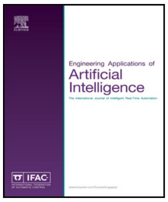
Engineering Applications of Artificial Intelligence 126 (2023) 106876

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



DeepCrackAT: An effective crack segmentation framework based on learning multi-scale crack features

Qinghua Lin [a], Wei Li [a], Xiangpan Zheng [b], Haoyi Fan [c,*], Zuoyong Li [d,*]

[a] School of Computer Science and Mathematics, Fujian University of Technology, 350118, Fuzhou, China
[b] College of Physics and Electronic Information Engineering, Minjiang University, 350121, Fuzhou, China
[c] School of Computer and Artificial Intelligence, Zhengzhou University, 450001, Zhengzhou, China
[d] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, College of Computer and Control Engineering, Minjiang University, 350121, Fuzhou, China

ARTICLE INFO

Keywords:
Crack detection
Attention mechanism
Hybrid dilated convolution
Deep learning

ABSTRACT

The detection of cracks is essential for assessing and maintaining building and road safety. However, the large appearance variations and the complex topological structures of cracks bring challenges to automatic crack detection. To alleviate the above challenges, we propose a deep multi-scale crack feature learning model called DeepCrackAT for crack segmentation, which is based on an encoder–decoder network with feature tokenization mechanism and attention mechanism. Specifically, we use hybrid dilated convolutions in the first three layers of the encoder–decoder to increase the network's receptive field and capture more crack information. Then, we introduce a tokenized multilayer perceptron (Tok-MLP) in the last two layers of the encoder–decoder to tokenize and project high-dimensional crack features into low-dimensional space. This helps to reduce parameters and enhance the network's ability of noise resistance. Next, we concatenate the features corresponding to the encoder–decoder layers and introduce the convolutional block attention module (CBAM) to enhance the network's perception of the critical crack region. Finally, the five-layer features are fused to generate a binary segmentation map of the crack image. We conducted extensive experiments and ablation studies on two real-world crack datasets, and DeepCrackAT achieved 97.41% and 97.25% accuracy on these datasets, respectively. The experimental results show that the proposed method outperforms the current state-of-the-art methods.

1. Introduction

The crack is one of the most common structural defects, widely present on the surface of structures such as residential buildings (Ai et al., 2023), pavements (Guo et al., 2023), and bridges (Li et al., 2023). Cracks that affect the safety of structures are mainly generated by insufficient load-bearing capacity. The continued development of these cracks will greatly affect the safety and service life of the structure. Therefore, the detection and analysis of cracks are significant for the evaluation and maintenance of building safety. Traditional crack detection relies mainly on manual inspection. However, for some geographically specific cracks in structures, manual crack detection faces high risks and costs, and there may be cases of missed detection. In addition, manual inspection of cracks depends on the professional level of surveyors and is difficult to standardize assessment. Therefore, the study of automatic crack detection methods has attracted widespread attention.

Researchers have previously suggested various solutions for crack detection using traditional image processing techniques (Kong and Li, 2019; Ogawa et al., 2019; Hsieh and Tsai, 2020). These include methods such as defect image segmentation and edge detection (Qingbo, 2016) for pavement crack detection, as well as crack detection based on central point (Lei et al., 2018). However, traditional crack detection methods rely on modeling specific noise fields, which results in reduced performance when the noise field changes. Consequently, when confronted with cracks in complex backgrounds, traditional methods often struggle to accurately identify crack features.

With the development of computer vision and convolutional neural networks, researchers have applied different deep learning methods to crack detection (Zhang et al., 2020, 2021; Han et al., 2021). For example, RCN (Wang et al., 2023) achieves precise segmentation of pavement defects by re-mapping the latent dependencies generated by the encoder at both local semantic and global detail levels. PCSN (Chen

* Corresponding authors.
E-mail addresses: Akametris@163.com (Q. Lin), liwei33660@163.com (W. Li), Zhengxp2018@163.com (X. Zheng), fanhaoyi@zzu.edu.cn (H. Fan), fzulzytdq@126.com (Z. Li).

https://doi.org/10.1016/j.engappai.2023.106876
Received 4 December 2022; Received in revised form 10 May 2023; Accepted 24 July 2023
Available online 9 August 2023

et al., 2020) employs a topless VGG16 encoder with the "Adadelta" optimizer to achieve precise detection of cracks in concrete pavements. The end-to-end trainable deep network DeepCrack (Zou et al., 2018) can further capture the line structure of cracks and achieve precise detection of bright crack images. Although deep learning-based crack detection methods have made promising progress, the segmentation of crack images still faces some challenges compared to common image segmentation tasks: (1) The background of crack images is complex, and background noise has a significant impact on crack segmentation performance. (2) Cracks are irregularly distributed in the image and may intersect or bifurcate, so complete segmentation of irregular cracks is still facing challenges. (3) Crack detection methods based on convolutional neural networks can well detect edge information of cracks. However, it exhibits poor segmentation performance for thick cracks.

To address these challenges, we propose a crack detection method called DeepCrackAT which is a five-layer encoder–decoder network. In the first three layers of encoder–decoder, we use the hybrid dilated convolutions to increase the receptive field of convolutional operations and capture more crack features, to achieve complete segmentation of irregular cracks. Then, in the last two layers of encoder–decoder, we introduce the Tok-MLP to tokenize high-dimensional crack features and project them into a low-dimensional space, reducing computational parameters and enhancing the model's ability of noise resistance. Next, we concatenate the extracted features from each layer in the attentional skip-layer fusion block and utilize CBAM to enhance the network's perception of critical crack regions, addressing the issue of information loss within thick cracks. Finally, the five-layer features are fused to generate a binary segmentation map of the crack image. Through these approaches, we learn multi-scale crack features and achieve accurate binary segmentation of crack images. The proposed method achieved accuracy of 97.41% and 97.25% on the two real-world crack datasets (i.e., Masonry Dais et al., 2021 and Rissbilder Pak and Kim, 2021), respectively. The experimental results demonstrate the effectiveness of DeepCrackAT.

In summary, our contributions can be summarized as follows:

- We propose a deep crack convolutional network, named DeepCrackAT, which utilizes an Attention mechanism to fuse multi-scale features from convolution and Tokenized MLP for crack segmentation.
- For irregularly distributed cracks, we use the hybrid dilated convolutions to increase the receptive field of convolutional operations and capture more crack features. Additionally, the proposed method employs a tokenized multilayer perceptron to project high-dimensional crack features into a low dimension space, enhancing the network's ability of noise resistance.
- We introduce the convolutional block attention module (CBAM) to construct an attentional skip-layer fusion block for multi-scale feature fusion. This helps to enhance the network's perception of the critical crack region and alleviate the problem of information loss in thick crack segmentation.
- We conduct extensive experiments and ablation studies on two real-world crack datasets, and the results show that our method outperforms other state-of-the-art methods.

2. Related work

2.1. Traditional methods

Many works apply traditional image processing techniques to crack detection, Qingbo (2016) from the angle of grey level, median filter, and image intensification to improve the pavement crack image algorithm. Lei et al. (2018) proposed a crack central point method combine with an unmanned aerial vehicle to identify the cracks under a small number of crack images. A vision-based non-contact method through image overlapping for fatigue crack detection was proposed by Kong and Li (2019). Yu et al. (2017) present a three-step method to identify and extract cracks from infrared images, which includes process images in the frequency domain, computing the conditional texture anisotropy, and connecting the crack subregion. There are many other crack detection methods based on traditional image processing technology. However, these unsupervised methods do not require the manually labeled ground truth, and performance is mediocre in complex crack backgrounds. Compared with the deep learning-based method, the crack detection method based on traditional image processing technology is weak in robustness.

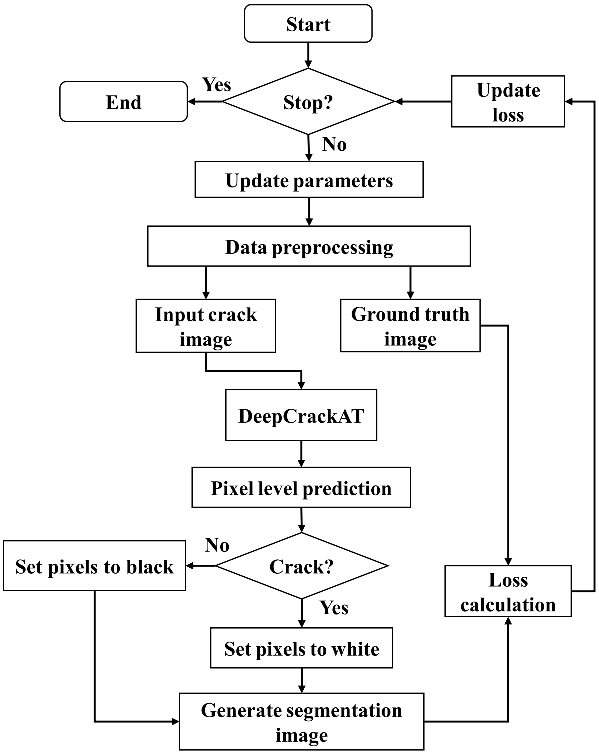


Fig. 1. The training flowchart of DeepCrackAT.

2.2. Deep learning-based methods

With the development of computer vision and convolutional neural networks, deep learning has achieved remarkable results in areas such as fault classification (Fantin Irudaya Raj and Balaji, 2021), anomaly detection (Pang et al., 2021) and object detection (Liu et al., 2020; Zhao et al., 2019). Given the good results achieved by deep learning technology in various fields, more and more researchers are applying deep learning to crack detection and achieving promising results. Crack detection can be regarded as a pixel-level semantic segmentation task. Zou et al. (2018) use deep convolution to capture the crack features and fuse the multi-scale crack feature maps to obtain a more detailed crack prediction map. Yuan et al. (2022) introduce residual detail attention to capture the line structure and accurately locate the crack position. Zhang et al. (2021) propose an Unet-based crack detection algorithm CrackUnet and adopt generalized dice loss to detect cracks more accurately. The fully convolutional network (FCN) is used to address the crack detection problem (Dung and Anh, 2019; Yang et al., 2018). Sun et al. (2022) enhance the DeepLabv3+ with a multi-scale attention module to combine multi-scale crack features. Zhang et al. (2020) propose CrackGAN based on generative adversarial learning to overcome the "All Black" issue. Inoue and Nagayoshi (2021) treat

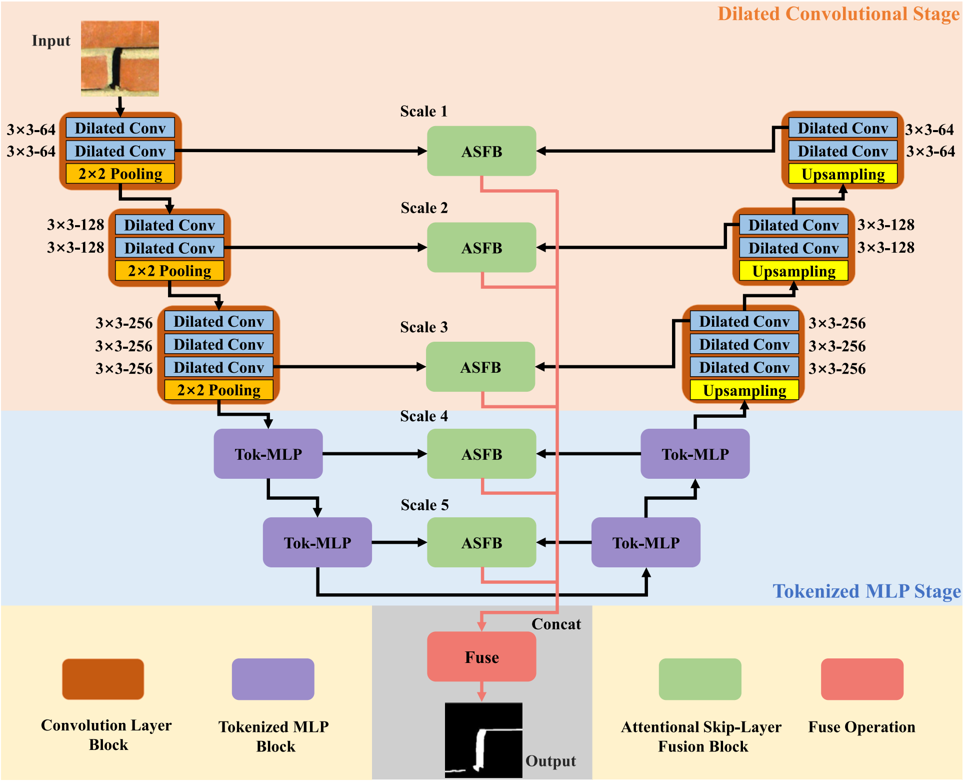


Fig. 2. An illustration of the DeepCrackAT network. The crack image is used as the input to the network. Firstly, the image passes through the first three layers of the encoder–decoder (Dilated Convolutional Stage) to extract low-dimensional features. Secondly, in the last two layers of the encoder–decoder (Tokenized MLP Stage), we utilize Tok-MLP to obtain high-dimensional features of the cracks. The features from the corresponding layers of the encoder–decoder are fused by Attentional Skip-Layer Fusion Block. Finally, the binary segmentation image of the crack is obtained by concatenating features from different scales.

crack detection as a weakly-supervised problem and propose a two-branched framework for crack detection. PCSN (Chen et al., 2020) is a fully convolutional neural network based on SegNet which could accept images of arbitrary size as input data. Generally, the crack detection method based on deep learning is usually superior to traditional crack detection methods in terms of results.

2.3. Attention mechanism

As one of the most critical perceptual contents of human beings, the attention mechanism has been popular in deep learning in recent years. Especially, the transformer is highly respected by researchers for its excellent performance. Hu et al. (2018) focus on the channel relationship and propose a "Squeeze-Excitation" block to boost the representational power of the network. The channel attention mechanism establishes the correlation between channel and feature by learning the weight of multi-dimension. Jaderberg et al. (2015) introduce a spatial transformer that allows spatial manipulation of data in the network and the spatial attention mechanism can effectively process multi-scale spatial features. Convolutional block attention module (Woo et al., 2018) as an extended application of channel attention and spatial attention, the weight is assigned to channel and space respectively, which not only reduces the number of parameters but also has good compatibility. We introduce the CBAM to help the network pay attention to crack features and enhance the crack segmentation performance.

3. Method

In this section, we first introduce the network structure of DeepCrackAT detailed. Then, we illustrate the hybrid dilated convolution and Tok-MLP respectively. The skip-layer fusion block based on the attention mechanism is expounded as followed. Finally, the loss function of this work is defined. Before detailed explanations, we also provide a detailed crack segmentation process flowchart as shown in Fig. 1 to give a preliminary understanding of the workflow.

3.1. Network architecture

In this work, DeepCrackAT dedicates to building a better-performance crack segmentation network. As shown in Fig. 2, DeepCrackAT is a pixel-level segmentation network based on an encoder–decoder structure. Traditional crack detection encoder network includes 13 convolution layers and 5 pooling layers, each layer in the decoder corresponds to a layer in the encoder. The difference between the encoder and decoder is that the first convolution layer of the encoder outputs a multi-channel feature map of the crack image, while the output of the last convolution layer in the decoder is a one-channel mapping.

To increase the receptive field of the convolution layer, we replace the convolution with the dilated convolution in the first 3 convolution stages and use the Tok-MLP block to reduce the number of parameters and computational complexity of the network. The Tok- MLP block first moves the feature window across width or height, then uses a depth-wise convolution for positional information encoding and GELU as an activation function, and finally passes the feature maps to the next module after layer normalization. In the Tok-MLP block, layer normalization performs better than batch normalization (Valanarasu and Patel, 2022). Furthermore, the attentional skip-layer fusion block (ASFB) is added to the encoder–decoder network to retain the crack features captured at different scales. Finally, we concatenate crack features of five scales and then pass a fuse operation with 1 × 1 convolution to output the crack segmentation result. Specific details of DeepCrackAT are shown in the following sections.

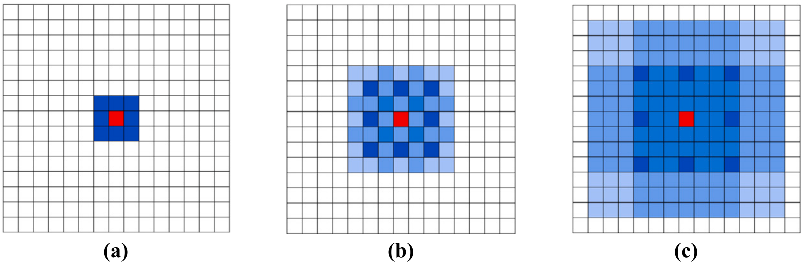


Fig. 3. The receptive field of successive convolution with different dilation rates, (a) Dilation rate = 1, (b) Dilation rate = 2, (c) Dilation rate = 3.

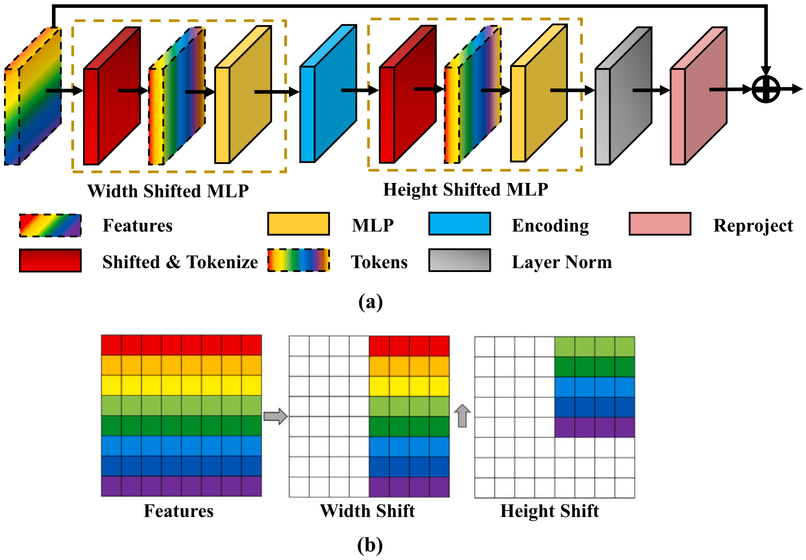


Fig. 4. The features are shifted in width and height to induce window locality, (a) Overview of the Tok-MLP block, (b) Shifting operation.

3.2. *Encoder–decoder network*

Multi-scale fusion map can effectively improve the performance of the crack segmentation (Zou et al., 2018). However, existing crack segmentation networks have not fully utilized the multi-scale crack semantic features. Therefore, we reconstructed the encoder–decoder network, learning more robust crack feature representations at different scales to improve the performance of our network in different crack scenarios. The motivation behind this is that low-dimensional features can capture local information such as crack details and textures in the image, but they are easily affected by background noise, which is common in crack images. High-dimensional features can describe the global structure of the crack image well and have strong noise resistance. Therefore, by combining crack features from different dimensions, we can achieve good noise reduction effects and achieve precise crack pixel segmentation.

3.2.1. *Convolution layer block*

The first three stages of the DeepCrackAT network are composed of continuous convolution layers, but in most cases, crack pixels account for a small proportion of the whole background pixels and are widely distributed. Ordinary convolution has the problem of a limited receptive field for crack detection, while dilated convolution has a poor effect on thin crack segmentation. Therefore, we adopted the hybrid dilated convolution as shown in Fig. 3. On the one hand, it retains more information while enlarging the receptive field, and on the other hand, it could avoid losing pixels due to the influence of the grid effect during continuous convolution operation.

In addition, batch normalization operation has been proven to be beneficial to prevent gradient disappearance or explosion (Ioffe and Szegedy, 2015). Therefore, the batch normalization layer is added to improve the training speed and generalization performance of the network after each convolution operation. Then to reduce the large number of parameters generated by deep convolution, the max-pooling operation is used to reduce the size of feature maps. When downsampling in the encoder, we use the maximum pooling index to obtain and save the boundary content of the feature maps for crack details preservation. In the decoder network, we use the maximum pooled index to perform nonlinear up-sampling to generate sparse feature maps.

3.2.2. *Tokenized MLP block*

In the initial and final blocks of DeepCrackAT, we use continuous convolution to capture crack features. In the bottleneck, we use Tok-MLP to project convolutional features into tokens to segment crack features. Shifting operation is included in the MLP to extract local information based on axial shifts. Compared with continuous convolution operation, MLP can significantly reduce the number of parameters and computational complexity while maintaining good performance.

The Tok-MLP block is shown in Fig. 4. The Tok-MLP block includes a Width Shifted MLP module, depth-wise convolution layer, Height Shifted MLP module, and layer normalization layer. The number of channels is changed to E which denotes the number of tokens for better tokenizing. Firstly, features are first sent to Width shifted MLP module. Before tokenizing, the axis of channels of features is shifted across the width, which helps MLP to add more locality. It is similar

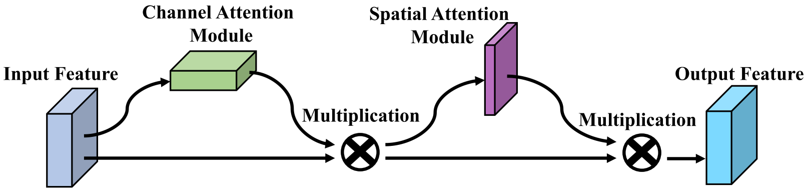


Fig. 5. An illustration of convolutional block attention module.

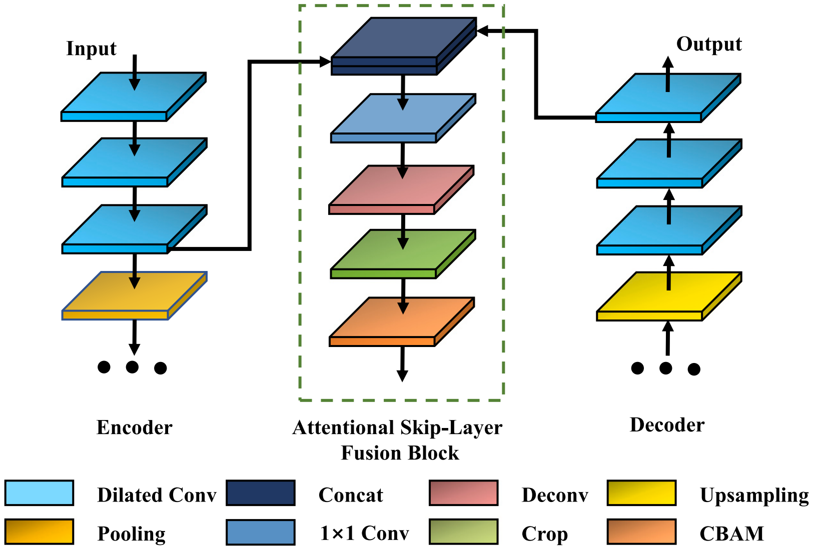


Fig. 6. An illustration of attentional skip-layer fusion block.

to Swin transformer (Liu et al., 2021) which introduced the window-based attention to add more locality (Valanarasu and Patel, 2022). Furthermore, depth-wise convolution contributed to encoding the MLP feature's positional information and improves efficiency by using fewer parameters. Then the GELU (Hendrycks and Gimpel, 2016) as the activation layer instead of ReLU for better performance. We then pass the features into the Height Shifted MLP module to shift across height. Finally, features are concatenated to the original features after layer normalization (Ba et al., 2016) and output to the next block.

3.3. Attentional skip-layer fusion block

In the traditional encoder–decoder network for crack segmentation, skip-layer fusion (Zou et al., 2018; Yuan et al., 2022) is often used to obtain the crack features from the shallow layer to the deep layer. Therefore, the output feature maps of each convolution stage in the encoder are concatenated with the corresponding feature maps in the decoder. However, in most cases, the distribution of crack pixels is a minority, so we introduce the attention mechanism into the skip-layer fusion block to better focus on the crack feature of fusion feature maps.

CBAM (Woo et al., 2018) is shown in Fig. 5, it contains a channel attention module and a spatial attention module. Firstly, the input features are passed into channel attention module to learn the dependencies of channels, then the features are multiplied with origin features as output to the next module. Secondly, the features are passed to spatial attention module to obtain weights of spatial dimensions. Finally, multiplication is used to features from spatial attention module and its input features to obtain the refined features. Segmentation of crack images can be considered as a binary classification task, with the goal of dividing the pixels in the crack image into crack and non-crack pixels. The channel attention module weights different channel features to enable the model to focus on the channel features that are useful for crack segmentation. Similarly, the spatial attention module weights the spatial location of cracks to enhance the model's attention to crack regions. By combining channel attention and spatial attention, CBAM can help the model achieve more accurate crack segmentation and the ablation study demonstrated its effectiveness.

Attentional skip-layer fusion block as shown in Fig. 6 includes a 1 × 1 convolution layer, deconvolution layer, crop layer, and CBAM. Firstly, the feature obtained by the encoder–decoder network are concatenated, then the 1 × 1 convolution layer is used to convert the feature that is multi-channel to 1 channel. Secondly, we use a deconvolution layer to upsample the feature and then crop it to the size of the original image. Then, the CBAM is applied to enhance the performance of crack segmentation. Finally, the crack feature is sent to the next block for generating the crack segmentation result.

3.4. Loss function

In crack segmentation, we define a training dataset $T = \{(I^n, M^n), n = 1, \ldots, N\}$, which contains N images, and each image has P pixels. $I^n = \left\{i_p, p = 1, \ldots, P\right\}$ represents the input crack image, where i_p denotes the pixel value of the crack image, and $M^n = \left\{l_p, p = 1, \ldots, P, l_p \in \{0, 1\}\right\}$ represents the corresponding label image, where l_p denotes the label of the pixel. Assuming the number of convolution stages is S, the feature map generated by the skip-layer in the s convolution stage is denoted as:

$$F_p^{(s)} = \left\{ f_p^{(s)}, p = 1, \ldots, P \right\} \tag{1}$$

where $f_p^{(s)}$ denotes the feature generated by the model for the pth pixel in the sth convolution stage, and $s = 1, \ldots, S$. Therefore, the fusion of

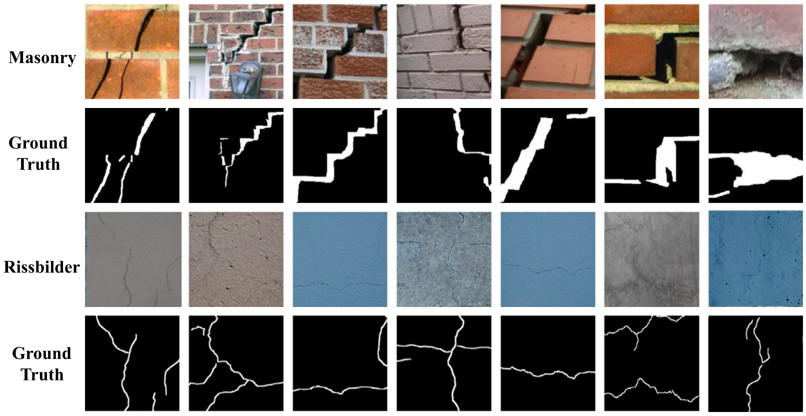


Fig. 7. Some typical crack samples and their corresponding ground truths in Masonry and Rissbilder datasets.

features at multiple scales can be expressed as:

$$F_p^{fuse} = Concat\left(F_p^{(1)}, F_p^{(2)}, \dots, F_p^{(S)}\right) \tag{2}$$

where $Concat(\cdot)$ denotes the concatenation operation.

The output of the crack segmentation has only two classes, which are the background pixel and the crack pixel. So, it can be treated as a binary classification problem. We use a cross-entropy loss to measure the prediction error. In most cases, the crack pixel takes a small proportion of the entire image which means an imbalance of segmentation. Some segmentation works (Liu et al., 2017; Zou et al., 2018; Yuan et al., 2022) to solve this problem by a weighted method. However, through the experiment, we find that the larger weights of the crack lead to the worst performance of crack segmentation. So, we define the loss function of pixel prediction as:

$$l\left(F_p; W\right) = \begin{cases} \log\left(1 - S\left(F_p; W\right)\right), & \text{if } y_p = 0, \\ \log\left(S\left(F_p; W\right)\right), & \text{otherwise}, \end{cases} \tag{3}$$

where F_p is the feature of network's output, W denotes the set of parameters in the network layer, and $S(F)$ is the sigmoid function which converts the feature map to prediction map. Finally, we define the total loss as:

$$\mathcal{L}(W) = \sum_{p=1}^{P}\left(\sum_{s=1}^{S} l\left(F_p^{(s)}; W\right) + l\left(F_p^{fuse}; W\right)\right) \tag{4}$$

4. Experiments

In this section, we first introduce the implement details, datasets, evaluation metrics, and comparison methods. Then, experimental results are presented to prove the effectiveness of the proposed DeepCrackAT, by comparing it with the advanced segmentation network on two crack image datasets.

4.1. Implement details

During the network training, the initial global learning rate is set to 1e-3 and in every iteration 10 K decays to 1/10 of the original value. The momentum decay is 0.9 and the weight decay is 0.0005, we set the batch size to 20 and each algorithm runs 100 epochs. Adaptive moment estimation (Adam) (Kingma and Ba, 2014) method is used to update the network parameters. Moreover, batch normalization is added after each convolution to prevent overfitting and accelerate convergence speed. We implement our method in Python 3.9 and PyTorch 1.12.1. An 8-core computer with an AMD Ryzen 7 5800X 3.8 GHz CPU (32 GB RAM) and NVIDIA GeForce RTX 3090 24G GPU is used for training and testing.

4.2. Datasets

To verify the performance of the network, we conducted experiments on two public datasets. To unify the size of the input images, we crop all the images in the dataset to 224 × 224 resolution randomly. Furthermore, we augment the images of these datasets to ensure the network performs best. Each dataset is divided into a training set, validation set, and test set in a ratio of 6:2:2.

Masonry (Dais et al., 2021) contains 240 masonry crack images captured with mobile phones or DSLR cameras from various masonry buildings in the Groningen region of the Netherlands. Each crack image with pixel-level annotated binary images. The original size of images is 224 × 224, and a total of 1440 images were obtained after 90°, 180°, and 270° rotation, flip, and brightness enhancement. The training dataset contains 864 images, validation, and test dataset each contains 288 images.

Rissbilder (Pak and Kim, 2021) contains 2736 concrete crack images captured with various devices in Florian, Germany. It contains many wall crack images and a few ground crack images. The original size of images is 448× 448 and we random crop size to 224 × 224. The training dataset contains 1642 images, validation, and test dataset each contains 547 images.

Fig. 7 shows some representative samples in the datasets we used including crack images and their corresponding labels. Cracks of the Masonry dataset are typical thick cracks and some cracks are accompanied by large width variations. The Rissbilder dataset is a challenging thin crack dataset, the crack structure is more complex and multiple cracks interlace.

4.3. Evaluation metrics

To evaluate the performance of networks, we introduce the evaluation index of semantic segmentation. Suppose k is the number of segmentation classes, P_{ij} denotes the prediction of class i pixels into class j pixels.

Mean Intersection over Union (MIoU) denotes the indicator quantifying the percentage overlap between the target mask and the predicted mask:

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{P_{ii}}{\sum_{i=0}^{k} P_{ij} + \sum_{i=0}^{k} P_{ji} - P_{ii}} \tag{5}$$

The Precision can be obtained by the ratio between the number of positive samples predicted by the model and the number of correct samples predicted:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

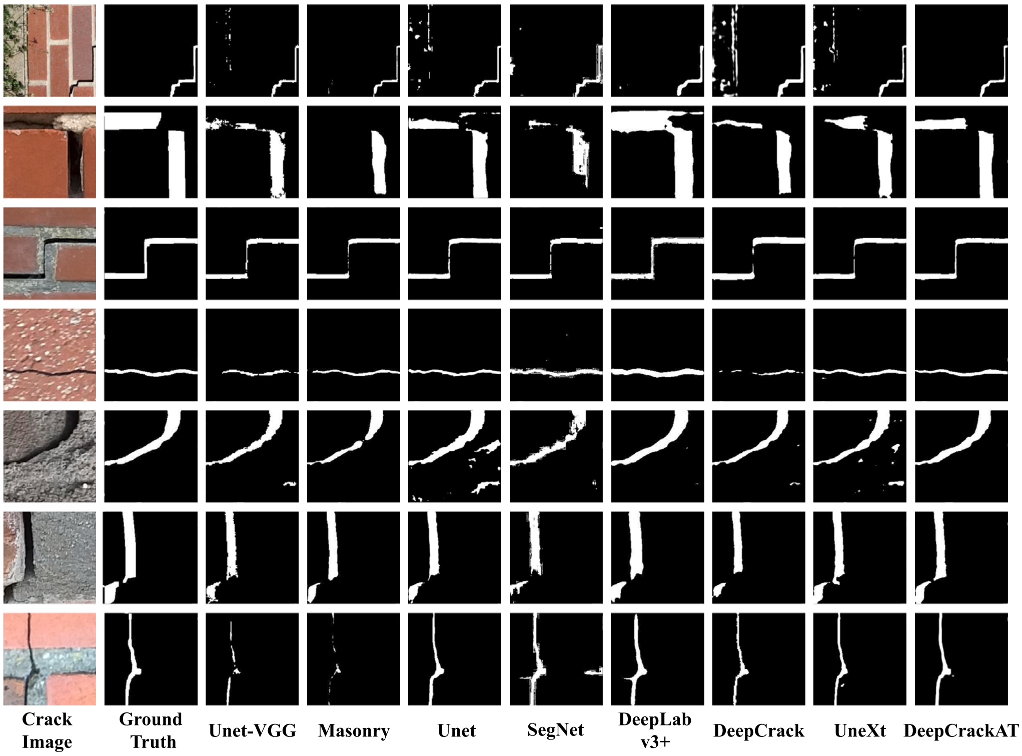


Fig. 8. Crack segmentation results of each network on the Masonry dataset.

Table 1
Comparison of different networks on the Masonry dataset.

Methods	Precision [%]	Accuracy [%]	F1-score [%]	Recall [%]	MIoU [%]
Masonry (Dais et al., 2021)	65.93	97.48	53.30	49.09	40.82
U-net-VGG (Ronneberger et al., 2015)	65.40	97.38	48.87	43.44	36.94
U-net (Ronneberger et al., 2015)	88.34	**97.56**	57.71	47.44	44.97
SegNet (Badrinarayanan et al., 2017)	75.02	80.56	65.43	62.87	50.38
DeepLabv3+ (Chen et al., 2018)	65.93	97.29	46.12	40.86	48.53
DeepCrack (Zou et al., 2018)	86.99	96.15	64.87	54.91	51.51
UneXt (Valanarasu and Patel, 2022)	86.06	96.74	68.47	62.34	56.41
DeepCrackAT	**88.62**	97.41	**76.67**	**71.24**	**64.44**

Table 2
Comparison of different networks on the Rissbilder dataset.

Methods	Precision [%]	Accuracy [%]	F1-score [%]	Recall [%]	MIoU [%]
Masonry (Dais et al., 2021)	41.20	33.01	33.23	30.48	35.80
U-net-VGG (Ronneberger et al., 2015)	45.80	34.24	28.96	22.75	30.03
U-net (Ronneberger et al., 2015)	65.16	92.26	39.25	31.97	26.76
SegNet (Badrinarayanan et al., 2017)	56.21	81.82	59.42	**67.82**	43.66
DeepLabv3+ (Chen et al., 2018)	28.59	96.91	36.18	52.72	22.44
DeepCrack (Zou et al., 2018)	64.37	97.15	55.43	50.56	42.26
UneXt (Valanarasu and Patel, 2022)	60.84	96.94	56.09	55.17	40.50
DeepCrackAT	**67.31**	**97.25**	**60.18**	55.90	**44.27**

where the true positive (TP) refers to the input data whose class is positive instance, and the model predicts it as positive instance. For the crack segmentation task, TP means that the pixel in the input image is a crack pixel, and the model predicts it as a crack pixel. Similarly, the false positive (FP) represent cases where the input data is negative, but the model predicts it as positive. In this task, FP means that the pixel in the input image is a non-crack pixel, but the model predicts it as a crack pixel.

The Accuracy is obtained by predicting the proportion of correct samples in all samples:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{7}$$

where the true negative (TN) refers to the input data being a negative instance, and the model predicts it as a negative instance. In this study, TN corresponds to non-crack pixels in the input image, and the model can predict them correctly. Conversely, the false negative (FN) indicates that the input data is a positive instance, but the model incorrectly identifies it as a negative instance. FN means that the model classifies crack pixels as non-crack pixels.

The Recall denotes the positive class proportion of the prediction image in the real crack label image:

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

The F1-score is used as a comprehensive evaluation indicator of Precision and Recall:

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{9}$$

4.4. Comparison methods

To verify the performance of the network proposed in this work, we compare it with the better segmentation methods in recent years, such as DeepCrack (Zou et al., 2018), Masonry (Ronneberger et al., 2015), and UneXt (Valanarasu and Patel, 2022). During the comparative experiment, we also compare classical deep learning semantic segmentation models, including Unet (Valanarasu and Patel, 2022),

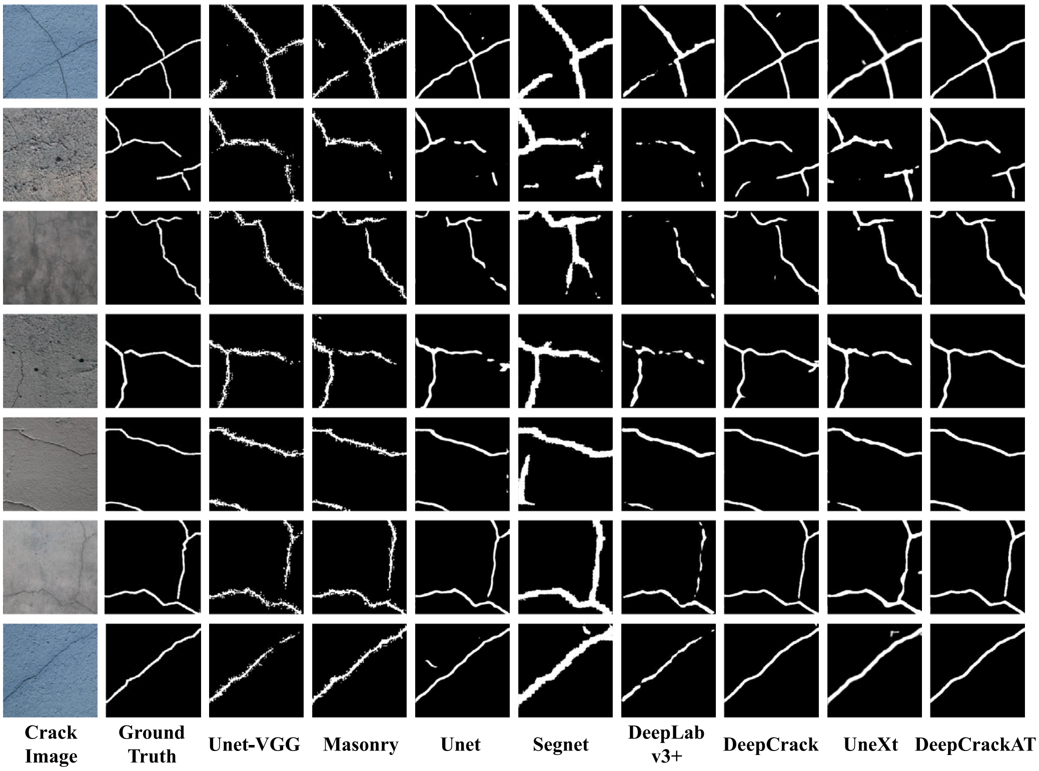


Fig. 9. Crack segmentation results of each network on the Rissbilder dataset.

SegNet (Badrinarayanan et al., 2017), and DeepLabv3+ (Chen et al., 2018). Details of those baselines are as follows:

- **DeepCrack** (Zou et al., 2018) is a deep supervised network base on the encoder–decoder network, it fuses multiple scale crack features to predict the crack.
- **UneXt** (Valanarasu and Patel, 2022) introduced the tokenized multilayer perceptron to Unet, UneXt reduced the number of parameters and computational complexity.
- **Masonry** (Dais et al., 2021) is a transfer learning base on Unet-moblienet (Ronneberger et al., 2015) which performs well in the crack segmentation of masonry.
- **Unet** (Ronneberger et al., 2015) is an end-to-end fully convolutional network, it has excellent performance on the dataset with a small number of images but a large image size.
- **DeepLabv3+** (Chen et al., 2018) uses atrous spatial pyramid pooling to extract features and a decoder is used to refine the segmentation result.
- **SegNet** (Badrinarayanan et al., 2017) is a segmentation network base on an encoder and decoder, its encoder part uses 13 convolution layers to capture the features. The decoder output is passed to softmax for prediction.

We conduct experiments on these datasets and obtain the evaluation metrics of each network. As shown in Table 1, the method proposed in this work outperforms other comparison methods on the Masonry dataset. DeepCrackAT has the best Precision, F1-score, Recall, and MIoU whose values are 88.62%, 76.67%, 71.24%, and 64.44%. Compared to UneXt and DeepCrack, there is an 8.2% and 11.8% improvement in the F1-score, respectively. Unet has the highest Accuracy with 97.56%, and the second rank is Masonry and DeepCrackAT with 97.48% and 97.41%. However, other segmentation methods do not perform well in this dataset. Although the Precision and MIoU of DeepLabv3+ reached 65.93% and 48.53%, other metrics are not satisfactory. The experimental results of Rissbilder are shown in Table 2. This dataset is very challenging and contains a large number of crack images in a complex background, which degrades the performance of the segmentation networks. DeepCrackAT has the best Precision, Accuracy, F1-score, and MIoU with 67.31%, 97.25%, 60.18%, and 44.27%, it still outperforms other methods. The second rank is DeepCrack, UneXt, and SegNet with F1-score 55.43%, 56.09%, and 59.42%, respectively.

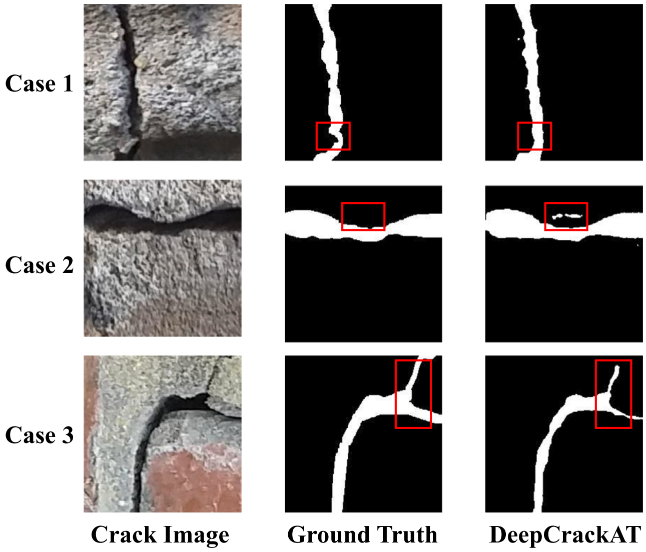


Fig. 10. Some failed segmentation examples.

Table 3
Ablation analysis. “Weight” denotes that a weighted loss function is used, “w/o CBAM” indicates that CBAM is eliminated from skip-layer, “w/o HDC” indicates that features are extracted using ordinary convolution, “w/o Tok-MLP” indicates that Tok-MLP is replaced by a continuous convolution structure.

Methods	Masonry			Rissbilder		
	Precision [%]	F1-score [%]	MIoU [%]	Precision [%]	F1-score [%]	MIoU [%]
Weight	61.39	66.34	53.37	31.73	38.78	24.58
w/o CBAM	83.06	67.57	56.27	50.85	33.98	23.90
w/o HDC	84.51	68.57	57.29	58.99	31.88	21.92
w/o Tok-MLP	82.80	46.80	36.36	63.37	40.95	28.47
DeepCrackAT	**88.62**	**76.67**	**64.44**	**67.31**	**60.18**	**44.12**

We train these comparison methods without pre-trained. Notably, Masonry, U-net-VGG, and DeepLabv3+ do not perform well on these two datasets. Too random weight of backbone results in weak feature extraction ability and poor network training effect. Moreover, the absence of fully utilized multi-scale crack features also is a reason for unsatisfactory segmentation performance.

Figs. 8 and 9 show the segmentation results of each network on the two datasets. The Unet, UneXt, DeepCrack, and DeepCrackAT all have good crack detection capability for thin cracks. SegNet tends to capture more non-crack pixels as crack pixels. DeepLabv3+ extracts more non-crack pixels when facing thick cracks but loses some crack pixels when facing thin cracks. DeepCrack extracts the edge information of a thick crack but loses the pixels inside. The segmentation of thick cracks by UneXt and DeepCrackAT is relatively complete, but some crack pixels are still lost. Furthermore, DeepCrackAT effectively eliminates noise in the complex crack background. However, SegNet, Unet, and UneXt are sensitive to the noise in the background of complex cracks.

Although our method performs well on most crack image segmentation, it did not achieve the expected results when dealing with continuous cracks with width variations. We provide some failed segmentation cases in Fig. 10. In Case 1, although DeepCrackAT achieved a relatively complete segmentation, the small changes in the red box were not well captured. In addition, when facing continuous cracks with width variations like in Case 2, our method made some errors in segmentation. This situation appeared more severe in Case 3, where the model lost some crack pixels when the width changed dramatically. The reasons for these failures may be: (1) the receptive field size is still not sufficient to handle crack images with large width variations, and (2) the lack of such images and corresponding ground truth labels in the training set. We will improve our method in future work to achieve better crack segmentation.

4.5. Ablation analysis

We set different weights for crack pixels and non-crack pixels to verify the effectiveness of the experimental weights. We redefine the weighted loss function as follows:

$$\mathcal{L}(F_p; W) = \begin{cases} w_0 \cdot \log(1 - S(F_p; W)), & if\ y_i = 0, \\ w_1 \cdot \log(S(F_p; W)), & otherwise, \end{cases} \tag{10}$$

where w_0 and w_1 are the weights of non-crack and crack pixels, where c_0 and c_1 denotes the number of non-crack pixels and crack pixels in the training set. When the pixel is negative $w_0 = 1$ otherwise $w_0 = c_0/c_1$.

The ablation study is shown in Table 3, results show that adding more weight to the non-crack background leads to the performance degradation of DeepCrackAT. Wrong predictions penalize the network more severely so that more pixels are predicted to be crack, and metrics down drastically. This deterioration is more obvious to a thin crack on the Rissbilder dataset. Thus, setting unbalanced weights for the loss function is not helpful for performance improvement.

Table 3 also provides the ablation results of CBAM, hybrid dilated convolution, and Tok-MLP, for which we demonstrated its effectiveness by replacing or removing these modules in the model. The experimental results show that CBAM and hybrid dilated convolution help to improve the crack segmentation ability of the model. Notably, when the continuous convolution structure replaces Tok-MLP to extract crack features, the Precision drops slightly while the F1-score and MIoU drop sharply.

Table 4
Ablation study of Tok-MLP deploys at a different location.

Location		Masonry			Rissbilder		
Scale 4	Scale 5	Precision [%]	F1-score [%]	MIoU [%]	Precision [%]	F1-score [%]	MIoU [%]
		82.80	46.80	36.36	63.37	40.95	28.47
	✓	82.39	55.26	43.98	**68.92**	29.85	19.75
✓	✓	**88.62**	**76.67**	**64.44**	67.31	**60.18**	**44.12**

Therefore, as shown in Table 4 we conducted additional ablation experiments on the Tok-MLP module. By using Tok-MLP to extract fracture features at different scales, we found that the model achieved better performance when using Tok-MLP instead of continuous convolution structure at Scale 4 and Scale 5.

4.6. Applications

To further evaluate the actual performance of DeepCrackAT, we applied it to crack detection in real-world scenarios. Fig. 11 shows some crack image acquisition equipment, as well as crack images collected by us using smartphones and DSLR cameras around Fuzhou, China. The cracks were found on surfaces such as walls, cement roads, brick structures, and marble, with varying lighting conditions. The results demonstrate that DeepCrackAT can effectively segment crack images in practical scenarios. In fact, in road maintenance, maintenance personnel typically use road repair vehicles to collect images of road surfaces, and our method can quickly detect surface cracks and damage, helping maintenance personnel carry out timely repairs and maintenance to ensure road safety and durability. Similarly, unmanned aerial vehicles can be used to collect comprehensive crack images of building surfaces, and automatic crack detection methods can be used to timely detect potential structural problems to ensure the safety of buildings. Furthermore, by combining some underwater image collection equipment, the proposed method can be applied to cracks in pipelines or dams, helping environmental protection personnel to timely detect and repair potential environmental problems. Additionally, crack detection can be applied to medical auxiliary diagnosis. It can be used to segment fractures or cracks in X-ray or MRI images to help doctors determine patient diagnoses and treatment plans. With the help of intelligent acquisition equipment, our model can be further applied to more industrial assessment scenarios in the future. We will improve the performance and generalization of the method in future work.

5. Conclusion

In this paper, we propose an effective crack segmentation framework based on learning multi-scale crack features, namely DeepCrackAT. Firstly, the hybrid dilated convolution is introduced to enlarge the receptive field of the convolution. Secondly, we introduce the Tok-MLP to tokenize high-dimensional crack features and project them into a low-dimensional space, reducing computational parameters and enhancing the model’s ability of noise resistance. Finally, the CBAM is used to construct an attentional skip-layer fusion block and enhance the network’s perception of the critical crack region. The proposed method achieves noise resistance in complex backgrounds, exhibiting good segmentation performance for both thin and thick cracks. However, the cracks with significant variations in width bring big challenges for deep learning methods. In the future, we will design a series of self-supervised pretext tasks for model pretraining to learn more semantic features, which is useful for the segmentation of big-width crack samples.

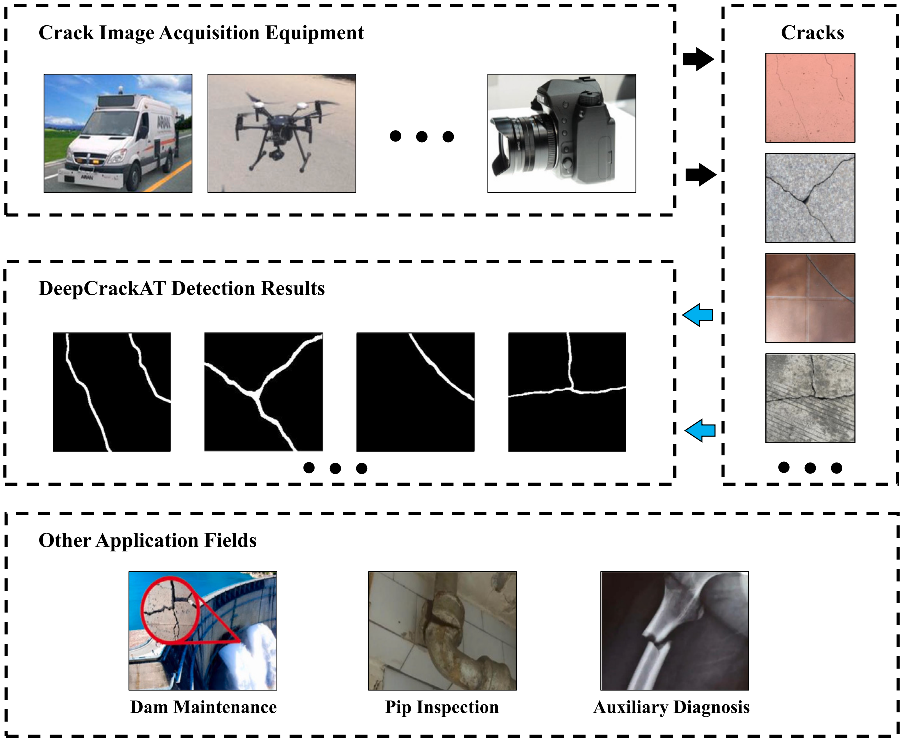


Fig. 11. Practical application of crack detection.

CRediT authorship contribution statement

Qinghua Lin: Conceptualization, Methodology, Software. **Wei Li:** Data collection, Data interpretation, Software. **Xiangpan Zheng:** Validation, Visualization. **Haoyi Fan:** Validation, Writing – reviewing. **Zuoyong Li:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Zuoyong Li reports financial support was provided by National Natural Science Foundation of China. Zuoyong Li reports financial support was provided by Natural Science Foundation of Fujian Province.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China (61972187, 62276146), Natural Science Foundation of Fujian Province, China (2020J02024, 2020Y4018), Fujian Provincial Department of Science and Technology Foreign Cooperation Project (2023I0039), Key Project of Colleges and Universities of Henan Province, China (23A52002), Science and Technology Innovation 2030-"New Generation of Artificial Intelligence" Major Project, China (2021ZD0111000).

References

Ai, Dihao, Jiang, Guiyuan, Lam, Siew-Kei, He, Peilan, Li, Chengwu, 2023. Computer vision framework for crack detection of civil infrastructure—a review. Eng. Appl. Artif. Intell. 117, 105478.

Ba, Jimmy Lei, Kiros, Jamie Ryan, Hinton, Geoffrey E., 2016. Layer normalization. arXiv preprint arXiv:1607.06450.

Badrinarayanan, Vijay, Kendall, Alex, Cipolla, Roberto, 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Chen, Tingyang, Cai, Zhenhua, Zhao, Xi, Chen, Chen, Liang, Xufeng, Zou, Tierui, Wang, Pan, 2020. Pavement crack detection and recognition using the architecture of segnet. J. Ind. Inf. Integr. 18, 100144.

Chen, Liang-Chieh, Zhu, Yukun, Papandreou, George, Schroff, Florian, Adam, Hartwig, 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. pp. 801–818.

Dais, Dimitris, Bal, Ihsan Engin, Smyrou, Eleni, Sarhosis, Vasilis, 2021. Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning. Autom. Constr. 125, 103606.

Dung, Cao Vu, Anh, Le Duc, 2019. Autonomous concrete crack detection using deep fully convolutional neural network. Autom. Constr. 99, 52–58.

Fantin Irudaya Raj, E., Balaji, M., 2021. Analysis and classification of faults in switched reluctance motors using deep learning neural networks. Arab. J. Sci. Eng. 46 (2), 1313–1332.

Guo, Feng, Qian, Yu, Liu, Jian, Yu, Huayang, 2023. Pavement crack detection based on transformer network. Autom. Constr. 145, 104646.

Han, Chengjia, Ma, Tao, Huyan, Ju, Huang, Xiaoming, Zhang, Yanning, 2021. CrackW-Net: A novel pavement crack image segmentation convolutional neural network. IEEE Trans. Intell. Transp. Syst. 23 (11), 22135–22144.

Hendrycks, Dan, Gimpel, Kevin, 2016. Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415.

Hsieh, Yung-An, Tsai, Yichang James, 2020. Machine learning for crack detection: Review and model performance comparison. J. Comput. Civ. Eng. 34 (5), 04020038.

Hu, Jie, Shen, Li, Sun, Gang, 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Inoue, Yuki, Nagayoshi, Hiroto, 2021. Crack detection as a weakly-supervised problem: towards achieving less annotation-intensive crack detectors. In: 2020 25th International Conference on Pattern Recognition. pp. 65–72.

Ioffe, Sergey, Szegedy, Christian, 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456.

Jaderberg, Max, Simonyan, Karen, Zisserman, Andrew, et al., 2015. Spatial transformer networks. Adv. Neural Inf. Process. Syst. 28.

Kingma, Diederik P., Ba, Jimmy, 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kong, Xiangxiong, Li, Jian, 2019. Non-contact fatigue crack detection in civil infrastructure through image overlapping and crack breathing sensing. Autom. Constr. 99, 125–139.

Lei, Bin, Wang, Ning, Xu, Pengcheng, Song, Gangbing, 2018. New crack detection method for bridge inspection using UAV incorporating image processing. J. Aerosp. Eng. 31 (5), 04018058.

Li, Ruoxian, Yu, Jiayong, Li, Feng, Yang, Ruitao, Wang, Yudong, Peng, Zhihao, 2023. Automatic bridge crack detection using unmanned aerial vehicle and faster R-CNN. Constr. Build. Mater. 362, 129659.

Liu, Yun, Cheng, Ming-Ming, Hu, Xiaowei, Wang, Kai, Bai, Xiang, 2017. Richer convolutional features for edge detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3000–3009.

Liu, Ze, Lin, Yutong, Cao, Yue, Hu, Han, Wei, Yixuan, Zhang, Zheng, Lin, Stephen, Guo, Baining, 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, Li, Ouyang, Wanli, Wang, Xiaogang, Fieguth, Paul, Chen, Jie, Liu, Xinwang, Pietikäinen, Matti, 2020. Deep learning for generic object detection: A survey. Int. J. Comput. Vis. 128, 261–318.

Ogawa, Shujiro, Matsushima, Kousuke, Takahashi, Osamu, 2019. Crack detection based on Gaussian mixture model using image filtering. In: 2019 International Symposium on Electrical and Electronics Engineering. ISEE, pp. 79–84.

Pak, Myeongsuk, Kim, Sanghoon, 2021. Crack detection using fully convolutional network in wall-climbing robot. In: Advances in Computer Science and Ubiquitous Computing. pp. 267–272.

Pang, Guansong, Shen, Chunhua, Cao, Longbing, Hengel, Anton Van Den, 2021. Deep learning for anomaly detection: A review. ACM Comput. Surv. 54 (2), 1–38.

Qingbo, Zhu, 2016. Pavement crack detection algorithm based on image processing analysis. In: 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics, Vol. 1. IHMSC, pp. 15–18.

Ronneberger, Olaf, Fischer, Philipp, Brox, Thomas, 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241.

Sun, Xinzi, Xie, Yuanchang, Jiang, Liming, Cao, Yu, Liu, Benyuan, 2022. DMA-Net: DeepLab with multi-scale attention for pavement crack segmentation. IEEE Trans. Intell. Transp. Syst.

Valanarasu, Jeya Maria Jose, Patel, Vishal M., 2022. UNeXt: MLP-based rapid medical image segmentation network. arXiv preprint arXiv:2203.04967.

Wang, Yanyan, Niu, Menghui, Song, Kechen, Jiang, Peng, Yan, Yunhui, 2023. Normal-knowledge-based pavement defect segmentation using relevance-aware and cross-reasoning mechanisms. IEEE Trans. Intell. Transp. Syst.

Woo, Sanghyun, Park, Jongchan, Lee, Joon-Young, Kweon, In So, 2018. CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. pp. 3–19.

Yang, Xincong, Li, Heng, Yu, Yantao, Luo, Xiaochun, Huang, Ting, Yang, Xu, 2018. Automatic pixel-level crack detection and measurement using fully convolutional network. Comput.-Aided Civ. Infrastruct. Eng. 33 (12), 1090–1109.

Yu, Tiantang, Zhu, Aixi, Chen, Yingying, 2017. Efficient crack detection method for tunnel lining surface cracks based on infrared images. J. Comput. Civ. Eng. 31 (3), 04016067.

Yuan, Genji, Li, Jianbo, Meng, Xianglong, Li, Yinong, 2022. CurSeg: A pavement crack detector based on a deep hierarchical feature learning segmentation framework. IET Intell. Transp. Syst. 16 (6), 782–799.

Zhang, Lingxin, Shen, Junkai, Zhu, Baijie, 2021. A research on an improved unet-based concrete crack detection algorithm. Struct. Health Monit. 20 (4), 1864–1879.

Zhang, Kaige, Zhang, Yingtao, Cheng, Heng-Da, 2020. CrackGAN: Pavement crack detection using partially accurate ground truths based on generative adversarial learning. IEEE Trans. Intell. Transp. Syst. 22 (2), 1306–1319.

Zhao, Zhong-Qiu, Zheng, Peng, Xu, Shou-tao, Wu, Xindong, 2019. Object detection with deep learning: A review. IEEE Trans. Neural Netw. Learn. Syst. 30 (11), 3212–3232.

Zou, Qin, Zhang, Zheng, Li, Qingquan, Qi, Xianbiao, Wang, Qian, Wang, Song, 2018. Deepcrack: Learning hierarchical convolutional features for crack detection. IEEE Trans. Image Process. 28 (3), 1498–1512.