DOI: 10.32604/sdhm.2022.018632

ARTICLE

Crack Segmentation Based on Fusing Multi-Scale Wavelet and Spatial-Channel Attention

Peng Geng[*], Ji Lu, Hongtao Ma and Guiyi Yang

School of Information Sciences and Technology, Shijiazhuang Tiedao University, Shijiazhuang, 050043, China

[*]Corresponding Author: Peng Geng. Email: gengpeng@stdu.edu.cn

Received: 07 August 2021 Accepted: 17 May 2022

ABSTRACT

Accurate and reliable crack segmentation is a challenge and meaningful task. In this article, aiming at the characteristics of cracks on the concrete images, the intensity frequency information of source images which is obtained by Discrete Wavelet Transform (DWT) is fed into deep learning-based networks to enhance the ability of network on crack segmentation. To well integrate frequency information into network an effective and novel DWTA module based on the DWT and scSE attention mechanism is proposed. The semantic information of cracks is enhanced and the irrelevant information is suppressed by DWTA module. And the gap between frequency information and convolution information from network is balanced by DWTA module which can well fuse wavelet information into image segmentation network. The Unet-DWTA is proposed to preserved the information of crack boundary and thin crack in intermediate feature maps by adding DWTA module in the encoder-decoder structures. In decoder, diverse level feature maps are fused to capture the information of crack boundary and the abstract semantic information which is beneficial to crack pixel classification. The proposed method is verified on three classic datasets including CrackDataset, CrackForest, and DeepCrack datasets. Compared with the other crack methods, the proposed Unet-DWTA shows better performance based on the evaluation of the subjective analysis and objective metrics about image semantic segmentation.

KEYWORDS

Attention mechanism; crack segmentation; convolutional neural networks; discrete wavelet transform

1 Introduction

In civil buildings and infrastructure, such as bridges, roads and so on, concrete cracks are one of the most common damages [1]. Manual inspection is used on crack detection in practice, but it shows low detection efficiency and subjective results. Continuously exposed to the environment, exterior cracks will lead to more severe damage to infrastructure. Therefore, it is significant to automatically detect concrete cracks based on computer vision technology.

In the past, various detecting cracks methods based on traditional computer vision technology have been proposed, such as edge detection methods [2], threshold-based methods [3], and spectral analysis methods [4]. But these conventional crack detection methods are easily affected by blurs, shadows, and scratches in the actual environment [1]. In recent years, with developing of deep learning, Convolutional Neural Networks (CNNs) are used in image recognition [5], object detection [6], and semantic segmentation [7].

Recently, CNNs are also applied in crack detection [8,9]. Cha et al. [10] used CNNs to detect crack damage patches in images. The result showed that CNNs was effective and precise compared with the conventional methods. However, this method only solved the problem of image classification and cannot complete the task of crack pixel classification.

In the methods of crack segmentation, there are some approaches to achieve crack segmentation by Fully Convolutional Neural Networks (FCNN) in which the fully connected layer is replaced by convolutional layers [7,11–13]. FCNN obtains diverse meaningful features in different stages. The low-level convolutional layers keep more structure information of crack. The high-level convolutional layers obtain abstract semantic information which is beneficial to identify the pixel category, but boundary and position information of crack is lost in the high-level convolutional layers. Therefore, it is difficult to obtain refined segmentation results only by the high-level convolutional feature maps. To tackle this problem, Wang et al. [14] proposed CrackFCN network where more low-level convolutional feature maps were used to preserve local detail, at the same time, they increased the network depth for larger input images. The CrackFCN network presented better performance than FCNN did. Subsequently, Unet was originally proposed for biomedical image segmentation by Ronneberger et al. [15]. It achieved excellent performance with less training images. Later, it is widely applied in semantic segmentation tasks [16–18]. To segment crack in concrete images, Liu et al. [19] firstly applied Unet model in crack segmentation and adopted FocalLoss as the loss function. The result showed that Unet had a better performance compared with DCNN-based approach [10]. Compared with conventional computer vision methods, FCNN-based methods can obtain rich image features through self-data learning. Meanwhile, FCNN-based methods present higher precision and stronger robustness than conventional methods. But there are weaknesses in deep learning-based segmentation network of the thin crack segmentation. Due to the features of thin crack are hardly extracted by networks, it is hard to accurately predict the thin crack by FCNN-based networks.

To tackle above problem, the method of wavelet feature extracted by DWT is introduced which can supplement the frequency information of crack into FCNN-based network. By DWT, the frequency information of crack boundary can effectively be obtained. At the same time, the background information (such as scratch, dirt and so on) which is not conducive to be predicted as crack is also extracted. To better integrate frequency information of crack into network and reduce the influence of the background information, a simple and effective Discrete Wavelet Transform and Attention (DWTA) module is proposed. The DWTA module by combining the DWTA and scSE attention module can balance the gap between frequency information from DWT and convolutional feature information from FCNN-based network, respectively. Besides, the DWTA module enhances the response of crack region and suppresses the response of background region in spatial dimension. By adding DWTA module in an encoder-decoder structures, an effective and novel crack segmentation Unet-DWTA network is proposed. The information of crack boundary and thin crack are well preserved in intermediate feature maps in Unet-DWTA network. The fused feature maps contain the information of crack boundary and the abstract semantic information which are beneficial to crack pixel classification. The organization of this article is as follows: In Section 2, related works are summarized. The proposed Unet-DWTA segmentation network is introduced in Section 3. In Section 4, the datasets are described and experimental results are discussed. In Section 5, the conclusion is provided.

2 Related Works

2.1 Crack Segmentation Network

In recent research, FCNN-based networks have been widely used in automatically cracks segmentation. The encoder-decoder architecture with the skip connection module is a popularity architecture in the field of semantic segmentation. The different scale feature maps are extracted in the encoder. In the task of detection crack, the low-level feature maps retain more boundary and position information of the crack, but more noises (scratches and dirt) also are contained in low-level feature maps. In contrast, high-level feature maps obtain high semantic information, but the position information and boundaries of crack are lost in high-level feature maps. To tackle this problem, skip connection is introduced in the network to fuse the different-level feature maps. The predicted results not only obtain refined crack boundaries but also obtain accurate pixel classification by the fused feature maps. Based on this structure, some approaches have been proposed to better deal with crack segmentation task [14,19]. Based on FCNN, Zou et al. [20] proposed a deep hierarchical feature learning architecture (DeepCrack) to extract crack area. The Deeply-Supervised Nets (DSN) training strategy is adopted to supervise the feature maps which are got from each convolutional stage in DeepCrack network. Predicted results of each stage are concatenated to achieve final prediction. Due to the low-level convolutional stage contains more noise information, Deepcrack mistakenly predicted some background pixels as crack pixels. Ren et al. [21] proposed CrackSegnet network based on VGG16 architecture for crack segmentation. The dilated convolution [22], spatial pyramid pooling [23] and skip connection module [15] were also applied in CrackSegnet network. CrackSegnet showed better performance on wide crack segmentation. Though these methods have achieved good results, thin crack cannot be predicted well, due to the thin crack information is lost in encoder part of the methods mentioned above. In recent works, DWT is applied to enhance the feature extraction ability of CNNs. Therefore, we decide to add it into deep learning-based segmentation network to improve the crack segmentation capability of network.

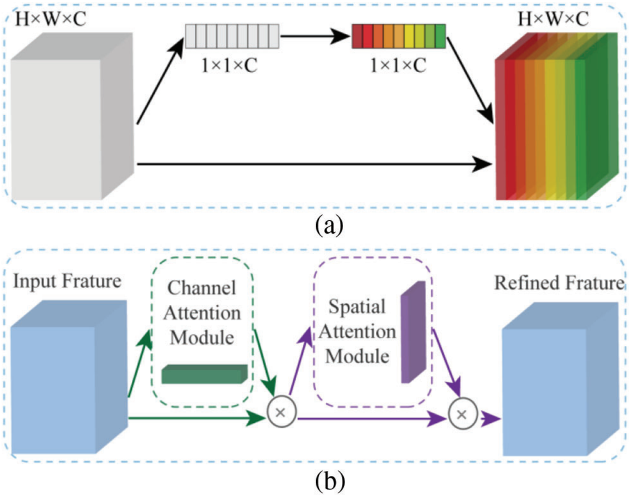
2.2 DWT

In recent years, there are some attempts using wavelet transform to extract image feature and then image features are fed into a classification to achieve better classification results. Wang et al. [24] combined anti-symmetrical bi-orthogonal wavelets and structural random forest method to achieve edge detection of cracks. The anti-symmetrical bi-orthogonal wavelet was applied to extract crack feature and the structure random forests are used to complete classification task. In addition, there are some methods that use wavelet transform to help CNN achieving better results in various vision tasks, including classification [25], segmentation [26,27], image restoration [28]. Liu et al. [29] found that CNNs can be regarded as a limited multi-resolution analysis method. This conclusion indicates that a great deal of frequency information would be lost in traditional CNNs. For image texture classification, Fujieda et al. [30] introduced multi-scale DWT into CNNs to make up the frequency information. Experiments showed that wavelet neural networks obtained higher classification accuracy. Duan et al. [26] proposed a method of SAR image segmentation based on convolutional-wavelet neural networks. The conventional pooling was replaced by a wavelet pooling layer in network. More structures information which is beneficial to SAR image segmentation is kept in the wavelet pooling layer. The experiment showed that this approach preserved more information of edge and obtained better performance. Fujieda et al. [31] added the high-frequency information from DWT into FCNN. This method obtained a more accurate segmentation result of road markings than FCNN obtained. Wang et al. [32] proposed a method that consisted of structural random forest methods [24] and FCNN to improve accuracy of crack segmentation and address the problems that local information was lost and the capacity of partial refinement in FCNN methods. DWT not only captures the frequency information of crack, but also obtains frequency feature maps from source image at different resolutions. To provide the network with accessing fully-spectral analysis of input images, multi-level DWT is applied. The high-frequency information including crack boundary is preserved in the FCNN-based network when DWT decomposition of source images is added into network. The feature extraction ability of network is increased by multi-level DWT. However, wavelet feature maps also contain background information (such as scratch and dirt) which is not conducive to predict crack. In addition, there is the gap between wavelet information and feature maps of convolutional stages. To better integrate wavelet information into segmentation network, an attention block is introduced in this paper.

2.3 Attention Mechanism

Recently, self-attention block has been extensively studied. Self-attention block is used to make up for the shortcomings of CNNs in computer vision tasks. A series of local information of object are got by CNNs. However different local features have different effects on representing scene information. CNNs cannot well describe the importance of different local features, so self-attention block is proposed to solve this problem effectively. Self-attention block is inspired by human visual characteristics which does not obtain all information of object at once. Instead, a series of local information of object are got, then salient local information is selected to get the perception of object. Self-Attention block allows CNNs to selectively focus on important information and suppress irrelevant information to improve the performance of the network.

Hu et al. [33] proposed Squeeze-and-Excitation (SE) module shown in Fig. 1a. The SE module which only extracts channel-information guides CNNs to enhance meaningful feature channels and suppress useless feature channels. The SE module achieved high performance in image classification task. Roy et al. [34] introduced Spatial Squeeze and Channel Excitation (cSE), Channel Squeeze and Spatial Excitation (sSE) and Concurrent Spatial and Channel Squeeze & Excitation (scSE) module to increase segmentation capability of FCNN. The cSE module is the same as the SE module. For segmentation task, the spatial information is more useful. Therefore, Roy proposed sSE module to strengthen spatial information. The spatial re-weighting map is obtained by 1 × 1 convolution to enhance useful information and suppress irrelevant information in spatial dimension. The scSE module consists of cSE module and sSE module. The feature maps are recalibrated by scSE module in spatial dimension and channel dimension. Woo et al. [35] proposed Convolutional Block Attention Module (CBAM) shown in Fig. 1b. The CBAM and scSE block can easily be added into CNNs aggregate channel and spatial information to refine input feature. For crack images, wavelet feature maps from DWT contain the information of crack boundary and the background information. The background information is not conducive to predict cracks. In this paper, scSE module is introduced to fuse convolution and wavelet features. The crack region response in fused feature maps is intensified and the influence of background information is suppressed in spatial dimension. At the same time, scSE module select channel-wise attention to selects more useful channel on wavelet feature and convolutional feature.



(a)

(b)

Figure 1: (a) Represents SE module; (b) Represents CBAM module

3 Proposed Method

In Section 3, DWTA module is introduced and the proposed deep learning-based segmentation network Unet-DWTA is described. Finally, the combining loss function is described.

3.1 Discrete Wavelet Transform and Attention Enhance Module

The core idea of the FCNN-based network is to extract image features by building multiple convolution layers and pooling layers. With the progress of convolution and pooling, the receptive field are gradually increasing and high-level semantic information is extracted. However, the information of thin crack is lost. In addition, the FCNN-based network processes the images in the spatial domain and the partial spectral domain. A lot of frequency information is lost in network. In 2D DWT, four filters $f_{ll}, f_{lh}, f_{hl}, f_{hh}$ are applied to convolve with an image X [36]. One low frequency subband F_{ll} and three high frequency subband F_{lh}, F_{hl} and F_{hh} are extracted by downsampling the convolutional results.

$$\begin{aligned}
F_{ll} &= (f_{ll} \otimes X) \downarrow 2 \\
F_{lh} &= (f_{lh} \otimes X) \downarrow 2 \\
F_{hl} &= (f_{hl} \otimes X) \downarrow 2 \\
F_{hh} &= (f_{hh} \otimes X) \downarrow 2
\end{aligned} \tag{1}$$

The approximation coefficients are continually computed by formula (1), and filters in multi-level DWT and the results are sub-sampled by a factor of 2. The resolution of the wavelet feature maps is 1/2 and 1/4 of input image in second-level DWT shown in Fig. 2, respectively. In this work, the Haar wavelet is used. In 2D Haar wavelet filters are defined as,

$$f_{ll} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, f_{lh} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, f_{hl} = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, f_{hh} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{2}$$

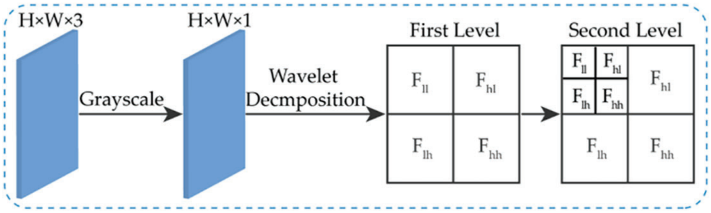


Figure 2: The different level DWT

Due to the feature maps are downsampled by a factor of 2 in MaxPooling layer, the wavelet feature maps can be concatenated with MaxPooling feature maps with the same resolution. In addition, because the crack boundary is presented in high frequency information of spectral domain, f_{lh}, f_{hl}, and f_{hh} are selected to form wavelet feature maps. The channel of wavelet feature maps is 3. Lost information is supplemented by integrating wavelet feature maps into FCNN-based network. The convolutional layer captures the local spatial patterns of all input channels by learning multiple filters and generates a new feature map group. Although the wavelet feature maps and the MaxPooling feature maps can be simply concatenated to supplement frequency information, the channel number of MaxPooling feature map is much greater than the channel number of wavelet feature map.

This leads to the dominance of MaxPooling feature maps. What's more, the high-frequency feature maps not only contain crack edge information but also include some local noises information which is useless. In order to eliminate the influence of background information and leverage better the high frequency information for crack segmentation, the proposed simple and effective DWTA module shown in Fig. 3 is proposed. The DWTA module consists of two inputs and one output. The inputs are feature maps of MaxPooling layer in Unet and wavelet feature maps. Let us suppose an input feature map $F_M \in R^{H\times W\times C}$ that passes through a MaxPooling layer and another input feature map $F_D \in R^{H\times W\times 3}$ which is obtained by DWT. The output feature map $F_U \in R^{H\times W\times 2C}$. H and W represent the height and width of the feature map, respectively. C represents channel of feature map. The wavelet feature map F_D are firstly fed into a 3×3 convolution layer to generate new feature map $F_D^{'} \in R^{H\times W\times C}$. The channel number of output maps is consistent with the channel number of MaxPooling feature maps to balance the channel gap between wavelet feature map F_D and MaxPooling feature map F_M. Then, the wavelet feature maps are concatenated with MaxPooling feature maps to generate new feature map $F_N \in R^{H\times W\times 2C}$,

$$F_N = [F_M, \phi(Cov^{3\times 3\times C}(F_D))] \tag{3}$$

where ϕ is Relu activation function, $Cov^{3\times 3\times C}$ is 3×3 convolutional layer and the output channel is C.

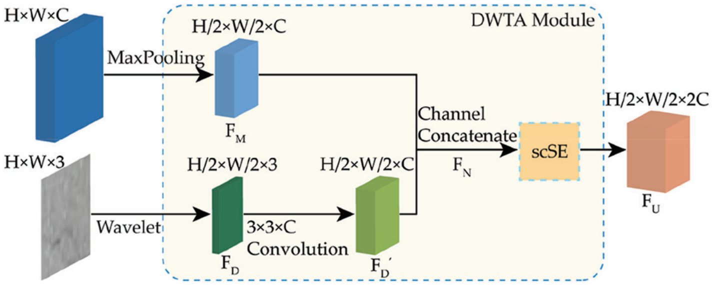


Figure 3: The proposed DWTA module

Finally, the feature map F_N is fed into scSE module. ScSE module is a combination of two attention mechanisms. As is shown in Fig. 4, the top part is the spatial attention module and the bottom part is the channel attention module.

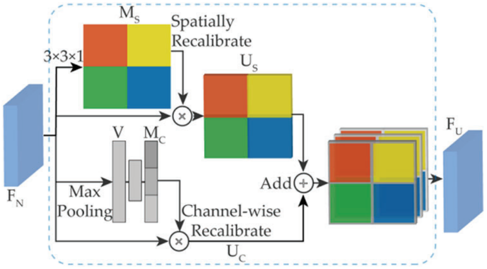


Figure 4: The architecture of scSE

Spatial attention map represents the inter-spatial relationship of feature maps. The spatial attention focuses on 'where' is crack area. The input feature map F_N passes through a convolutional layer which kernel size is 3×3 to generate the spatial attention map $M_S \in R^{H\times W\times 1}$. Unlike original spatial attention module, the 1×1 convolutional layer is replaced by 3×3 convolutional layer. A large kernel size is effective to capture long-range contextual information and it can effectively eliminate the local noise of feature map F_N. The spatial attention map M_S is calculated by,

$$M_S = \delta\left(Cov^{3\times3\times1}(F_N)\right) \tag{4}$$

where $Cov^{3\times3\times1}$ is 3 * 3 convolutional layer and the output channel is 1, δ is Sigmoid activation function. M_s is adaptively adjusted to enhances the crack area and ignore background area.

The recalibrated feature map $U_S \in R^{H\times W\times 2C}$ is calculated by,

$$U_S = \left[M_S{}^{1,1}F_N{}^{1,1}, M_S{}^{1,2}F_N{}^{1,2}, \cdots, M_S{}^{\mathrm{i},j}F_N{}^{i,j}, \ldots, M_S{}^{h,w}F_N{}^{h,w}\right] \tag{5}$$

where $i \in$ and $j \in (1, H)$, $M_S{}^{\mathrm{i},j}$corresponds to (i,j) position information of M_S, $F_N{}^{i,j}$ corresponds to (i,j) position information of F_N.

Channel attention map represents the inter-channel relationship of feature maps. As each feature map is considered as a feature detector [37], channel attention selects 'which' is useful in input feature map. Global MaxPooling is performed to get the vector $V \in R^{1\times1\times2C}$ in channel attention module. The vector V is transformed to V' by two fully-connected layers. The first weight is $W_1 \in R^{2C\times C}$ and the second weight is $W_2 \in R^{C\times 2C}$. The V' is activated by sigmoid activation function. The channel attention vector $Mc \in R^{1\times1\times2C}$ is calculated by,

$$M_C = \delta\left(W_2\left(\phi\left(W_1(MaxPool(F_N))\right)\right)\right) \tag{6}$$

where MaxPool is Global MaxPooling, ϕ is Relu activation function, δ is Sigmoid activation function. M_C is adaptively adjusted to enhances the useful channel and ignore irrelevant channels.

The recalibrated feature map $U_C \in R^{H\times W\times 2C}$ is calculated by,

$$U_C = \left[M_C{}^{1}F_N{}^{1}, M_C{}^{2}F_N{}^{2}, \cdots, M_C{}^{\mathrm{i}}F_N{}^{i}, \cdots, M_C{}^{2C}F_N{}^{2C}\right] \tag{7}$$

where $i \in (1, 2C)$, $M_S{}^{\mathrm{i}}$ corresponds to $i-$th value of M_S which indicates the importance of the $i-$th channel. $F_N{}^{i}$ corresponds to $i-$th channel of F_N.

The final recalibrated feature map F_U is obtained by element-wise addition of feature map U_S and feature map U_C.

$$F_U = U_S + U_C \tag{8}$$

In this way, the thin crack information lost in MaxPooling layer is supplemented to intermediate feature maps. At the same time, the intermediate feature maps get better representation. The ability of feature expression which boosts the crack segmentation ability of FCNN is improved.

3.2 The Proposed Unet-DWTA Model

At present, most semantic segmentation networks adopt encoder-decoder structures. In this paper, a Unet-like model is adopted as the basic network architecture. The feature maps in the encoder have a resolution similar to that of the decoder and is integrated into the decoder through skip connections. By fusing the hierarchical features of the encoder, the decoder gradually increases the spatial resolution and fills the missing details. Fig. 5 shows the overall framework of the proposed Unet-DWTA model. The Unet-DWTA network includes three parts: encoder part, decoder part and DWTA module. There are 4 steps in each of the encoder and decoder, and the encoder and decoder are connected through the central layer. In the encoder, each downsampling step is composed of two convolution blocks, a DWTA block and a downsampling convolution, a batch normalization (BN) layer, and a rectified linear unit (ReLU). Due to the thin crack feature is hardly extracted in encoder and the frequency information is lost in convolutional layers and MaxPooling layers of encoder. To deal with these problems, the wavelet feature maps of source images obtained by Multi-level DWT and feature maps of MaxPooling layer are fed into DWTA module. DWTA module is seamlessly added to the middle of the two convolution stages of Unet network. DWTA module guides wavelet feature maps to better fuse the frequency information with feature maps from encoder. By adding DWTA module, the information of crack boundary and thin crack are well preserved in intermediate feature maps. Moreover, each upsampling step in the decoder uses the nearest neighbor sampling to double the upsampling first, and the following two convolution blocks and attention modules are the same as the downsampling step in the encoder. In the central layer, there are only two convolution blocks.

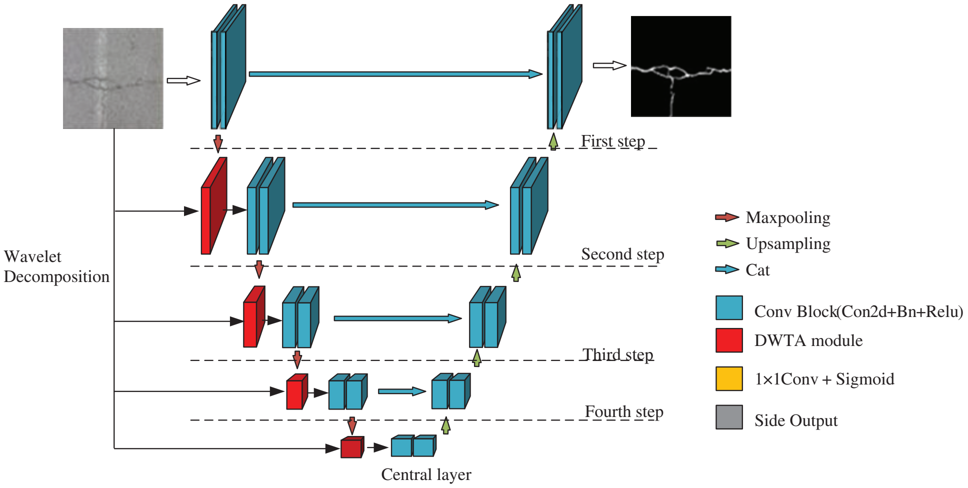


Figure 5: The proposed Unet-DWTA model

3.3 Loss Function

The loss function is used to compute the distance between the current output of the algorithm and the expected output. It is a method to evaluate how the algorithm models the data. A suitable loss function chosen is crucial to train deep learning-based network. For semantic segmentation tasks, loss functions such as cross entropy loss function, Dice loss function, Focal loss function and so on, are selected as training loss function. In this paper, Dice loss function and modified cross-entropy loss function are integrated into a combining loss function to solve the unbalanced classification problem. For a binary classification problem, typical cross-entropy loss function can be expressed as,

$$L(W) = -\frac{1}{N}\sum\nolimits_{\mathrm{n}=1}^{N} y_n \log \hat{y}(x_n, W) + (1 - y_n)(1 - \log \hat{y}(x_n, W)) \tag{9}$$

where N represents the number of image pixel, x_n is the input pixel value, $y_n \in \{0, 1\}$ corresponds to the true label, $\hat{y} \in \{0, 1\}$ corresponds to the prediction probability, W is trainable weight matrix and $L(\cdot)$ represents the loss function.

However, most pixels in one piece of concrete image belong to the non-crack category for crack segmentation. It causes the unbalanced classification problem. The pixels of image tend to be predicted as non-crack pixels when the cross-entropy loss function is adopted. Therefore, the typical cross-entropy loss function is modified as the weight cross-entropy loss function. When crack pixel is predicted as no-crack pixel, the loss value will be improved. This is achieved by introducing the parameter λ into the cross-entropy loss function. The weight cross-entropy loss function is defined as,

$$L(W) = -\frac{1}{N}(\lambda \sum_{n=1}^{N} y_n \log \hat{y}(x_n, W) + \sum_{n=1}^{N} (1 - y_n) \log(1 - \hat{y}(x_n, W))) \tag{10}$$

Dice coefficient (Dice) is similarity measurement function usually used to calculate the similarity of two samples. The formula is as follows:

$$Dice = \frac{2|\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X}| + |\mathrm{Y}|} \tag{11}$$

In semantic segmentation, X is Ground Truth image, and Y is predicted image. Dice is one of the evaluation indexes for image segmentation. The intersection ratio between the predicted area and the ground truth area is calculated by Dice. All the pixels in the same category are together calculated loss value in Dice loss function which is also used to solve the unbalanced classification problem. Dice loss function is defined as,

$$D\mathrm{iceLoss} = 1 - \frac{2|\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X}| + |\mathrm{Y}|} \tag{12}$$

In the end, the proposed combining loss function is as follows:

$$L\mathrm{oss} = \mathrm{L}_W + L_{\mathrm{Dice}} \tag{13}$$

Loss is adopted in the train stage. L_W is the weight cross-entropy loss function and L_{Dice} is Dice loss function.

4 Experiments and Discussion

4.1 Dataset

To verify the effectiveness of the proposed Unet-DWTA, 118 images in CrackForest [38] dataset, 537 images in DeepCrack dataset [20] and 408 images in CrackDataset dataset [39] are selected in this paper. CrackForest dataset includes more thin crack images, DeepCrack dataset and CrackDataset are common crack dataset consisting of various crack image including shadows, small holes, and other interference noise. In CrackForest dataset, 100 images of CrackForest dataset are used as training images and 18 images are used as the testing dataset. In DeepCrack dataset, the training dataset includes 459 images and the testing dataset contains 78 images. The size of the crack image in CrackForest and CrackForest dataset is 448 × 448 × 3. The size of the label image named as ground truth is 448 × 448 × 1. 80% images of CrackDataset dataset are used as the training dataset and almost 20% images of CrackDataset dataset are used as the testing dataset. The size of the crack image in CrackDataset dataset is 768 × 768 × 3 and the size of the label image named as ground truth is 768 × 768 × 1. Some images and ground truth in three crack datasets are shown in Fig. 6. Rows 1, 3 and 5 show the original images of CrackForest dataset, DeepCrack dataset and CrackDataset dataset, respectively. Rows 2, 4 and 6 show the Ground truth of CrackForest dataset, DeepCrack dataset and CrackDataset dataset, respectively. They are used to compare the performance of different approaches in the following experiments.

4.2 Implementation

In experiments, the publicly available Keras library is used to build networks. The Adam [40] optimization function is used in experiments. The size of the batch is 4 and the learning rate is 0.0001. The size of input images is resized to 256 × 256 × 3 in train stage. The networks are trained on a single Nvidia 1080ti GPU. Because the concrete surface image is simple, the pre-trained VGG16 [41] is not adopted in experiments. Global threshold is set as 0.5 to binarize the probability maps in experiments.

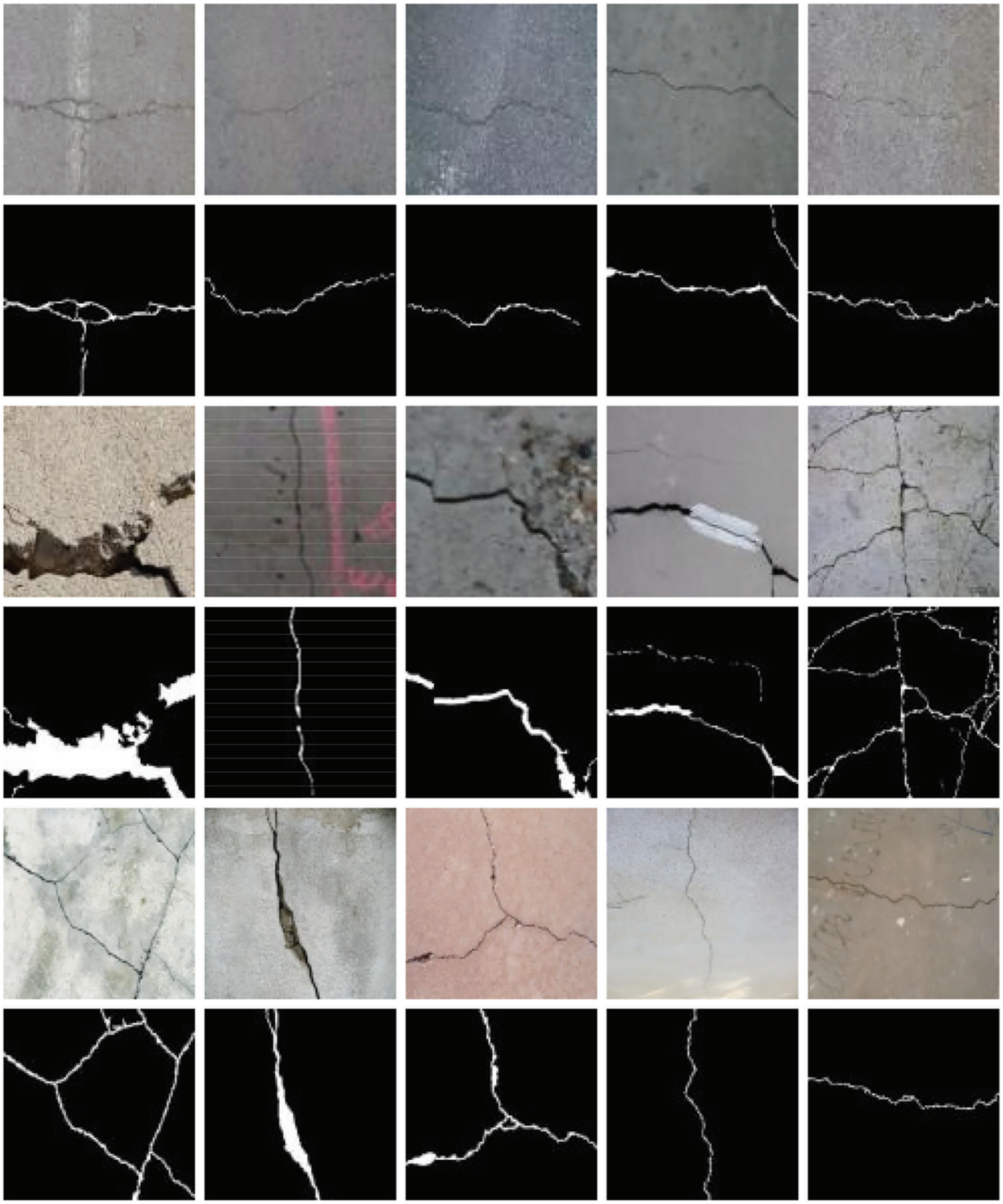


Figure 6: Some images on three crack datasets

4.3 Evaluation Mertric

In crack segmentation task, image pixels are considered to belongs to two categories (crack and non-crack). For evaluating the performance of different networks in the experiment. Global accuracy (Accuracy), Mean accuracy, Mean intersection over union (Mean IOU), Recall, Precision, and F1-score are used as the metrics.

The calculation formula of Accuracy is,

$$Acc = \sum_i (n_{ii} / \sum_i t_i) \tag{14}$$

The calculation formula of Mean accuracy is,

$$Meanacc = \frac{1}{n_{cls}} \sum \frac{n_{ii}}{t_i} \tag{15}$$

where n_{cls} is the number of category , and t_i is the total number of pixels of the class i, n_{ii} is the number of category i pixels predicted as the category j.

The calculation formula of Mean IOU is,

$$\mathrm{mean}IOU = \frac{1}{\mathrm{n}_{cls}} \sum_i n_{ii} / (t_i + \sum_j n_{ji} - n_{ii}) \tag{16}$$

Recall represents the percentage of crack pixels classified correctly in all crack pixels and Precision represents the percentage of crack pixels classified correctly in all detected crack pixels.

The Recall and Precision are calculated as,

$$Recall \ = \ T\mathrm{rue}Positives / (TruePositives + FalseNegatives) \tag{17}$$

$$Precision \ = \ TruePositives / (TruePositives + FalsePositives) \tag{18}$$

The F1-score is defined as the weighted harmonic mean of the test precision and recall. The F1-score is calculated as,

$$F1 - score = \ 2 * Precision * Recall / (Recall + Precision) \tag{19}$$

4.4 Experimental Results

The performance of the proposed model is evaluated by comprehensive experiments. The ablation experiments are firstly conducted to investigate the influence of DWT, DWTA module and the combining loss function. The effect of multi-level DWTA module is investigated on Unet network. We further compare the performance of different attention mechanisms in DWTA module. In the end, the effectiveness of proposed model also is verified on FCN-8 and two state-of-the-art crack detection algorithms.

4.4.1 Ablation Experiments

For verifying the influence of DWT, DWTA module and the combining loss function (Combining) in crack segmentation, the ablation experiments are conducted to compare the performance after using these modules. The typical cross-entropy loss function is used to train different network. In addition, the Haar wavelet base is used in the DWT and DWTA module. The loss curves and accuracy curves with epochs on training and testing datasets of two datasets are shown in Figs. 7 and 8. With training iterations increasing, the loss values of train datasets and test datasets both are lower than the original Unet, when DWTA module and DWT module are introduced into the Unet network. At the same time, the accuracy rate is improved by introducing DWT and DWTA module. The quantification results of network using different modules in CrackForest dataset and DeepCrack dataset are shown in Tables 1 and 2, respectively.

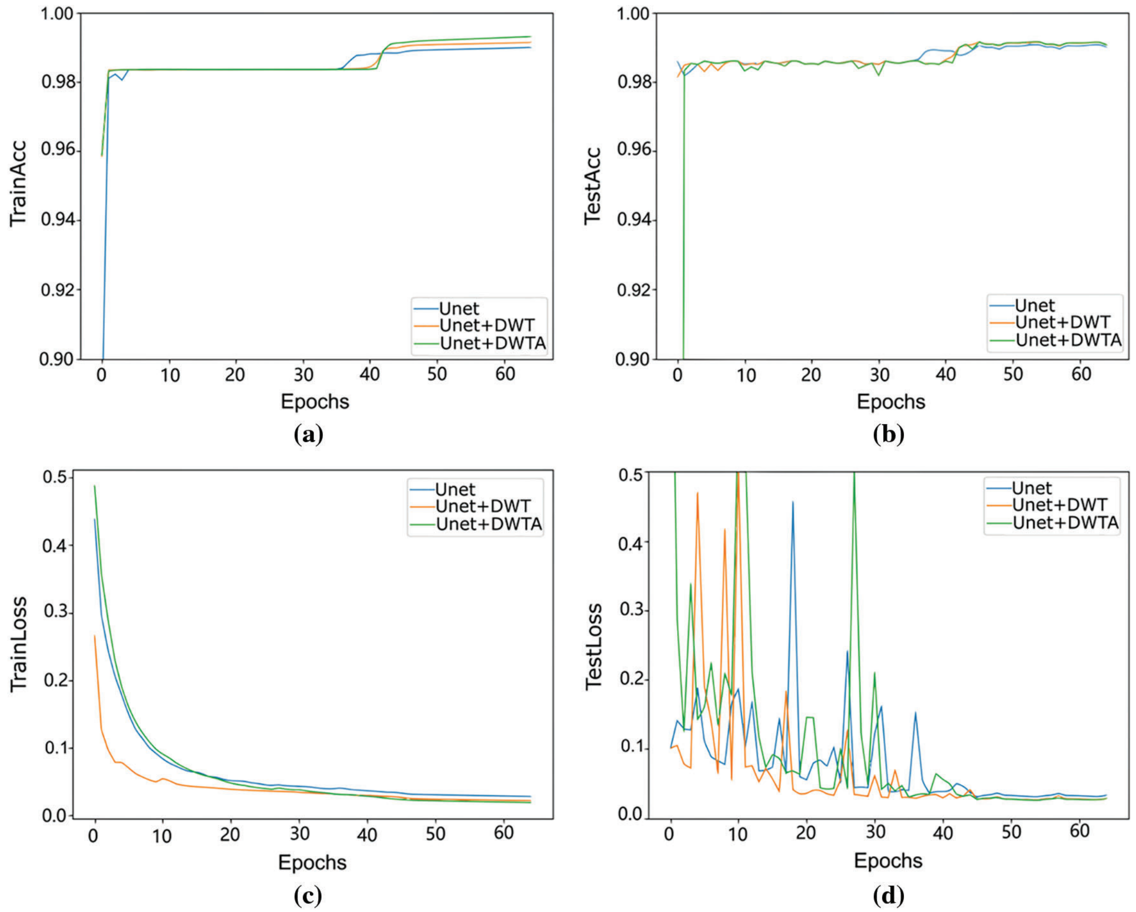
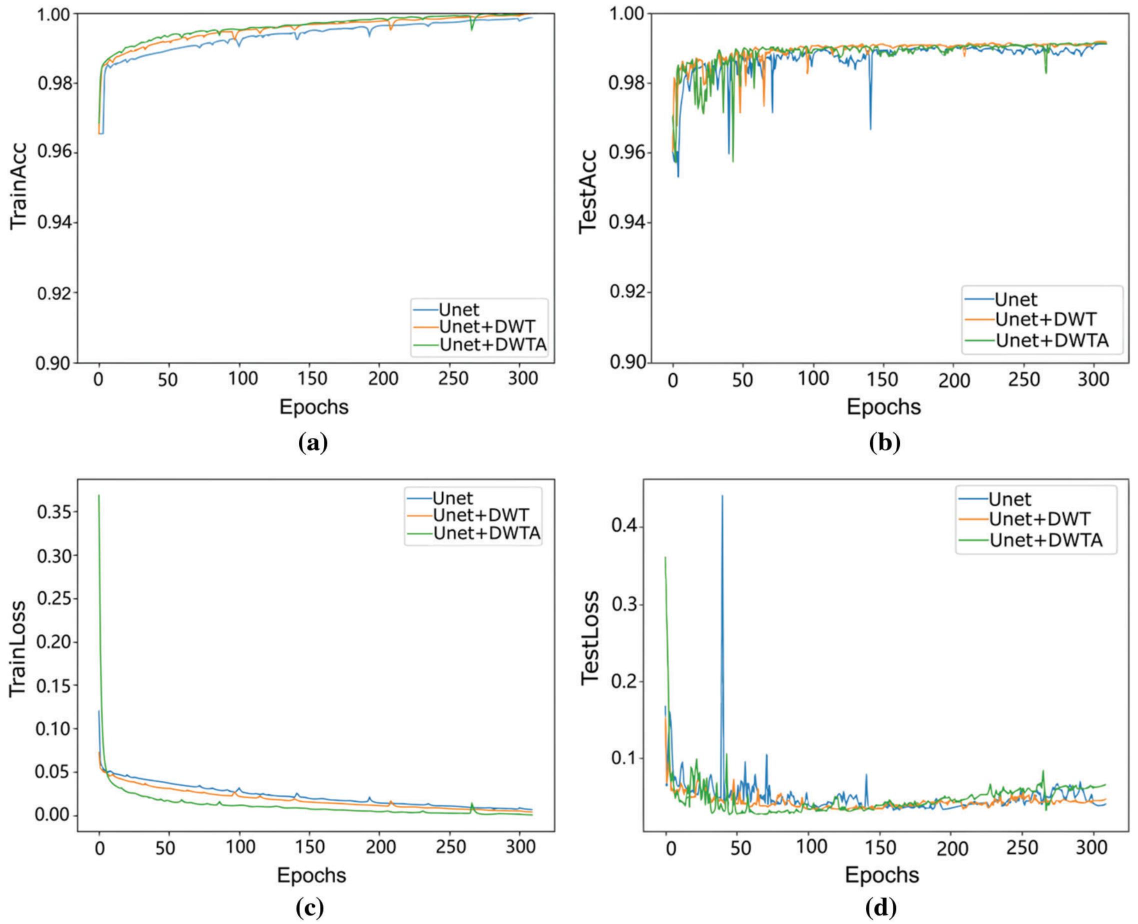


Figure 7: The accuracy curves and loss curves on CrackForest datasets. (a) and (b) represent training and testing accuracy curve of CrackForest dataset, respectively. (c) and (d) represent training and testing loss curve of CrackForest dataset, respectively

The performance of fundamental Unet can reach 83.31% and 70.00% in Mean IOU on DeepCrack and CrackForest, respectively, and F1-score can reach 82.83% and 59.06%. Because it is difficult for encoder of Unet to extract the features information of thin crack, it can be seen from the third column Fig. 10 that Unet cannot predict continuous and complete cracks. This problem can be solved by introducing frequency feature maps which are extracted by DWT. The crack boundary information is obtained by DWT. For crack segmentation, the texture and edge information are necessary. Therefore, the wavelet feature maps by four level DWT are fed into Unet to supplement the crack frequency information into network and improve the crack segmentation ability of Unet. As it can be seen from Tables 1 and 2, although the indicator of Unet enhanced by DWT slightly declines on DeepCrack dataset, the performance is significantly improved by DWT on CrackForest dataset. As shown in the third column of Fig. 11, some background pixels are also predicted as crack pixel by DWT-Unet. To better integrate wavelet information into Unet, the DWTA module is applied. As it can be seen from Tables 1 and 2, Unet including DWTA module reaches the higher value on Recall, Mean IOU, F1-score, Accuracy and Mean accuracy than DWT-Unet on two datasets. At the same time, the indicators of DWTA-Unet are almost the same as the indicators of Unet on DeepCrack dataset. Due to introducing additional module, the computational cost will be increase. The Precision-Recall (PR) curve of testing datasets is shown in the first raw of Fig. 9. The Receiver Operating Characteristic (ROC) curve of testing dataset is shown in the second raw of Fig. 9. Fig. 9 show that the proposed Unet-DWTA model achieves better performances than original Unet in thin crack.



(a)

(b)

(c)

(d)

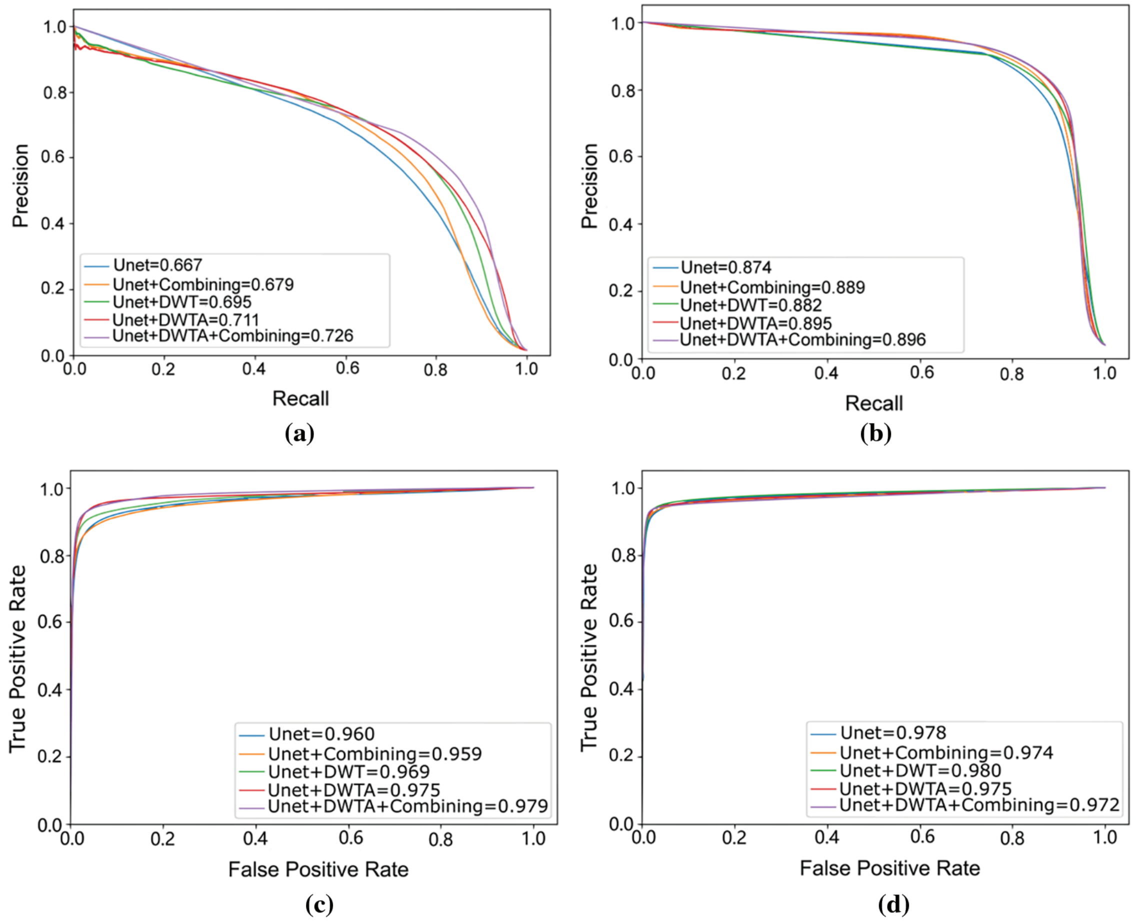
Figure 8: The accuracy and loss curves with epochs on DeepCrack datasets. (a) and (b) represent training and testing accuracy curve on DeepCrack dataset, respectively. (c) and (d) represent training and testing loss curve on DeepCrack dataset, respectively

Table 1: Comparison between DWT, DWTA and combining loss function on DeepCrack test dataset

Method	DWT	DWTA	Combining	Accuracy	Recall	Precision	Mean IOU	F1-score	Mean accuracy	Time (ms)
Unet				98.70%	78.69%	87.44%	83.31%	82.83%	88.90%	**41**
			√	98.80%	80.25%	**88.55%**	83.92%	84.19%	89.72%	41
	√			98.75%	80.87%	86.85%	83.77%	83.75%	90.04%	42
		√		98.85%	83.17%	87.51%	84.35%	85.29%	91.23%	49
		√	√	**98.86%**	**83.70%**	87.25%	**84.41%**	**85.44%**	**91.51%**	48

Table 2: Comparison between DWT, DWTA and combining loss function on CrackForest test dataset

Method	DWT	DWTA	Combining	Accuracy	Recall	Precision	Mean IOU	F1-score	Mean accuracy	Time (ms)
Unet				99.03 %	48.05 %	**76.62 %**	70.00%	59.06 %	73.64 %	**41**
			√	99.09 %	59.41 %	72.99 %	74.84 %	65.51 %	79.41 %	41
	√			99.11%	59.74%	74.22%	74.08%	66.20%	79.57%	44
		√		**99.11%**	62.55%	72.44%	74.87%	67.13%	81.00%	50
		√	√	99.08%	**72.92%**	67.03%	**76.51%**	**69.85%**	**86.26%**	51



(a)

(b)

(c)

(d)

Figure 9: PR curve and ROC curve on CrackForest and DeepCrack dataset. (a) and (b) are PR curve on CrackForest and DeepCrack dataset. (c) and (d) are ROC curve on CrackForest and DeepCrack dataset

When the combining loss function is used to address the unbalanced category problem, the DWTA-Unet reaches the highest value of Recall, Mean IOU, F1-score, Accuracy and Mean accuracy on two datasets. Compared with Unet, the value of F1-score increases 10.79% and 2.61% on CrackForest dataset and DeepCrack dataset, respectively. The seventh column of Figs. 10 and 11 shows that DWTA-Unet can obtain more precise crack segmentation. Therefore, the combining loss function is adopted in later experiments.

Image	Ground truth	original	Combining	DWT	DWTA	DWTA +Combining
		TP: 512 FP: 122 FN: 1455 TN: 63447	TP: 604 FP: 262 FN: 981 TN: 63307	TP: 732 FP: 202 FN: 1235 TN: 63269	TP: 986 FP: 262 FN: 981 TN: 63307	TP: 1153 FP: 300 FN: 814 TN: 63269

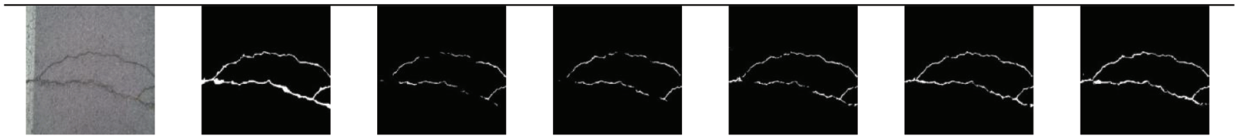


Figure 10: Crack segmentation results of network with different modules on CrackForest dataset

Image	Ground truth	original	Combining	DWT	DWTA	DWTA +Combining
		TP: 2802 FP: 146 FN: 1129 TN: 61459	TP: 2442 FP: 72 FN: 1489 TN: 61533	TP: 3379 FP: 612 FN: 552 TN: 60993	TP: 3039 FP: 191 FN: 892 TN: 61414	TP: 3507 FP: 329 FN: 424 TN: 61276

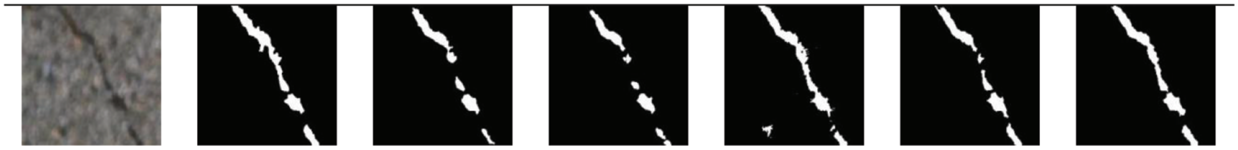


Figure 11: Crack segmentation results of network with different modules on Deepcrack dataset

4.4.2 The Effect of Multi-Level DWTA

The proposed DWTA module can be integrated with the five locations after five level of convolution layer and pooling layer of Unet encoder, respectively. Hence, there are five locations of Unet which can be embedded with DWTA. Because the DWT high-frequency coefficients by one level of DWT decomposition can be integrated with one scSE block to a DWTA module. Similarly, the high-frequency coefficients obtained by second level of wavelet decomposition can be combined with the second scSE module to form the second DWTA module. Hence, the high-frequency coefficients obtained by fifth level of wavelet decomposition can be combined with the fifth scSE module to form the fifth DWTA module. Table 3 lists the run time and performance of different number of DWTA embedded from top to bottom in the proposed Unet-DWTA model on CrackForest dataset. The run time slightly increases as the number of DWTA module increases, but the performance is improved by using multi-level DWTA. Table 3 indicates that the best performance is obtained by introducing four DWTA modules. The five DWTA modules fed into Unet lead to bad predictions. This may be which is caused by the gap between wavelet feature and convolutional feature semantic information. Therefore, the network of embedding four DWTA modules combined with four level DWT and four scSE modules into UNET is selected as the proposed network in this paper.

Table 3: Comparison between different level DWTA on CrackForest test dataset

Different level DWT	Accuracy	Recall	Precision	Mean IOU	F1-score	Mean accuracy	Time (ms)
	99.12	64.93	71.71	74.94	68.15	82.21	**45**
1,2	99.10	68.73	69.44	76.07	69.08	84.13	49
1,2,3	99.10	70.65	68.36	76.12	69.49	85.11	50
1,2,3,4	99.08	**72.92**	**67.03**	**76.51**	**69.85**	**86.26**	51
1,2,3,4,5	99.12	67.00	70.87	75.81	68.88	83.26	53

4.4.3 The Effect of Different Wavelet

We also compared the influence with different wavelet bases on the DWTA module. The experiments are performed on the CrackForest dataset. The Precision-Recall curve of different wavelet bases in the Unet-DWTA model are shown in Fig. 12. In Fig. 12, haar, bior, rbio, sym2, dmey and coif1 represent Haar, Biorthogonal, Reverse biorthogonal, Symlets, Discrete Meyer and Coiflets wavelet base, respectively. The dyadic means Dyadic wavelet transform. Compared with the Unet method without DWTA module, the methods with the different wavelets bases achieve better performance. The best performance is obtained by haar wavelet base. Therefore, the haar wavelet base is adopted in the proposed Unet-DWTA model.

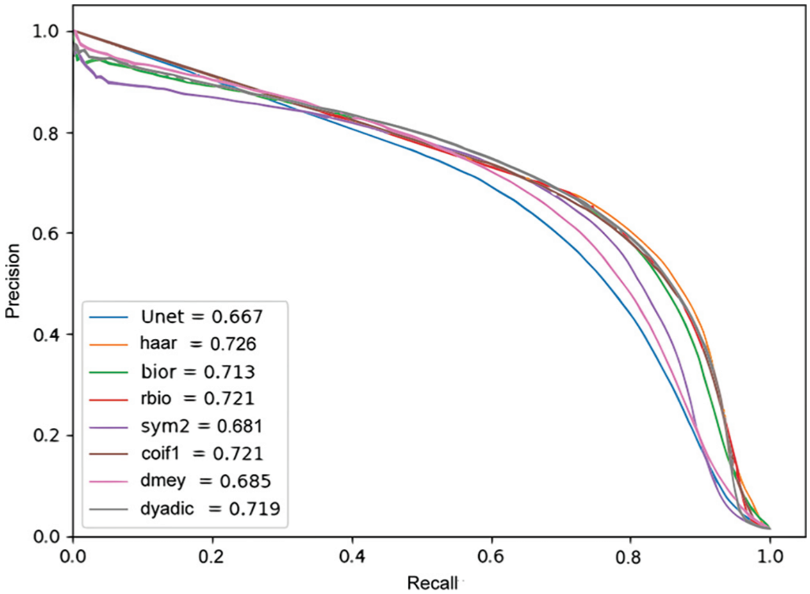


Figure 12: The results of DWTA with different wavelet on CrackForest dataset

4.4.4 Comparison with Other Attention Block

The scSE attention mechanism is applied in DWTA module. In this section, CBAM and scSE with different convolution kernel is investigated. Spatial 1 × 1 and spatial 3 × 3 represent 1 × 1 convolutional kernel adopted and 3 × 3 convolutional kernel adopted in spatial attention part in scSE module, respectively. Both CBAM module and scSE module both calibrate feature maps in spatial dimension and channel dimension. They enhance the response of crack region and suppress the response of background region in spatial dimension. In addition, they also select useful channels of feature maps. As it can be seen from Tables 4 and 5, Unet with scSE module obtains higher precision than CBAM. But when these attention modules are applied in DWTA module, CBAM module obtains higher precision than original scSE in DeepCrack dataset. When the kernel size of convolution layer is changed from 1 × 1 to 3 × 3, using scSE and CBAM obtains the equivalent performance on DeepCrack datasets. However, using scSE module shows higher values of Mean IOU, F1-score and Mean accuracy than using CBAM on CrackForest dataset. In addition, the computational cost of CBAM module is higher than scSE. Therefore, in the proposed Unet-DWTA model, scSE (spatial 3 × 3) is adopted in DWTA module.

Table 4: Comparison between CBAM and scSE module on Deepcrack test dataset

Network	Attention	Accuracy	Mean IOU	F1-score	Mean accuracy	Time (ms)
Unet	CBAM	98.77	83.98	84.60	91.87	44
	scSE (spatial 1 × 1)	98.85	84.41	85.55	**92.34**	43
	scSE (spatial 3 × 3)	**98.90**	84.34	**85.91**	91.78	**42**
Unet+ DWTA	CBAM	98.87	84.68	85.37	91.06	50
	scSE (spatial 1 × 1)	98.82	**84.75**	84.31	90.66	46
	scSE (spatial 3 × 3)	98.86	84.41	85.44	91.51	48

Table 5: Comparison between CBAM and scSE module on CrackForest test dataset

Network	Attention	Accuracy	Mean IOU	F1-score	Mean accuracy	Time (ms)
Unet	CBAM	99.00	73.14	64.45	80.76	44
	scSE (spatial 1 × 1)	99.04	74.14	66.23	81.93	43
	scSE (spatial 3 × 3)	99.06	73.91	66.27	81.41	**42**
Unet+ DWTA	CBAM	**99.11**	75.68	68.90	83.62	53
	scSE (spatial 1 × 1)	99.06	75.67	68.26	84.64	50
	scSE (spatial 3 × 3)	99.08	**76.51**	**69.85**	**86.26**	51

4.4.5 Comparison with Other Methods

The fourth experiment is designed to prove the effectiveness of the proposed Unet-DWTA. The performance of the Unet-DWTA is also compared with FCN-8 semantic segmentation network, Unet model, and the crack detection algorithms DeepCrack and CrackSegnet.

Tables 6–8 show the proposed Unet-DWTA achieves better performance than the other four models do on three datasets, especially for CrackForest dataset which contains more thin cracks. Fig. 13 shows the crack detection results by different methods on three datasets. The first and second column are CrackForest dataset, the third and fourth columns are Deepcrack dataset and the last two columns are CrackDataset dataset. As seen from the fifth and sixth row of Fig. 13, FCNN fuses less low-level feature information so that the boundary information of cracks and thin crack cannot be predicted well. Though more low-level feature maps are fused by various ways to retain more boundary and position information of crack in segmentation head of Unet, CrackSegnet and Deepcrack, they still cannot predict thin crack well even show worse anti-noise ability. The proposed model obtains better performance than other networks on CrackForest dataset and CrackDataset dataset. Compared with Unet, Unet including DWTA module increases 4.34% and 2.00% in F1-score on CrackForest dataset and CrackDataset dataset, respectively. In addition to the metric, the other three indexes of Mean IOU, F1-score, and Mean accuracy on these three datasets by the proposed model obtain the highest value compared with the other three methods. Because the frequency information of crack is added into segmentation network to enhance the crack feature extraction ability of network, the proposed Unet-DWTA can predict crack better than the other four models as seen from Fig. 13.

Table 6: Comparison between different models on DeepCrack test dataset

Network	Accuracy	Mean IOU	F1-score	Mean accuracy	Time (ms)
FCN-8	98.14%	77.42%	76.27%	86.94%	**18**
Unet	98.80%	83.92%	84.19%	89.72%	43
CrackSegNet	**98.61%**	82.50%	81.22%	87.08%	62
DeepCrack	98.77%	84.22%	84.20%	90.74%	28
Proposed	98.86%	**84.41%**	**85.44%**	**91.51%**	47

Table 7: Comparison between different models on CrackForest test dataset

Network	Accuracy	Mean IOU	F1-score	Mean accuracy	Time (ms)
FCN-8	98.46%	66.32%	50.75%	76.88%	**18**
Unet	**99.09 %**	74.84 %	65.51 %	79.41 %	43
CrackSegNet	90.08%	72.63%	61.94%	75.39%	62
DeepCrack	99.04%	69.74%	58.87%	73.15%	28
Proposed	99.08 %	**76.51%**	**69.85 %**	**86.26 %**	47

Table 8: Results of different models on CrackDataset test dataset

Network	Accuracy	Mean IOU	F1-score	Mean accuracy	Time (ms)
FCN-8	97.53%	63.19%	54.62%	78.32%	**18**
Unet	98.39%	72.11%	67.20%	81.65%	43
CrackSegNet	**98.48%**	70.92%	66.04%	78.06%	62
DeepCrack	98.13%	64.87%	57.20%	73.57%	28
Proposed	98.44%	**74.13%**	**69.20%**	**83.58%**	47

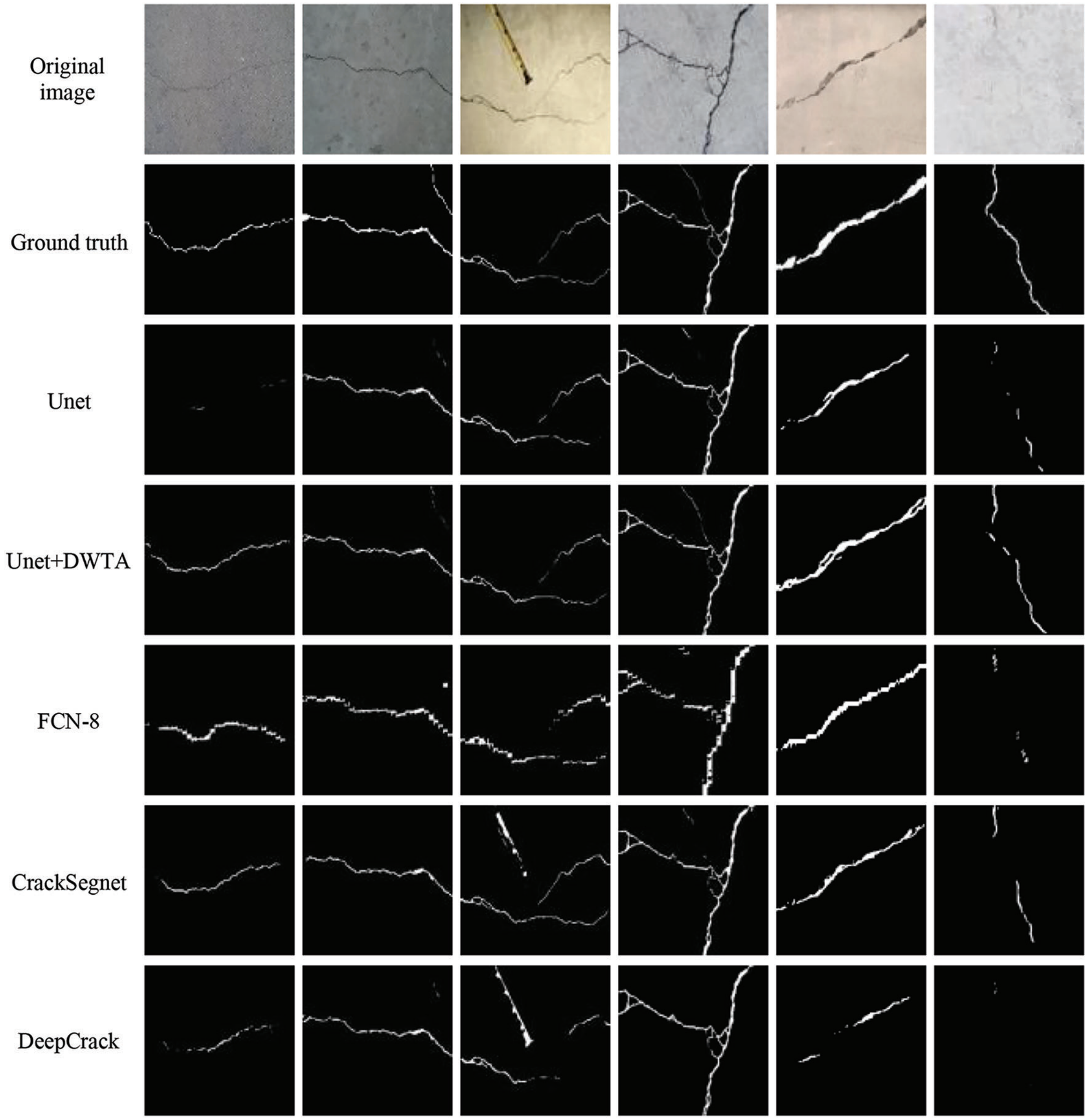


Figure 13: Crack segmentation results by different methods

5 Conclusion

It is hard to accurately predict the thin crack by FCNN-based networks, because the thin crack feature is hardly extracted in encoder and the frequency information is lost in Convolutional layers and MaxPooling layers of encoder. To tackle above problem, multi-scale wavelet feature from DWT is introduced to supplement the frequency information of crack in FCNN-based network and improve the crack feature expression ability of networks. At the same time, the scSE attention mechanism is combined with the DWT to construct the DWTA module as a new attention mechanism. The DWTA module can integrate frequency information of crack into a deep neutral network and reduce the influence of the background information. By DWTA module, the response of crack information is enhanced and unimportant information is suppressed in feature maps. In addition, the gap between frequency information and convolutional information from network is balanced to better integrate the frequency information into network. At the same time, the scSE attention mechanism is combined with the DWT to construct the DWTA module as a new attention mechanism to better integrate frequency information of crack into a deep neutral network and reduce the influence of the background information. By adding DWTA module in an encoder-decoder structures, the information of crack boundary and thin crack is well preserved in intermediate feature maps. The fused feature maps contain the information of crack boundary and the abstract semantic information which are beneficial to crack pixel classification. For crack detection task, weight cross-entropy loss function and Dice loss function are combined as a new loss function to solve unbalanced classification problem. The experiments are conducted on three datasets. One is the pixel-level thin crack dataset. Another two are common crack datasets. The results show that DWTA module improves the crack segmentation ability of different networks. The Mean IOU, F1-score and Mean accuracy is slightly improved with slight more inference time on three crack datasets by adding the proposed DTWA module in FCN, Unet, DeepCrack, and CrackSegnet.

Acknowledgement: The authors are grateful for the financial support provided by the National Natural Science Foundation of China, the National Natural Science Foundation of Hebei Province, the University Science Research Project of Hebei Province. Meanwhile, sincere thanks are given to the previous researchers who contributed to the development of deep learning technology, which lay the foundation for this research. Moreover, the authors would like to thank the anonymous reviewers for providing their invaluable comments.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grant 61972267, the National Natural Science Foundation of Hebei Province under Grant F2018210148, the University Science Research Project of Hebei Province under Grant ZD2021334.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Mohan, A., Poobal, S. (2018). Crack detection using image processing: A critical review and analysis. *Alexandria Engineering Journal, 57(2),* 787–798. https://doi.org/10.1016/j.aej.2017.01.020
2. Shan, B. H., Zheng, S. J., Ou, J. P. (2016). A stereovision-based crack width detection approach for concrete surface assessment. *KSCE Journal of Civil Engineering, 20(2),* 803–812. https://doi.org/10.1007/s12205-015-0461-6
3. Zhang, Y. Y. (2014). The design of glass crack detection system based on image preprocessing technology. *Proceedings of Information Technology and Artificial Intelligence Conference*, pp. 39–42. Chongqing, China.
4. Adhikari, R. S., Moselhi, O., Bagchi, A. (2014). Image-based retrieval of concrete crack properties for bridge inspection. *Automation in Construction, 39(1),* 180–194. https://doi.org/10.1016/j.autcon.2013.06.011
5. Krizhevsky, A., Ilya, S., Geoffrey, E. H. (2012). Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, vol. 25, no. 2, pp. 1097–1105.
6. Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*, vol. 28, pp. 91–99.
7. Jonathan, L., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440. Boston, USA.
8. Lei, Z., Fan, Y., Zhang, D., Zhu, Y. (2016). Road crack detection using deep convolutional neural network. *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3708–3712. Phoenix, USA.
9. Dung, C. V., Hidehiko, S., Suichi, H., Takayuki, O., Chitoshi, M. (2019). A vision-based method for crack detection in gusset plate welded joints of steel bridges using deep convolutional neural networks. *Automation in Construction, 102(2),* 217–229. https://doi.org/10.1016/j.autcon.2019.02.013

10. Cha, Y. J., Choi, W., Büyüköztürk, O. (2017). Deep learning-based crack damage detection using convolutional neural networks. *Computer Aided Civil & Infrastructure Engineering, 32(5),* 361–378. https://doi.org/10.1111/mice.12263
11. Dung, C. V. (2019). Autonomous concrete crack detection using deep fully convolutional neural network. *Automation in Construction, 99(4),* 52–58. https://doi.org/10.1016/j.autcon.2018.11.028
12. Wang, S., Wu, X., Zhang, Y., Chen, Q. (2018). Fully convolutional network image crack detection based on deep learning. *Journal of Computer Aided Design and Graphics, 30(5),* 115–123.
13. Ji, J., Wu, L., Chen, Z., Yu, J., Lin, P. et al. (2018). Automated pixel-level surface crack detection using U-Net. *International Conference on Multi-Disciplinary Trends in Artificial Intelligence*, pp. 69–78. Hanoi, Vietnam.
14. Wang, S., Wu, X., Zhang, Y., Chen, Q. (2018). Deep learning based full convolutional network image crack detection. *Journal of Computer-Aided Design & Computer Graphics, 30(5),* 859–867.
15. Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–341. Munich, Germany.
16. Zhang, Z., Liu, Q., Wang, Y. (2018). Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters, 15(5),* 749–753. https://doi.org/10.1109/LGRS.2018.2802944
17. Iglovikov, V., Mushinskiy, S., Osin, V. (2017). Satellite Imagery Feature Detection using deep convolutional neural network: A Kaggle competition. https://arxiv.org/abs/1706.06169.
18. Gao, X., Cai, Y., Qiu, C., Cui, Y. (2018). Retinal blood vessel segmentation based on the Gaussian matched filter and U-net. *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pp. 1–5. Shanghai, China.
19. Liu, Z., Cao, Y., Wang, Y., Wang, W. (2019). Computer vision-based concrete crack detection using U-net fully convolutional networks. *Automation in Construction, 104,* 129–139. https://doi.org/10.1016/j.autcon.2019.04.005
20. Zou, Q., Zhang, Z., Li, Q., Qi, X., Wang, Q. et al. (2019). DeepCrack: Learning hierarchical convolutional features for crack detection. *IEEE Transactions on Image Processing, 28(3),* 1498–1512. https://doi.org/10.1109/TIP.2018.2878966
21. Ren, Y., Huang, J., Hong, Z., Lu, W., Yin, J. et al. (2020). Image-based concrete crack detection in tunnels using deep fully convolutional networks. *Construction and Building Materials, 234,* 117367. https://doi.org/10.1016/j.conbuildmat.2019.117367
22. Chen, L. C., Papandreou, G., Schroff, F. (2017). Rethinking atrous convolution for semantic image segmentation. https://arxiv.org/abs/1706.05587.
23. He, K., Zhang, X., Ren, S. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9),* 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824
24. Wang, S., Liu, X., Yang, T. (2018). Panoramic crack detection for steel beam based on structured random forests. *IEEE Access, 99*. https://doi.org/10.1109/ACCESS.2018.2812141
25. Williams, T., Robert, L. (2018). Wavelet pooling for convolutional neural networks. *International Conference on Learning Representations*, Vancouver, Canada.
26. Duan, Y., Liu, F., Jiao, L., Zhao, P., Zhang, L. (2017). SAR Image segmentation based on convolutional-wavelet neural network and markov random field. *Pattern Recognition, 64(11),* 255–267. https://doi.org/10.1016/j.patcog.2016.11.015
27. Wang, S., Wu, X., Zhang, Y., Liu, X. (2020). Structural crack segmentation method based on full convolutional neural network and structured forest. *Chinese Journal of Scientific Instrument, 41(8),* 172–181.
28. Azimi, S. M., Fischer, P., Korner, M. (2018). Aerial LaneNet: Lane-marking semantic segmentation in aerial imagery using wavelet-enhanced cost-sensitive symmetric fully convolutional neural networks. *IEEE Transactions on Geoscience & Remote Sensing, 57(5),* 1–19.
29. Liu, P., Zhang, H., Zhang, K., Lin, L., Zuo, W. (2018). Multi-level wavelet-CNN for image restoration. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 773–782. Salt Lake City, USA.

30. Fujieda, S., Takayama, K., Hachisuka, T. (2018). Wavelet convolutional neural networks. https://arxiv.org/abs/1805.08620.
31. Fujieda, S., Takayama, K., Hachisuka, T. (2017). Wavelet convolutional neural networks for texture classification. https://arxiv.org/abs/1707.07394.
32. Wang, S., Wu, X., Zhang, Y. (2020). A neural network ensemble method for effective crack segmentation using fully convolutional networks and multi-scale structured forests. *Machine Vision and Applications, 31(7–8),* 60. https://doi.org/10.1007/s00138-020-01114-0
33. Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141. Salt Lake City, USA.
34. Roy, A. G., Nassir, N., Christian, W. (2018). Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 421–429. Granada, Spain.
35. Woo, S., Park, J., Lee, J. Y., Kweon, I. S. (2018). Cbam: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision*, pp. 3–19. Munich, Germany.
36. Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(7),* 674–693. https://doi.org/10.1109/34.192463
37. Zeiler, M. D., Fergus, R. (2014). Visualizing and understanding convolutional networks. *Proceedings of the European Conference on Computer Vision*, pp. 818–833. Zurich, Switzerland.
38. Yong, S., Cui, L., Qi, Z., Fan, M., Chen, Z. (2016). Automatic road crack detection using random structured forests. *IEEE Transactions on Intelligent Transportation Systems, 17(12),* 3434–3445. https://doi.org/10.1109/TITS.2016.2552248
39. Yang, L., Li, B., Li, W., Jiang, B., Xiao, J. (2018). Semantic metric 3D reconstruction for concrete inspection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1543–1551. Salt Lake City, USA.
40. Kingma, D., Ba, J. (2014). Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980.
41. Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556.