Theory and Practice of Science and Technology
2025, VOL. 6, NO. 6, 30-34
DOI: 10.47297/taposatWSP2633-456906.20250606

Research on Road Pothole Detection System Based on Improved YOLOV11

Jialin Liu

School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou, Henan, 450046, China

ABSTRACT
Pothole detection on road surfaces is a critical component for ensuring road traffic safety. Aiming at the insufficient detection accuracy of the traditional YOLOv11 model in complex environments, this study introduces the CBAM (Convolutional Block Attention Module) and DFF (Dynamic Feature Fusion) module respectively for comparative experiments. Specifically, the CBAM module enhances the model's channel and spatial dimension attention mechanism for pothole features, significantly improving the recognition ability of small-scale potholes. The DFF module optimizes the dynamic fusion strategy of multi-scale feature maps, effectively enhancing the detection robustness of targets in complex backgrounds. Experimental results show that after adding the CBAM module, the model's precision increases by 0.2% and recall rate improves by 4.6%. When introducing the DFF module, the precision remains unchanged while the recall rate increases by 1.6%, both outperforming the detection performance of the original YOLOv11 model. This research provides two feasible lightweight improvement schemes for road disease detection, offering a reference for the optimization of automated detection technologies in intelligent transportation systems.

KEYWORDS
Road pothole detection; YOLOv11 model; Object detection; Deep learning

1 Introduction

With the modernization of society, transportation networks continue to expand and extend, making preventive maintenance of roads[1] particularly important. As a vital infrastructure for transportation, the condition of roads directly affects traffic safety and efficiency. Road potholes, as a typical type of road disease, not only accelerate vehicle mechanical wear but also may trigger serious traffic accidents. In the past, the detection of these potholes mainly relied on manual inspection and other methods. However, these traditional approaches are extremely inefficient and susceptible to various factors such as environmental lighting changes and background interference.

In recent years, with the continuous development of deep learning technologies, object detection algorithms represented by YOLO and TOLAKT[2] have demonstrated detection efficiency far exceeding traditional methods in road disease identification scenarios, thanks to their end-to-end automated detection capabilities and multi-scale feature processing advantages. Nevertheless, there remains significant room for improvement in road pothole detection—challenges such as small-scale target features being easily submerged in complex road textures and unstable target representation under dynamic lighting lead to insufficient environmental adaptability of existing models in practical engineering applications.

Therefore, this paper improves the latest YOLOv11 architecture by introducing a lightweight attention mechanism and adaptive feature fusion strategy, aiming to enhance the robustness and accuracy of pothole detection and provide a more engineering-applicable technical solution for intelligent road maintenance.

2 YOLOV11 algorithm

2.1 Introduction to YOLOV11 algorithm

As the latest iteration of the YOLO series of models, YOLOv11 employs a Transformer-based backbone in its architecture, which enables it to better capture long-range dependencies and significantly enhances the ability to detect small objects. The dynamic head design can adaptively adjust according to the complexity of images, optimize resource allocation, and achieve a faster and more efficient processing process. The neck network adopts the PAN architecture and integrates the C3K2 unit, which helps to integrate features from multiple scales and improve feature transmission efficiency.

2.2 Dynamic feature fusion module

In response to the static weight issue of traditional feature pyramids during multi-scale feature fusion, the DFF module[3] introduces an adaptive dynamic fusion strategy based on feature characteristics. By constructing a cross-level feature interaction mechanism, it enhances the model's ability to represent potholes of varying scales. The DFF module

dynamically assigns weights to achieve an optimal combination of features by calculating the contribution of features from different hierarchical levels to the object detection task. The dynamic feature fusion process can be formalized as the following equation:

$$F_f = \sum_{i=1}^{n} \omega_i \cdot F_i \tag{1}$$

Among them, F_i represents the feature map of the i-th layer, and ω_i is the dynamic weight corresponding to the layer, satisfying $\sum_{i=1}^{n} \omega_i = 1$. Through this mechanism, the model can adaptively adjust the contribution of features from different layers according to the specific characteristics of the input image. Specifically, for small - scale pothole features, the DFF module enhances the weights of low - level features to improve the ability to perceive details; for large - scale potholes affected by background interference, it strengthens the weights of high - level semantic features, effectively balancing the detection performance for multi - scale targets. In addition, the DFF adaptively fuses local feature maps of different scales based on global information, enabling the network to efficiently combine multi - scale information with a larger receptive field. Through dynamic fusion, this module can not only better preserve the details of local features, but also enhance the effective utilization of global information. This method is mainly applicable to segmentation networks and can also be used in object detection. The following is a schematic diagram of the DFF network structure:

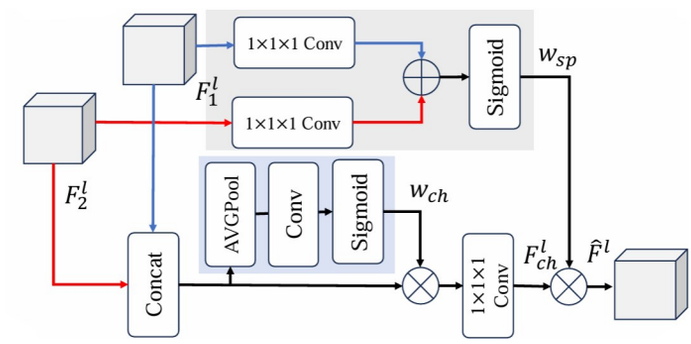

Figure 1 Schematic Diagram of Dynamic Feature Fusion (DFF) Module

In the DFF module, there are two input features F_1^l and F_2^l. On the one hand, they are respectively processed by a $1 \times 1 \times 1$ convolution and then added together. After that, through the Sigmoid activation function, the spatial weight W_{sp} is generated. On the other hand, these two input features are first concatenated, and then successively go through global average pooling, convolution, and Sigmoid operations to obtain the channel weight W_{ch}. Subsequently, the fused feature F_{ch}^l adjusts the channel information through a $1 \times 1 \times 1$, and then performs an element - wise multiplication operation with the feature weighted by spatial attention, ultimately producing the fused feature $\hat{F}^l$. By means of this operational mechanism, the DFF module can both strengthen the ability to retain the details of local features and integrate global information, thereby improving the segmentation accuracy and robustness of the model.

2.3 Convolutional block attention module

The essence of CBAM[4] lies in enhancing the network's representational ability by focusing on key features and suppressing redundant ones. In practical operation, this module follows a specific procedure: first, it implements the channel attention mechanism, whose purpose is to accurately screen out those "crucial" features, thereby discriminating and strengthening features from the channel dimension; subsequently, it introduces the spatial attention mechanism, which focuses on the "key positions" of the above - mentioned key features in the image, further locating and highlighting valid information from the spatial dimension. Through this sequential application of channel - first and then spatial attention, CBAM can effectively guide the network to highly concentrate its attention on the key information within the image. This not only significantly enhances the expressive strength of features in both channel and spatial dimensions, but also enables the network to more efficiently extract and utilize valuable information when processing images, thus optimizing its overall performance. The following figure shows the principle structure diagram of CBAM:

In the experiment of detecting road potholes using YOLOv11, considering the complex and diverse shapes of road

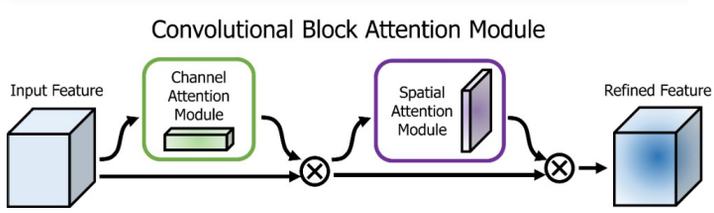

Figure 2 The overview of CBAM

potholes and the potential class imbalance in the data, we introduced the CBAM module to improve the model's detection accuracy. Specifically, the CBAM module was embedded between the backbone network and the neck network of YOLOv11 to function during the feature extraction stage. First, the module uses the channel attention mechanism. Based on global average pooling and global max pooling operations, it obtains feature information in the channel dimension. Through calculations by a multi - layer perceptron and an activation function, it generates channel attention weights, screening out the feature channels that are more critical for pothole detection, suppressing redundant channel information, and highlighting key features related to potholes. Subsequently, taking the feature map processed by the channel attention as input, the spatial attention mechanism aggregates information in the spatial dimension using average pooling and max pooling. Through a convolution operation and an activation function, it obtains spatial attention weights, accurately locating the spatial position of potholes in the image and enhancing the feature representation of the pothole area. In this way, the CBAM module enables the model to focus more on the key features and positions of road potholes, reducing the influence of factors such as background interference, and effectively improving the accuracy of the model in detecting road potholes.

3 Experiments and results

3.1 Experimental setup and training environment

The training dataset used in this experiment is a public road pothole dataset, and the research is carried out in the YOLOV-11 format. The experimental platform currently consists of two main parts. The hardware platform includes an i7-14700HX CPU and an NVIDIA GeForce RTX4060 GPU. The software platform configurations are specifically: Cuda12.4, Python3.8, PyTorch2.4.0, and Opencv4.11.0.86.

3.2 Experimental results and analysis

To verify the effectiveness of the improved model, the dataset covers crack images in different environments, including various lighting, weather and road conditions. The dataset is divided into training set, validation set and test set, with the ratio of 70%, 15% and 15%. To evaluate the results of integrating the lightweight attention mechanism and adaptive feature fusion strategy, two metrics, Precision and Recall, are used for assessment. During the model training process, data augmentation strategies were used to achieve better experimental results. The calculation formulas are as follows, and Table 1 shows the meanings of the symbols in each formula.

$$\mathrm{Re}call = \frac{TP}{TP + FN} \times 100\% \tag{2}$$

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \times 100\% \tag{3}$$

Table 1 Meanings of Formula Symbols

Formula symbols	meaning
Recall	The proportion of samples correctly predicted as positive by the model among all actual positive samples
Precision	The proportion of samples correctly predicted as positive by the model among all samples predicted as positive
TP	The number of positive samples correctly predicted by the model
FP	The number of negative samples incorrectly predicted as positive by the model
FN	The number of positive samples incorrectly predicted as negative by the model

The maximum number of training epochs set in the experiment was 300, and an early stopping strategy was added. When the indicators did not improve for 50 consecutive epochs, the training was stopped to prevent overfitting.Figures 3-5 show the confusion matrix results of the original model, the model injected with the CBAM module, and the model added with the DFF module after training. From the results, the improved model proposed in this paper is obviously superior to the original YOLOv11 model in correctly identifying targets, and the number of false positives and false negatives is significantly reduced.

The recall rate of the YOLOV11 training results is 62.4% with a precision of 69.6%. By adding the lightweight attention mechanism and adaptive feature fusion strategy, the recall rates reach 67% and 64% respectively, while the precisions achieve 69.9% and 69.6%. Although the improvement in precision is minimal, the introduction of the lightweight attention mechanism and adaptive feature fusion strategy has effectively enhanced the model's detection capability for road potholes. In particular, the former demonstrates a more significant improvement in recall rate, indicating its obvious advantage in reducing target missed detections.

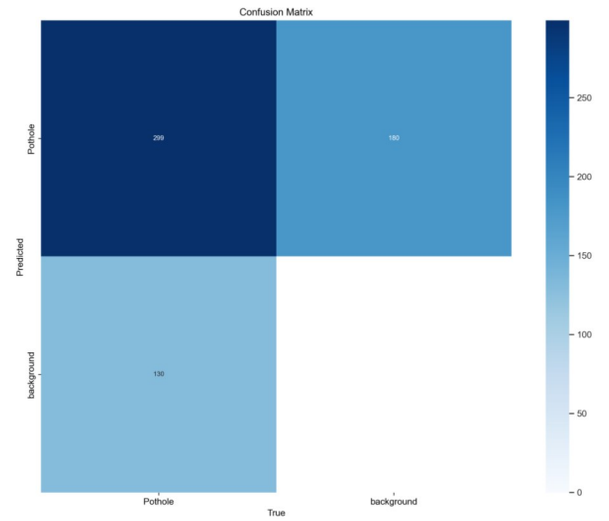

Figure 3 YOLOV11 Module

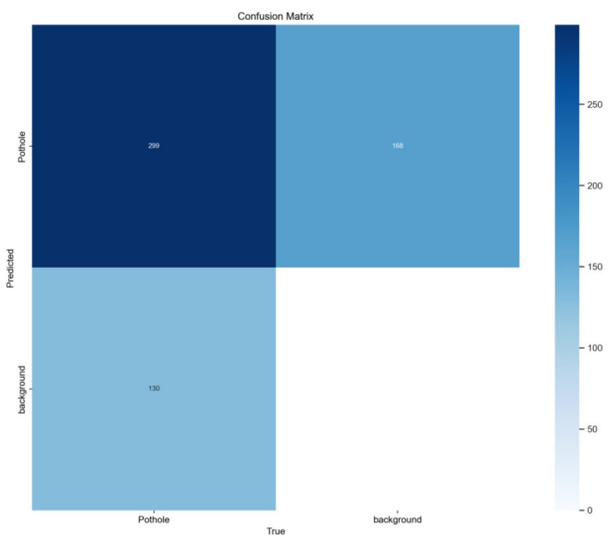

Figure 4 YOLOV11+DFF Module

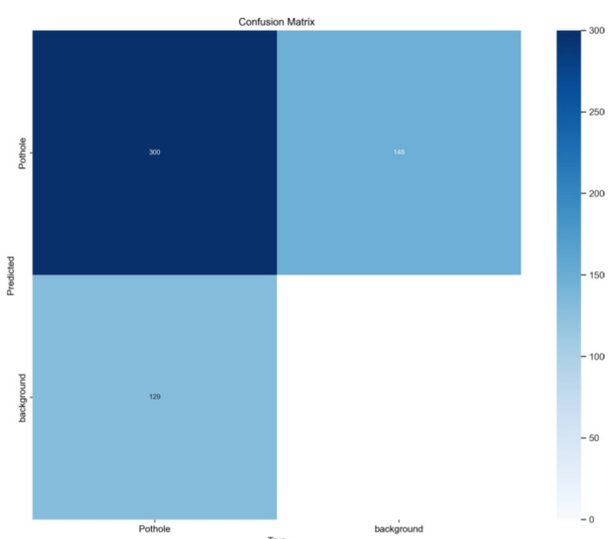

Figure 5 YOLOV11+CBAM Module

4 Conclusion

This study proposes an improved YOLOv11 model for road pothole detection by integrating the CBAM and DFF modules, addressing the challenges of low detection accuracy in complex environments. The CBAM module enhances

channel and spatial attention to optimize feature representation, particularly for small-scale potholes, while the DFF module improves multi-scale feature fusion robustness against background interference. Experimental results demonstrate that the CBAM model achieves a 0.2% precision increase and 4.6% recall improvement, and the DFF model maintains precision while boosting recall by 1.6%, outperforming the original YOLOv11. These lightweight improvement strategies not only enhance the adaptability of road pothole detection in practical engineering but also provide a feasible technical reference for intelligent transportation systems. Future work will focus on further optimizing model efficiency, expanding dataset diversity, and exploring real-time detection applications in complex scenarios to promote the practical implementation of automated road maintenance technologies.

Acknowledgements : In this research, I would like to particularly thank my supervisor for the valuable suggestions provided in the design of the research framework and the revision of the thesis. These suggestions were like a lighthouse, illuminating my research path. Meanwhile, I would like to extend sincere gratitude to Yundi Han from North China University of Water Resources and Electric Power for providing equipment support for the experiment. Looking back on the entire research process, every experiment and improvement has been a sublimation of myself, pushing me towards the destination of success. Finally, I would like to thank my parents for their constant support and encouragement along the way.

About the Author

* Corresponding Author: Jialin Liu, linlinaididi@outlook.com.

References

[1] Li Wei. Application of Preventive Highway Maintenance Technology in Modern Highway Maintenance[J]. Jushe, 2020, (5): 49.

[2] Yuan Wenhao, Yin Junyu, Fang Lina, et al. Road Crack Extraction Based on Crack-YOLACT [J/OL]. Journal of Nanjing University of Information Science and Technology, 1-14 [2025-05-29].

[3] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[4] Yang J, Qiu P, Zhang Y, et al. D-net: Dynamic large kernel with dynamic feature fusion for volumetric medical image segmentation[J]. arXiv preprint arXiv:2403.10674, 2024.