Pavement Crack Extraction Based on Multi-scale Convolutional Neural Network

ZHAN Biheng[1,2], *SONG Xiangyu*[1,2,3*], *CHENG Jianrui*[4], *QIAO Pan*[5,6], *WANG Tengfei*[7]

1. Key Laboratory of Road and Railway Engineering Safety and Security of the Ministry of Education (Shijiazhuang Tiedao University), Shijiazhuang 050043, P. R. China; 2. School of Civil Engineering, Shijiazhuang Tiedao University, Shijiazhaung 050043, P. R. China; 3. Key Laboratory of Polar Environment Monitoring and Public Governance of the Ministry of Education (Wuhan University), Wuhan 430079, P.R. China; 4. Hebei Provincial Institute of Cartography, Shijiazhuang 050031, P. R. China; 5. Hebei Traffic Planning Institute Ruizhi Transportation Technology Consulting Co., Ltd., Shijiazhuang 050090, P. R. China; 6. Hebei Innovation Center for Intelligent Monitoring and Operation Maintenance Technology of Road Engineering, Baoding 071799, P. R. China; 7. China Railway Tunnel Bureau Group Road and Bridge Engineering Co., Ltd., Tianjin 300450, P. R. China

(Received 1 June 2025; revised 9 September 2025; accepted 11 October 2025)

Abstract: Cracks represent a significant hazard to pavement integrity, making their efficient and automated extraction essential for effective road health monitoring and maintenance. In response to this challenge, we propose a crack automatic extraction network model that integrates multi-scale image features, thereby enhancing the model's capability to capture crack characteristics and adaptation to complex scenarios. This model is based on the ResUNet architecture, makes modification to the convolutional layer of the model, proposes to construct multiple branches utilizing different convolution kernel sizes, and adds a atrous spatial pyramid pooling module within the intermediate layers. In this paper, comparative experiments on the performance of the basic model, ablation experiments, comparative experiments before and after data augmentation, and generalization verification experiments are conducted. Comparative experimental results indicate that the improved model exhibits superior detail processing capability at crack edges. The overall performance of the model, as measured by the F1-score, reaches 71.03%, reflecting a 2.1% improvement over the conventional ResUNet.

Key words: road engineering; neural networks; multi-scale convolution; pavement cracks

CLC number: TP391.4 **Document code:** A **Article ID:** 1005-1120(2025)06-0749-18

0 Introduction

Pavement cracks, recognized as one of the most prevalent hazards on roads[1-2], contribute to the structural degradation of pavements, reduce the lifespan of highways, and pose significant safety risks, thereby impacting traffic safety. Consequently, the detection and repair of pavement cracks are critical yet challenging tasks in highway maintenance. Traditional crack detection methods predominantly rely on manual visual inspections, which not only require considerable human and material resources but also exhibit a high degree of subjectivity, complicating the fulfillment of large-scale, high-precision monitoring demands. In recent years, with the ongoing advancements in computer hardware and related technologies, the process of pavement crack extraction has progressively transitioned from traditional manual visual interpretation to computer-aided automatic recognition[3-5]. Currently, automatic recognition methods can be broadly

*Corresponding author, E-mail address: songxiangyu@stdu.edu.cn.

How to cite this article: ZHAN Biheng, SONG Xiangyu, CHENG Jianrui, et al. Pavement crack extraction based on multi-scale convolutional neural network[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2025, 42(6): 749-766.

http://dx.doi.org/10.16356/j.1005-1120.2025.06.004

categorized into two main types: Traditional image processing-based methods and neural network-based extraction techniques.

Among traditional image processing methods that emerged relatively early, threshold segmentation[6] is one of the earliest and most widely adopted methods for target segmentation in image processing and pattern recognition. The fundamental concept involves partitioning an image into target and background regions by selecting an appropriate grayscale threshold. Otsu's method[7], one of the earliest global thresholding techniques, determines the optimal segmentation threshold by maximizing the between-class variance. Since its proposal, numerous scholars have developed improved methods based on this classical approach[8-10]. A growing number of image processing methods based on the principle of global threshold segmentation have been proposed[11-13]. However, with the increasing complexity of crack detection scenarios, the effectiveness of global threshold methods significantly diminishes in crack images characterized by uneven illumination or intricate backgrounds. This limitation has shifted research focus towards local threshold methods[14-16], which operate on the principle of dividing the image into multiple local regions and calculating an independent threshold for each region.

Another major category of traditional image processing methods is edge detection[17], whose core lies in detecting grayscale discontinuities. This technique is used to identify regions in an image where significant grayscale changes occur, such as object contours, texture boundaries, or junctions between different regions in the image. The Canny algorithm[18], initially applied to asphalt pavement images, serves as the foundation for numerous subsequent improved methods[19-20]. Nowadays, more commonly used edge detection operators have been proposed[21-25]. In addition to the aforementioned primary image processing techniques, various image processing-based methods[26-29] have been employed for road crack detection, offering multiple avenues for future research.

Traditional image recognition algorithms have the advantage of not requiring large amounts of image data for training; however, their drawbacks are significant: They require manual assistance, exhibit poor robustness, and are susceptible to various forms of noise, such as variations in lighting and debris. With the continuous advancements in computer hardware and software in recent years, more intelligent detection algorithms have emerged, particularly those based on neural networks for image classification, which have gradually gained prominence. Neural network models can extract effective data features more accurately through extensive sample learning. Zhang et al.[30] were pioneers in employing a deep learning-based crack extraction method utilizing convolutional neural networks (CNN) in the realm of road crack detection, marking a significant development for subsequent neural network models. Since its proposal, CNN has quickly attracted widespread attention. Based on its unique network structure, scholars in the academic community have continuously pursued innovations, and propose new technical methods to integrate with CNN, thereby promoting the in-depth development of related research fields[31-34]. In recent years, although classical network models such as CNN have still received attention and further development from scholars in various fields, and have achieved considerable progress in the direction of crack extraction[35-39], the rise of Transformer technology has led researchers to explore detection frameworks that combine Transformer with CNN. This integration aims to fully leverage the local feature extraction capability of CNN and the global modeling advantage of Transformer. Such hybrid detection methods have become a research focus in the field and also provided new solutions for pavement crack detection[40-42]. Although neural network-based image extraction algorithms demonstrate higher accuracy compared with traditional image recognition methods and can be applied to various complex scenarios, numerous challenges still remain. Existing models predominantly emphasize local feature extraction while neglecting the correlation between multi-scale features and global structural information[43-44]. Furthermore, these models demonstrate limited adaptability to complex environments and exhibit insufficient generalization capabilities in scenarios characterized by noise interference, weak signals, or small targets[30, 45-46]. Consequently, there is an urgent necessity to develop an intelli-

gent recognition framework that effectively balances efficiency, robustness, and high precision to address the recognition challenges existing in complex underground environments.

Among various convolutional neural network models, U-Net[47] has been widely applied to diverse image segmentation tasks due to its superior performance and straightforward design. Its success in the medical imaging field[48] has inspired numerous researchers to enhance this architecture. ResUNet, a network model developed from UNet, was first utilized by Zhang et al.[49] in 2018 for road extraction from remote sensing images and has since been applied to other domains[50-51]. ResUNet effectively integrates the advantages of ResNet[52] (Residual Network) and UNet by incorporating residual connections from ResNet into UNet. This methodology not only mitigates the gradient vanishing issue encountered during deep UNet training but also preserves UNet's characteristic "U"-shaped structure and efficient encoder-decoder design, thereby significantly improving the network's ability to learn deep image features. Therefore, considering the inherent advantages of ResUNet and its flexible basic architecture, this study adopts this network model as the base model. Meanwhile, recognizing that the atrous spatial pyramid pooling (ASPP) module exhibits good compatibility with ResUNet, an attempt is made to integrate this module into the original architecture.

ASPP was first introduced in DeepLabv2[53] and later refined in DeepLabv3[54]. Its core design uses parallel atrous convolutions with varying dilation rates to process input feature maps, and expands the receptive field without sacrificing the feature map resolution, thereby effectively capturing multi-scale image information and boosting segmentation accuracy. Specifically, DeepLabv2 added ASPP based on DeepLabv1[55] to improve recognition of objects of different sizes; DeepLabv3 optimized it by integrating global average pooling into ASPP for stronger global context capture; and DeepLabv3+[56], built on DeepLabv3, incorporated an encoder-decoder structure to handle local details, balanced global semantics and local information, and achieved an optimal performance-efficiency trade-off. However, unlike existing studies that simply integrate the ASPP module directly into the bottleneck layer or decoder of ResUNet, the innovations of this study are as follows: (1) A multi-scale convolution branch structure is redesigned in the encoder. Specifically, three types of convolution kernels (3×3, 5×5, and 7×7) are used for parallel operations, followed by feature fusion. This design fully extracts crack texture information at different scales and enables the interaction of multi-scale features from the feature extraction stage onward; (2) on this basis, a modified ASPP module is introduced in the intermediate layer, allowing the fused features output by multi-scale convolution to further undergo collaborative learning with global contextual features. This realizes a hierarchical feature representation that progresses from "local texture enhancement" to "global semantic aggregation". This collaborative structure not only addresses the deficiency of traditional ResUNet in capturing multi-scale features but also overcomes the limitations of existing ResUNet-ASPP combination methods in fine-grained crack recognition. Ultimately, the proposed collaborative fusion strategy of multi-scale convolution and ASPP effectively enhances the accuracy and robustness of the model for crack extraction under complex backgrounds.

In summary, this paper adopts the ResUNet network architecture as its foundation, preserving the residual connections and UNet structure of the original model. First, the convolutional layers in the encoder are modified to perform three parallel convolutions with different kernel sizes simultaneously, followed by summation to fully extract multi-scale image features of roads and provide richer feature representations for the model. Second, an ASPP module is incorporated in the middle layer to capture image information at multiple scales through atrous convolutions with varying dilation rates, thereby further enlarging the receptive field. By integrating local detail features with global semantics, the model ultimately achieves improved accuracy in crack extraction.

1 Method

1.1 ResUNet model architecture

ResUNet inherits the "U"-shaped architecture of UNet, comprising a symmetric feature extraction path and an expansion path. The feature extraction

path (encoder) extracts features from the input image while progressively reducing the size of the feature maps. The expansion path (decoder) gradually restores the spatial resolution of the image and integrates the features extracted by the encoder to perform pixel-wise predictions. At each stage of the encoder, feature maps are directly conveyed to the corresponding layers in the decoder via skip connections, ensuring that high-resolution features are preserved during the encoding process. Additionally, ResUNet incorporates residual connections from ResNet. The input data undergoes a 1×1 convolution, bypassing intermediate convolutional operations, and is then directly added to the output of the convolutional layers. The resulting output forms the network structure, as illustrated in Fig.1. In the encoder structure, the size of feature maps gradually decreases as the number of convolutional operations increases, while the number of channels gradually increases. Specifically, the dimensions are 320×640×16, 160×320×32, 80×160×64, 40×80×128, and 20×40×256 in sequence. Subsequently, the feature maps are gradually restored to their original sizes in the decoder structure.

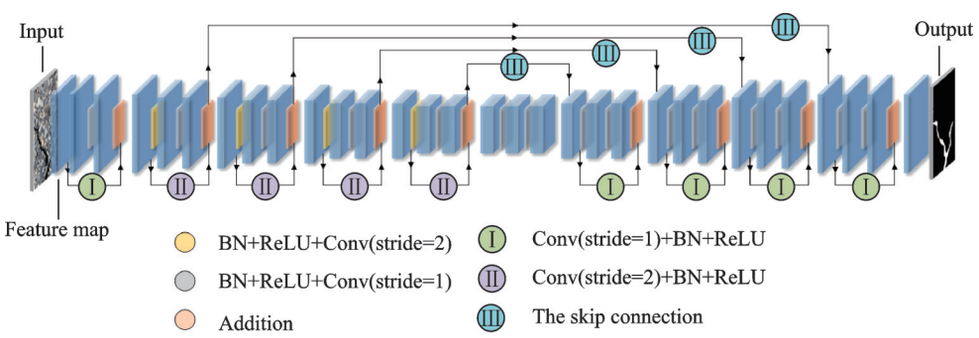


Fig.1 Structure of the ResUNet model

In the convolutional layers of the encoder structure within the ResUNet network model, three primary operations are executed: (1) The input image undergoes batch normalization (BN), followed by activation through the ReLU function, and is then convolved with a 3×3 kernel. The first convolution employs a stride of 2, while the second convolution utilizes a stride of 1. (2) The input image is convolved once with a 1×1 kernel using a stride of 2 to establish the residual connection. (3) The results from the first two steps are summed to produce the output of this layer. Two convolutional sequences with different orders are adopted in the convolutional layers: The skip connections use the sequence of "convolution+normalization+activation function", while the network backbone uses the sequence of "normalization+activation function+convolution". The main reason is that the former is used for feature dimension transformation and alignment, and the latter for improving gradient flow and training stability. This differentiated design not only retains the optimization advantages of deep networks but also ensures the intuitiveness of the feature fusion module and its consistency with classical structures. The entire convolutional layer can be represented by

$$x_{n+1} = h_2(x_n) + g(f(x_n)) \tag{1}$$

where x_n and x_{n+1} denote the input and output of the convolutional layer, respectively; f represents a convolution operation with a 3×3 kernel and stride 2, g a convolution with a 3×3 kernel and stride 1, and h_2 a convolution with a 1×1 kernel and stride 2.

1.2 Multi-scale convolution

Considering that the traditional ResUNet model structure cannot fully satisfy the extraction of multi-scale feature information from images, this study aims to enhance both the robustness of the network model and its ability to capture complex feature information in images by integrating multi-scale convolution into the ResUNet framework. The convolutional layers of ResUNet are improved by transforming them into multiple branches with varying receptive fields while maintaining the original residual connection structure. This feature fusion results in richer feature representations, thereby enhancing the extraction performance of crack features.

Since the objective of multi-scale convolution is to strengthen the network's capability to extract crack features, the enhanced convolutional layer structure is applied solely to the encoder portion of the ResUNet model and is excluded from the decoder structure. Additionally, since the size of the output result is related to the stride of the convolutional kernel, the larger the stride, the smaller the size. To ensure that the image size remains unchanged and can be directly summed after multi-scale convolution, all convolutional kernels are configured with a stride of 1. Following the summation of the parallel branches, a max pooling operation with a pool size of 2 (MaxPooling2D) is performed to ensure that the resulting feature map size aligns with that of the residual connection feature map, which utilizes a 1×1 convolutional kernel with a stride of 2. This facilitates subsequent addition, and the weights of all parties involved in the summation during the convolution process are equal. The specific operation is illustrated in Fig.2.

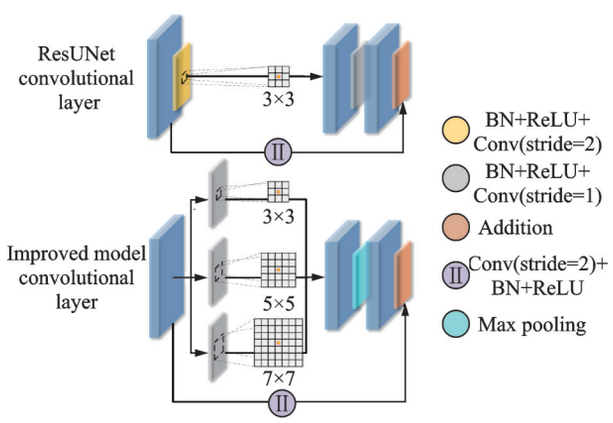


Fig.2 Comparison of convolutional layer structures

This improved model optimizes the convolution stride parameter by changing the original stride from 2 to 1, compensating for the resulting change in feature map size through max pooling. This enhancement is predicated on the observation that reducing the stride to 1 augments the convolutional layer's capacity to extract local features and filter essential characteristics, with max pooling employed to ensure feature map size compatibility. In comparison to the stride-2 configuration, this approach markedly enhances detail preservation and model accuracy. The reason for selecting convolutional kernels with different receptive fields (3×3, 5×5, and 7×7) is as follows: 3×3 convolutions can preserve local texture and detailed features, which is conducive to the accurate characterization of crack edges; 5×5 convolutions expand the receptive field and can capture the contextual relationships of adjacent regions; 7×7 convolutions provide a larger receptive field, enhancing the model's ability to perceive wider or complex crack regions. Through the fusion of these three scales, the model can not only capture the boundary information of small cracks but also acquire large-scale structural features. Additionally, the numerical continuity among the three kernel sizes prevents excessive dimensional jumps between features. The improved convolutional layer can be represented by

$$x_{n+1}=h_1(x_n)+i(g_3(x_n)+g_5(x_n)+g_7(x_n)) \quad (2)$$

where g_3, g_5 and g_7 denote the convolution operations with kernel sizes of 3, 5, and 7, respectively, all with a stride of 1; h_1 denotes a convolution operation with a 1×1 kernel and stride of 1 and i the max pooling operation.

1.3 ASPP module

To further enhance the capability of capturing multi-scale feature information from images and to strengthen the connection between global context and local feature details, this study integrates the ASPP module into the intermediate layer of the ResUNet model, as illustrated in Fig.1. The first convolution operation in this intermediate layer is replaced with the ASPP module. Embedding the ASPP module in the intermediate layer is a pivotal design choice that balances feature representation capacity with computational efficiency. This is due to the fact that the feature maps at the network's intermediate layer already contain rich semantic information while maintaining relatively low resolution. The ASPP module utilizes atrous convolutions with varying dilation rates to establish a feature pyramid structure without further diminishing resolution, thereby enlarging the receptive field and effectively capturing multi-scale contextual information. As for why the ASPP module is not inserted into the encoder or decoder, the main considerations are as follows: (1)

When the ASPP module is placed in the early stage of the encoder, the network is still in the shallow feature extraction phase. The obtained features mainly contain low-level texture and edge information, and multi-scale features have not yet been fully formed. Although introducing atrous convolution at this point can spatially expand the receptive field, the weak semantic expression ability of the input features prevents the ASPP module from fully exerting its multi-scale sampling effect. It may even introduce redundant spatial information, which impairs the effectiveness of subsequent feature extraction. (2) When the ASPP module is located in the decoder phase, the input features are mainly derived from the shallow network. While this helps restore image details, the semantic information at this stage is relatively insufficient. The contextual information extracted by the ASPP module's atrous convolution in this low-semantic feature space is limited, resulting in a weakened effect of global receptive field expansion. In addition, the multi-scale feature fusion of the ASPP module may overlap with the upsampling operation in the decoder phase, which affects the efficiency of feature fusion and the overall accuracy of the model. Atrous convolutions with dilation rates of 6, 12, and 18 are employed to form a progressively expanding receptive field from small to large: A dilation rate of 6 tends to capture medium-range contextual information, while 12 and 18 provide broader global structural perception. This design avoids both insufficient context due to excessively small receptive fields and loss of details caused by overly large ones, thereby striking a balance between multi-scale context modeling and preservation of crack details.

The specific operations within the ASPP module are as follows: (1) A convolution with a kernel size of 1 and a stride of 1 is employed to reduce feature dimensionality and computational cost. (2) Three parallel convolutions with a kernel size of 3, a stride of 1, and dilation rates of 6, 12, and 18, respectively, are executed to capture feature information at different scales. (3) Global average pooling (GlobalAveragePooling2D, an operation that takes the average value in the spatial dimension) is applied to the input feature map X (assumed to be $H\times W\times C$) to extract global contextual features, resulting in a feature vector ($1\times1\times C$). This feature vector is subsequently processed through a 1×1 convolution to adjust the number of channels (with a stride of 1) for dimensionality reduction, before being upsampled to the original spatial size ($H\times W$) using bilinear interpolation, which ensures spatial dimension alignment with other atrous convolution branches. The core function of this module is to compress spatial information into a global semantic representation, which makes the model lighter, more stable, and less prone to overfitting. (4) The outputs from all branches are concatenated and integrated into a unified feature representation via a 1×1 convolution. The specific operation is illustrated in Fig.3.

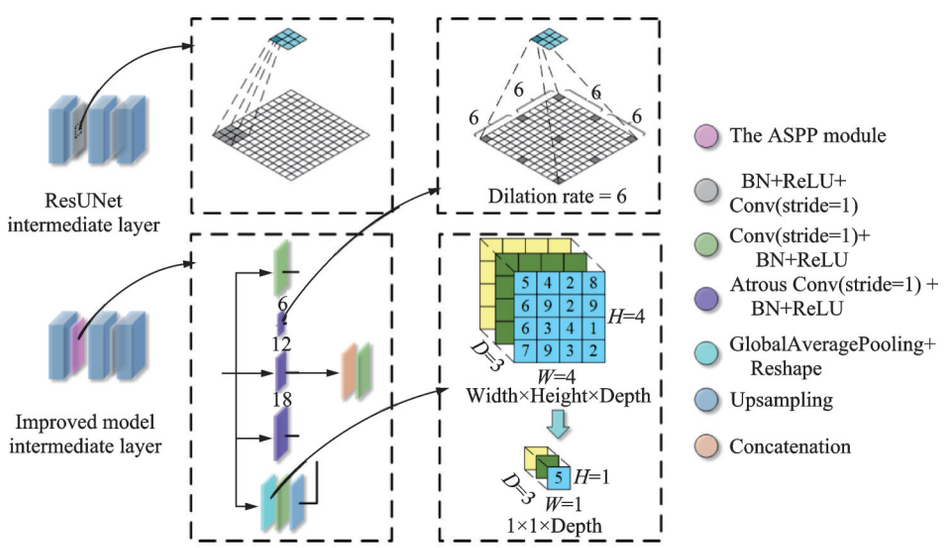


Fig.3 Comparison of intermediate layer structures

2 Experimental Setup and Evaluation Metric

2.1 Experimental environment and parameter settings

All experiments in this study were conducted within the TensorFlow 2.10.1 and Python 3.10.14 environment, utilizing an NVIDIA GeForce RTX 4060 Ti GPU with CUDA 11.2 as the computing platform. In the model accuracy comparison experiments, to highlight the contrast among different models, all settings were kept consistent except for differences in network architecture. Specific consistent settings include: The initial learning rate was set to 0.001 for all models, which adaptively changed based on the validation set loss value with a minimum of 10^{-6}; the activation function was ReLU for all; the optimizer was Adam for all; and the loss function used was Dice loss for all models. The specific differences in parameters between models are detailed in Table 1. The comparison models include UNet, ResUNet, PSPNet (Pyramid scene parsing network)[57], DeepLabv3+, ALP-UNet[58], FlexiCrackNet[59] and the improved model proposed in this study. Based on the training loss trends of each model, as well as a comprehensive consideration of training cost and computational resources, the number of training epochs was set to 800.

Table 1 Comparison of various models

Training model	Convolution	Residual connection	ASPP	Feature description
UNet	Single-size convolutional kernel	None	None	Basic UNet structure
ResUNet	Single-size convolutional kernel	Yes	None	Introduce residual units to improve training stability
DeepLabv3+	Single-size convolutional kernel	None	Yes	ASPP extracts global contextual information
PSPNet	Multi-size convolutional kernels; multi-scale concatenation	None	Pyramid pooling	Strong global context modeling capability
ALP-UNet	Multi-size convolutional kernel	Yes	Laplacian pyramid	Fuse multi-scale features and improve fine-crack detection with minimal model complexity
FlexiCrackNet	Single-size convolutional kernel; multi-scale addition	Yes	None	Flexible feature pipeline with SAM-transferred features
Improved model	Multi-size convolutional kernels; multi-scale addition	Yes	Yes	Enhance multi-scale feature extraction capability

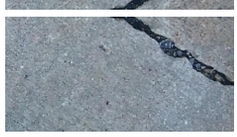
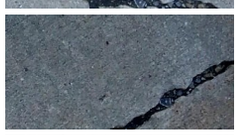
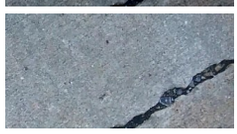
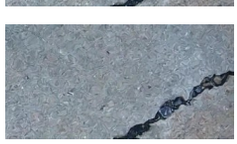
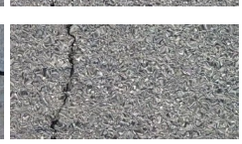
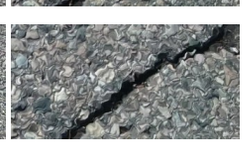
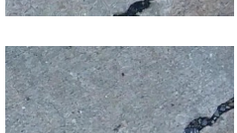
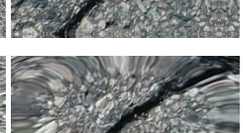
2.2 Dataset

The training and validation data utilized in this study were derived from Crack500[44], a dataset comprising pavement images captured from the asphalt roads of the main campus of Temple University, USA. However, the test data adopted two datasets: One was Crack500, and the other was a self-made dataset. The images of the self-made dataset were collected from the concrete pavements of the main roads in Shijiazhuang Tiedao University. After comparing the original pavement crack images with their corresponding annotated images, it was found that some samples in the Crack500 dataset exhibit low matching accuracy between crack images and their labels. Specifically, the annotated images contained crack regions that were identified based on annotators' experience but were difficult to verify in the original images through pixel comparison or visual inspection. Such samples would severely compromise the model's judgment. To ensure the quality of model training, samples with excessively large contrast differences were excluded from the

original dataset, and data samples with high matching accuracy were selected as the dataset for this study. The refined training set consists of 559 images, each sized at 640 pixel×320 pixel; the validation set contains 179 images, while the test set comprises 321 images, with each pavement image paired with its corresponding label maps. The self-made dataset includes 215 images with a pixel size of 640×320 and their corresponding label maps.

To further expand the training set, image augmentation was applied to the original samples before each training epoch. Random modifications were performed one or multiple times on the original images, ensuring that each training iteration utilized different images, thereby effectively expanding the training samples. The augmentation operations included flipping, rotation, brightness adjustment, and optical distortion, among others. A comparison of the effects before and after augmentation is presented in Table 2. Data augmentation operations not only enhance the diversity of data samples, sufficiently meeting the training volume requirements of all models in this study, but also improve the model generalization ability, thereby reducing the occurrence of overfitting to a certain extent.

Table 2 Examples of enhancement effects

Augmentation types	Parameter description	Group 1	Group 2	Group 3
Original image	None			
Flip	Horizontal or vertical flipping			
Rotation	Random rotation angle range: ±30°			
Brightness adjustment	Brightness variation range: (−0.2, 0.2), contrast variation range: (−0.2, 0.2)			
Random gamma transformation	γ value range: (80, 120) (for adjustment in relatively bright or dark scenarios)			
Elastic deformation	Alpha = 120 (deformation intensity), sigma = 6 (Gaussian smoothing intensity), alpha_affine = 0 (affine transformation)			
Grid distortion	Num_steps = 5 (dividing the image into a 5×5 grid), distort_limit = 0.3 (controlling the maximum offset of each grid point)			
Optical distortion	Distort_limit = 2 (radial distortion coefficient range), shift_limit = 0.5 (center point translation range)			

2.3 Evaluation metrics

This study evaluates and compares model performance based on the prediction outcomes of TP (True positive), FP (False positive), TN (True negative), and FN (False negative). Using these metrics, the mean intersection over union (mIoU),

Precision, Recall, and F1-score are calculated for assessment.

TP, FP, TN, and FN are typically represented by the number of pixels and are used to compute various evaluation metrics, as shown in Table 3.

Table 3 Meanings of TP, FP, TN, and FN

Index	Definition
TP	Samples that are actually positive are correctly classified as positive by the model
FP	Samples that are actually negative are wrongly classified as positive by the model
TN	Samples that are actually negative are correctly classified as negative by the model
FN	Samples that are actually positive are wrongly classified as negative by the model

The intersection over union (IoU) refers to the ratio of the intersection to the union between the predicted result and the true annotation. For each category, the intersection and union ratio of that category is calculated, and then the average of the intersection and union ratios of all categories is taken to obtain mIoU, as shown in

$$\mathrm{mIoU}=\frac{1}{k}\sum_{i=1}^{k}\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \tag{3}$$

Precision represents the proportion of samples that are actually in the positive class among those predicted by the model, reflecting the prediction accuracy of the network model, as shown in

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{4}$$

Recall represents the proportion of samples that are actually in the positive class and are correctly predicted by the model to be in the positive class. It mainly measures the network model's ability to recognize positive class samples, as shown in

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{5}$$

F1-score is the harmonic average of precision and recall, and used to comprehensively evaluate the performance of the model, as shown in

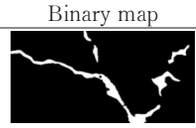
$$\text{F1-score}=\frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{6}$$

3 Experimental Results and Analysis

3.1 Comparison of evaluation metrics before and after data augmentation

To demonstrate the effectiveness of the data augmentation operation, this study trained the improved model both with and without data augmentation, and compared their performance results. Specific data are presented in Table 4. By comparing the model's prediction results before and after augmentation, it can be observed that all evaluation metrics of the improved model increase. This result fully verifies the importance and effectiveness of the data augmentation operation. As a key preprocessing step, data augmentation not only effectively expands the scale of the training set and alleviates the problem of model overfitting, but also helps improve the model robustness, ultimately achieving the optimization of model segmentation accuracy.

Table 4 Comparison of evaluation metrics for the improved model before and after data augmentation

Model	Original image	Ground truth map	Precision	Recall	F1-score	mIoU	Binary map
Improved model before augmentation			0.531 3	0.706 6	0.514 4	0.385 9	
Improved model after augmentation			**0.734 7**	**0.724 1**	**0.710 3**	**0.568 7**	

3.2 Ablation experiments

To verify the effectiveness of different modules in the proposed improved model, this section investigates the impacts of the multi-scale convolution and ASPP module on model performance through ablation experiments. Based on the basic ResUNet framework, three experimental settings were designed as follows: (1) Introducing only multi-scale convolu-

tion; (2) introducing only the ASPP module; (3) introducing multi-scale convolution and ASPP. All experiments were conducted under the Crack500 dataset and training strategy, with mIoU, Precision, Recall, and F1-score adopted as evaluation metrics. The specific comparisons are presented in Table 5.

Table 5 Comparison of evaluation metrics in ablation experiments

Model	Precision	Recall	F1-score	mIoU
ResUNet	0.688 0	**0.748 6**	0.695 8	0.552 1
Introducing only multi-scale convolution	0.702 0	0.742 7	0.701 2	0.556 9
Introducing only ASPP module	0.718 7	0.719 1	0.701 2	0.557 5
Improved model	**0.734 7**	0.724 1	**0.710 3**	**0.568 7**

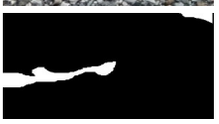
As can be seen from the data comparison in the table, both the introduction of multi-scale convolution alone and the ASPP module alone achieved a certain degree of performance improvement compared with the basic ResUNet. This indicates that both improvement strategies play a positive role in feature extraction and contextual information fusion. Furthermore, after combining multi-scale convolution with the ASPP module, all metrics are relatively balanced, and all metrics except Recall reach the optimal values. This demonstrates that the two modules have complementary advantages, which can effectively improve the segmentation accuracy and robustness of crack regions, thus verifying the rationality and effectiveness of the design of the proposed improved model.

3.3 Comparison of model prediction results

In the comparative experiments of this study, four typical semantic segmentation models, such as UNet, ResUNet, DeepLabv3+ and PSPNet, and two novel models proposed in recent years, namely ALP-UNet and FlexiCrackNet, were selected. As a classic encoder-decoder structure, UNet serves as a commonly used baseline model in medical imaging and crack segmentation. ResUNet introduces a residual structure on this basis, enhancing feature extraction capability. DeepLabv3+ achieves multi-scale context modeling through atrous convolution and the ASPP module, making it a widely applied high-precision method in current research. Meanwhile, PSPNet acquires global contextual information via the pyramid pooling module (PPM), representing another type of multi-scale feature aggregation approach. Incorporating PSPNet into the comparison not only helps demonstrate the differences among various multi-scale mechanisms, but also contributes to comprehensively verifying the effectiveness and generalization of the improved model proposed in this study. In addition, as crack segmentation models proposed in recent years, ALP-UNet and FlexiCrackNet also incorporate multi-scale feature fusion mechanisms in their structures. Specifically, ALP-UNet enhances its ability to capture fine cracks through adaptive pyramid feature extraction and a lightweight attention module; FlexiCrackNet, on the other hand, improves the model's adaptability to crack morphologies of different scales by leveraging multi-scale convolution and a feature reorganization strategy. Incorporating these two models with multi-scale fusion capabilities into the comparison can more comprehensively evaluate the advantages and improvement effects of the proposed method under the same multi-scale modeling framework.

As demonstrated by the partial model prediction results presented in Table 6, all five original images exhibit varying degrees of noise, characterized by relatively rough crack edges. The models exhibit distinct feature extraction performances when processing such images. UNet and ResUNet demonstrate significant deficiencies in handling noisy crack images: Their prediction results contain speckle noise, and the noise at the boundaries severely impairs the accuracy of crack shape extraction. PSPNet outperforms the aforementioned two models; however, certain crack regions in its prediction results appear blurred, and the precision of edge details is lower compared with the proposed improved model, leading to a reduction in overall accuracy. Although

Table 6 Test results of different models on the Crack500 dataset

Group	1	2	3	4	5
Original image					
Ground truth map					
UNet					
ResUNet					
PSPNet					
DeepLabv3+					
ALP-UNet					
FlexiCrackNet					
Improved model					

DeepLabv3+ effectively suppresses noise with minimal speckles in the generated results, it is prone to partial crack loss or incomplete extraction. FlexiCrackNet still suffers from noise interference, while the extraction performance of ALP-UNet is second only to the improved model with a clean background. The improved model proposed in this study not only adeptly mitigates noise interference but also preserves crack edge details with higher precision, demonstrating superior comprehensive performance.

In Group 3 images, a comparison of the cracks within the red bounding boxes reveals that the extraction result of the improved model is most consistent with the original image. In Group 5 images, by comparing the crack branches within the red bounding boxes, it can be observed that the improved model not only extracts complete branches but also avoids generating extraneous noise at the boundaries, exhibiting the highest similarity to the ground truth map.

In summary, when noise is present in the images, the improved model in this study is almost unaffected by noise and can capture more precise details at crack edges, better restoring the crack boundary details and achieving higher accuracy.

3.4 Comparison of evaluation metrics

As shown in Table 7, the improved model proposed in this study outperforms other models, achieving the highest scores in both mIoU and F1-score. In contrast, DeepLabv3+ and ALP-UNet attain the highest Precision and Recall, respectively. The improved model achieves an mIoU of 0.568 7, surpassing UNet's score of 0.539 0, ResUNet's 0.552 1, PSPNet's 0.566 7, Deep-

Labv3+'s 0.553 9, ALP-UNet's 0.533 1, and FlexiCrackNet's 0.558 6. This indicates that the improved model more accurately delineates crack regions, demonstrating superior overall segmentation accuracy and better overlap between predicted and actual crack areas.

The Precision of the improved model is 0.734 7, which exceeds UNet's 0.674 1, ResUNet's 0.688 0, PSPNet's 0.711 1, ALP-UNet's 0.660 2, and FlexiCrackNet's 0.696 4, although it is slightly lower than DeepLabv3+'s 0.784 6. This suggests that the improved model generates relatively fewer false positives in crack identification, indicating that a higher proportion of predicted crack areas are indeed cracks. The marginally lower Precision compared with DeepLabv3+ may be attributed to the latter's use of a lightweight backbone network MobileNetv2, which utilizes depthwise separable convolution to effectively reduce the computational cost during feature extraction while maintaining high discriminative ability, thus gaining certain advantages in distinguishing between positive and negative samples. In addition, the lightweight design achieves a good balance between accuracy and efficiency, which is also a possible reason for the slightly better Precision, and meanwhile provides a certain improvement direction for the model in this paper.

The Recall of the improved model is 0.724 1, which is lower than ResUNet's 0.748 6, PSPNet's 0.741 3, UNet's 0.747 2, ALP-UNet's 0.755 6, and FlexiCrackNet's 0.745 5, but higher than DeepLabv3+'s 0.660 3. This finding may suggest that the improved network has established a stricter decision boundary, which reduces the number of samples predicted as positive, thereby lowering the risk of false positives; however, this also results in a greater number of positive samples being missed. The F1-score of the improved model is 0.710 3, surpassing UNet's 0.685 1, ResUNet's 0.695 8, PSPNet's 0.709 2, DeepLabv3+'s 0.693 2, ALP-UNet's 0.679 6, and FlexiCrackNet's 0.702 4. Although the improved model does not achieve the highest Precision or Recall individually, its superior F1-score indicates a better balance between Precision and Recall compared with the other models, effectively identifying cracks accurately while maximizing the detection of all cracks. Although ALP-UNet is a novel model proposed in recent years, its performance is poor, and there are several potential reasons for this phenomenon: (1) All the settings of the original model are not fully reproduced during the implementation process; (2) its lightweight design results in suboptimal performance on the Crack500 dataset; (3) the detail enhancement component of the model amplifies noise components. Overall, the improved model in this study demonstrates certain advancements over other models. It achieves significant gains relative to UNet and ResUNet and exhibits balanced performance when compared with PSPNet and DeepLabv3+, resulting in partial improvements as well.

Table 7 Performance comparison of various models on the Crack500 dataset

Model	Precision	Recall	F1-score	mIoU
UNet	0.674 1	0.747 2	0.685 1	0.539 0
ResUNet	0.688 0	0.748 6	0.695 8	0.552 1
PSPNet	0.711 1	0.741 3	0.709 2	0.566 7
DeepLabv3+	0.784 6	0.660 3	0.693 2	0.553 9
ALP-UNet	0.660 2	0.755 6	0.679 6	0.533 1
FlexiCrackNet	0.696 4	0.745 5	0.702 4	0.558 6
Improved model	0.734 7	0.724 1	0.710 3	0.568 7

3.5 Model generalization validation

To further verify the generalization ability of the improved model, supplementary experiments are conducted on the self-made dataset in this paper. Compared with the Crack500 dataset, the self-made dataset has significant differences in image acquisition environment, pavement base material, and crack morphological features, which can more comprehensively test the adaptability of the model in non-training scenarios. The experiment maintains consistent control variables: All comparison models are trained based on the Crack500 dataset and only perform prediction tests on the self-made dataset. All other parameter settings are fully consistent with those of the aforementioned comparative experiments. Selected visual results of the tests are presented in Table 8, while the detailed quantitative evaluation metrics are summarized in Table 9.

Table 8 Test results of different models on the self-made dataset

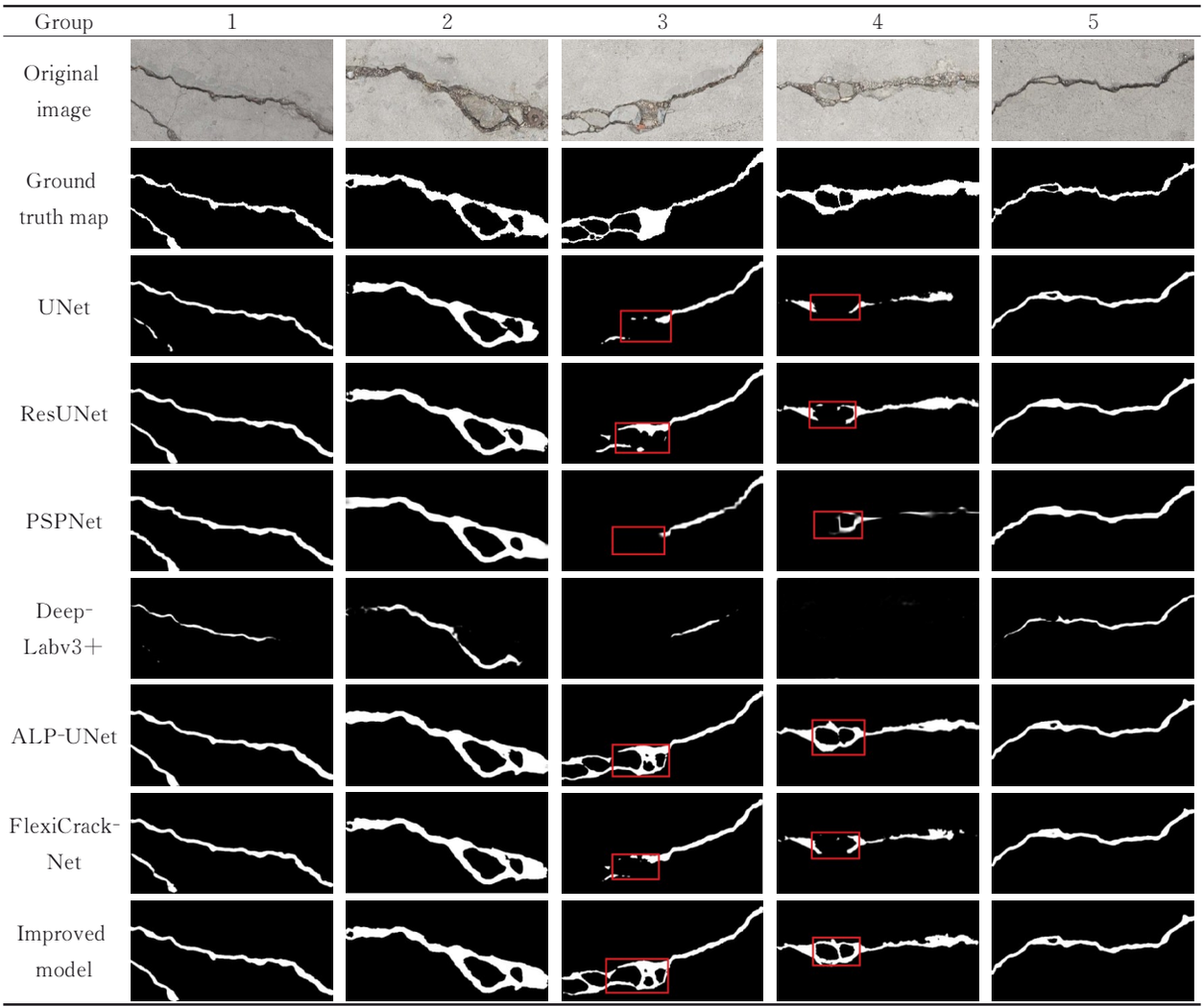


Table 9 Performance comparison of various models on the self-made dataset

Model	Precision	Recall	F1-score	mIoU
UNet	0.882 7	0.567 0	0.654 2	0.528 8
ResUNet	0.881 9	0.708 2	0.761 6	0.645 0
PSPNet	0.833 8	0.699 5	0.727 5	0.621 0
DeepLabv3+	0.854 6	0.188 0	0.269 4	0.167 3
ALP_UNet	0.876 9	0.639 0	0.711 3	0.586 2
FlexiCrackNet	0.858 8	0.656 2	0.714 7	0.590 5
Improved model	0.873 1	0.779 4	0.807 3	0.690 3

Through a comparative analysis of the extraction results across all image groups, it can be observed that only DeepLabv3+ exhibits inferior performance, characterized by incomplete crack extraction with some images even failing to detect any cracks. Thus, DeepLabv3+ is deemed to have poor generalization capability. For the remaining models, comparative evaluations reveal no significant differences in their extraction performance when processing images with distinct crack-pavement contrast such as Groups 1 and 5. However, notable discrepancies emerge among the models when handling images with blurred cracks such as Groups 3 and 4. In Group 3, only ALP-UNet and the proposed improved model successfully extract cracks consistent with the ground truth within the red bounding boxes while the other models yield incomplete results, and the cracks extracted by the improved model are more aligned with the ground truth compared to those from ALP-UNet. In Group 4, ALP-UNet and the improved model again demonstrate superior extraction performance within the red bounding boxes with the improved model achieving the optimal results.

The experimental results show that the overall performance of the improved model on the self-made dataset is outstanding: Not only its comprehensive performance is better than the test results of the model on the Crack500 dataset, but also its advantage over other comparison models is more significant. It can be seen from Table 9 that except for the Precision of the improved model being 0.873 1, which is slightly lower than that of UNet (0.882 7), all other values are the highest, namely mIoU is 0.690 3, Recall is 0.779 4, and F1-score is 0.807 3. Based on a comprehensive analysis of data characteristics and model structure, the causes of the above results can be summarized into the following three points. (1) Difference in data annotation accuracy: The ground truth maps of the self-made dataset have a higher matching degree with the original images, and the annotation errors are smaller. However, some samples in the Crack500 dataset have large deviations between the original images and the ground truth maps. This difference directly leads to better test accuracy of the model on the self-made dataset. (2) Advantage of feature extraction in the improved model: The improved model fuses crack features of different dimensions through multi-scale convolution, and can fully learn the local details and global structural information of cracks during training, making it with stronger feature transfer ability. It is able to adapt to test sets with significant differences from the training set, highlighting good generalization performance. (3) Although DeepLabv3+ performs well on the Crack500 dataset, its performance drops significantly on the self-made dataset. It is speculated that its model structure has a strong dependence on the style of training data, its generalization ability has certain limitations, and its adaptability is insufficient in crack segmentation tasks across scenarios and materials. (4) While PSPNet achieves comparable performance to the improved model proposed in this study on the Crack500 dataset, significant discrepancies arise when tested on the self-made dataset. This indicates that despite the similar test results of the two models on the Crack500 dataset, there exists a substantial gap in their generalization performance, undoubtedly demonstrating the superiority of the improved model in this work. (5) ALP-UNet and FlexiCrackNet exhibit only moderate performance on the self-made dataset. By analyzing their evaluation metric results and combining this with an examination of their model structures, it is inferred that the potential cause for their underwhelming performance lies in the accuracy degradation induced by their lightweight design.

In conclusion, the improved model can still maintain a higher accuracy level than most comparison models on the new test dataset, and shows better segmentation performance in datasets with higher annotation quality and significant scene differences. This result fully proves that the improved model proposed in this paper has strong generalization ability and higher application value in actual complex scene crack segmentation tasks.

4 Conclusions

This study implements the automatic extraction of pavement cracks using neural networks, with the resulting classification outcomes serving as a foundation for subsequent pavement crack processing. Seven network models were compared in terms of classification performance, including UNet, ResUNet, PSPNet, DeepLabv3+, ALP-UNet, FlexiCrackNet, and the improved model proposed in this paper, to evaluate their effectiveness in crack detection tasks. In addition, the effectiveness of data augmentation and the two proposed improvements was verified separately. The comparison of results before and after data augmentation demonstrates the necessity of data augmentation for enhancing model robustness, while the ablation experiments confirm that both the multi-scale convolution and the ASPP module can indeed improve the accuracy of the original model to a certain extent, with the combined effect of the two being more favorable. Finally, by comparing the test results on the self-made dataset with those on the Crack500 dataset, the excellent generalization ability of the improved model is verified. Through the experimental

analysis of comparing the data results of different models, significant differences among the models in crack extraction capabilities were demonstrated. Among the representative overall accuracy metrics, the F1-scores indicated that the proposed improved model achieved the best performance. This improvement is attributed to the network structure, which extracts features simultaneously through multiple branches with varying convolution kernel sizes, thereby enhancing the model's ability to capture crack features at various scales and improving its effectiveness in delineating boundaries and fine cracks. Additionally, the incorporation of the ASPP module with atrous convolutions of different sizes further enhances the model's ability to extract details at multiple scales, while the combination of local details with global semantics significantly boosts accuracy. Unlike other neural network models that simply add the ASPP module, the model proposed in this study achieves early multi-scale feature fusion through a multi-scale parallel convolution structure in the encoding phase. Furthermore, it leverages the ASPP module in the intermediate layer to further capture cross-scale contextual information. Through the organic integration of multi-scale convolution and ASPP, a bottom-up hierarchical feature representation framework that balances details and semantics is constructed. This not only leads to a certain improvement in the overall accuracy of the model but also significantly optimizes its classification performance. This series of comparative experiments not only validates the superiority of the improved model but also provides direction for future enhancements, making the new model design more targeted.

The improved model proposed in this study has achieved certain results in crack segmentation accuracy and robustness, but there is still room for further improvement. First, to address the issues of large model parameter size and high computational cost, lightweight attention mechanisms or structural reparameterization techniques can be explored in future research to reduce redundant computations and improve inference efficiency. Second, regarding the insufficient detection of fine cracks, pyramid feature fusion or multi-scale attention mechanisms can be integrated to enhance the model's ability to perceive targets of different scales. In addition, self-supervised learning and multi-modal data fusion remain promising research directions, for instance, using unlabeled data to improve the generalization of feature learning, or combining image data with structural sensing data to achieve more comprehensive crack identification. Through in-depth research on the above directions, it is expected that while ensuring segmentation accuracy, the lightweight level of the model and its small-target detection performance can be further improved, thereby better adapting to the practical needs of complex road scenarios.

References

[1] WANG Zhi, WANG Jinbin, GUAN Minjie, et al. Common crack disease and prevention measures for concrete or asphalt pavements[J]. Urban Roads Bridges & Flood Control, 2018(11): 64-68, 11, 72. (in Chinese)

[2] DEME D. A review on effect of pavement surface failure on road traffic accident[J]. American International Journal of Sciences and Engineering Research, 2020, 3(1): 14-19.

[3] MA Jian, ZHAO Xiangmo, HE Shuanhai, et al. Review of pavement detection technology[J]. Journal of Traffic and Transportation Engineering, 2017, 17(5): 121-137. (in Chinese)

[4] KHERADMANDI N, MEHRANFAR V. A critical review and comparative study on image segmentation-based techniques for pavement crack detection[J]. Construction and Building Materials, 2022, 321: 126162.

[5] WANG W X, WANG M F, LI H X, et al. Pavement crack image acquisition methods and crack extraction algorithms: A review[J]. Journal of Traffic and Transportation Engineering (English Edition), 2019, 6(6): 535-556.

[6] WU Yiquan, MENG Tianliang, WU Shihua. Research progress of image thresholding methods in recent 20 years (1994—2014)[J]. Journal of Data Acquisition and Processing, 2015, 30(1): 1-23. (in Chinese)

[7] OTSU N. A threshold selection method from gray-level histograms[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62-66.

[8] HUANG M X, YU W J, ZHU D H. An improved

image segmentation algorithm based on the otsu method[C]//Proceedings of the 2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. Kyoto, Japan: IEEE, 2012: 135-139.

[9] TALAB A M A, HUANG Z C, XI F, et al. Detection crack in image using otsu method and multiple filtering in image processing techniques[J]. Optik, 2016, 127(3): 1030-1033.

[10] WANG Zhuo, GE Bin, TU Mingyu, et al. Image segmentation algorithm based on improved otsu algorithm and artificial fish swarm optimization[J]. Packaging Journal, 2019, 11(2): 81-86. (in Chinese)

[11] KIRSCHKE K R, VELINSKY S A. Histogram-based approach for automated pavement-crack sensing[J]. Journal of Transportation Engineering, 1992, 118(5): 700-710.

[12] WEI Xiuling. Application of global threshold of genetic algorithm in image segmentation[J]. Digital Technology & Application, 2012, 30(6): 128-129. (in Chinese)

[13] QIAO Haojie, WANG Xiaoyun. Fruit image segmentation algorithm based on lab and global threshold segmentation[J]. The Magazine on Equipment Machinery, 2024(2): 12-16, 64. (in Chinese)

[14] OLIVEIRA H, CORREIA P L. Automatic road crack segmentation using entropy and image dynamic thresholding[C]//Proceedings of the 2009 17th European Signal Processing Conference. Glasgow, UK: IEEE, 2009: 622-626.

[15] SAFAEI N, SMADI O, SAFAEI B, et al. Efficient road crack detection based on an adaptive pixel-level segmentation algorithm[J]. Transportation Research Record: Journal of the Transportation Research Board, 2021, 2675(9): 370-381.

[16] WEI Chuntao, YU Junchen, ZHAO Ping, et al. Automatic detection method of small cracks and micro grayscale difference cracks based on adaptive threshold[J]. Journal of China & Foreign Highway, 2019, 39(1): 58-63. (in Chinese)

[17] DUAN Ruiling, LI Qingxiang, LI Yuhe. Summary of image edge detection[J]. Optical Technique, 2005, 31(3): 415-419. (in Chinese)

[18] CANNY J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, PAMI-8(6): 679-698.

[19] SUBIRATS P, DUMOULIN J, LEGEAY V, et al. Automation of pavement surface crack detection using the continuous wavelet transform[C]//Proceedings of the 2006 International Conference on Image Processing. Atlanta, GA, USA: IEEE, 2006: 3037-3040.

[20] WANG Shifang, CHE Yanli, LI Nan, et al. Asphalt pavement crack detection algorithm based on multi-scale ridges[J]. China Journal of Highway and Transport, 2017, 30(4): 32-41. (in Chinese)

[21] ROSENFELD A. The max roberts operator is a hueckel-type edge detector[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1981, 3(1): 101-103.

[22] LANG Y, ZHENG D. An improved sobel edge detection operator[C]//Proceedings of the 2016 6th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2016). Shenyang, China: Atlantis Press, 2016: 590-593.

[23] YANG L, WU X Y, ZHAO D W, et al. An improved Prewitt algorithm for edge detection based on noised image[C]//Proceedings of the 2011 4th International Congress on Image and Signal Processing. Shanghai, China: IEEE, 2011: 1197-1200.

[24] ULUPINAR F, MEDIONI G. Refining edges detected by a LoG operator[J]. Computer Vision, Graphics, and Image Processing, 1990, 51(3): 275-298.

[25] LI E S, ZHU S L, ZHU B S, et al. An adaptive edge-detection method based on the canny operator[C]//Proceedings of the 2009 International Conference on Environmental Science and Information Application Technology. Wuhan, China: IEEE, 2009: 465-469.

[26] HU Y, ZHAO C X. A novel LBP based methods for pavement crack detection[J]. Journal of Pattern Recognition Research, 2010, 5(1): 140-147.

[27] ZOU Q, CAO Y, LI Q Q, et al. CrackTree: Automatic crack detection from pavement images[J]. Pattern Recognition Letters, 2012, 33(3): 227-238.

[28] SALMAN M, MATHAVAN S, KAMAL K, et al. Pavement crack detection using the Gabor filter[C]//Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013). The Hague, Netherlands: IEEE, 2013: 2039-2044.

[29] AMHAZ R, CHAMBON S, IDIER J, et al. Automatic crack detection on two-dimensional pavement images: An algorithm based on minimal path selection[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(10): 2718-2729.

[30] ZHANG L, YANG F, ZHANG Y D, et al. Road crack detection using deep convolutional neural network[C]//Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP). Phoenix, AZ, USA: IEEE, 2016: 3708-3712.

[31] KRIZHEVSKY A, SUTSKEVER I, HINTON G

E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[32] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. International Journal of Artificial Intelligence and Machine Learning in Engineering, 2014, 12(8): 301-307.

[33] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015: 1-9.

[34] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015: 3431-3440.

[35] MA D, FANG H Y, WANG N N, et al. A real-time crack detection algorithm for pavement based on CNN with multiple feature layers[J]. Road Materials and Pavement Design, 2022, 23(9): 2115-2131.

[36] LEE T, YOON Y, CHUN C J, et al. CNN-based road-surface crack detection model that responds to brightness changes[J]. Electronics, 2021, 10(12): 1402.

[37] DI BENEDETTO A, FIANI M, GUJSKI L M. U-Net-based CNN architecture for road crack segmentation[J]. Infrastructures, 2023, 8(5): 90.

[38] ZHANG Tianjie, WANG Yangyang, HAN Haihang, et al. A method for recognizing asphalt pavement cracks based on modifiedtwo-step convolutional neural network[J]. Journal of Highway and Transportation Research and Development, 2022, 39(10): 1-8, 58. (in Chinese)

[39] YUAN Lin. Image feature extraction andapplication based on deep learning[D]. Changchun: Jilin University, 2022. (in Chinese)

[40] YANG Yulong. Research on classification and detection of asphaltpavement crack based on diffusion model andvision transformer[D]. Xi'an: Chang'an Univrsity, 2024. (in Chinese)

[41] GUO F, QIAN Y, LIU J, et al. Pavement crack detection based on transformer network[J]. Automation in Construction, 2023, 145: 104646.

[42] WANG J, ZENG Z G, SHARMA P K, et al. Dual-path network combining CNN and transformer for pavement crack segmentation[J]. Automation in Construction, 2024, 158: 105217.

[43] ZOU Q, ZHANG Z, LI Q Q, et al. DeepCrack: Learning hierarchical convolutional features for crack detection[J]. IEEE Transactions on Image Processing, 2018, 28(3): 1498-1512.

[44] YANG F, ZHANG L, YU S J, et al. Feature pyramid and hierarchical boosting network for pavement crack detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(4): 1525-1535.

[45] ZHANG A, WANG K C P, FEI Y, et al. Automated pixel-level pavement crack detection on 3D asphalt surfaces with a recurrent neural network[J]. Computer-Aided Civil and Infrastructure Engineering, 2019, 34(3): 213-229.

[46] SHI Y, CUI L M, QI Z Q, et al. Automatic road crack detection using random structured forests[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(12): 3434-3445.

[47] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]//Proceedings of the Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Cham: Springer International Publishing, 2015: 234-241.

[48] YIN Xiaohang, WANG Yongcai, LI Deying. Suvery of medical image segmentation technology based on U-Net structure improvement[J]. Journal of Software, 2021, 32(2): 519-550. (in Chinese)

[49] ZHANG Z, LIU Q, WANG Y. Road extraction by deep residual U-Net[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 15(5): 749-753.

[50] ZHANG Qianwen, CHEN Ming, QIN Yufang, et al. Lung nodule segmentation based on 3D ResUNet network[J]. Chinese Journal of Medical Physics, 2019, 36(11): 1356-1361. (in Chinese)

[51] ZHAI Chenfei, DONG Wenhan, ZHANG Xiaoming, et al. Prediction of strong cumulus clouds by improved U-Net network and self-attention mechanism [J]. Computer Engineering and Applications, 2022, 58(22): 297-304. (in Chinese)

[52] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, Nevada: IEEE, 2016: 770-778.

[53] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[54] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image

segmentation[J]. arXiv preprint arXiv: 170605587, 2017.

[55] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFS[J]. arXiv preprint arXiv:14127062, 2014.

[56] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of Computer Vision—ECCV 2018. Cham: Springer International Publishing, 2018: 833-851.

[57] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 6230-6239.

[58] ZHANG Y, GAO X, ZHANG H Z. Deep learning-based semantic segmentation methods for pavement cracks[J]. Information, 2023, 14(3): 182.

[59] WAN X, JIANG X, LUO G, et al. FlexiCrackNet: A flexible pipeline for enhanced crack segmentation with general features transfered from SAM[J]. arXiv preprint arXiv:250118855, 2025.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China (No. 42401166), the Open Fund of Key Laboratory of Polar Environment Monitoring and Public Governance, Ministry of Education (No.202405), and the Key Research and Development Program of Hebei Province (No.23375405D).

Authors

The first author Mr. **ZHAN Biheng** received the B.S. degree in School of Civil Engineering at Shijiazhuang Tiedao University in 2020. He is currently pursuing M.S. degree in engineering at Shijiazhuang Tiedao University. His research focuses on road crack extraction based on neural networks.

The corresponding author Dr. **SONG Xiangyu** received his Ph.D. degree in geodesy and surveying engineering from Wuhan University in 2021.In the same year, he joined School of Civil Engineering at Shijiazhuang Tiedao University as a master's supervisor. His main research directions are remote sensing and the construction of neural network models.

Author contributions Dr. **SONG Xiangyu** designed the study. Mr. **ZHAN Biheng** complied the models, conducted the analysis, interpreted the results, and wrote the manuscript. Mrs. **CHENG Jianrui** and Mr. **WANG Tengfei** contributed to the discussion and background of the study. Mr. **QIAO Pan** offered suggestions on the structure and grammar writing of this manuscript. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: XU Chengting)

基于多尺度卷积神经网络的路面裂缝提取

詹必恒[1,2], 宋翔宇[1,2,3], 程建蕊[4], 乔 盘[5,6], 王腾飞[7]

(1.道路与铁道工程安全保障省部共建教育部重点实验室(石家庄铁道大学), 石家庄 050043, 中国; 2.石家庄铁道大学土木工程学院,石家庄 050043, 中国; 3.极地环境监测与公共治理教育部重点实验室(武汉大学),武汉 430079,中国; 4.河北省制图院,石家庄 050031, 中国; 5.河北交规院瑞志交通技术咨询有限公司,石家庄 050090, 中国; 6.河北省道路工程智能监测与运维技术创新中心,保定 071799, 中国; 7.中铁隧道局集团路桥工程有限公司,天津 300450, 中国)

摘要:裂缝作为道路的主要危害之一,实现高效自动化的提取对于道路健康监测和维护而言至关重要。为此提出一种融合图像多尺度特征的裂缝自动提取网络模型,旨在提升模型对裂缝特征的捕获能力和对复杂场景的适应性。该模型以ResUNet网络模型架构为基础,针对模型卷积层做出修改,提出构建多条不同卷积核尺寸的支路,并在中间层添加空洞空间金字塔池化模块。本文进行了基础模型性能对比实验、消融实验、数据增强前后对比实验和泛化性验证实验。对比实验结果显示,本文提出的改进模型在处理裂缝边缘部分具备更加准确的细节处理能力,模型综合精度指标F1分数达到71.03%,相比传统ResUNet提高2.1%。

关键词:道路工程;神经网络;多尺度卷积;路面裂缝