Mathematical Biosciences and Engineering

MBE, 20(8): 14920–14937.
DOI: 10.3934/mbe.2023668
Received: 10 June 2023
Revised: 03 July 2023
Accepted: 04 July 2023
Published: 11 July 2023

http://www.aimspress.com/journal/MBE

Research article

Multi-scale feature fusion for pavement crack detection based on Transformer

Yalong Yang[1,2,3], Zhen Niu[1,2,3], Liangliang Su[1,2,3,*], Wenjing Xu[1,2,3] and Yuanhang Wang[1,2,3]

[1] Anhui Province Key Laboratory of Intelligent Building and Building Energy Saving, Anhui Jianzhu University, Hefei 230022, China

[2] Anhui Institute of Strategic Study on Carbon Dioxide Emissions Peak and Carbon Neutrality in Urban-Rural Development, Hefei 230022, China

[3] School of Electronic and Information Engineering, Anhui Jianzhu University, Hefei 230601, China

* **Correspondence:** Email: llsu_yz@ahjzu.edu.cn; Tel: +86 13866123115.

Abstract: Automated pavement crack image segmentation presents a significant challenge due to the difficulty in detecting slender cracks on complex pavement backgrounds, as well as the significant impact of lighting conditions. In this paper, we propose a novel approach for automated pavement crack detection using a multi-scale feature fusion network based on the Transformer architecture, leveraging an encoding-decoding structure. In the encoding phase, the Transformer is leveraged as a substitute for the convolution operation, which utilizes global modeling to enhance feature extraction capabilities and address long-distance dependence. Then, dilated convolution is employed to increase the receptive field of the feature map while maintaining resolution, thereby further improving context information acquisition. In the decoding phase, the linear layer is employed to adjust the length of feature sequence output by different encoder block, and the multi-scale feature map is obtained after dimension conversion. Detailed information of cracks can be restored by fusing multi-scale features, thereby improving the accuracy of crack detection. Our proposed method achieves an F1 score of 70.84% on the Crack500 dataset and 84.50% on the DeepCrack dataset, which are improvements of 1.42% and 2.07% over the state-of-the-art method, respectively. The experimental results show that the proposed method has higher detection accuracy, better generalization and better crack detection results can be obtained under both high and low brightness conditions.

Keywords: crack detection; Transformer; feature map; multi-scale feature; feature fusion

1. Introduction

Crack is one of the common pavement diseases, and pavement cracks will reduce the efficiency of road traffic, and can even lead to serious traffic accidents and endanger life safety. Therefore, timely detection, accurate evaluation and repair of cracks are one of the key tasks in pavement maintenance. Figure 1 shows some examples of pavement cracks. In the face of the huge stock of domestic roads, the traditional manual crack detection method has been unable to meet the current demand, so the intelligent crack detection method, based on pavement image, has gradually attracted wide attention. However, pavement cracks have complex topological structure, uneven light and noisy texture background [1], making effective crack detection a significant challenge.

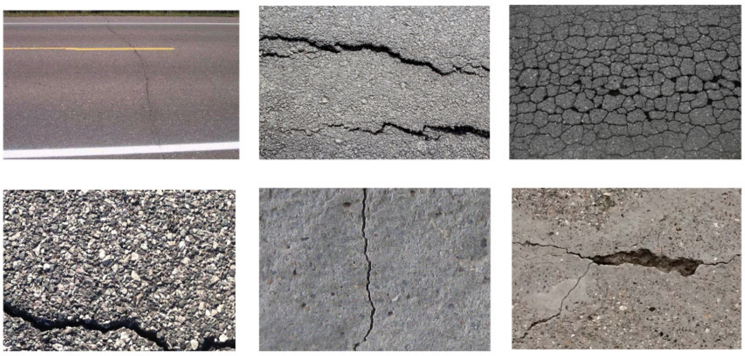

Figure 1. Examples of pavement crack.

In recent years, deep learning has made significant breakthroughs in various computer vision tasks, and new deep neural network models are constantly emerging [2]. Deep convolutional neural networks (CNNs) [3,4] and Transformer neural network [5] are two representative models for semantic segmentation. Although CNN-based models have dominated this field since the advent of the fully convolutional network (FCN), the recent segmentation Transformer (SETR) [6] replaced the CNN encoder with a pure Transformer structure encoder, which altered the architecture of the current semantic segmentation models. However, the output feature map of Transformer has low resolution and lacks detailed information about cracks, and the decoding structure proposed by SETR does not effectively solve this problem, resulting in poor performance for detecting slender cracks.

In this paper, we propose a multi-scale feature fusion network for pavement crack detection based on Transformer, including the vision Transformer (ViT) model, to extract crack features, dilated convolution to expand receptive field and upsampling to restore resolution and multi-scale feature fusion. Our method globally models the feature map and recovers the detailed information of cracks by fusing multi-scale features, so that we can achieve accurate segmentation of slender cracks and have a better generalization.

The contributions of this work are as follows:

(1) We propose an automatic crack segmentation method based on Transformer. Compared with existing pavement crack detection networks, our method has higher detection accuracy and better generalization.

(2) To address the challenge that the ViT model can only output feature maps at a fixed resolution, we introduce a multi-scale feature fusion module as a solution.

(3) A study of the influence of different brightness levels on the crack segmentation model performance.

(4) We demonstrate the effectiveness of various blocks and their combinations, such as dilated convolution blocks and feature fusion blocks.

2. Related work

2.1. Traditional crack detection methods

Early researchers usually use digital image processing technology to realize the automatic detection of pavement cracks. The gray value of the crack and the gray value of the background often have obvious differences, and the crack can be effectively separated from the background by selecting the appropriate threshold [7–9]. However, the thresholding method is sensitive to noise and usually requires the addition of preprocessing and postprocessing operations. Edge detection is a method to segment an image based on the abrupt change and discontinuity of image gray level. At the boundary of the object, the gray value of the pixel often changes significantly [10]. Edge detection operators, such as Sobel [11] and Canny [12,13], can separate the crack from the background by calculating the gradient change of the crack boundary. However, edge detection is an ill-posed problem, and no edge detector can respond only to the features of the target object in the image [14].

Although the crack detection methods based on digital image processing can achieve good detection results, the quality of the input image is required to be high, and the performance of these detection methods will be seriously affected when the crack and background contrast is not obvious, the light is uneven, or there is noise interference.

2.2. Crack detection using deep convolutional neural networks

In recent years, deep learning, especially deep convolutional neural networks, has been widely used in image classification, object detection, semantic segmentation and other fields, automatic extraction of image features by deep neural network and back propagation.

Ronneberger et al. [4] proposed U-Net based on FCN [3] structure. The network is composed of encoder and decoder, has simple structure and fast training speed, and has been widely used in biomedical image segmentation field. Liu et al. [15] first applied U-Net in the field of crack segmentation, and achieved good results in small datasets. After the success of U-Net network in the field of crack detection, some scholars improved the U-Net network by adding the attention mechanism to enhance the identification of crack, so as to obtain more accurate semantic information [16–18]. Chen et al. [19] proposed an encoder-decoder network based on the SegNet [20] model and initialized with pretrained weights, which has high crack detection performance and generalization ability. Liu et al. [21] proposed a deep hierarchical convolutional neural network named DeepCrack, which consists of a fully convolutional network and a deeply supervised net (DSN), and the final feature map aggregates the multi-scale and multi-level features of different convolutional layers. The features of different convolution stages are directly supervised by DSN, and the end-to-end pixel-level crack segmentation is realized. Ren et al. [22] proposed a deep full convolutional neural network CrackSegNet, in which a dilated convolution and pyramid pooling module [23] were added to the network, and the context information was obtained by increasing the receptive field, thus realizing the effective segmentation of concrete cracks.

However, due to the limited size of convolutional kernel, it is difficult for convolutional neural networks to obtain a larger receptive field. Although this problem can be solved by deepening the number of layers in the network, it will make the model too complicated and increase the calculation cost.

2.3. Crack detection using Transformer neural networks

Transformer model first emerged in the field of natural language processing, and was first applied to the task of image classification in [24], where the ViT model was proposed. In 2020, Zheng et al. [6] proposed the SETR segmentation algorithm based on the Transformer model, which uses the ViT to extract image features, providing a sequence-based image semantic segmentation perspective for subsequent researchers. CrackFormer [25] model adopts the self-attention mechanism to encode and decode the feature map, and combines the output of the corresponding encoder and decoder by scaling-attention block to obtain the clear crack boundary. SegFormer [26] model abandons position embedding and designs a layered Transformer encoder, which uses overlapping patch merging to downsample the feature map to obtain multi-scale features. The decoder only uses linear layers, which reduces the complexity of the network and achieves good segmentation results. SegCrack [27] model adopts the same encoder as SegFormer. When decoding, it uses lateral connection to restore the feature map scale layer by layer, and then fuses the feature maps of all scales to form a multi-scale feature map, which presents a more powerful representation by combining local features with global features. Feng et al. [28] used swin Transformer [29] to encode the crack image, and input the features of different encoder stages into the multi-layer perceptron (MLP) layer to unify the channel and size. Efficient and accurate segmentation of pavement cracks can be achieved by fusing the features of different stages. The TransMF [30] model uses a symmetric encoder and decoder structure, and added the fusion module. The encoder uses a hybrid model of convolution and Swin Transformer to model the crack from a local and global perspective, and the fusion module fuses both encoding and decoding features. The influence of noise can be reduced and the correlation between contexts can be strengthened.

Compared with convolution operation, Transformer uses self-attention mechanism dynamic modeling to discover the importance of feature sequences, and adopts a sequence-to-sequence learning method with less inductive bias [31]. It can achieve better results in the detection of cracks, and has gradually become one of the mainstream methods of crack detection.

3. The proposed method

As shown in Figure 2, the crack detection process consists of two main parts: training the network model and crack detection. First, the original crack image is inputted into the trained network model, and then the detected crack prediction map is outputted.

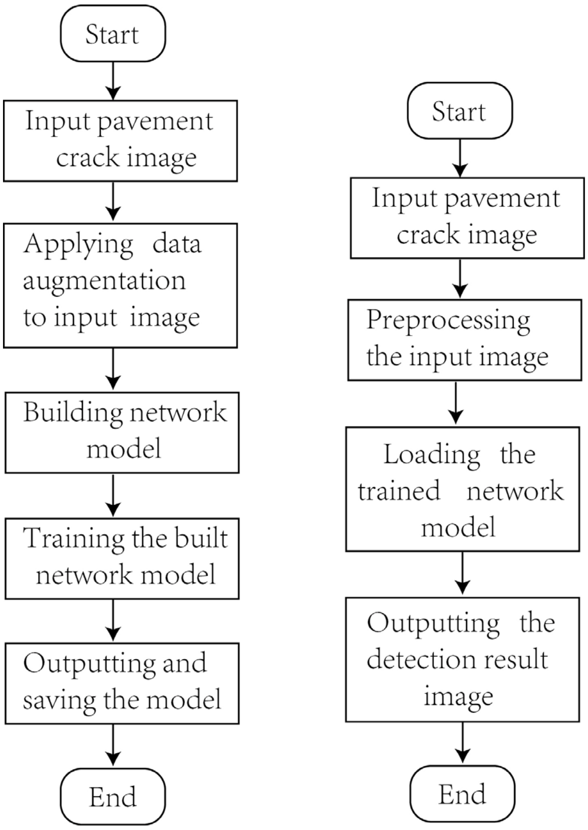
Start
Input pavement crack image
Applying data augmentation to input image
Building network model
Training the built network model
Outputting and saving the model
End
Start
Input pavement crack image
Preprocessing the input image
Loading the trained network model
Outputting the detection result image
End

Figure 2. Flow chart of model training (left) and crack detection (right).

3.1. Network architecture

In this paper, we proposed a multi-scale feature fusion network for pavement crack detection based on Transformer. The network architecture is shown in Figure 3, which is composed of encoder and decoder network. The encoder network adopts ViT model as the backbone, which is composed of Patch Embedding, Position Embedding and Transformer Encoder. The dilated convolution block with different combinations of dilation rates is added after the backbone, which can capture multi-scale information. The decoder consists of upsampling block, convolution block and the multi-scale feature fusion block proposed in this paper. The decoder network recovers the resolution of the feature map output by the encoder network layer by layer and fuses the multi-scale features to obtain the final segmentation result.

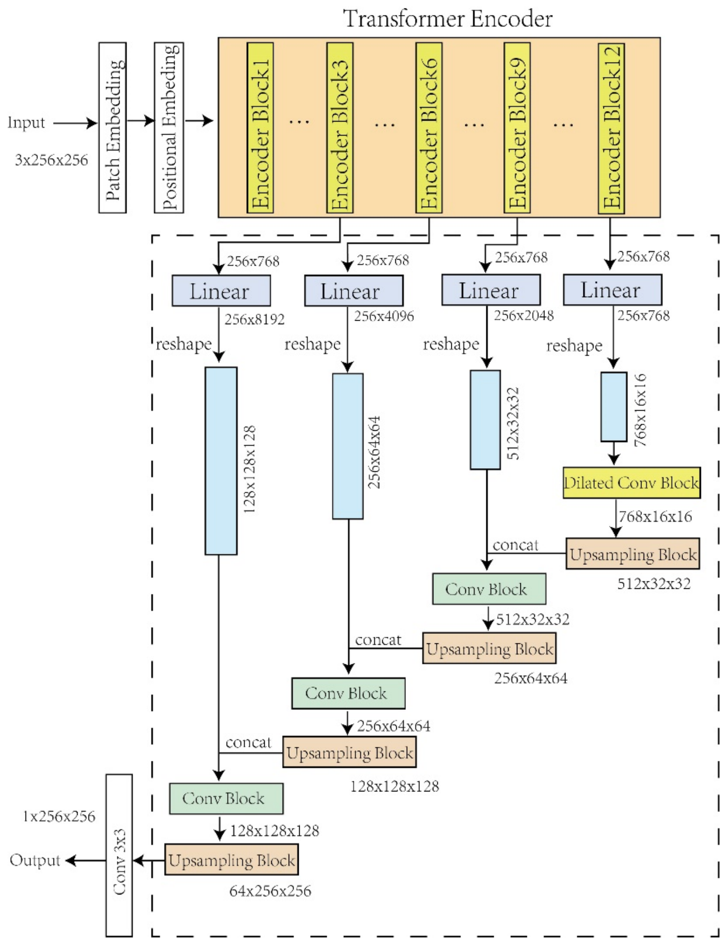

Figure 3. Network structure of proposed method.

3.2. Encoder network

3.2.1. Backbone

The Transformer encoder of ViT adpots the Pre-LN Transformer architecture [32], this structure puts the layer normalization block of Transformer model inside the residual structure, which can improve the convergence speed. In the encoding stage, first, the input image is split into 16 × 16 patches, and each patch is flattened into a one-dimensional vector, which named token. Then, the token is layer normalized, activated by the GeLU function and add position embeddings. Finally, the tokens are input into Transformer encoder to extract crack image features. The encoder is stacked with 12 identical encoder blocks. Each encoder block is composed of layer norm, multi-head self-attention, dropout and MLP block. Inside the encoder block, the input data distribution is unified into Gaussian distribution through the Layer Norm layer at first, then self-attention and full connection operation are performed. The self-attention formula [5] is as follows:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

where, Q , K , V are three matrices, obtained from the input features through three linear transformations, Q refers to the feature matrix of the crack area that needs to be attended to, K refers to the feature matrix of all locations in an entire image, V refers to the feature matrix of the location that corresponds to K. The matrix K is used to compute similarity with matrix Q, in order to perform a weighted average of the corresponding matrix V based on attention score. When the similarity between the matrix Q and matrix K is high, the corresponding matrix V is given a higher weight, and vice versa. This allows crack features to stand out while suppressing unnecessary information. d_k is the dimension of the matrix K.

3.2.2. Dilated convolution block

In order to improve the detection effect of slender cracks, it is necessary to enlarge the receptive field to obtain long-distance dependence. The receptive field of the feature map can be increased without reducing the resolution by employing the dilated convolution, thereby capturing the context information [33].

As shown in Figure 4, the dilated convolution block combines convolutions with different dilation rates, and the receptive fields of each layer are 3, 7 and 15, respectively. The feature sequence output from the backbone network is dimensionally transformed to obtain the feature map of size 16 × 16 × 768. In the dilated convolution block, the receptive field can cover the main part of the feature map. By fusing the feature maps of different receptive fields, it is helpful to obtain the context information and further extract the crack features.

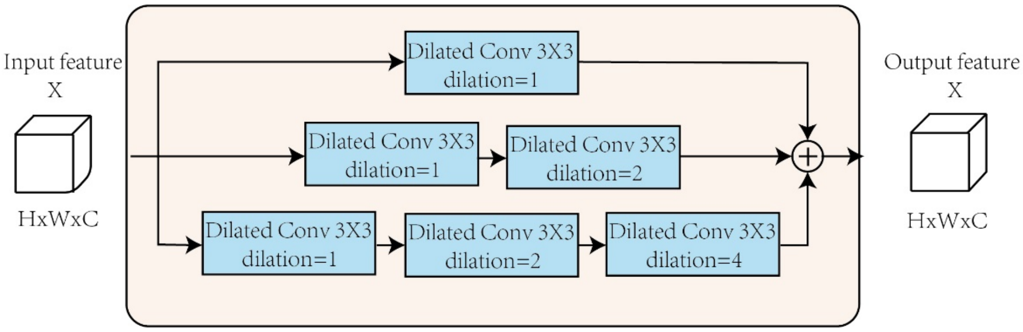

Figure 4. Dilated convolution block.

3.3. Decoder network

Low-level features contain more deatil information, which is helpful to restore the crack boundary and improve segmentation accuracy. This paper designed a simple and efficient feature fusion block. A linear layer is added after the output of the encoder block of the 3rd, 6th and 9th layers of Transformer encoder, respectively. The length of the feature sequence is changed by setting different numbers of neurons, which are 256 × 8192, 256 × 4096 and 256 × 2048, respectively. After converting the dimension of the feature sequence, three feature maps with different scales are obtained, and the sizes of the feature maps are 128 × 128 × 128, 64 × 64 × 256 and 32 × 32 × 512, respectively.

In the upsampling block, the resolution of the feature map is increased to two times and the channel number is reduced to one half of the original feature map, and fused with the feature map of the same scale, then the fused feature map is input to the convolution block for twice convolution operation, and batch norm and RuLU activation function operations are added after each convolution.

Repeat the above operations to recover the resolution of the feature map layer by layer and fuse the multi-scale features, so as to obtain the final crack segmentation map.

4. Experimental and results

To validate the proposed methodology and conduct comparative analysis with other approaches, we selected two standard crack datasets, namely Crack500 [34] and DeepCrack [21], for our experimentation.

(1) The Crack500 dataset contains a variety of complex pavement backgrounds and various types of asphalt pavement cracks, including 3368 pavement crack images with 640 × 360 pixels, each of which has a corresponding binary image labeled with pixel-level cracks. Among them, 2244 images are used for training and 1124 images are used for testing.

(2) The DeepCrack dataset contains multi-scale and multi-scene concrete pavement cracks, including 537 concrete pavement crack images with 544 × 384 pixels and their corresponding crack labels. 300 images are used for training and 237 images are used for testing.

Before training the model, the dataset was extended. Each image in the Crack500 dataset was clipped at the circumference and center with 256 × 256. After clipping, we count the total number of pixels of cracks in each crack image and crack pixels less than 1000 were deleted, then the crack image was rotated at 4 different angles, 90° each time. Due to the small amount of DeepCrack dataset, in addition to the above operation, horizontal flip operation has been added. When testing, only crop and delete operations were performed. After data expansion, there are 21704 images for training and 3278 images for testing on Crack500 dataset, and 7888 images for training and 900 images for testing on DeepCrack dataset.

The experimental environment is Nvidia GeForce RTX3090 GPU, implemented on PyTorch framework. The unified size of input image for network model training and testing is 256 × 256 × 3.

4.1. Loss function

Pavement crack segmentation is a binary classification task, the binary cross entropy (BCE) loss function is used to measure the model’s ability to correctly classify each pixel at the pixel level. The BCE loss is defined as follows:

$$L_{BCE} = -\frac{1}{N}\sum_{i}^{N}[y_i \log(\widehat{y_i}) + (1 - y_i)\log(1 - \widehat{y_i})] \tag{2}$$

where N represents the total number of pixels contained in the image, y_i and $\widehat{y_i}$ represent the ground truth and prediction of point i, respectively.

Due to the small proportion of crack pixels, the BCE loss function will bias the network towards learning background features. The Dice loss function [35] focuses more on preserving the detailed information of cracks in the image, especially for detecting crack boundaries with good performance. The Dice loss is defined as follows:

$$L_{Dice} = 1 - \frac{2\sum_{i}^{N} y_i\widehat{y_i}}{\sum_{i}^{N} y_i + \sum_{i}^{N}\widehat{y_i}} \tag{3}$$

We combine the Dice loss function and BCE loss function. The combined loss function pays more attention on crack region, and can effectively eliminate the problem caused by the imbalance of positive and negative samples [36]. The loss function of combination [37] is defined as:

$$L_{Total} = L_{Dice} + \alpha L_{BCE} \tag{4}$$

where α is the weighting factor to balance the importance between the BCE loss function and the Dice loss function.

4.2. Evaluation indicators

In this paper, precision (Pr), recall (Re), and F1 score are used to evaluate the results of pavement crack detection, which are defined as:

$$Pr = \frac{TP}{TP+FP} \quad Re = \frac{TP}{TP+FN} \quad \mathrm{F1} = \frac{2Pr \times Re}{Pr+Re} \tag{5}$$

where TP refers the number of pixels correctly detected and classified as crack in the detection results, FP refers the number of background pixels misclassified as cracks, FN refers the number of pixels in the cracks misclassified as background.

4.3. Experimental details

In order to verify the effectiveness of the proposed method, the following three sets of experiments are designed: 1) Model comparison experiments: Each model was trained and tested on DeepCrack and Crack500 datasets respectively, and the F1 score and other indicators of each model were compared; 2) Comparison experiment of generalization: The models were trained on the Crack500 dataset, and then the DeepCrack testset was copied three times, one was enhanced the brightness by 1.5 times, one was reduced the brightness to 0.5 times, and the other remained the same. Then the three testsets were tested separately to compare the generalization of each model; 3) Ablation experiment: DeepCrack dataset was used to train and test the models that removed the feature fusion block and the dilated convolution block, and the influence of each block on the model was evaluated.

During the experiment, batchsize is set to 16, the epoch is set to 50, the initial learning rate is set to 1e-5 reduced by the decay rate 0.5 after every 5 epochs, the optimizer uses Adam, α is set to 0. 2.

4.4. Experimental results and analysis

To illustrate the effectiveness of the proposed method, other state-of-the-art ones were selected as comparative methods, including CNN-based methods, such as U-Net [15] and CrackSegNet [22], and Transformer-based methods, such as SETR [6], SegFormer [26] and SegCrack [27]. The SETR adopts two different decoder designs named SETR-MLA and SETR-PUP, respectively, the backbone of them are both ViT-B/16. The backbone of SegCrack and SegFormer are MiT-B2.

4.4.1. Model comparison experiment

Tables 1 and 2 present the quantitative results of seven models on the Crack500 dataset and the DeepCrack dataset. According to the results, we have the highest Pr and F1 score, although *Re* is slightly lower than CrackSegNet and U-Net, it is still higher than the other methods. Since U-Net and CrackSegNet sacrifice Pr for higher Re, objects in a wider range will be identified as crack in the actual detection process, it will lead to serious false positive problems. Our method comprehensively considers the importance of Pr and Re, achieves the highest F1 score of 70.84% and 84.50% on the two datasets, respectively, which are 1.42% and 2.07% higher than the second method.

Table 1. Comparison results of various methods on Crack500 dataset.

Methods	Pr	Re	F1
U-Net	64.49%	**81. 25%**	69.08%
CrackSegNet	65.64%	78.30%	68.43%
SegCrack	67.12%	77.64%	69.42%
SegFormer	64.53%	77.15%	67.44%
SETR-PUP	65.69%	78.48%	68.90%
SETR-MLA	65.69%	72.53%	66.29%
Ours	**68.06%**	79.11%	**70.84%**

Table 2. Comparison results of various methods on DeepCrack dataset.

Methods	Pr	Re	F1
U-Net	83.92%	85.92%	82.43%
CrackSegNet	74.59%	**91.63%**	79.94%
SegCrack	83.34%	82.21%	80.28%
SegFormer	82.61 %	82.54%	80.29%
SETR-PUP	82.50%	80.21%	79.35%
SETR-MLA	69.48%	75.31%	69.99%
Ours	**85.98%**	86.20%	**84.50%**

The segmentation results of various methods on the Crack500 dataset and DeepCrack dataset are shown in Figures 5 and 6. It can be seen that Transformer-based models, such as SegFormer, SETR-MLA and SETR-PUP are less affected by noise, because of larger receptive field, however, satisfactory results cannot be achieved when detecting slender crack. Benefited from its special structure, U-Net has the better segmentation result in slender crack by fusing feature maps of different layers, however, due to the small receptive field of the low-level feature map, there is still much noise in the segmentation results in Figures 4 and 5. Our method uses Transformer to extract crack features and fuses multi-scale features, which can effectively eliminate the negative impact of noise on performance and realize accurate segmentation of slender crack under complex background. It can be seen from precision-recall (P-R) curves in Figure 7, our method outperforms the other compared methods.

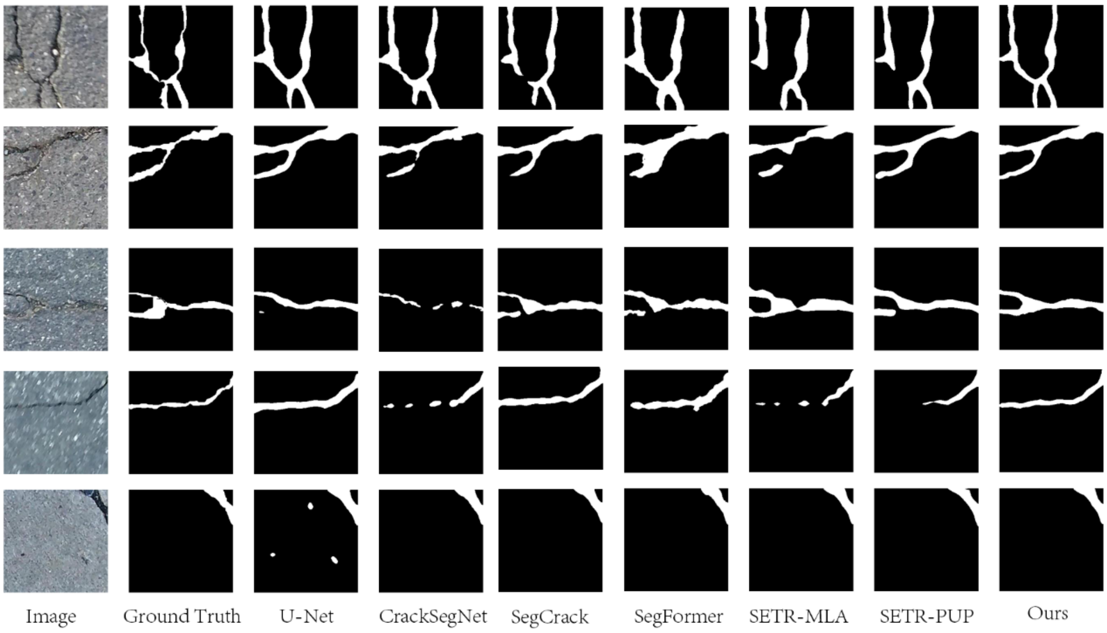

Figure 5. Detection results of different methods on the Crack500 dataset.

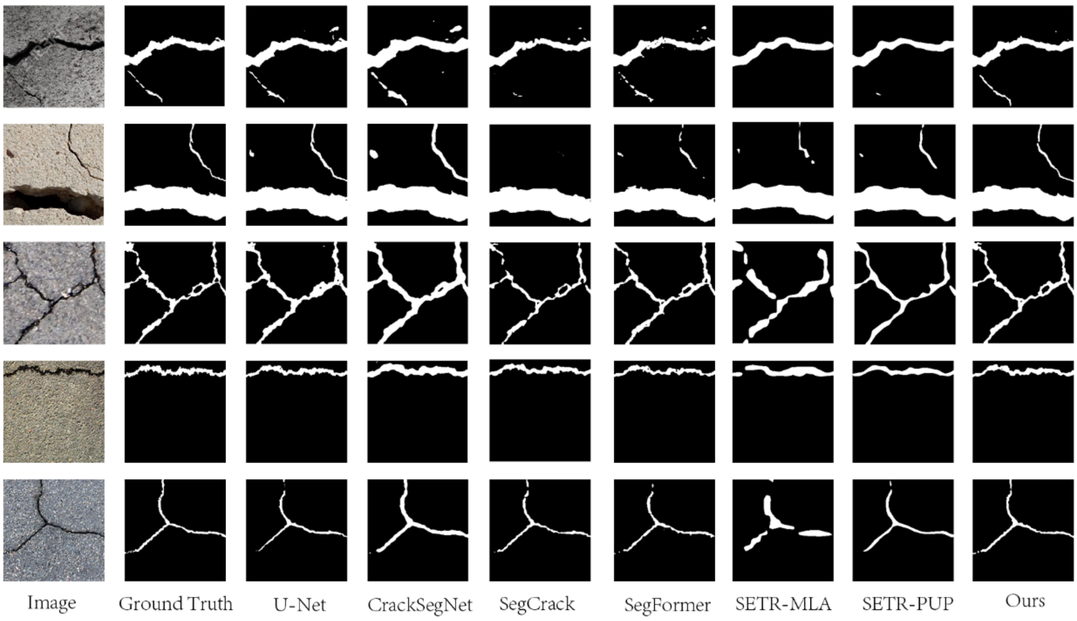

Figure 6. Detection results of different methods on the DeepCrack dataset.

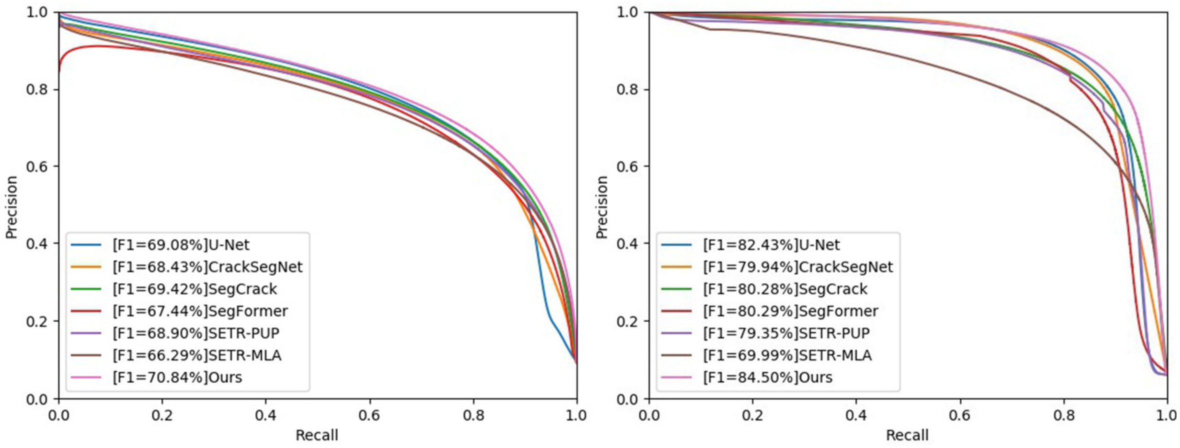

Figure 7. P-R curves on Crack500 dataset (left) and DeepCrack dataset (right).

4.4.2. Comparison experiment of generalization

To test the generalization ability of each model, after training the model on the Crack500 trainset, they are tested separately on the DeepCrack testset. The experimental results are shown in Table 3.

Table 3. Generalization experimental results on DeepCrack testset.

Methods	Pr	Re	F1
U-Net	56.11%	86.40%	61.75%
CrackSegNet	63.81%	81.44%	66.91%
SegCrack	65.24%	85.31%	70.32%
SegFormer	66.59%	85.01%	70.06%
SETR-PUP	65.13%	86.84%	71.93%
SETR-MLA	57.23%	81.77%	64.42%
Ours	**68.22%**	**90.25%**	**75.19%**

According to the results of generalization experiment on DeepCrack dataset, the Pr、Re and F1 score of the our method are 1.63%, 3.41% and 3.26% higher than the second method, respectively, demonstrating that the proposed model has high generalization ability.

In addition, considering crack detection is often affected by light in the actual work. In order to simulate different light conditions, the testset of DeepCrack is processed with brightness enhancement and brightness decrease, respectively, the processed testsets are shown in Figure 8. The two testsets are tested separately, and the experimental results are shown in Tables 4 Table 5.

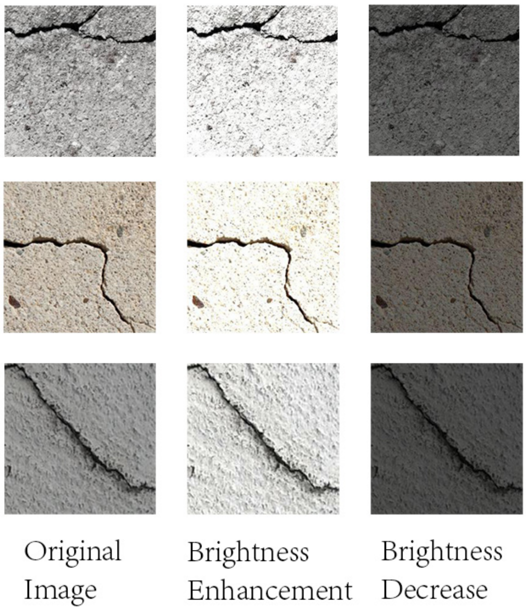

Figure 8. Crack images at different brightness.

Table 4. Results on DeepCrack testset after brightness enhancement.

Methods	Pr	Re	F1
U-Net	60.55%	72.69%	59.77%
CrackSegNet	52.84%	75.54%	56.71%
SegCrack	72.89%	74.98%	67.52%
SegFormer	**78.26%**	70.58%	67.98%
SETR-PUP	67.50%	83.37%	71.59%
SETR-MLA	53.89%	83.39%	62.21%
Ours	72.11%	**86.54%**	**75.78%**

Table 5. Results on DeepCrack testset after brightness decrease.

Methods	Pr	Re	F1
U-Net	38.82%	**89.24%**	48.05%
CrackSegNet	49.22%	44.01%	39.46%
SegCrack	**76.08%**	70.55%	65.70%
SegFormer	61.98%	75.20%	59.05%
SETR-PUP	49.81%	87.81%	59.94%
SETR-MLA	69.55%	53.13%	53.86%
Ours	70.29%	82.48%	**71.44%**

It can be seen from the experimental results that the CrackSegNet performs poorly when the brightness is high, while other models are less affected. In the case of low image brightness, all models are affected to some extent when detecting cracks. U-Net has the highest Re, however, its Pr and F1 score are also the lowest among all models. CrackSegNet is difficult to distinguish crack from the background when the image brightness is low, so all evaluation indicators are the lowest. Our method is the least affected, F1 score can reach more than 70%, satisfactory crack segmentation results can still be obtained under this condition. As shown in Figure 9, our method can accurately detect cracks from the background in both brightness enhancement and dark conditions, it is further proved that our method has better generalization and can achieve better detection results under different lighting conditions.

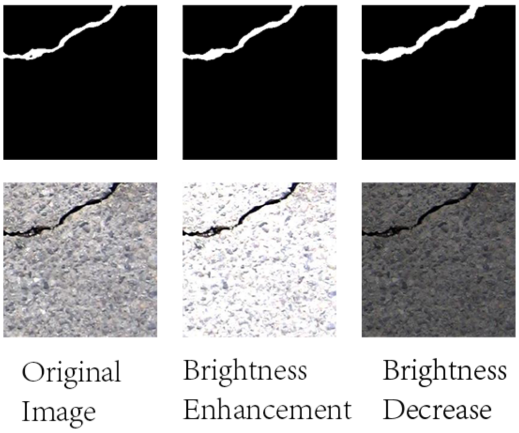

Figure 9. Crack segmentation results at different brightness.

4.4.3. Ablation experiment

To ascertain the effectiveness of the feature fusion and dilated convolution blocks, we performed ablation experiments on the DeepCrack dataset. The dilated convolution block is abbreviated as "DC", and the feature fusion block is denoted by "FF", The experimental results are detailed in Table 6.

Table 6. The results of ablation experiment on the DeepCrack dataset.

Methods	Pr	Re	F1
Ours	74.64%	**91. 21%**	80. 40%
Ours(DC)	84.53%	84 23%	82. 92%
Ours(FF)	83.77%	86. 94%	83. 49%
Ours(FF+DC)	**85. 98%**	86. 20%	**84. 50%**

Table 6 presents the F1 score of the model, which amounts to 80.40% when the feature fusion block and dilated convolution block are not integrated. However, the inclusion of these blocks individually results in an increase of 3.09% and 2.52% in the F1 score, respectively. Remarkably, when both blocks are integrated, the F1 score rises by 4.10%. These findings serve as compelling evidence to support the effectiveness of dilated convolution and multi-scale feature fusion for crack detection.

According to the above three experiments, it has been fully demonstrated that the proposed method in this paper has the advantages of high accuracy and high generalization, and has the ability to identify road cracks under different lighting conditions.

5. Conclusions

In this paper, a multi-scale feature fusion method for pavement crack detection based on Transformer is proposed. In the encoding stage, the ViT model is adopted as the backbone, modeling the crack images from a sequence-to-sequence perspective. Compared to convolutional neural networks, it has a larger receptive field and can capture global information, which can better extract the crack features, and the dilated convolution block is added to increase the receptive field of the feature map, to further obtain the context information. In the decoding stage, we propose a multi-scale feature fusion module. The linear layer is employed to adjust the length of the feature sequence output by different encoder blocks of Transformer model, and then it is converted into feature maps of different scales. By fusing multi-scale semantic features, the detailed information can be recovered, and the accuracy of crack detection can be improved. We compare our method with other methods on Crack500 and DeepCrack datasets, and F1 scores of our method are 70.84% and 84.50%, respectively, which are better than other methods. In addition, in the generalization experiments, the proposed method has better generalization ability and can accurately identify cracks under different light conditions. The effectiveness of the dilated convolution block and feature fusion block is verified by setting ablation experiments.

The proposed method can reduce labor costs, improve the detection efficiency and the accuracy of crack detection and can be extended to biomedical field such as the identification of retinal diseases from optical coherence tomography [38,39]. Although the proposed method has good performance, it also has some limitations. On the one hand, Transformer encoder needs a large amount of data for training. Manual annotation is time-consuming and subject to subjective influence, which will inevitably generate errors. On the other hand, the parameters and calculations of Transformer are very large, and we spent 10 hours training the model on the Crack500 dataset and 3.5 hours on the DeepCrack dataset, which took more time compared to training other models. Next, further investigation will be performed to improve the accuracy of crack detection, reduce the model's complexity and improve the detection efficiency of the model. In addition, exploration on other detection fields by using this method will be performed too.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62001004), the Key Research and Development Program of Anhui Province (202104A07020017), the Academic Funding Program for Top-Notch Talents of Disciplines (Majors) in Universities of Anhui Province (gxbjZD2022028), the Research Project Reserve of Anhui Jianzhu University (2020XMK04), the New Era Education Quality Project of Anhui Province (2022cxcysj143 and 2022cxcysj145). All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. Li, D. Song, Y. Liu, Automatic pavement crack detection by multi-scale image fusion, *IEEE Trans. Intell. Transp. Syst.*, **20** (2018), 2025–2036. https://doi.org/10.1109/TITS.2018.2856928
2. L. Hong, L. Mu, Survey on new progresses of deep learning based computer vision, *J. Data Acquis. Process.*, **37** (2022), 247–278.
3. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2015), 640–651. https://doi.org/10.1109/TPAMI.2016.2572683
4. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
5. A. Vaswani, N. Shazeer, N. Parmar, Attention is all you need, *Adv. Neural Inf. Process. Syst.*, **30** (2017), 5998–6008.
6. S. Zheng, J. Lu, H. Zhao, Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 6877–6886. https://doi.org/10.1109/CVPR46437.2021.00681
7. J. W. Wang, C. Li, X. Zhang, Surface crack detection of rubber insulator based on machine vision, in *14th International Conference on Intelligent Robotics and Applications (ICIRA)*, (2021), 175–185. https://doi.org/10.1007/978-3-030-89098-8_17
8. H. F. Li, Z. L. Wu, J. J. Nie, An automatic fine crack recognition algorithm for airport pavement under significant noises, *Comput. Eng. Sci.*, **42** (2020), 2020–2029.
9. C. Peng, M. Q. Yang, Q. H. Zheng, A triple-thresholds pavement crack detection method leveraging random structured forest, *Constr. Build. Mater.*, **263** (2020), 120080. https://doi.org/10.1016/j.conbuildmat.2020.120080
10. N. Kheradmandi, V. Mehranfar, A critical review and comparative study on image segmentation-based techniques for pavement crack detection, *Constr. Build. Mater.*, **321** (2022), 126162. https://doi.org/10.1016/j.conbuildmat.2021.126162
11. J. L. Chen, Concrete detection method based on image processing, *Chin. J. Liq. Cryst. Disp.*, **35** (2020), 395–401. https://doi.org/10.3788/YJYXS20203504.0395
12. Z. Othman, S. Zukfily, S. S. S. Ahmad, Road crack detection using modification of threshold values in Canny algorithm, in *7th Mechanical Engineering Research Day (MERD)*, (2020), 184–186.
13. J. Luo, H. Z. Lin, X. X. Wei, Adaptive canny and semantic segmentation networks based on feature fusion for road crack detection, *IEEE Access*, **11** (2023), 51740–51753. https://doi.org/10.1109/ACCESS.2023.3279888
14. S. Konishi, A. Yuille, J. Coughlan, Statistical edge detection: learning and evaluating edge cues, *IEEE Trans. Pattern Anal. Mach. Intell.*, **25** (2003), 57–74. https://doi.org/10.1109/TPAMI.2003.1159946

15. Z. Liu, Y. Cao, Y. Wang, Computer vision-based concrete crack detection using U-net fully convolutional networks, *Autom. Constr.*, **104** (2019), 129–139. https://doi.org/10.1016/j.autcon.2019.04.005
16. X. N. Cui, Q. C. Wang, J. P. Dai, Intelligent crack detection based on attention mechanism in convolution neural network, *Adv. Struct. Eng.*, **24** (2021), 1859–1868. https://doi.org/10.1177/1369433220986638
17. F. Liu, J. F. Wang, Z. Y. Chen, Parallel attention based UNet for crack detection, *J. Comput. Res. Dev.*, **58** (2021), 1718–1726.
18. X. W. Gao, B. R. Tong, MRA-UNet: Balancing speed and accuracy in road crack segmentation network, *Signal Image Video Process.*, **17** (2023), 2093–2100. https://doi.org/10.1007/s11760-022-02423-9
19. T. Chen, Z. Cai, X. Zhao, Pavement crack detection and recognition using the architecture of segNet, *J. Ind. Inf. Integr.*, **18** (2020), 100144. https://doi.org/10.1016/j.jii.2020.100144
20. V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2017), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615
21. Y. Liu, J. Yao, X. Lu, DeepCrack: A deep hierarchical feature learning architecture for crack segmentation, *Neurocomputing*, **338** (2019), 139–153. https://doi.org/10.1016/j.neucom.2019.01.036
22. Y. Ren, J. Huang, Z. Hong, Image-based concrete crack detection in tunnels using deep fully convolutional networks, *Constr. Build. Mater.*, **234** (2020), 117367. https://doi.org/10.1016/j.conbuildmat.2019.117367
23. H. Zhao, J. Shi, X. Qi, Pyramid scene parsing network, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2017), 2881–2890.
24. A. Dosovitskiy, L. Beyer, A. Kolesnikov, An image is worth 16 x 16 words: Transformers for image recognition at scale, preprint, arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929
25. H. Liu, X. Miao, C. Mertz, Crackformer: Transformer network for fine-grained crack detection, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 3783–3792. https://doi.org/10.1109/ICCV48922.2021.00376
26. E. Xie, W. Wang, Z. Yu, SegFormer: Simple and efficient design for semantic segmentation with transformers, *Adv. Neural Inf. Process. Syst.*, **34** (2021), 12077–12090.
27. W. Wang, C. Su, Automatic concrete crack segmentation model based on transformer, *Autom. Constr.*, **139** (2022), 104275. https://doi.org/10.1016/j.autcon.2022.104275
28. F. Guo, Y. Qian, J. Liu, Pavement crack detection based on transformer network, *Autom. Constr.*, **145** (2023), 104646. https://doi.org/10.1016/j.autcon.2022.104646
29. Z. Liu, Y. Lin, Y. Cao, Swin transformer: Hierarchical vision transformer using shifted windows, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 10012–10022.
30. X. Ju, X. Zhao, S. Qian, TransMF: Transformer-based multi-scale fusion model for crack detection, *Mathematics*, **10** (2022), 2354. https://doi.org/10.3390/math10132354
31. S. Khan, M. Naseer, M. Hayat, Transformers in vision: A survey, *ACM Comput. Surv.*, **54** (2022), 1–41. https://doi.org/10.1145/3505244

32. R. Xiong, Y. Yang, D. He, On layer normalization in the transformer architecture, in *International Conference on Machine Learning*, (2020), 10524–10533.
33. Q. Zhong, C. Wen, Concrete pavement crack detection based on dilated convolution and multi-features fusion, *Comput. Sci.*, **49** (2022), 192–196.
34. F. Yang, L. Zhang, S. Yu, Feature pyramid and hierarchical boosting network for pavement crack detection, *IEEE Trans. Intell. Transp. Syst.*, **21** (2019), 1525–1535. https://doi.org/10.1109/TITS.2019.2910595
35. R. Shamir, Y. Duchin, J. Kim, Continuous dice coefficient: A method for evaluating probabilistic segmentations, preprint, arXiv: 1906.11031. https://doi.org/10.1101/306977
36. G. Shu, L. Xiao, W. Xue, Crack detection based on enhanced semantic information and multi-channel feature fusion, *Comput. Eng. Appl.*, **57** (2021), 204–210.
37. C. Hui, R. Zhi, L. Yu, Research on tunnel crack segmentation algorithm based on improved U-Net network, *Comput. Eng. Appl.*, **57** (2021), 215–222.
38. R. K. Meleppat, K. E. Ronning, S. J. Karlen, In vivo multimodal retinal imaging of disease-related pigmentary changes in retinal pigment epithelium, *Sci. Rep.*, **11** (2021). https://doi.org/10.1038/s41598-021-95320-z
39. R. K. Meleppat, C. R. Fortenbach, Y. F. Jian, In vivo imaging of retinal and choroidal morphology and vascular plexuses of vertebrates using swept-source optical coherence tomography, *Translational Vision Sci. Technol.*, **11** (2022). https://doi.org/10.1167/tvst.11.8.11