Track slab crack detection based on full convolutional neural network

Wenju Li¹,*, Huiling Chen¹, Yaoxing Zhang², Yanjiao Shi¹

¹School of Computer Science and Information Engineering and Shanghai Institute of Technology, Shanghai, China
²School of Railway Transportation and Shanghai Institute of Technology, Shanghai, China

*Corresponding author’s e-mail: wjli@sit.edu.cn

Abstract. In recent years, with the rapid development of railway industry, railway safety has become an issue of increasing concern to people. The crack is the biggest threat to the safety of the railway. Because the pictures can only be collected at night, the pictures collected have many problems such as noise and uneven illumination. In order to solve these problems, this paper proposed a method of track slab crack image segmentation based on full convolutional neural network. In this method, VGG-16 is used as the backbone network, and the dilated convolution is added into the backbone network. Then the spatial pyramid pooling method is used to extract more global and local feature information. In addition, the HED network is used to obtain multiple output results. Finally, in order to make the model more discriminative to the characteristics of each channel, SE module is added. Experimental results show that the proposed method reaches 81.84%, 67.68% and 84.55% in precision, IoU and F1-score respectively in the self-made data set, which is superior to the traditional segmentation method and U-Net method, and can meet the needs of practical engineering.

1. Introduction

In recent years, the track building has gradually entered the maintenance stage. The traditional crack detection method is mostly observed by human eyes. Due to the factors of night collection, the detection accuracy is low and the workload is large. With the development of image processing technology, support vector machine⁷, Gabor filtering, edge detection method, spectral analysis method⁸ and other methods have solved the detection problem of some images with single background. Faced with most images with complex background and uneven illumination, researchers have started to study deep learning models.

In recent years, with the development of artificial intelligence, many researchers have applied deep learning models to various industries such as medicine and transportation, and crack detection is no exception. Zhang L et al.³ applied deep learning to pavement crack detection for the first time and proposed an automatic detection method based on deep convolutional neural network. Literature⁴ adopts deep convolutional neural network to detect tunnel cracks, which improves detection speed and accuracy to a certain extent. Literature⁵ explored the automatic detection of pixel-level cracks, proposed a surface crack detection method based on convolutional neural network, and compared it with some existing segmentation networks. Experiments have proved that its segmentation effect is good. Literature⁶ proposed a concrete apparent crack detection method using the Hu-ResNet model, and the accuracy was further improved. More and more deep learning methods have been applied in the
field of image segmentation, and all of them have achieved good results. However, for crack images with complex background, the accuracy and accuracy of detection still need to be improved.

In this paper, a crack segmentation model based on full convolutional neural network is proposed. In this model, dilated convolution with different dilation rate is added into the backbone VGG16 network, followed by improved spatial pyramid pooling, and multi-scale features are obtained by using HED network. At the same time, SE module is added into spatial pyramid pooling and HED network in order to make the model more distinguishable to the characteristics of each channel.

2. Network structure design

2.1. Network architecture

The overall network architecture of the track slab crack segmentation network model proposed in this paper is shown in Figure 1. Firstly, the network is improved on the basis of VGG-16 encoder. Each block consists of two 3×3 convolutional layers and a 2×2 max pooling layer. After each convolutional layer, there is a ReLU function and a BN layer, which is used to extract the texture information and spatial information features in the crack image of track slab. After four times of convolution and three times of max pooling, the feature vectors are input into a four-layer dilated convolution, and different dilation rate are set for the four-layer dilated convolution to obtain a larger reception field and capture more features. Then a decoding module of the spatial pyramid is connected to obtain the global and local feature information, and each layer in the pooling is up-sampled. In addition, the HED network is used to output the feature map convolved from each block of the backbone network from different scales, which is compared with the output results of the backbone network. Finally, the output results of the spatial pyramid are connected with the output results of the three convolution layers to finally get the pixel-level prediction results.

![Figure 1. Network architecture diagram](image)

2.2. Analysis of correlation methods

2.2.1. Dilated Convolution

The dilated convolution adds 0 on the basis of the convolution kernel to obtain the reception field by adjusting the dilation rate. When the dilation rate is 1, the dilated convolution is the ordinary convolution; when the dilation rate is 2, the convolution kernel skips one pixel at each input; when the dilation rate is 3, the convolution kernel skips two pixels at each input, and so on. Different dilation rate are used to realize the dilated convolution according to the actual situation.
In this paper, a four-layer dilated convolution is connected after the backbone network to deepen the network. The dilation rate of the four-layer dilated convolution are 2, 2, 4 and 4 respectively, which can better extract the image features while maintaining the spatial resolution of the feature map.

![Dilated convolution reception field with different dilation rates](image1)

**Figure 2.** The dilated convolution reception field with different dilation rates

### 2.2.2. Spatial Pyramid Pooling

Spatial pyramid pooling can transform feature maps of any size into feature vectors of the same size, which is a very successful multi-scale image fusion method. Spatial pyramid pooling aggregates multi-level features into the output, and divides the target image from coarse to detailed, thus effectively obtaining global and local information.

In this paper, in the spatial pyramid pooling, feature maps with length and width of 1, 1/2, 1/4 and 1/8 are respectively taken to output feature maps of 1×1, 2×2, 4×4, 8×8, and then the feature maps are upsampled and connected.

### 2.2.3. SE Module

The SE module of attention mechanism is introduced to solve the unbalanced problem of crack image. According to the importance of features, different weights are applied to capture more advanced semantic information and improve the precision of segmentation. The SE module is composed of both Squeezer and Excitation. The feature map for the convolutional layer input is compressed as a 1×1×C vector (C is the number of channels). Then the computational load is reduced by reducing the number of channels; Finally, the weight value of each channel is multiplied by the two-dimensional matrix of the original characteristic channel, and the result is obtained.

In this paper, SE modules are added after each operation of spatial pyramid pooling above, and one SE module is connected separately in HED network. See Figure 3. In order to improve the semantic segmentation accuracy of crack images, weight coefficients are adjusted in space and channel, and important features are enhanced.

![SE Model](image2)

**Figure 3.** SE Model

### 3. Experiment

#### 3.1. Data Set

In this paper, the data set of track slab cracks obtained by field shooting was adopted. Through screening, images of track slab cracks without obvious shooting problems were retained, totaling more than 800 images of 1200×1400 pixels. Then, Labelme tool is used to locate the crack location in the data set. Labelme is used to hook out the crack area in the original image to obtain the generated label file, which is used to label the image. Finally, each image is randomly rotated, clipped, translated and inverted using image enhancement technology, and more than 50,000 images of 256×256 pixels are finally generated as data sets. The batch size was set to 12, the epoch was 50, and the learning rate was set to 0.0001.
The experimental environment of this experiment is PyTorch framework, and it is run under Windows system. The experiments are tested and trained on NVIDIA GeForce RTX 2080TI.

3.2. Evaluation indicators
In order to better quantitatively evaluate the performance of the model, Precision, IoU, Recall and F1-score were used in this paper. The calculation formulas are as follows:

\[ \text{Precision} = \frac{TP}{TP + FP} \]  (1)

\[ \text{IoU} = \frac{TP}{TP + FP + FN} \]  (2)

\[ \text{Recall} = \frac{TP}{TP + FN} \]  (3)

\[ F1\text{-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]  (4)

Where TP is the the pixel number of true positives; FP is the pixel number of false positives. Fn is the number of false negative. Recall is used as a standard to evaluate missed detection, highlighting the proportion of missed detection. F1-score comprehensively considers Precision and Recall, and the higher the value of F1-score is, the better the performance of the model.

3.3. Experimental analysis
On the same test data set, this experiment compares the traditional crack segmentation method based on image processing and the U-Net segmentation method. The three evaluation indexes of the three methods were compared, such as Precision, IoU, Recall and F1-score, as shown in Table 1.

| Methods  | IoU    | Precision | Recall  | F1-score |
|----------|--------|-----------|---------|----------|
| HED-Net  | 48.28% | 59.02%    | 57.13%  | 61.08%   |
| U-Net    | 57.84% | 72.46%    | 75.42%  | 74.10%   |
| Ours     | 62.68% | 81.84%    | 85.46%  | 84.55%   |

In contrast, the image segmentation method based on deep learning is higher than the traditional method in IoU, Precision, Recall and F1-score. The IoU of the crack segmentation method based on the full convolutional neural network is 62.68%, the Precision is 81.84%, and the F1-score is 84.55%. The proposed method surpasses the U-Net method in both Precision and F1-score, and obtains more accurate segmentation results, and has good generalization ability.
In the figure above, the first column is the original image, the second column is Ground truth, followed by the fracture segmentation results of the traditional edge detection, U-Net, and our method. According to the comparison between the segmentation result and the label, the traditional segmentation method is not very good for the image with complex background and large contrast, and there is a case of missing and wrong detection. U-Net method can basically detect the cracks in the image, but in the detection results, it can be seen that some small cracks are not segmented, while the crack background with noise is segmented. The method in this paper can accurately segment small cracks with the Precision of 81.84%, and there are basically no noisy pixels. Compared with U-Net, the segmentation results of cracks are significantly improved, and the detection result of cracks is more similar to the label, showing good robustness to the interference factors in the image background.

4. Conclusion

In this paper, a pixel-level semantic segmentation network based on the full convolutional neural network is proposed. First, the dilated convolution with different dilation rate is added into the backbone VGG16 network, then the improved spatial pyramid pooling is connected, and the HED network is used to obtain multi-scale features. In this method, the main network is used to extract the features, and the dilated convolution is used to obtain a larger reception field and extract rich semantic information. In order to convert the feature map of different sizes obtained by convolution to the same size, the space pyramid pooling is connected after the dilated convolution. In order to improve the accuracy of the model, the features of different scales are extracted, and the feature connections after the spatial pyramid pooling are made through skip connections. Finally, in order to make the model more distinguishable to the characteristics of each channel, SE modules are added into the spatial pyramid pooling and HED network. The model has good generalization and robustness, and the experimental results show that the model has been greatly improved in the speed and precision of segmentation, which can meet the needs of engineering.

Acknowledgments

This work had supported by the National Natural Science Foundation of China (61806126).

References

[1] Dai B, Gu C, Zhao E, Zhu K, Cao W, Qin X. (2019) Improved online sequential extreme learning machine for identifying crack behavior in concrete dam[J]. Advances in Structural Engineering, 22(2):402-412.
[2] ADHIKARI R S, MOSELHI O, BAGCHI A. (2014) Image-based retrieval of concrete crack properties for bridge inspection[J]. Automation in Construction, 39(4): 180-194.
[3] Zhang L, Yang F, Zhang Y D, et al. (2016) Road crack detection using deep convolutional neural network[C] IEEE, 3708-3712.
[4] Liu X, Chen Y, Zhu A, et al. (2018) Tunnel crack identification based on deep learning[J]. Journal of Guangxi University (Natural Science Edition), (6):2243-2251.
[5] Feng C, Zhang H, Wang H, et al. (2020) Automatic Pixel-Level Crack Detection on Dam Surface Using Deep Convolutional Network. 20(7).
[6] XU G, LIAO Cg, CHEN J, et al. (2020) Extraction of apparent crack information of concrete based on HU-ResNet [J]. Computer Engineering, 46(11) : 279-285.
[7] Yu F , Koltun V. (2016) Multi-Scale Context Aggregation by Dilated Convolutions[J].