Smart Structures and Systems, *Vol. 29, No. 1 (2022) 29-39*
https://doi.org/10.12989/sss.2021.29.1.029

A three-stage deep-learning-based method for crack detection of high-resolution steel box girder image

Shiqiao Meng [1a], Zhiyuan Gao [2b], Ying Zhou*[1], Bin He [3c] and Qingzhao Kong [1d]

[1] *State Key Laboratory of Disaster Reduction in Civil Engineering, Tongji University, Shanghai 200092, China*
[2] *College of Civil Engineering, Tongji University, Shanghai 200092, China*
[3] *College of Electronic and Information Engineering, Tongji University, Shanghai 200092, China*

(*Received March 27, 2021, Revised September 8, 2021, Accepted September 9, 2021*)

Abstract. Crack detection plays an important role in the maintenance and protection of steel box girder of bridges. However, since the cracks only occupy an extremely small region of the high-resolution images captured from actual conditions, the existing methods cannot deal with this kind of image effectively. To solve this problem, this paper proposed a novel three-stage method based on deep learning technology and morphology operations. The training set and test set used in this paper are composed of 360 images (4928 × 3264 pixels) in steel girder box. The first stage of the proposed model converted high-resolution images into sub-images by using patch-based method and located the region of cracks by CBAM ResNet-50 model. The *Recall* reaches 0.95 on the test set. The second stage of our method uses the Attention U-Net model to get the accurate geometric edges of cracks based on results in the first stage. The *IoU* of the segmentation model implemented in this stage attains 0.48. In the third stage of the model, we remove the wrong-predicted isolated points in the predicted results through dilate operation and outlier elimination algorithm. The *IoU* of test set ascends to 0.70 after this stage. Ablation experiments are conducted to optimize the parameters and further promote the accuracy of the proposed method. The result shows that: (1) the best patch size of sub-images is 1024 × 1024. (2) the CBAM ResNet-50 and the Attention U-Net achieved the best results in the first and the second stage, respectively. (3) Pre-training the model of the first two stages can improve the *IoU* by 2.9%. In general, our method is of great significance for crack detection.

Keywords: crack detection; high-resolution image; steel box girder; three-stage method

1. Introduction

The detection of fatigue crack damage, which usually caused by the initial flaws in the welding parts and live loads, plays an essential role in the maintenance and protection of steel box girder (de Freitas *et al*. 2012). The development of crack will shorten the service life and decrease the reliability of bridges. Therefore, it is crucial to detect cracks accurately and promptly.

In the past few decades, the early researchers implemented damage detection by adopting advanced signal processing techniques, such as blind feature extraction, sparse representation classification and compressive sensing algorithm (Yang and Nagarajaiah 2014, 2016). However, it is difficult for the traditional structural health monitoring method to precisely capture the fatigue cracks in time. Besides, the frequently employed ultrasonic technique (Mutlib *et al*. 2016) and acoustic emission technique (Han *et al*. 2015) are complex and expensive.

Crack diseases have strong visual characteristics, so detection methods based on computer vision and image processing are proposed by researchers, such as outlier elimination algorithm (Canny 1986) or edge detection algorithm (Ong *et al*. 2015), to detect cracks automatically. For example, Sun and Qiu (2007) proposed a pavement crack recognition algorithm based on mathematical morphology, including eliminating Impulse noise and Gaussian noise, using multi-scale morphological edge detection methods to extract crack edges and labelling algorithms to segment cracks in images. However, the results of these methods are greatly affected by the environment, and the detection result is not robust. Therefore, it is difficult to utilize their proposed techniques in actual working conditions.

In recent years, breakthroughs in the research of convolutional neural networks (CNN) (Krizhevsky *et al*. 2017) improved the ability of a computer to process images. Methods that proposed based on CNN can significantly enhance the robustness of the result of crack detection. For example, Chen and Jahanshahi (2017) proposed a technique of using CNN to analyze a single video frame for crack detection, which combined with the Naive Bayes data fusion scheme; Cha *et al*. (2018) used RCNN model, which was proposed by Girshick *et al*. (2014), to detect various concrete diseases such as cracks and rust, and establishes a framework model of apparent diseases. Xu *et al*. (2019) proposed a deep fusion convolutional neural network to detect the area in the picture of steel box girder cracks collected by consumer cameras, and obtain a more accurate crack location. However, these models can only get the

*Corresponding author, Ph.D., Professor,
E-mail: yingzhou@tongji.edu.cn

ISSN: 1738-1584 (Print), 1738-1991 (Online)

position instead of the geometric edge of the crack in the original picture. Therefore, it is difficult for these methods to quantitatively describe the geometric boundaries of the crack quantitatively.

With the development of image semantic segmentation technology based on deep learning, crack geometric edge detection efficiency and accuracy have been significantly improved. In 2014, Long *et al*. (2015) proposed an effective model named fully convolutional networks (FCN), which is composed of two modules: encoder and decoder. Later, Ronneberger *et al*. (2015) proposed the U-Net model which is designed based on FCN, and further improves the segmentation effect. In terms of the segmentation of cracks, Ye *et al*. (2019) and Dung and Anh (2019) applied semantic segmentation technology to concrete crack detection and successfully obtained the geometric edges and the width of cracks. Zhang *et al*. (2018) proposed an improved CrackNet to obtain better results in the segmentation of road cracks. Later, researchers proposed various improved semantic segmentation models for pixel-level automatic crack detection of low-resolution images (Yang *et al*. 2018, Li *et al*. 2019, Bang *et al*. 2019, Bao *et al*. 2019, Bao and Li 2021, Spencer *et al*. 2019). Since Vaswani *et al*. (2017) brought up the idea of Self-Attention layers in the deep neural network, it is widely used in computer vision tasks. This is because the attention mechanism has a good effect in obtaining the details of the image. What's more, cracks account for a small proportion of the pixels in the picture, so there are many related researches in crack detection to improve the effect of segmentation (Qiao *et al*. 2021, Song *et al*. 2019, 2020, Wan *et al*. 2021). However, these methods can only be effectively applied to images where the crack accounts for a large proportion. Besides, limited by the size of computer memory size, it is arduous to implement these methods on high-resolution images. Since the steel box girder cracks are relatively small, high-resolution images, such as 3264 × 4926, are often used for crack detection. Thus, using these methods in actual conditions is unlikely to have effective and acceptable results.

To address these limitations, this paper proposes a three-stage method that can enhance the accuracy and effectiveness of pixel-level crack detection on high-resolution images. In the first stage, we employ a patch-based method for image classification, which converts a large high-resolution image into several small sub-images of a fixed size. Then, the classification model based on a convolutional neural network with self-attention mechanism is applied to classify the region where crack exists. In the second stage, by using the Attention U-Net model to segment sub-images, the veracious crack edges will be obtained. The Attention module is integrated into these models to filter out high-value information rapidly from a large amount of data with accurate classification and segmentation of crack images. After stitching sub-images, the geometric edges of the cracks in the high-resolution images can be accurately segmented. In the third stage, we apply the outlier elimination algorithm and the dilation algorithm to post-process the predicted results of the second stage. Based on retaining the segmented cracks, we successfully remove most of the noise pixels in the predicted result, which makes it more conform to the ground truth. Several high-resolution crack images of the steel box girder were used to verify the performance of our method. The results show that our method has high identification accuracy of crack images and successfully achieved accurate crack segmentation on high-resolution images.

2. Methodology

Since cracks have self-similarity on two orders of magnitude (Saouma *et al*. 1990), the probability of only obtaining part of the crack by simply utilizing object detection algorithms such as Faster R-CNN (Ren *et al*. 2016) is exceptionally high. In the problem of crack detection, the target object detection can realize the detection of fragments of cracks, which may cause the detected cracks to appear discontinuous (Cha *et al*. 2018). Inspired by the patch-based method (Hou *et al*. 2016), we first converted the original image into several small sub-images. A classification model is proposed to judge whether the crack exists in each sub-images, which will effectively locate the region of cracks. Then, by utilizing a segmentation model on the area where the crack exist, all of the crack areas will be processed. Since the classification and segmentation model has errors inevitably, we use the dilate operation and outlier elimination algorithm to eliminate the noise in the prediction results. Generally speaking, the process is divided into three stages. The flow chart is shown in Fig. 1.

The convolutional neural network model used in the first two stages can be replaced by any classification and segmentation model according to the actual engineering situation. The remainder of this section will describe the detail of the three-stage method individually.

2.1 The first stage

Deep residual networks (ResNet) (He *et al*. 2016) is a widely used and recognized deep learning architecture in the field of image classification, which shows compelling accuracy and excellent convergence behaviours. To be capable of classifying slight cracks, we employ a simple yet effective attention module called convolutional block attention module (CBAM) (Woo *et al*. 2018) to improve the performance of ResNet.

2.1.1 Data preprocessing

For a high-resolution image, we assume that the size of the patch is $s \times s$, while the overlap length is $s/2$. The overlapping patch is aimed at improving the accuracy of judging the fracture area. After acquiring the small image, we normalize each pixel to between 0 and 1 by dividing the value of each pixel in the image by 255 and standardize each pixel to between -1 and 1. The expression of image standardize is shown as Eq. (1).

$$x^{`} = \frac{x - \mu}{\sigma} \tag{1}$$

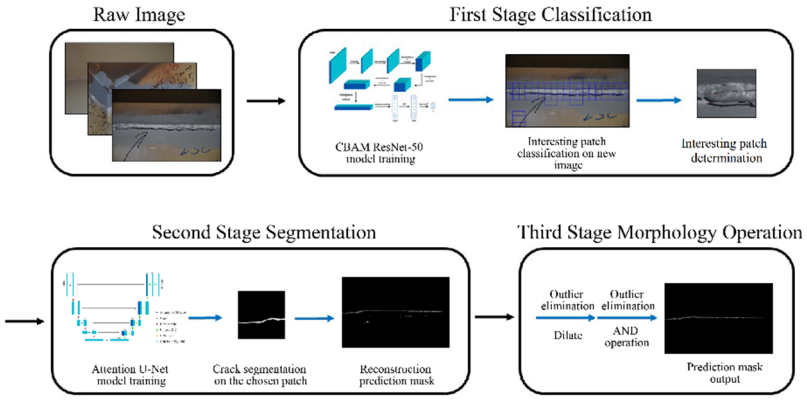


Fig. 1 Flowchart for the three-stage crack segmentation strategy

where x' is the standardized pixel value; x is the normalized pixel value; μ and σ are the average and standard deviation of all pixels on each image channel, respectively.

2.1.2 Model architecture

The model we selected for classification in the first stage is CBAM ResNet-50. Among them, the role of the CBAM module is to quickly filter out high-value information from a large amount of data, which refers to the information of cracks. Therefore, the ability of the model to identify cracks can be significantly improved after adding this module.

The CBAM module has two sequential sub-modules, which are the channel module and the spatial module. For the convolutional layer of a neural network at any depth, the two modules of CBAM can adaptively refine the intermediate feature maps.

The aim of channel attention module aims to render a one dimension channel map ($C \times 1 \times 1$). The module will generate a channel attention map using the relationship between channels of features. Since each channel of the feature map is treated as a feature detector, the channel attention map focuses on what is meaningful for a given input image. The average-pooling is widely used to gather spatial information, and max-pooling is used to collect different object characteristics from inferring better channel attention. To combine their advantages, the module uses both average-pooling and max-pooling features. After the input feature map passes through two pooling layers, the two feature maps are obtained through the shared multi-layer perceptron (MLP). The final feature vectors are combined by element-wise summation. The Sigmoid function shown as Eq. (2) is used at the end of the channel attention module to distinguish the feature more prominent.

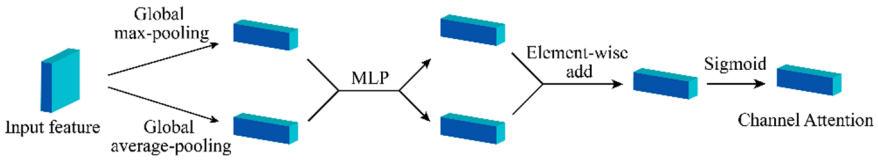
$$\sigma = \frac{1}{1 + e^{-z}} \tag{2}$$

where z is the input value of Sigmoid function. To sum up, the channel attention is computed as Eq. (3).

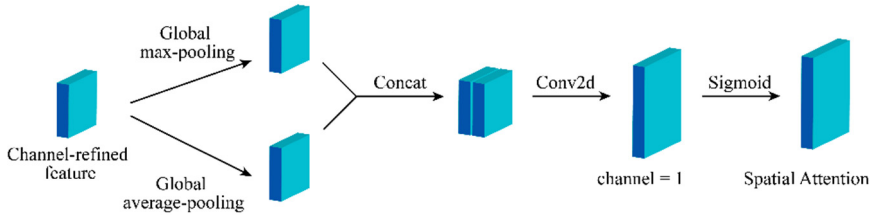
$$Mc(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{3}$$

where **F** represents the input feature. The whole process of the channel attention module is shown in Fig. 2.

The aim of the spatial attention module is to render a two-dimension channel map ($1 \times H \times W$). The module uses the spatial relationship between features to generate a spatial attention map. To supplement channel attention, spatial attention focuses on the location of information. When calculating the spatial attention, the average-pooling

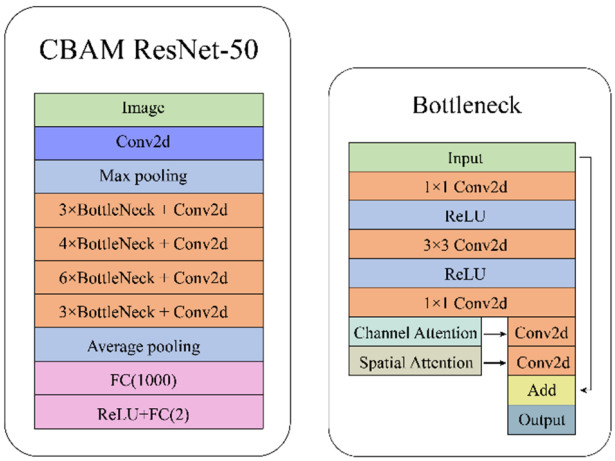


(a) Diagram of channel attention module



(b) Diagram of spatial attention module

Fig. 2 Diagram of module in CBAM



(a) The detailed structure of ResNet-50 (b) The detailed structure of Bottleneck

Fig. 3 The architecture of ResNet-50

and the max-pooling are carried out along the channel axis. Then, a convolutional layer is used to generate the spatial attention map. Finally, a two-dimensional spatial attention map is generated after calculating by the Sigmiod function. To sum up, the spatial attention is computed as Eq. (4).

$$Ms(F) = \sigma(Conv([AvgPool(F); Maxpool(F)])) \quad (4)$$

where **F** represents the input feature. The whole process of the spatial attention module is shown in Fig. 2.

The architecture of ResNet-50 is shown in Fig. 3. The bottleneck part can be easily replaced by the bottleneck that contains the CBAM module by implementing this structure. The detailed structure of the bottleneck implemented in the model is shown in Fig. 3, which contains both channel attention module and spatial attention module.

2.2 The second stage

2.2.1 Data preprocessing

The size of the input image for the model in stage two is conformed with stage one. Since the regions where crack exists were located, the segmentation model only needs to identify sub-images from crack regions, which means only the sub-image judged to contain cracks in the first stage should be processed. Thus, the sub-images preprocessed in stage one can be directly used by the segmentation model in stage two.

2.2.2 Model architecture

The Attention U-Net model we employed in the second stage is implemented based on U-Net (Ronneberger *et al.* 2015). It has skip connections between the encoder and decoder. The whole model is consists of four modules: Attention CNNLayer, CNNLayer, Downsample module and Upsample module. The overall architecture of the model is shown in Fig. 4.

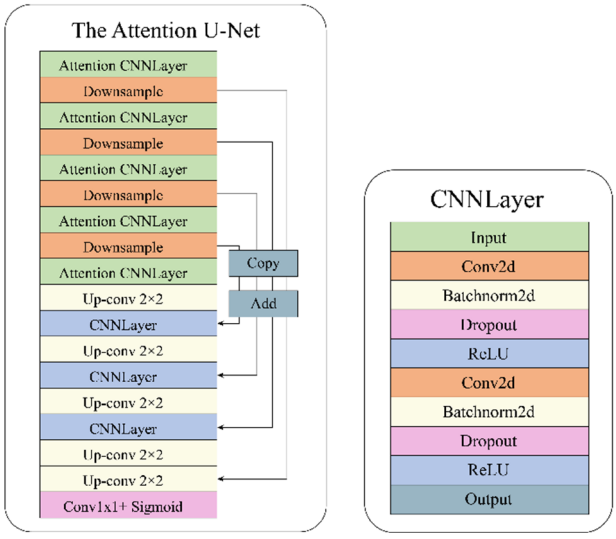
The skip connection between encoder and decoder can make full use of the multi-scale features in the crack image. At the same time, five Attention modules were implemented in the model to improve the performance further. They can make it capable of focus on practical crack information. The final layer of the model only contains one channel. Thus each pixel in the output result represents the probability of whether the point belongs to a crack.

The Attention U-Net consists of four kinds of substructures: Attention CNNLayer, CNNLayer, Downsample module and Upsample module. The function of the CNNLayer is to perform continuous convolution operations and increase or decrease the number of the channel of the feature maps. The detailed network structure diagram is shown in Fig. 4. The function of the Downsample module is to reduce the size of the extracted feature maps, which is composed of a convolutional layer and a ReLU layer. As for the Upsampling module, it is capable of restoring the size of the input image by applying a bilinear interpolation layer. After processed by the Upsampling module, the feature maps will be concatenated channel-wise with the corresponding feature maps extracted by the encoder.

The Attention CNNLayer is designed mainly based on CNNLayer, which contains three sub-blocks, namely Channel Attention block, Spatial Attention block and CNNLayer. The overall structure diagram is shown in Fig. 5. By adding the Channel Attention module and Spatial Attention module on CNNLayers, the model gains the ability to concentrate on high-value information on the channel and spatial dimensions. The Spatial Attention block is entirely the same as the spatial attention module of CBAM, which is introduced in the previous section. To sum up, the channel attention is computed as Eq. (5).

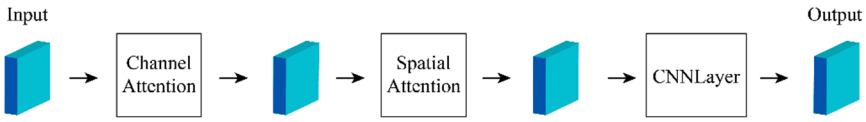
$$Mc'(F) = \sigma\begin{pmatrix} Conv[Conv(ReLu[AvgPool(F)])] \oplus \\ Conv[Conv(ReLu[Maxpool(F)])] \end{pmatrix} \otimes F \quad (5)$$

where **F** represents the input feature. The detailed structure of the Channel Attention module is shown in Fig. 5.

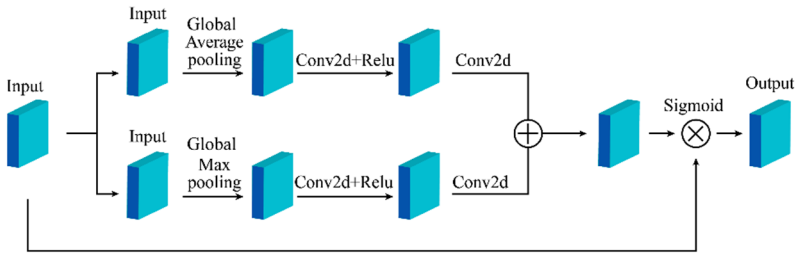


(a) The detailed structure of the Attention U-Net (b) The detailed structure of CNNLayer

Fig. 4 The overall architecture of the Attention U-Net



(a) The detailed structure of the attention module



(b) The detailed structure of channel attention module employed in the attention module

Fig. 5 The architecture of the attention module

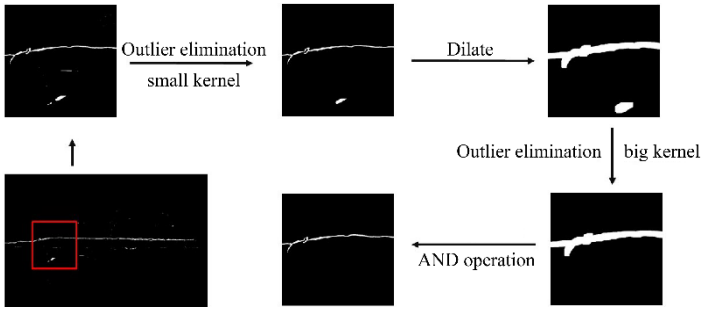


Fig. 6 The flow chart of the third stage

2.3 The third stage

All of the operations utilized in the third stage are designed to eliminate the noise in the results predicted by the model of the second stage. Since the images of the input model overlap each other in the first two stages, the maximum value of the pixels in the overlapping area should be taken as the prediction result.

Due to the inevitable errors in the prediction results of the previous model, if the prediction results of the second stage output directly, the prediction results will contain some isolated points. To eliminate these points, the outlier elimination algorithm can be utilized. However, it is difficult for the algorithm to eliminate large isolated points, especially when the crack is not consecutive, making the size of the crack similar to isolated points. Therefore, morphological processing is required to connect the fractured crack temporarily.

The flow chart of the processing method is shown in Fig. 6. First, adjust the threshold to a deficient level and eliminate small outliers through the outlier elimination algorithm. Then expand the picture to connect the discontinuous cracks and increase the threshold of the outlier elimination algorithm to eliminate large isolated points. Finally, to not damage the crack boundary in the predicted image, it is necessary to perform a logical AND operation between the generated image and the original predicted image. After processing these operations, the predicted results without isolated points will be obtained.

The steps to remove outliers are as follows: By using a connected domain labelling algorithm, the total number of the connected regions and the number of pixels in each connected regions can be calculated. If the number is less than the preset threshold, the corresponding regions are regarded as outliers and removed.

3. Experiment

3.1 Dataset description

The dataset used in this article is from the 1st International Project Competition for Structural Health Monitoring (Bao *et al*. 2021), which is composed of 360 images taken in a steel box girder. This dataset consists 120 original images with corresponding labels and 240 additional images. A total of 180 images were randomly selected as the training set, and 20 images were prescribed as a validation set, while the remaining 160 images were selected as a test set. The dataset consists of RGB images with 3264 × 4928 pixel resolution and corresponding mask, as shown in Fig. 7.

The dataset has the following characteristics, which makes it challenging to distinguish cracks:

- Cracks in the image have various shapes, sizes and strikes.
- The region that contains cracks is extremely small compares to the whole image. Since the width of the

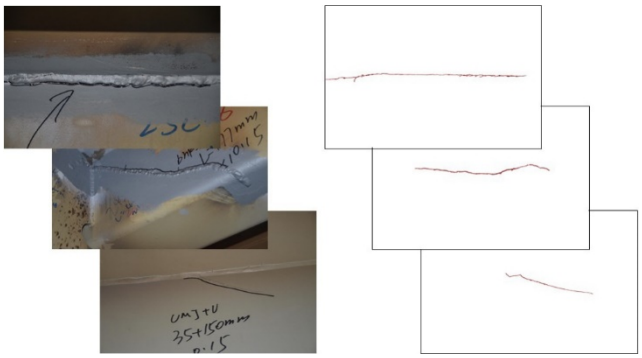


Fig. 7 Partial dataset presentation. The left side is the original image, and the right side is the corresponding mask. The black background has been removed for display convenience

cracks is mostly within 10 pixels, the relevant information may be lost after the general convolution operation.

- There are interference factors such as handwriting and scratches in the steel box girder, making it very difficult to distinguish the cracks

To enable the semantic segmentation model to learn more valuable features, we selected images which pixels belong to crack is more than 500 as a dataset.

3.2 Implementation details

3.2.1 Loss function

For the loss function in the first stage, we adopt balanced cross entropy loss. Cross entropy loss is commonly used in the classification task. However, under this circumstance, the number of crack images in the dataset is far less than that of images without cracks, which means using the standard cross entropy loss function in the training process will make it difficult for the model to identify crack images. Therefore, we employ balanced cross entropy loss, which introduced a weight factor α into on the cross entropy loss. α is multiplied by the loss value of crack images, while $(1-\alpha)$ should be multiplied by the loss value of the non crack category. The α is set to 0.9 in this paper. The expression of the balanced cross entropy loss is shown as Eq. (6).

$$\begin{aligned} BCE(p_t) &= -\alpha_t \log(p_t) \\ \alpha_t &= \begin{cases} \alpha & if \quad y=1 \\ 1-\alpha & otherwise \end{cases} \\ p_t &= \begin{cases} p & if \quad y=1 \\ 1-p & otherwise \end{cases} \end{aligned} \tag{6}$$

where p is the prediction probability of the model for category $y = 1$.

As for the loss function employed in the second stage, since the crack only occupy a relatively small area, using binary cross entropy loss alone will affect the efficiency of the training process. Therefore, the combination of dice loss and binary cross entropy loss are used as loss function in the training process. The expression of the dice loss function is shown as Eq. (7).

$$DL_2 = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{7}$$

where X is the real crack label; Y represents the crack prediction result output by the model; $|X \cap Y|$ indicates the intersection of set X and set Y, that is, the number of pixels whose prediction result is a crack and whose real label is also a crack; $|X|$ and $|Y|$ represent the number of elements of sets X and Y, that is, the number of pixels in the groundtruth and the number of pixels predicted as cracks.

Using dice loss combined with binary cross entropy loss in the training process, the values calculated through these two loss functions are added together to get the final loss value. Then, the loss function is minimized to derive the loss function from optimizing the network parameters and the optimal prediction effect of the model can be achieved.

3.2.2 Learning rate

In the first stage and the second stage of training, the cosine learning rate method is used to value the learning rate. The expression is shown as Eq. (8).

$$\begin{aligned} LR = &\frac{LR_{init} - LR_{\min}}{2} \\ &\cdot cos\left(\frac{mod(x-1, cycle)\cdot \pi}{cycle} + 1\right) + LR_{\min} \end{aligned} \tag{8}$$

In Eq. (8), LR_{init} is the starting value of learning rate, LR_{min} is the minimum value of learning rate, x is the iteration number of training, mod indicates remainder operation, *cycle* is the iteration number of a training cycle. In this paper, *cycle* is equal to 8000.

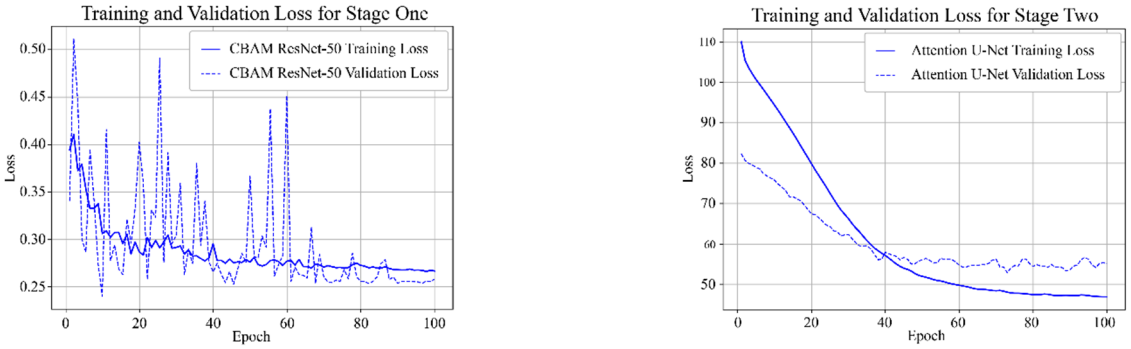
By using the method that fluctuates the value of learning rate similar to the cosine function, the model can realize a more extensive learning rate and find the better potential region of learning rate. At the same time, it can use a puny learning rate to quickly converge the model quickly. Therefore, the cosine learning rate can make the model easier to optimize to a better value.

In terms of specific numerical value, the initial learning rate of the CBAM ResNet-50 model in the first stage is set to 0.01, while the minimum value is 0.0001. In the second stage, the initial learning rate of the Attention U-Net is set to 10^{-5}, while the minimum value is 0. The choice of these values is due to the attempt of several possible optimal learning rates, and it is found that they can best adapt to this problem while a slight change in this parameter does not have a significant effect on the final result.

3.2.3 Other parameters

The optimizer of the models used in the first and the second stages is Adam optimizer (Kingma *et al.* 2015), which parameters are as follows: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $EPS = 10^{-8}$, *weight decay* $= 5 \times 10^{-4}$.

In the first stage of training, the CBAM ResNet-50 loads the parameters of the ResNet model, which is fine-tuned on the training set after pre-trained on the Imagenet dataset (Deng *et al.* 2009). Both CBAM ResNet-50 and the Attention U-Net were trained for 100 epochs.



(a) Loss curve for the first stage (b) Loss curve for the second stage

Fig. 8 The loss curve for training and validation part

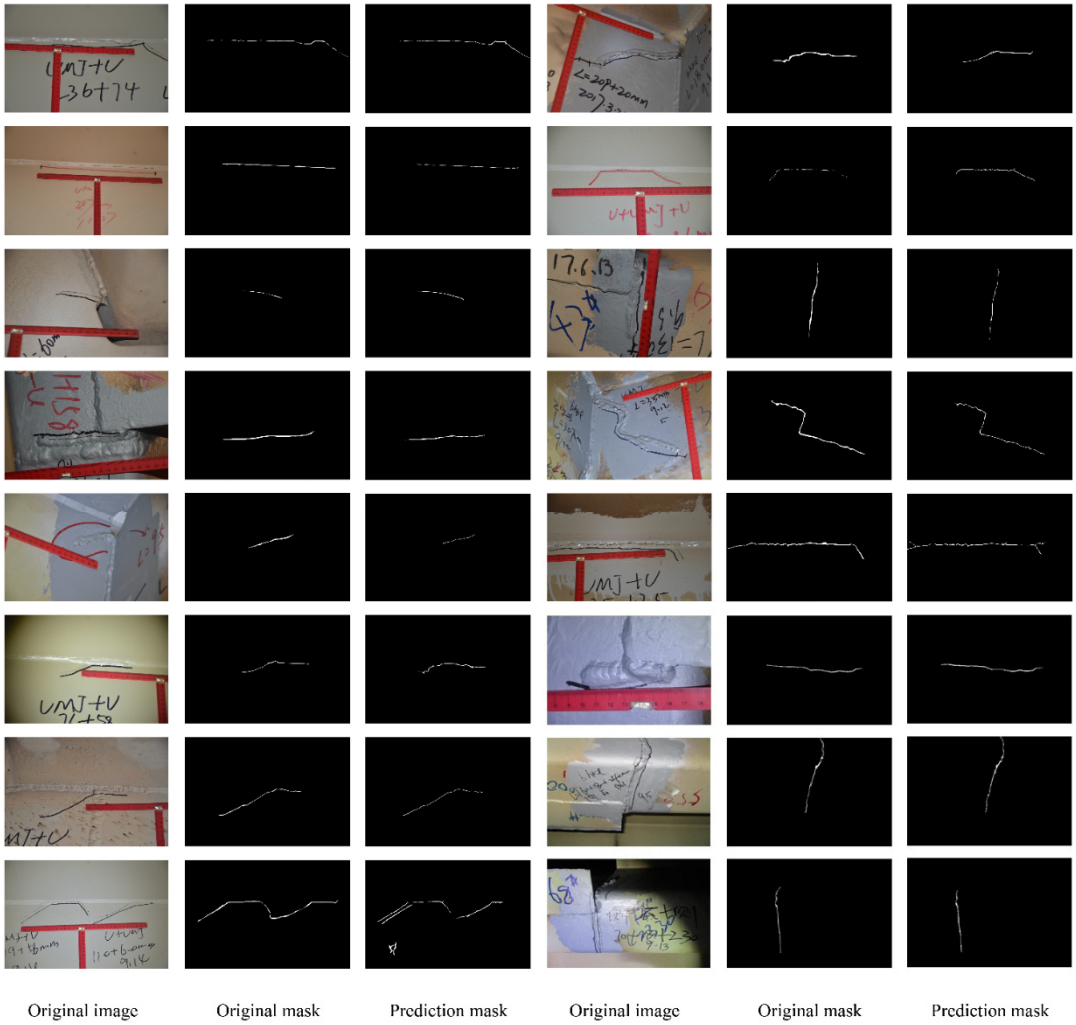


Fig. 9 The visual display of the detection result

3.2.4 Evaluating indicator

In the first stage, the purpose of the model is to identify areas with cracks in high-resolution images successfully. Thus, we choose the *Recall* as the evaluating indicator, which is equal to the number of images identified to crack images correctly divided by the total number of crack images.

In the second stage and the third stage, *IoU* is selected as the evaluation standard. The *IoU* equals the ratio of intersection and union of ground truth and prediction results. This method can effectively evaluate the performance of semantic segmentation. If *IoU* is greater than or equal to 0.5, segmentation is considered to be successful. Since most of the pixel in the images are background, only the *IoU* of the crack (not the *IoU* of the background) was a significant indicator.

3.3 Train and evaluate process

To verify the effect of the proposed model of the first stage, we used CBAM ResNet-50 to conduct an experiment on the training set, and validate the result on the test set. The loss curve of CBAM ResNet-50 in the training set and validation set is shown in Fig. 8. Correspondingly, we also experiment on the second stage by using the Attention U-Net. The loss curve of the Attention U-Net on the training set and validation set is also shown in Fig. 8. The experiment results indicate that the $Recall$ of the first stage model reaches 0.95 on the test set, while the *IoU* of the second stage model achieves 0.48.

To verify the effectiveness of the third stage of image processing, we compared the *IoU* of the input image and the output image of the third stage. Since the image processing operation in the third stage has a solid ability to obliterate the noise generated by the previous model, the *IoU* will be effectively improved. The experiment result shows that after processed by the third stage of our method, the *IoU* of the test set reaches 0.70, which is 45.8% higher than the output of the second stage model. Thus, the effectiveness of the third stage can be confirmed. The visual display of the prediction results is shown in Fig. 9.

3.4 Ablation study

To do further research on the parameters of the three-stage method and propose strategies to improve the effectiveness of our method further, we conducted ablation experiments on the following three most influential aspects: the patch size of the first stage, the selection of the classification and segmentation model, and whether the model is pre-trained.

3.4.1 The size of patch

For an entire high-resolution image, the strategy we adopted is to use a patch that retains a fix-sized sliding window to preprocess each sub-images separately. In this section, the identification results for different size of patches are presented. Since the models we use in the first and second stages are fully convolutional neural networks, the size of the input images of these two models is variable. Therefore, we selected the following five patch sizes for ablation experiments. The results of the experiments are shown in Table 1. It can be seen that 1024 × 1024 patch has the highest *IoU* in the final result.

Table 1 Ablation study for the size of patch

Patch size	*Recall* of the first stage	*IoU* of the second stage	Final *IoU*
64 × 64	0.95	0.24	0.50
128 × 128	0.94	0.41	0.54
256 × 256	0.94	0.38	0.58
512 × 512	0.94	0.40	0.63
1024 × 1024	0.95	0.48	0.70

3.4.2 Model architecture

To verify the effect of the classification and segmentation model we proposed, we selected a variety of different models to conduct ablation experiments for the first and the second stage. The test set we used in this ablation study is sub-images instead of entire high-resolution images. And the size of the sliding window we selected in this experiment is 1024 × 1024. The types of models used in the experiment and the results of the experiments are shown in Tables 2 and 3 below. It can be seen from the experimental results that the CBAM ResNet-50 model presents the best results for the classification part, and the Attention U-Net we proposed is the best for the second stage.

3.4.3 Pre-training

Pre-training is an effective technique for training neural network models. This method is widely adopted in the training phase of various neural network models. The pre-training method is to obtain prior knowledge by pre-training the model on heterogeneous datasets and then using the trained parameters as the initial parameters of the formal training model. To improve the identification effect of our method, we pre-trained the CBAM ResNet-50 and the Attention U-Net respectively. The dataset used for pre-training includes 10000 crack images, which size varies from 227 × 227 to 3456 × 5184, as shown in Fig. 10. And the size of the sliding window we selected in this experiment is 1024 × 1024. The results of the experiments are shown in Table 4. It can be seen that implementing pre-training can improve the accuracy of our crack detection model by 2.9%.

Table 2 Ablation study for the selection of classification model

Model	*Recall*
CBAM ResNet-50	0.95
SE ResNet-50	0.92
ResNet-50	0.81

Table 3 Ablation study for the selection of segmentation model

Model	*Recall*
Attention U-Net	0.67
CBAM U-Net	0.52
SE U-Net	0.52
DenseUNet-161	0.58
Attention DenseUNet-161	0.60

Table 4 Ablation study for the pre-training of model, using CBAM ResNet-50 and the Attention U-Net

Pre-trained	*IoU* of the second stage	Final *IoU*
Yes	0.49	0.72
No	0.48	0.70

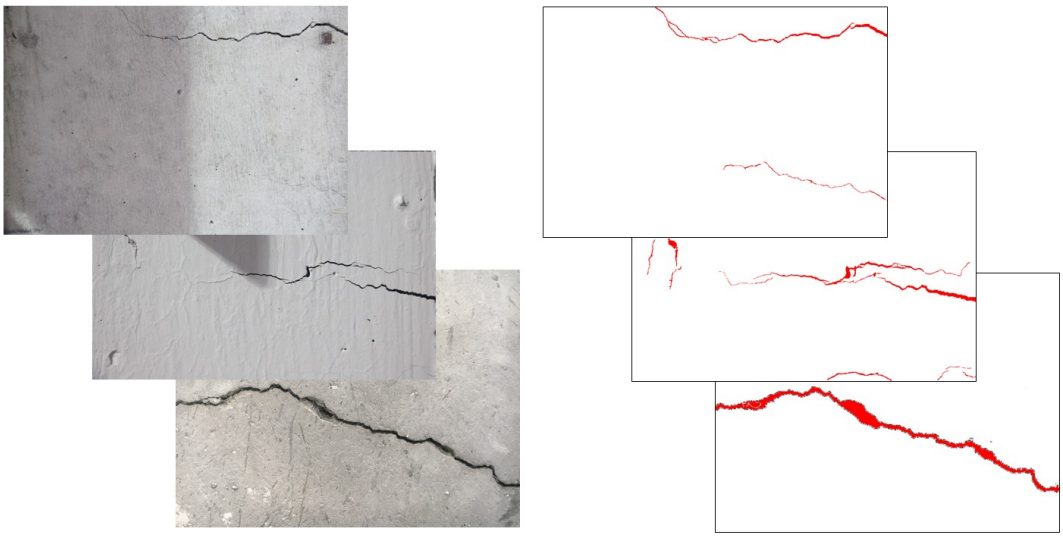


Fig. 10 Partial pre-trained dataset presentation. The left side is the original image, and the right side is the corresponding mask. The black background has been removed for display convenience

4. Conclusions

This paper proposed a novel three-stage method based on CBAM ResNet-50, the Attention U-Net and morphology operations. The first stage of the proposed model converted high-resolution images into several sub-images using a patch-based method. Then a CBAM ResNet-50 model was implemented to classify each sub-image and judge whether cracks exist. The second stage of the proposed method can detect the geometric edges of the crack images by employing the Attention U-Net. The third stage of the proposed method contains dilate operation and outlier elimination algorithm, which can remove the wrong-predicted isolated point in the predicted image of previous steps. Based on the proposed three-stage method, this paper realizes automatic identification and segmentation of the cracks in high-resolution images that captured from the inner side of the steel girder box. To verify the performance of the proposed method, ablation experiments were conducted. The conclusions are summaries as follows:

(1) Using CBAM ResNet-50, the first stage of the proposed method automatically located the regions of cracks and converted the original images into several sub-images, which greatly enhance the efficiency of the segmentation model to detect the geometric edges of cracks. As a result, the *Recall* of the trained model reaches 0.95, and the generated images form a dataset for the second stage detection.

(2) Using Attention U-Net, the second stage of the proposed method realized pixel level segmentation with a high *IoU* of 0.48 on the test set. The geometric edges of cracks can be obtained through the second stage.

(3) By using dilate operation and outlier elimination algorithm, the third stage of the proposed method can obliterate the wrong-predicted isolated points in the predicted results of the second stage model, which will improve the performance of the final output significantly while retaining the geometric edges of cracks indicated by the second stage. The final *IoU* reaches 0.70.

(4) Ablation experiments were conducted to obtain the optimized parameters and further improve the performance of the proposed method. The experiment results show that the best *IoU*, which is 0.70, is obtained when the path size is 1024 × 1024. Besides, the CBAM ResNet-50 we used performed best in the first stage, which *Recall* achieves 0.95. As for the second stage, the proposed Attention U-Net also outperform other models on the test set that composed of sub-images, which *IoU* reaches 0.67. In addition, the experiment result has proved that pre-training can improve the proposed method for crack detection, and the improvement effect can reach 2.9%.

Acknowledgments

The authors acknowledge the financial support from National Natural Science Foundation of China (Grant No. 52025083). The authors would like to thank the organizations of the International Project Competition for SHM (IPC-SHM 2020) ANCRiSST, Harbin Institute of Technology (China), and University of Illinois at Urbana-Champaign (USA) for their generously providing the invaluable data from actual structures. The authors also would like to thank the chairs of IPC-SHM 2020 Prof. Hui Li, and Prof. Billie F. Spencer Jr for their leadership on the competition.

References

Bang, S., Park, S., Kim, H. and Kim, H. (2019), "Encoder–decoder network for pixel-level road crack detection in black-box images", *Comput.-Aided Civil Infrastr. Eng.*, **34**(8), 713-

727. https://doi.org/10.1111/mice.12440
Bao, Y. and Li, H. (2021), "Machine learning paradigm for structural health monitoring", *Struct. Health Monitor.*, **20**(4), 1353-1372. https://doi.org/10.1177/1475921720972416
Bao, Y., Chen, Z., Wei, S., Xu, Y., Tang, Z. and Li, H. (2019), "The state of the art of data science and engineering in structural health monitoring", *Engineering*, **5**(2), 234-242. https://doi.org/10.1016/j.eng.2018.11.027
Bao, Y., Li, J., Nagayama, T., Xu, Y., Spencer, B.F. and Li, H. (2021), "The 1st International Project Competition for Structural Health Monitoring (IPC-SHM, 2020): A summary and benchmark problem", *Struct. Health Monitor.*, **20**(4), 2229-2239. https://doi.org/10.1177/14759217211006485
Canny, J. (1986), "A computational approach to edge detection", *IEEE Transact. Pattern Anal. Mach. Intell.*, (6), 679-698.
Cha, Y.-J., Choi, W., Suh, G., Mahmoudkhani, S. and Büyüköztürk, O. (2018), "Autonomous Structural Visual Inspection Using Region-Based Deep Learning for Detecting Multiple Damage Types", *Comput.-Aided Civil Infrastr. Eng.*, **33**(9), 731-747. https://doi.org/10.1111/mice.12334
Chen, F.-C. and Jahanshahi, M.R. (2017), "NB-CNN: Deep learning-based crack detection using convolutional neural network and Naïve Bayes data fusion", *IEEE Transact. Indust. Electro.*, **65**(5), 4392-4400. https://doi.org/10.1109/TIE.2017.2764844
de Freitas, S.T., Kolstein, H. and Bijlaard, F. (2012), "Parametric study on the interface layer of renovation solutions for orthotropic steel bridge decks", *Comput.-Aided Civil Infrastr. Eng.*, **27**(2), 143-153. https://doi.org/10.1111/j.1467-8667.2010.00693.x
Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L. (2009), "Imagenet: A large-scale hierarchical image database", *Proceedings of the IEEE conference on computer vision and pattern recognition*, Miami Beach, FL, USA, June.
Dung, C.V. and Anh, L.D. (2019), "Autonomous concrete crack detection using deep fully convolutional neural network", *Automat. Constr.*, **99**, 52-58. https://doi.org/10.1016/j.autcon.2018.11.028
Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014), "Rich feature hierarchies for accurate object detection and semantic segmentation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, June.
Han, Q., Xu, J., Carpinteri, A. and Lacidogna, G. (2015), "Localization of acoustic emission sources in structural health monitoring of masonry bridge", *Struct. Control Health Monitor.*, **22**(2), 314-329. https://doi.org/10.1002/stc.1675
He, K., Zhang, X., Ren, S. and Sun, J. (2016), "Deep residual learning for image recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, VA, USA, June.
Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E. and Saltz, J.H. (2016), "Patch-based con- volutional neural network for whole slide tissue image classification", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, VA, USA, June.
Hu, J., Shen, L. and Sun, G. (2018), "Squeeze-and-excitation networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June.
Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. (2017), "Densely connected convolutional networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July.
Kingma, D.P. and Ba, J. (2014), "Adam: A method for stochastic optimization", *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, May.
Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017), "ImageNet classification with deep convolutional neural networks", *Commun. ACM*, **60**(6), 84-90. https://doi.org/10.1145/3065386
Li, S., Zhao, X. and Zhou, G. (2019), "Automatic pixel-level multiple damage detection of concrete structure using fully convolutional network", *Comput.-Aided Civil Infrastr. Eng.*, **34**(7), 616-634. https://doi.org/10.1111/mice.12433
Lim, R.S., La, H.M. and Sheng, W. (2014), "A robotic crack inspection and mapping system for bridge deck maintenance", *IEEE Transact. Automat. Sci. Eng.*, **11**(2), 367-378. https://doi.org/10.1109/TASE.2013.2294687
Long, J., Shelhamer, E. and Darrell, T. (2015), "Fully convolutional networks for semantic segmen- tation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June.
Mutlib, N.K., Baharom, S.B., El-Shafie, A. and Nuawi, M.Z. (2016), "Ultrasonic health monitoring in structural engineering: buildings and bridges", *Struct. Control Health Monitor.*, **23**(3), 409-422. https://doi.org/10.1002/stc.1800
Oh, J.-K., Jang, G., Oh, S., Lee, J.H., Yi, B.-J., Moon, Y.S., Lee, J.S. and Choi, Y. (2009). "Bridge inspection robot system with machine vision", *Automat. Constr.*, **18**(7), 929-941. https://doi.org/10.1016/j.autcon.2009.04.003
Ong, E.P., Lee, J.A., Cheng, J., Xu, G., Lee, B.H., Laude, A., Teoh, S., Lim T.H., Wong D.W.K. and Liu, J. (2015), "A robust outlier elimination approach for multimodal retina image registration", *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, *Munich*, Germany, October.
Pan, Y., Zhang, G. and Zhang, L. (2020). "A spatial-channel hierarchical deep learning network for pixel-level automated crack detection", *Automat. Constr.*, **119**, 103357. https://doi.org/10.1016/j.autcon.2020.103357
Qiao, W., Liu, Q., Wu, X., Ma, B. and Li, G. (2021), "Automatic Pixel-Level Pavement Crack Recognition Using a Deep Feature Aggregation Segmentation Network with a scSE Attention Mechanism Module", *Sensors*, **21**(9), 2902. https://doi.org/10.3390/s21092902
Ren, S., He, K., Girshick, R. and Sun, J. (2016), "Faster R-CNN: Towards real-time object detection with region proposal networks", *arXiv:1506.01497 [cs]*. http://arxiv.org/abs/1506.01497
Ronneberger, O., Fischer, P. and Brox, T. (2015), "U-net: Convolutional networks for biomedical image segmentation", *Proceedings of International Conference on Medical Image Computing and Computer- Assisted Intervention*, Munich, Germany, October.
Saouma, V.E., Barton, C.C. and Gamaleldin, N.A. (1990), "Fractal characterization of fracture surfaces in concrete", *Eng. Fract. Mech.*, **35**(1-3), 47-53. https://doi.org/10.1016/0013-7944(90)90182-G
Sofia, T.D.F., Henk, K. and Frans, B. (2012), "Parametric study on the interface layer of renovation solutions for orthotropic steel bridge decks", *Comput.-Aided Civil Infrastr. Eng.*, **27**(2), 143-153. https://doi.org/10.1111/j.1467-8667.2010.00693.x
Song, W., Jia, G., Jia, D. and Zhu, H. (2019), "Automatic Pavement Crack Detection and Classification Using Multiscale Feature Attention Network", *IEEE Access*, **7**, 171001-171012. https://doi.org/10.1109/ACCESS.2019.2956191
Spencer Jr, B.F., Hoskere, V. and Narazaki, Y. (2019), "Advances in computer vision-based civil infrastructure inspection and monitoring", *Eng.*, **5**(2), 199-222. https://doi.org/10.1016/j.eng.2018.11.030
Sun, B.-C. and Qiu, Y. (2007). "Automatic identification of pavement cracks using mathematic morphology", *Proceedings of International Conference on Transportation Engineering 2007*, pp. 1783-1788.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017), "Attention is all you need", *arXiv preprint arXiv*:*1706.03762*.

Wan, H., Gao, L., Su, M., Sun, Q. and Huang, L. (2021), "Attention-Based Convolutional Neural Network for Pavement Crack Detection", *Adv. Mater. Sci. Eng.* https://doi.org/10.1155/2021/5520515

Woo, S., Park, J., Lee, J.Y. and Kweon, I.S. (2018), "Cbam: Convolutional block attention module", *Proceedings of the European Conference on Computer Vision*, Munich, Germany, September.

Xu, Y., Bao, Y., Chen, J., Zuo, W. and Li, H. (2019), "Surface fatigue crack identification in steel box girder of bridges by a deep fusion convolutional neural network based on consumer-grade camera images", *Struct. Health Monitor.*, **18**(3), 653-674. https://doi.org/10.1177/1475921718764873

Yang, Y. and Nagarajaiah, S. (2014), "Blind identification of damage in time-varying systems using independent component analysis with wavelet transform", *Mech. Syst. Signal Process.*, **47**(1-2), 3-20. https://doi.org/10.1016/j.ymssp.2012.08.029

Yang, Y. and Nagarajaiah, S. (2016), "Harnessing data structure for recovery of randomly missing structural vibration responses time history: Sparse representation versus low-rank structure", *Mech. Syst. Signal Process.*, **74**, 165-182. https://doi.org/10.1016/j.ymssp.2015.11.009

Yang, X., Li, H., Yu, Y., Luo, X., Huang, T. and Yang, X. (2018), "Automatic pixel-level crack detection and measurement using fully convolutional network", *Comput.-Aided Civil Infrastr. Eng.*, **33**(12), 1090-1109. https://doi.org/10.1111/mice.12412

Ye, X.-W., Jin, T. and Chen, P.-Y. (2019), "Structural crack detection using deep learning–based fully convolutional networks", *Adv. Struct. Eng.*, **22**(16), 3412-3419. https://doi.org/10.1177/1369433219836292

Zhang, A., Wang, K.C.P., Fei, Y., Liu, Y., Tao, S., Chen, C., Li, J.Q. and Li, B. (2018), "Deep learning–based fully automated pavement crack detection on 3D asphalt surfaces with an improved CrackNet", *J. Comput. Civil Eng.*, **32**(5), 04018041. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000775