inspection systems.

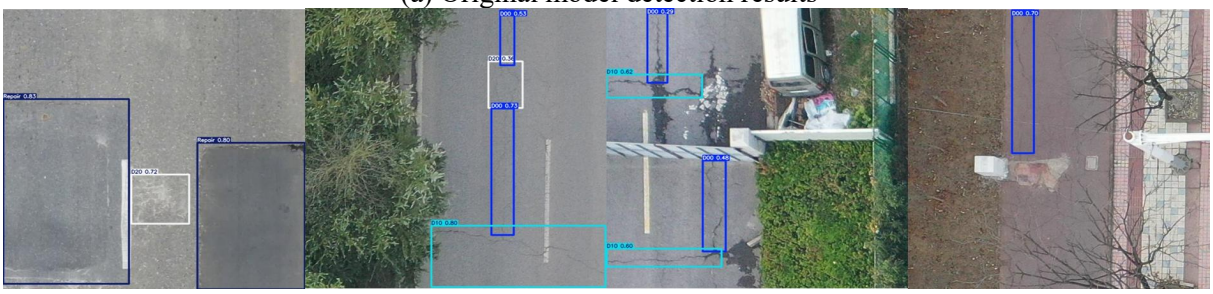
3.6 Results Analysis

To visually corroborate the efficacy of the proposed model enhancement, individual road damage images were evaluated, and Figure 6 illustrates the actual detection outcomes of the original and improved models on the RDD2022 dataset. In the first images of Figures 6(a) and 6(b), the boundaries of the repair area (Repair) exhibit coloration closely resembling that of the surrounding normal pavement, causing the original model to fail in effective feature extraction and resulting in missed detections. The fourth image demonstrated insufficient sensitivity to elongated and minute longitudinal cracks (D00), likewise leading to undetected instances, thereby exposing the limitations of the original model in feature capture. In contrast, the improved model successfully identified these previously missed cracks, signifying an enhanced small-target detection capability. The second and third images involve complex scenarios featuring leaf occlusion and repair defects with visual similarity to the pavement surface. The detection accuracy of the original model falls short of that achieved by the improved model, further substantiating the superior performance of the enhanced model under occlusion conditions.

In summary, the improved model demonstrates significant performance enhancements, enabling more precise identification of road defects, more accurate determination of their spatial locations and categories, and consequently a more effective response to the challenges encountered in practical road inspection applications.



(a) Original model detection results



(b) Improved model detection results

Figure 6. Visualized Detection Results

4. Conclusion

Targeting the challenges of micro-crack features being readily overlooked and complex topological damages being highly susceptible to background noise interference and false detections in real open-road scenarios, a lightweight E-YOLO26 road damage detection model built upon enhancements to YOLO26 is proposed in this study. The model implements an orthogonal 1D pooling encoding strategy through the integration of the EMA cross-spatial attention mechanism within the deep network layers. The enhanced E-YOLO26 model achieved a detection Precision of 78.0% on the RDD2022_China dataset. The parameter count of the E-YOLO26 model amounts to merely 2.51M, and the inference computational cost stands at only 5.9 GFLOPs, realizing a harmonious balance between lightweight design and high precision. This work provides a viable solution for highly accurate and lightweight intelligent detection of complex road damages. Future research directions will explore the incorporation of feature decoupling and dynamically adaptive threshold generation networks, enabling the model to dynamically modulate the rigor of feature filtering in response to the illumination intensity and texture granularity of the input image. Additionally, extended experiments will be conducted on multi-source and multi-scenario road damage datasets, incorporating a broader range of detection and segmentation frameworks for systematic benchmarking to verify the cross-scenario generalization stability of the model.

References

[1] Zhang J Q, Yang X, Wang W, et al. Automated guided vehicles and autonomous mobile robots for recognition and tracking in civil engineering. Automation in Construction, 2023, 146: 104699.

[2] Luo Z, Jiang Y, Li W C. Road Defect Detection Model Based on Improved YOLO11n. Microelectronics & Computer, 2025, 42(11): 25-36.

[3] Liu D D, Xue J J, Zhao X K, et al. YOLOv8 Road Crack Detection Method Based on Frequency-Aware Feature Fusion. Journal of Spatiotemporal Information, 2025, 32(06): 675-684.

[4] Feng C H, Xu H Y, Tao J Q, et al. SSBS-YOLO: Optimization of Small Target Pavement Damage Detection in YOLOv8 Based on Lightweight Road Inspection. Journal of Zhejiang Normal University (Natural Sciences), 2026:1-12[2026-05-10].

[5] Yavuz B. Scale-Dependent Performance Analysis of YOLO26 and YOLOv11 for PPE Detection. Electronics, 2026, 15(6): 1146.

[6] Su W G, Wang J X. Research on Road Crack Recognition Model Based on YOLO v3 Deep Learning Algorithm. Journal of China & Foreign Highway, 2023, 43(2): 58-63.

[7] Yuan H, Li Q, Wang Y. Road Crack Detection Based on YOLO-CD. Science Technology and Engineering, 2025, 25(9):3888-3895

[8] Li G. Road Defect Identification and Location Method Based on an Improved ML-YOLO Algorithm. Sensors (Basel, Switzerland), 2026, 24(21):6783.

[9] Liu Y Y, Zhu K, Gu Z H, et al. Lightweight Pavement Crack Detection Method with Adaptive Features. Optics and Precision Engineering, 2026, 34, (2):336-351.

[10] Ning Z P, Wang H, Li S L, et al.YOLOv7-RDD: A lightweight efficient pavement distress detection model. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(7):6994-7003.

[11] Xuan Y G, Yu C B, Jiang Q C, et al. Road Crack and Pothole Detection Algorithm Based on Improved YOLOv7. Science Technology and Engineering, 2024, 24(17): 7205-7213.

[12] Arya D, Maeda H, Ghosh S K, et al. RDD2022: a multi-national image dataset for automatic road damage detection. Geoscience Data Journal, 2024, 11(4):846-862.

