Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 15, No. 3, 2026

Research on Pavement Crack Identification Method Based on Improved YOLOv11

Xuanyu Fang

College of Metropolitan Transportation, Beijing University of Technology, Beijing, 100124, China

Abstract: Pavement cracks are the most predominant form of distress in asphalt pavements; therefore, pavement crack detection constitutes a critical component of pavement maintenance. To address the issues of low recognition accuracy, false detection, and missed detection in highway pavement crack identification, this paper proposes a pavement crack detection model based on YOLOv11.First, Ghost convolution and DynamicConv are integrated into the C3K2 module to reduce computational load and enhance feature extraction capabilities. Second, the CGA attention mechanism is introduced in combination with C2PSA to strengthen feature fusion and contextual information extraction. Finally, the Adaptive Threshold Focal Loss function is employed to address the issue of class imbalance and improve the detection capability for difficult samples. Experiments conducted on a self-constructed dataset demonstrate that the improved model outperforms existing methods in both accuracy and efficiency. The proposed method provides an efficient and accurate solution for pavement crack detection.

Keywords: Pavement Crack Detection; Dynamic Convolution; Attention Mechanism; Loss Function.

1. Introduction

With the rapid development of China's economy, significant progress has been made in highway infrastructure construction. By the end of 2023, the total length of China's highway network had reached 5.4368 million kilometers.As the most widespread mode of transportation carrying the largest number of passengers, highway transportation plays a vital role in supporting national production and people's livelihoods.Over time, highway pavements are prone to various distresses, such as cracks and potholes, due to factors including vehicle overloading, extreme weather, and natural aging.Among these, pavement cracking is the most common type of distress and requires the most timely treatment, as improper handling can lead to its propagation into more severe structural damage, thereby affecting driving comfort and safety. Consequently, timely crack detection is of paramount importance for maintenance operations.

Traditional pavement distress detection methods primarily rely on manual inspection, an approach that not only consumes substantial manpower and financial resources but is also highly dependent on individual experience and judgment, making the results prone to bias and even errors.This makes it difficult to meet the modern requirements for the early detection and timely treatment of pavement distresses.Although advancements in technology have led to the introduction of modern equipment and techniques, such as specialized highway distress inspection vehicles that use onboard cameras for rapid patrolling and photographing of maintenance sections—thereby significantly improving inspection efficiency—challenges remain.However, even with these advanced methods, the subsequent step still requires professional technicians to classify and process the collected distress images, a process that is time-consuming and hinders overall work efficiency.

With the advancement of computer technology, pavement crack detection has transitioned from traditional manual inspection to intelligent detection.In recent years, deep learning technology has developed rapidly and become a research hotspot.Deep learning techniques can autonomously learn crack features to achieve automatic identification, demonstrating significant practical application value.Liu et al. [1] proposed a multi-scale attention mechanism called MsCGA, which learns both local detail features and global features from high-level representations to compensate for the loss of detailed information, thereby improving the accuracy of multi-scale crack detection.Geng et al. proposed a feature fusion method to compensate for the feature information lost in the path aggregation network, thereby improving detection accuracy.Zhang et al. addressed the issue of discontinuous crack feature extraction caused by inaccurate pixel recovery during upsampling by proposing a crack-aware attention module, which enhances the model's accuracy in extracting details of pavement cracks.Current methods for pavement crack detection based on deep learning models exhibit high detection accuracy and speed; however, further improvements are needed to address the challenges of weak feature information in asphalt pavement cracks and the lack of significant contrast with the background, which lead to difficulties in feature extraction and suboptimal model learning.

The YOLOv11 object detection algorithm incorporates optimizations and improvements based on YOLOv8, achieving higher detection accuracy and faster detection speed compared to previous models. This paper presents improvements based on the YOLOv11 model, with the main contributions as follows:(1) Ghost dynamic convolution is integrated into the C3K2 module to reduce computational load and enhance feature representation capability.(2) The CGA attention mechanism is combined with C2PSA to strengthen feature fusion capability.(3) The ATFL loss function is introduced to address the issue of class imbalance.

2. Research Methodology

2.1. YOLOv11 Basic Architecture

YOLOv11 is a new generation object detection algorithm developed by Ultralytics, featuring significant architectural and training methodology improvements over previous YOLO versions.It integrates improved model architecture

design, enhanced feature extraction techniques, and optimized training methods.Through these design improvements, YOLOv11 provides better feature extraction—the process of identifying important patterns and details from images—enabling more accurate capture of complex aspects even in challenging scenarios.

The specific structure of the YOLOv11 model is shown in Figure 1.The backbone section is responsible for feature extraction, employing a series of convolutional and deconvolutional layers, while utilizing residual connections and bottleneck structures to reduce network size and improve performance. The core of the backbone network is the C3K2 module, which optimizes the information flow within the network by splitting feature maps and applying a series of smaller kernel convolutions; the smaller 3x3 kernels allow for more efficient computation while preserving the model's ability to capture essential features in images.Compared to the C2f module in YOLOv8, the C3K2 module uses fewer parameters to improve feature representation.

The C2PSA module employs two PSA (Partial Spatial Attention) modules that operate on different branches of the feature map and are then concatenated, similar to the structure of the C2f module.This configuration ensures that the model focuses on spatial information while maintaining a balance between computational cost and detection accuracy.The C2PSA module refines the model's ability to selectively focus on critical regions by applying spatial attention to the extracted features.This enables YOLOv11 to outperform previous versions in scenarios requiring precise detection.

The Neck network is situated between the backbone network and the head network, and its function is to perform feature fusion and enhancement.The Head network is the decision-making part of the object detection model, responsible for generating the final detection results.

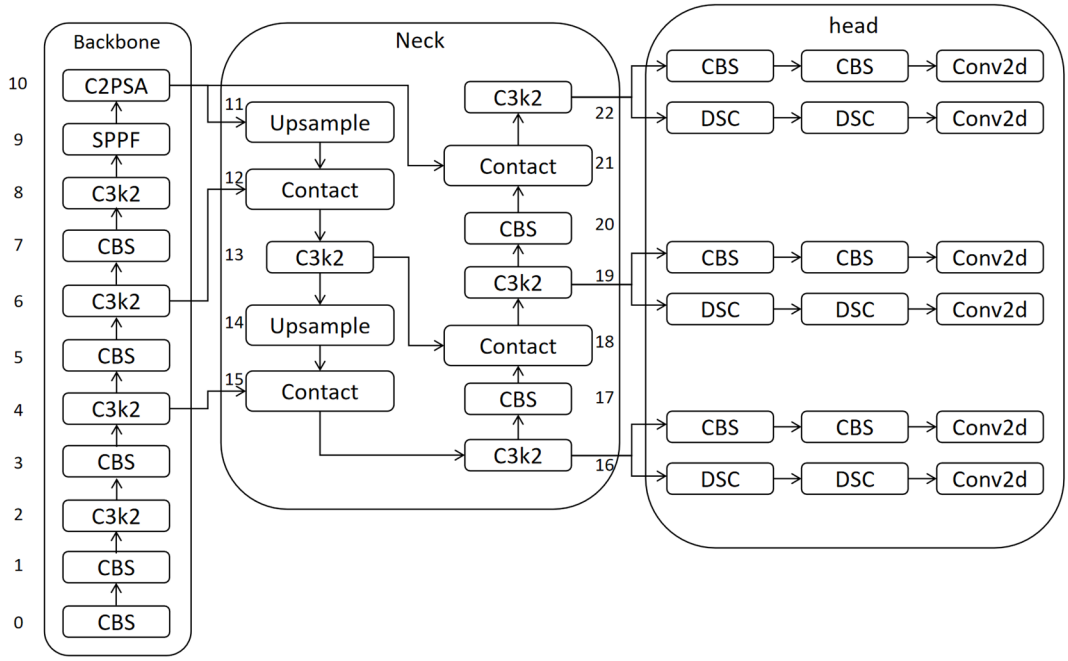


Fig 1. YOLOv11 network architecture diagram

2.2. Integration of Ghost Convolution and DynamicConv into the C3K2 Module

The fusion of Ghost[4] Convolution and DynamicConv[5] in the Neck section (C3k2 module) of the YOLOv11 model significantly enhances the model's feature fusion capability and multi-scale object detection ability by dynamically adjusting convolution kernels and generating redundant feature maps.This fusion improves model accuracy while maintaining low computational cost, with particularly outstanding performance in scenarios involving complex backgrounds and multi-scale objects.

The Ghost module reduces computational load and enhances the model's representational capacity by generating redundant feature maps, thereby decreasing the number of parameters and computational complexity without significantly compromising model performance, and improving inference speed, particularly on resource-constrained devices.It reduces the computational burden of convolution operations by generating "ghost" feature maps through linear transformations.

DynamicConv enhances the model's representational capacity by dynamically aggregating multiple convolution kernels while maintaining low computational cost.By selectively combining multiple convolution kernels through an attention mechanism, it enables the model to adaptively adjust kernel weights based on the input, thereby improving the model's expressive power. Dynamically adjusting the weights of convolution kernels allows the model to better capture complex features of the input data; particularly when handling objects of different scales and shapes, it can adaptively adjust the kernels, enhancing the model's generalization ability.Adaptively adjusting convolution kernels enables better accommodation of diverse input data, thereby improving model accuracy; especially in object detection tasks, it facilitates better handling of multi-scale objects and those in complex backgrounds.

The Neck section of YOLOv11 is primarily responsible for multi-scale feature fusion, where the C3k2 module fuses feature maps of different scales through convolution operations.With the introduction of Ghost+DynamicConv, the C3k2 module can fuse features more efficiently, reducing computational load while retaining more feature information, thereby improving the effectiveness of feature fusion.The fusion of the Ghost module and DynamicConv significantly enhances the multi-scale object detection capability of the

C3k2 module.The Ghost module preserves rich detail information while reducing computational load by generating redundant feature maps, particularly enhancing the multi-scale feature fusion capability of the Neck section. DynamicConv, on the other hand, dynamically adjusts convolution kernel weights through an attention mechanism, enabling the model to adaptively select the optimal kernels for feature extraction, effectively improving the model's representational capacity. This fusion mechanism achieves a favorable balance between computational efficiency and model complexity: the Ghost module reduces computational cost, while DynamicConv enhances the adaptability of feature extraction.Experimental results demonstrate that this design significantly improves the model's capability in handling complex detection tasks without substantially increasing the computational burden, particularly achieving notable improvements in multi-scale object detection accuracy.

2.3. Combination of the CGA Attention Mechanism and C2PSA

Against the backdrop of high computational costs and slow inference speeds in deep learning models, the Cascaded Group Attention (CGA) [6] mechanism was developed. By splitting input features into different parts, feeding them into respective attention heads for self-attention computation, and cascading the outputs, it addresses the issue of low computational efficiency caused by redundant attention heads in multi-head self-attention. This reduces the waste of computational resources, increases the diversity of attention maps, and enhances model capacity without introducing excessive additional parameters or latency overhead, thereby improving the overall computational efficiency and performance of the model.

Upon entering the CGA module, the input features are evenly split into several parts, with each part serving as the input to one attention head. This allows each head to focus on a different portion of the input features, enabling parallel feature processing while laying the foundation for improved attention diversity.Each attention head contains three linear projection layers (for mapping input features into different subspaces) and a unit for computing self-attention.During computation, through the cascading operation between attention heads, the output of the preceding head is added to the input of the subsequent head, achieving progressive feature refinement.After the cascading computation, the output features from all attention heads are merged into a complete feature map through a concatenation operation. This is then passed through a linear projection layer for dimensionality transformation, ensuring the output feature dimensions are consistent with the input features, facilitating further processing by subsequent network layers.

By introducing the Cascaded Group Attention (CGA) mechanism, which provides different splits of input features to each attention head and explicitly decomposes the attention computation across heads, redundant calculations are reduced. This improves computational efficiency, enabling the model to handle more tasks under the same computational resources or complete tasks in a shorter time.The CGA mechanism enables each attention head to focus on different parts of the input features, thereby increasing the diversity of attention maps. This allows the model to better capture various feature patterns within the input data, subsequently enhancing the model's representational capacity and performance. Employing a cascading operation, where the output of the preceding head is added to the input features of the current head, progressively refines feature representation. This also increases network depth without introducing additional parameters, further enhancing model capacity. This enables the model to learn more complex feature representations, improving its accuracy across various vision tasks.

2.4. Adaptive Threshold Focal Loss (ATFL) Function

The Adaptive Threshold Focal Loss (ATFL)[7] is a loss function that dynamically adjusts loss weights. It aims to enhance model performance in object detection and segmentation tasks, particularly under class imbalance, by reducing the influence of easily classified samples and increasing focus on hard-to-classify samples.

The design of ATFL is inspired by Focal Loss, which concentrates the model's attention on hard-to-classify samples by reducing the loss weight of easily classified samples.ATFL adaptively adjusts the loss weights based on the characteristics of each sample and the model's output.It can dynamically adjust the threshold based on the discrepancy between the prediction and the ground truth, causing the model to pay more attention to hard-to-classify samples during training.By introducing a focusing mechanism, ATFL reduces the influence of easily classified samples while enhancing the focus on hard-to-classify samples.This mechanism helps improve the overall performance of the model, particularly in situations with class imbalance.ATFL introduces an adaptive threshold, capable of calculating an appropriate threshold for each sample to determine how to weight the loss.This approach enables the loss function to better reflect the importance of individual samples.Through its optimized loss calculation method, ATFL enables the model to learn valuable features more quickly, thereby improving both the efficiency and effectiveness of training.

In terms of specific implementation, ATFL typically includes the following steps:1. Calculate the predicted probability for each sample.2. Compute the loss based on the predicted probability and the ground truth label, and adjust the loss weight according to the adaptive threshold.3. Update the model parameters through backpropagation.This method is particularly suitable for handling datasets with highly imbalanced classes and can significantly improve model performance on difficult samples.For different tasks, ATFL can be adjusted and optimized according to specific requirements.

This paper uses ATFL to replace the traditional cross-entropy loss in YOLOv11.By introducing ATFL, YOLOv11 can focus more on hard-to-detect samples during the training process, thereby improving overall detection accuracy. Furthermore, the adaptivity of ATFL allows the loss function to automatically adjust according to the data distribution, further enhancing the model's generalization capability across diverse scenarios.

3. Experimental Analysis

3.1. Dataset and Experimental Setup

3.1.1. Data Acquisition and Annotation

The dataset used in this study consists of real pavement images captured by a test vehicle with the camera angle perpendicular to the ground; the original image size is 2528 × 6349 pixels.To facilitate training, the original images were split horizontally and resized to 640× 640 pixels.After data

screening, the original dataset comprises 1060 images. The crack categories are divided into alligator cracking and general cracks, which are named "alligatoring" and "cracks" respectively, as shown in Figure 2.

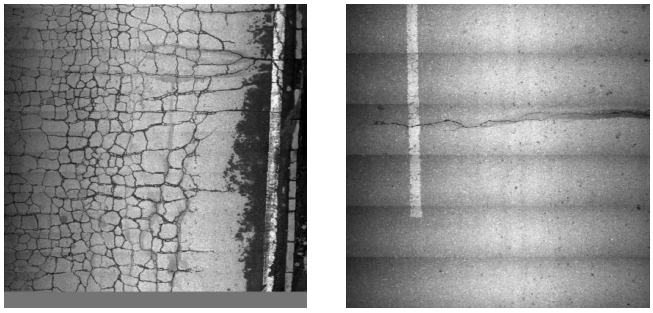


Fig 2. Examples of crack types

The dataset in this paper was manually annotated using LabelImg software. During the process, care was taken to ensure that the annotation boxes closely adhered to the crack edges. For elongated cracks, a segment-by-segment annotation approach was adopted to ensure complete coverage, as shown in Figure 3.

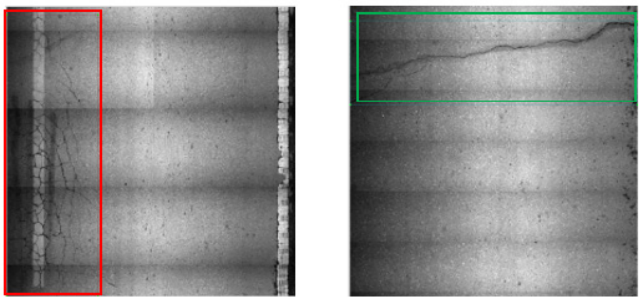


Fig 3. Data annotation

3.1.2. Data Augmentation

Data augmentation diversifies the data distribution, exposing the model to various data forms. This enables the model to learn the essential characteristics of the data more easily rather than features specific to the training data, reducing reliance on local features of the training set and effectively preventing overfitting. Furthermore, augmented data allows the model to perform more consistently when encountering data from different environments, angles, and forms, thereby effectively increasing the model's robustness.

Given the characteristics of the existing dataset, namely its monotone color and single crack type, this paper employs data augmentation methods including horizontal flipping, brightness enhancement, random cropping, and Mosaic.Horizontal flipping involves mirroring the image left to right along its vertical central axis.Brightness enhancement is achieved by adjusting the brightness channel of the image pixels.Random cropping involves randomly selecting a region from the original image data and cropping it to obtain a new image.Mosaic combines parts of four randomly selected images by stitching them together into a single new image.Performing the aforementioned data augmentation increases the complexity of the dataset, which helps enhance the model's robustness and anti-interference capability, thereby improving the model's ability to identify cracks in complex scenes.The forms of data augmentation are shown in Figure 4.

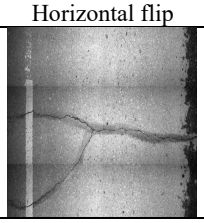
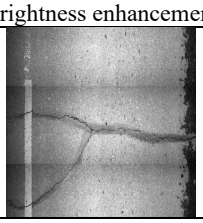
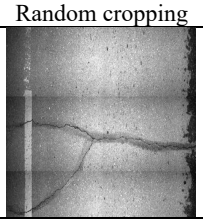
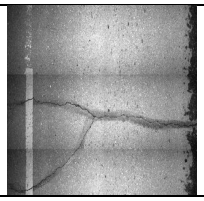
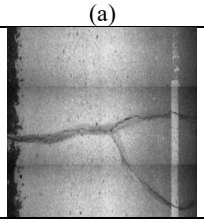
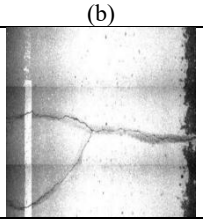
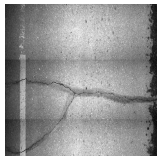
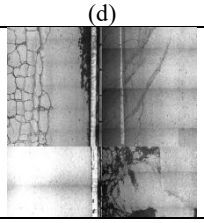
method	Horizontal flip	Brightness enhancement	Random cropping	Mosaic	
original					
	(a)	(b)	(c)	(d)	
result					
	(e)	(f)	(g)	(h)	

Fig 4. Data augmentation

After data augmentation, the total number of images in the dataset is 4646, which are divided into training set, validation set, and test set in an 8:1:1 ratio. The training set is used to train the model; the model learns from the data in the training set and adjusts its parameters to discover patterns and features within the data.The validation set is used to evaluate model performance during the training process, helping to adjust the model's hyper parameters, such as learning rate and batch size.During training, the model is validated on the validation set; metrics such as accuracy and loss on the validation set are used to determine whether the model is over_fitting or under_fitting.After training is complete, the final performance of the model is evaluated using the test set data. The test set is independent of the training and validation sets and can truly reflect the model's performance in practical applications.

3.1.3. Experimental Environment

The experiments in this paper utilize the PyTorch training framework, with Windows 10 as the system environment. The GPU model is an NVIDIA GeForce RTX 4060, the CPU model is an Intel Xeon E5-2586 v4, the video memory size is 16GB, and GPU acceleration is provided by CUDA

12.1.Before model training, the training parameters were set as follows: input image size of 509×640 pixels, batch size of 8, SGD optimizer, training for 300 epochs, and a learning rate of 0.01.

3.2. Evaluation Metrics

This paper adopts Precision, Recall, and mean Average Precision (mAP) as metrics to evaluate model performance, as shown in Equations (1) to (3).Here, Precision represents the number of samples correctly identified as positive among those predicted as positive; Recall represents the number of positive samples correctly classified out of all actual positive samples. Higher values for both indicate better trained model performance. The mAP value is used to evaluate the overall performance of different models; a higher mAP value indicates a higher Intersection over Union (IoU) between the ground truth boxes and predicted boxes during detection, meaning the area under the Precision-Recall curve is closer to 1.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$precision = \frac{TP}{TP + FP} \tag{2}$$

In the formulas, within TP, FP, TN, and FN, P and N respectively indicate prediction as a Positive sample and prediction as a Negative sample; T and F respectively indicate whether the prediction is True or False.

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{3}$$

In the formula, k represents the number of categories; APi represents the area under the PR curve for the i-th category.

3.3. Ablation Experiment

The enhancement algorithm proposed in this paper combines CGA attention, the fusion of Ghost convolution and DynamicConv modules, and the introduction of the ATFL loss function.To verify the effectiveness of this algorithm, corresponding ablation experiments were designed to evaluate the impact of each improved method on model performance under identical experimental conditions.The results of the ablation experiments are shown in Table 1. In Table 1, v11 represents the original YOLOv11 model; C represents the use of the CGA attention mechanism; G represents the fusion of the Ghost convolution module; T represents the introduction of the ATFL loss function; √ and × indicate whether the specific improvement method is used or not, respectively.

Table 1. Ablation Experiment

Model	CGA	Ghost	ATFL	precision	recall	mAPAll
v11	×	×	×	80.0	73.4	81.4
v11+C	√	×	×	79.1	80.9	83.0
v11+G	×	√	×	80.2	80.1	83.7
v11+T	×	×	√	78.3	80.3	83.6
v11+C+G	√	√	×	82.0	80.7	84.2
v11+G+T	×	√	√	77.7	80.6	82.8
v11+C+T	√	×	√	80.2	78.4	83.8
v11+C+G+T	√	√	√	81.7+1.7	80.6+7.2	84.7+3.3

It can be seen from Table 1 that as the algorithm is continuously improved, the detection accuracy correspondingly increases.The CGA attention mechanism, the Ghost+DynamicConv module, and the ATFL loss function each contribute significant performance improvements to the model in different aspects.Particularly when these improvement methods are used in combination, the overall performance of the model is significantly enhanced, with the most prominent improvement observed in the mAPAll metric.After adding the CGA attention mechanism, the model's precision slightly decreased to 79.1%, but the recall rate significantly increased to 80.9%, and mAPAll also improved to 83.0%.This indicates that the CGA attention mechanism has a significant effect on improving the model's recall rate.After adding the Ghost convolution module, the model's precision and recall rate increased to 80.2% and 80.1%, respectively, with mAPAll reaching 83.7%.This demonstrates that the Ghost convolution module can effectively enhance the overall model performance while maintaining precision.After introducing the ATFL loss function, the model's precision slightly decreased to 78.3%, but the recall rate increased to 80.3%, and mAPAll reached 83.6%.This suggests that the ATFL loss function plays a positive role in balancing the model's precision and recall rate.When any two of the improvement methods were used simultaneously, the mAP metric still showed improvement, indicating that combining two methods can further enhance model performance.After applying all improvement methods simultaneously, the model's precision, recall rate, and mAPAll reached 81.7%, 80.6%, and 84.7%, respectively, representing increases of 1.7%, 7.2%, and 3.3% compared to the original model. This indicates that the combination of all improvement methods maximizes the enhancement of model performance.The ablation experiment demonstrates that simultaneously incorporating improvements from CGA, Ghost, and ATFL yields the optimal performance.

3.4. Comparative Experiment

Table 2. Comparative Experiment

Model	Precision	Recall	mAPAll
yolov5	0.76	0.73	0.79
yolov7	0.746	0.756	0.79
yolov8	0.732	0.731	0.75
yolov10	0.769	0.753	0.769
yolov11	0.80	0.734	0.814
Faster-RCNN	0.468	0.881	0.809
SSD	0.815	0.437	0.619
CGT-yolov11	**0.817**	**0.806**	**0.847**

To verify the effectiveness of the algorithm proposed in this study in enhancing detection accuracy, mainstream object detection models including YOLOv5, YOLOv7, YOLOv8,

YOLOv10, YOLOv11, Faster-RCNN, and SSD were selected for comparison with the enhanced algorithm presented in this paper. All models were trained and tested on the self-constructed dataset, ensuring that each model reached convergence. The experimental results are shown in Table 2. From the experimental results, it can be observed that the improved YOLOv11 model achieves higher precision, recall rate, and mAP values compared to the other models, demonstrating a significant advantage in detection accuracy.

3.5. Visualization Results

To verify the effectiveness of the improved algorithm proposed in this paper for asphalt pavement crack detection, a comparison was made between the original images, the detection results of the original YOLOv11 model, and the detection results of the improved YOLOv11 model. The experimental results are shown in the figure.Figure 5(a) shows the original image, Figure 5(b) shows the detection result of the original YOLOv11 model, and Figure 5(c) shows the detection result of the improved YOLOv11 model. "cracks" refers to general cracks, and "alligatoring" refers to alligator cracking.It can be seen from Figure 5 that the original YOLOv11 model exhibited issues of false detection and missed detection. The algorithm proposed in this paper effectively mitigates these issues, significantly improves recognition accuracy, and better meets the requirements of real-time detection tasks.

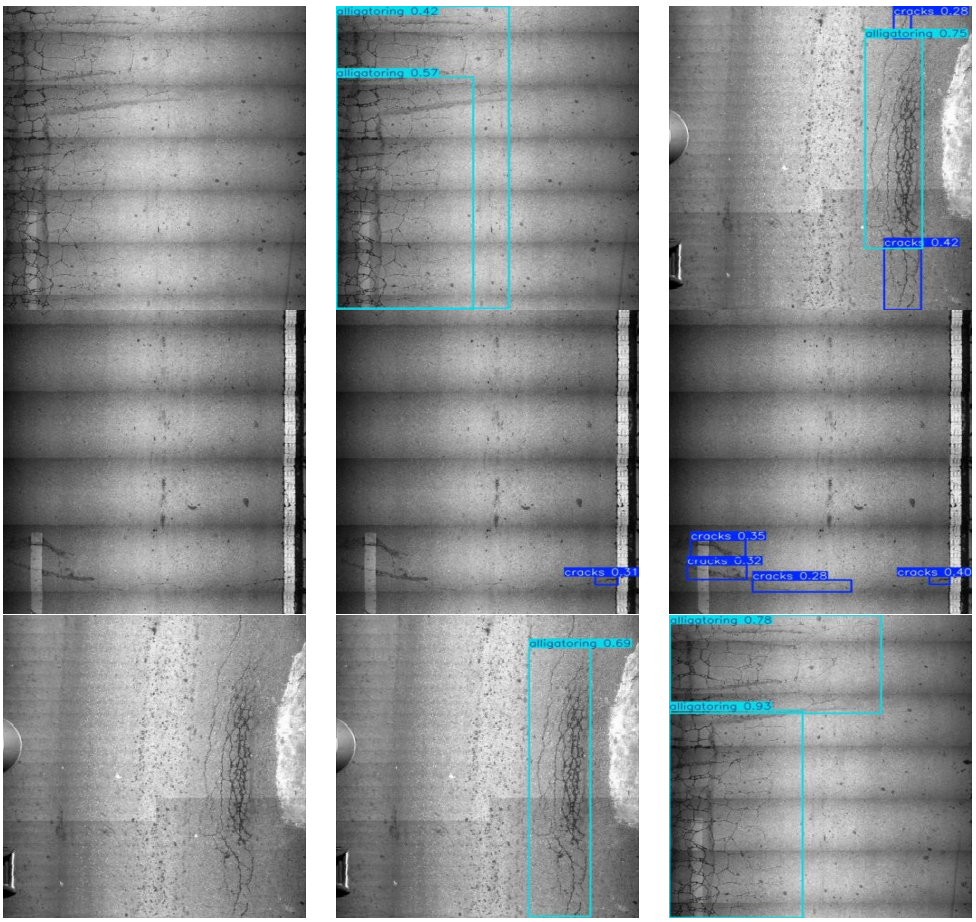


Fig 5. Comparison chart of detection results

4. Conclusion

To meet the accuracy and efficiency requirements of highway pavement crack detection, this paper proposes a pavement crack detection method based on an improved YOLOv11. By integrating Ghost+DynamicConv, combining the CGA attention mechanism with C2PSA, and introducing the ATFL loss function, the detection accuracy and efficiency are significantly enhanced. This method can provide efficient and accurate technical support for pavement maintenance.

References

[1] Liu G, Wu X, Dai F, et al. Crack-MsCGA: a deep learning network with multi-scale attention for pavement crack detection[J]. Sensors, 2025, 25(8): 24-46.

[2] Geng H, Liu Z, Wang Y, et al. SDFC-YOLO: a YOLO-based model with selective dynamic feature compensation for pavement distress detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(2): 1842-1856.

[3] Zhang J, Sun S, Song W, et al. Automated pavement distress detection based on convolutional neural network[J]. IEEE Access, 2024, 12: 105055-105068.

[4] Han K, Wang Y, Tian Q, et al. GhostNet: more features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 1580-1589.

[5] Chen Y, Dai X, Liu M, et al. Dynamic convolution: attention over convolution kernels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 11030-11039.

[6] Zhao Y, Miao J, Li Z, et al. CGA-ViT: channel-guided additive attention for efficient vision recognition[J]. Applied Sciences, 2026, 16(4): 1740.

[7] Yang B, Zhang X, Zhang J, et al. EFLNet: enhancing feature learning network for infrared small target detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-11.

[8] Sunkara R, Luo T. No more strided convolutions or pooling: a new CNN building block for low-resolution images and small objects[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer, 2022: 443-459.

[9] Zong F, Zhao K, Jiang S, et al. Detecting highway pavement diseases by developing an improved YOLOv5 algorithm[C]// International Conference on Artificial Intelligence and Autonomous Transportation. Singapore: Springer, 2024: 142-151.

[10] Li Z, Peng Y, Liu M, et al. Asphalt pavement raveling identification based on machine learning[C]//Fourth International Conference on Image Processing and Intelligent Control. Bellingham: SPIE, 2024, 13250: 214-219.

[11] Huang Z, Chen X, Liu H, et al. Pavement diseases detection using improved YOLOv5[C]//2023 IEEE International Conference on Mechatronics and Automation. Piscataway: IEEE, 2023: 1786-1791.

[12] Liang Z, Di R, Tan F, et al. Fig-YOLO: an improved YOLOv11-based fig detection algorithm for complex environments[J]. Foods, 2025, 14(23): 41-54.