IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 4, 2020, accepted November 9, 2020, date of publication November 16, 2020,
date of current version November 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3037667

Convolutional Neural Network-Based Pavement Crack Segmentation Using Pyramid Attention Network

WENJUN WANG AND CHAO SU
College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China
Corresponding author: Wenjun Wang (wenjunwang@hhu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 51579089.

ABSTRACT Cracks are the most common road pavement damage. Due to the propagation of cracks, the detection of early cracks has great practical significance. Traditional manual crack detection is extremely time-consuming and labor-intensive. Researchers have turned their attention to automated crack detection. Although automated crack detection has been extensively researched over the past decades, it is still a challenging task due to the intensity inhomogeneity of cracks and complexity of the pavement environment, e.g. To solve these problems, we propose an efficient pavement crack segmentation model based on deep learning. The model uses pre-trained DenseNet121 as an encoder to extract pavement features. Feature Pyramid Attention module fuses features under different pyramid scales and provides precise pixel-attention. The Global Attention Upsample module which is a combination of convolutional neural network and pyramid module acts as a decoder. The sum of Cross-entropy loss and Dice loss is selected as loss function. We use poly policy to tune learning rate. In order to verify the effectiveness of the proposed method, we conduct training and testing on the Crack500 dataset and MCD dataset. Our method achieves a Dice coefficient of 0.7681, an IoU of 0.6235 on the Crack500 dataset and 0.6909, 0.5278 on the MCD dataset. We perform ablation study to verify the effectiveness of the loss function on improving the performance of our model.

INDEX TERMS Convolutional neural network, deep learning, DenseNet121 network, pyramid attention network, pavement crack segmentation.

I. INTRODUCTION

Cracks are an important indicator reflecting the safety of pavement. The formation of cracks will accelerate the aging of road and affect the strength and stability of roadbeds. If the crack can't be maintained in time, it will give rise to more serious defects, even affects traffic safety or road life, causing casualties or waste of materials. According to the 2017 American Society of Civil Engineers (ASCE) Infrastructure Report Card, the road infrastructure in the United States received a "D" grade [1], which is largely due to maintenance delay. Cracks have great influence and diffusion. Traditional manual detection methods have problems such as low detection efficiency, affecting normal traffic, time-consuming and unsafe [2]. Intelligent crack detection method has become the focus of research.

The associate editor coordinating the review of this manuscript and approving it for publication was jiju Poovvancheri.

As image processing technologies (IPTs) develop, researchers have proposed a series of automated detection methods for pavement cracks. Akagic *et al.* [3] proposed a crack image segmentation method based on histogram and Ostu's thresholding. The method divides the input image into four sub-images of the same size, then performs a crack search on each sub-image, and finally recombines the sub-images into a predicted image. This unsupervised learning method is suitable for rough estimation of asphalt pavement cracks under low signal-to-noise ratio images. Ayenu-Prah *et al.* [4] combined Bidimensional Empirical Mode Decomposition (BEMD) with Sobel edge detector for pavement crack detection. Firstly, they use BEDM to filter images for the purpose of removing noise, and then use Sobel edge detector to analyze the remaining images. But edge detection is susceptible to noise. Subirats *et al.* [5] used continuous wavelet transform for automated crack detection and tested on pavement images. The wavelet-based crack

image processing method does not work well for images with wide number of textures. Hu *et al.* [6] proposed a crack detection method based on texture analysis and shape descriptors. This method uses shape descriptors to distinguish between irregular texture and uneven brightness features, and finally uses SVM classifier to output prediction results. Compared with edge detection, this method has improved the accuracy of crack detection in complex background. Hizukuri *et al.* [7] used machine learning methods to classify pavement cracks. The classification accuracy of pavement cracks is only about 0.85, and there is still a gap for practical engineering applications. Hoang *et al.* [8] built a model suitable for asphalt pavement crack detection and classification tasks based on machine learning algorithms. The highest classification accuracy of this model achieves 87.50%, which can be used to assist professionals in evaluating road conditions. Crack detection methods based on image processing technology are sensitive to noise in crack images. When applied to practical tasks, the performance of these methods is not very satisfactory due to variations in image source, the complexity of the crack background, the diversity of textures, and the unevenness of illumination. The automated detection of pavement cracks is still a challengeable task for researchers.

In recent years, deep learning becomes popular. The outstanding performance of convolutional neural networks in the field of computer vision has aroused the interest of researchers in automated pavement crack detection. The essence of deep learning [9] is feature learning, by means of building a deep-level machine learning architecture, and then extracting features from a large amount of input data, layer by layer abstraction. As the depth of the network increases, deep learning models can learn more abstract features, and finally make classifications and predictions. Compared with traditional machine learning, the features of each layer in deep learning do not need to be manually designed, but are obtained from data using a general learning method.

At present, researchers have proposed a series of segmentation algorithms based on deep learning. Cheng *et al.* [10] used U-Net encoder-decoder structure to realize the pixel-level detection of pavement cracks. The crack segmentation method achieves an accuracy over 92%, which has obvious advantage in comparison to other methods. This is the first study that a deep learning-based method is used to process the crack images as a whole and directly generate crack segmentation. Jenkins *et al.* [11] used an encoder to compress the input image into a low-level feature map, and then uses the up-sampling layer in the decoder to restore the feature map to its original resolution. However, the robustness of the model is not very good. Kim *et al.* [12] proposed a model suitable for surface crack segmentation of concrete infrastructures based on deep learning. The model uses a two-stage image processing pipeline to obtain crack texture features, and the accuracy of crack classification and segmentation is over 90%. Mazzini *et al.* [13] proposed a data augmentation method based on Generative Adversarial Network (GAN) in order to solve the problem of dataset expansion of pavement crack segmentation. This method is very helpful in improving model performance. Wang *et al.* [14] proposed a novel large-scale irregular masks image inpainting method. Compared with state-of-the-art models, this method has better performance. It is a good choice for image denoising to embed multistage attention module into a neural network. Lee *et al.* [15] proposed a semantic segmentation model to detect infrastructural cracks and measure the maximum crack width. A shape-sensitive kernel and a modified deep module are the core part of the model. Cai *et al.* [16] proposed a novel cross-attention mechanism and graph convolution integration algorithm. This method overcomes the shortcomings of traditional attention mechanism that may lose feature information. Wang *et al.* [17] proposed a deep learning algorithm suitable for the classification of small samples of hyperspectral remote sensing images. Experiments were performed on three public hyperspectral datasets, and the results proved the effectiveness of the algorithm. Benefited from the development of computer vision, Rubio *et al.* [18] used a fully convolutional neural network to realize the multi-class damage segmentation of bridge. This model can be used as an automated damage judgment system for bridge deck inspection. Bang *et al.* [19] used the residual network as encoder for feature extraction. The decoder is composed of a deconvolution module. This method obtains an IoU value of 59.65% when testing images in black-box. This method is not ideal for the detection of very fine cracks, nor can it quantify the number and types of road cracks. Huang *et al.* [20] used compressed sensing with a generative model to decompress images for crack segmentation tasks. Compared with the traditional compressed sensing method, the calculation cost of this method is greatly reduced. Li *et al.* [21] proposed a semi-supervised pavement crack semantic segmentation model. This method can generate supervision signals for unlabeled road images to make up for the shortcomings of manual labeling, and use a full convolution discriminator to distinguish ground truth and predicted output images. You *et al.* [22] proposed an improved SRNN and attention-treated GCN-based parallel (SAGP) model to improve the accuracy of image recognition. This method makes full use of the contextual semantic relationship between features in pixels, and makes features with high probability weight related to each other. Choi *et al.* [23] proposed a real-time concrete cracks segmentation model. The encoder consists of a standard convolution module, a depth separable convolution module, and a modified atrous spatial pyramid pooling module. This model structure enables the model to improve the performance of the model while reducing the amount of parameters. Hoskere *et al.* [24] proposed a multi-task semantic segmentation model to detect multiple damages in infrastructures with different material. Dong *et al.* [25] proposed a semantic segmentation network to obtain the spatial and topological information of cracks in building materials. This method greatly reduces the amount of manual labeling and effectively avoids the

subjectivity of labeling, but the ability to detect small cracks needs to be further improved. The proposed model can make structural inspection more autonomous and flexible. Wang *et al.* [26] combined fully convolutional neural networks and multi-scale structured forests, and proposed a crack segmentation model. This network solves the problem of poor utilization of local information in complex backgrounds and overcomes the limitations of edge detection, but the robustness of the classification method needs to be further improved. Kim *et al.* [27] proposed an automated detection method for concrete surface cracks based on AlexNet network. They increased the robustness of sliding window detection through the proposed probability map of the softmax layer and proved that the method has good applicability in field crack detection tasks. However, this method unable to classify the cracks at the pixel level, so the texture characteristics of the cracks cannot be well described. Dung *et al.* [28] used VGG16 which is superior to InceptionV3 and ResNet in crack classification task as the feature extraction backbone of the full convolutional neural network (FCN), and then conducted end-to-end training of the entire network and verified the performance of the method. Their method can well detect cracks and accurately assess crack density. Santos *et al.* [29] realized the detection of concrete cracks with biological stains through processing and analyzing hyper-spectral images. By combining the number of clusters with the original hyper-spectral images, the crack detection effect is improved. This method takes the surface cracks with biological stains into consideration, which is a great improvement for the field application of automatic detection. The above research clearly shows that deep learning, more specifically, deep convolutional neural networks, is becoming the main choice for automated crack detection. There are few learnable features of crack images, and most of the existing crack segmentation methods attempt to combine the features of adjacent stages to enhance the low-level features, but ignore their different representations and global context information. The results of crack image segmentation are not very satisfactory. To solve this problem, we propose a crack image segmentation model using Pyramid Attention Network.

In this paper, our contributions are as follows:

We propose a crack segmentation model, which combines attention mechanism and spatial pyramid to extract accurate dense crack features for pixel labeling, and achieves good performance on the benchmark dataset.

We introduce a loss function, which consists of two parts: Dice loss and Cross-entropy loss, and proves the effect of this loss function on improving model performance through ablation study.

The remaining of this paper is structured as follows. Section II focuses on the architecture of pavement crack segmentation model, the loss function, and various steps during training procedure. Section III describes the Crack500, DeepCrack, GAPS384, MCD datasets. Section IV presents experimental results and discussion. Section V shows ablation study on loss function and encoder. Section VI delivers the conclusion of this paper.

II. NETWORK STRUCTURE

In this paper, we propose a new model suitable for pixel-level detection of cracks. The model is based on Pyramid Attention Network (PAN) [30] and uses DenseNet121 [31] (pre-trained on ImageNet, its last two layers removed) as the encoder. The model structure is shown in Fig. 1. Firstly, the encoder DenseNet121 is used to extract features of the input image, a Feature Pyramid Attention (FPA) module is inserted between the encoder and the decoder to collect the dense pixel-level attention information extracted by the encoder to guide the classification and positioning of pixels. The decoder module is Global Attention Upsample (GAU) module which combines low-level and high-level feature information accurately. Finally, the model restores the image resolution through an up-sampling operation.

A. DENSE BLOCKS

DenseNet121 achieves better results and fewer parameters through the ultimate use of features. It is used in our model as an encoder to extract crack feature, which plays a significant role in mitigating the disappearance of gradients and enhancing feature transfer. The DenseNet121 encoder starts with a convolution operation with a kernel size of 7×7 and stride of 2. Next comes the Dense Block, the core module of DenseNet121. Fig. 2 shows a 5-layer dense block. The dense block uses dense connection, and each layer uses all preceding feature maps as its own input:

$$x_l = H_l\left([x_0, x_1 \ldots., x_{l-1}]\right) \tag{1}$$

where $[x_0, x_1 \ldots., x_{l-1}]$ is the concatenation of feature maps produced by layer $0, 1, \ldots l-1$. H is a compound function, with three consecutive operations, batch normalization (BN), rectified linear unit (ReLU), and 3×3 convolution.

So that, each layer and input are directly connected to loss, which can account for the fact that it can alleviate the problem of gradient disappearance.

The structure of the layer in the Dense Block is illustrated in Fig. 3. Firstly, batch normalization and ReLU operations are performed on the input, and then the number of channels is reduced through a 1×1 convolution operation. Therefore, the number of feature maps output by each convolutional layer in the dense block is small, the model width is limited, making the model easier to train. Then perform batch normalization, ReLU, and 3×3 convolution operations on the model in turn to complete all operations in a layer.

B. TRANSITION LAYERS

The transition layer is located between the two dense blocks to perform dimensionality reduction operations. As shown in Fig. 4, this module includes a 1×1 convolution layer and a 2×2 average pooling layer with a stride of 2.

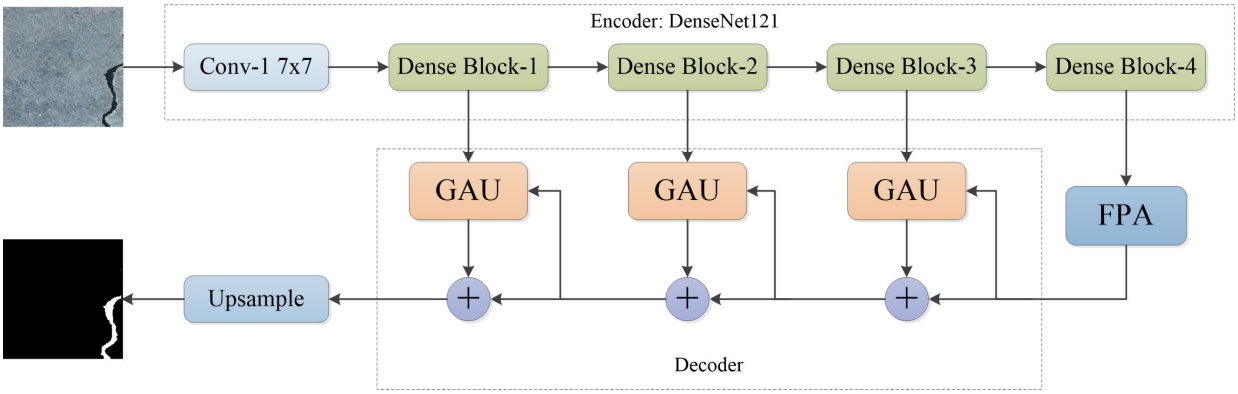


FIGURE 1. The overall structure of proposed model.

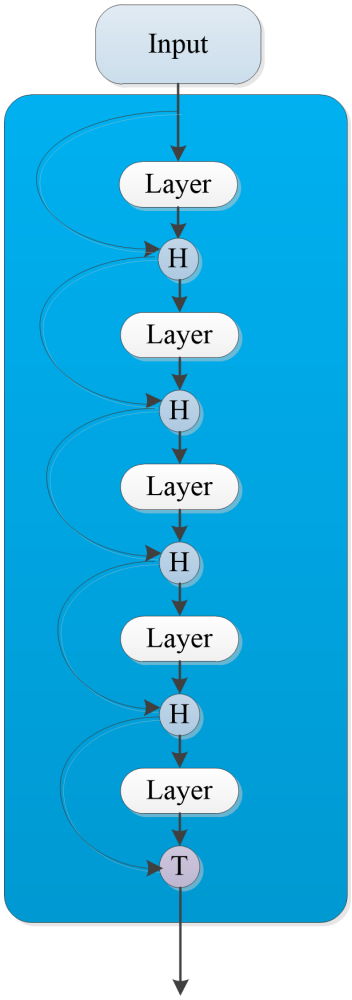


FIGURE 2. A 5-layer dense block.

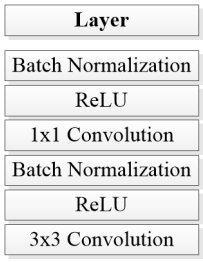


FIGURE 3. The structure of the layer in the dense block.

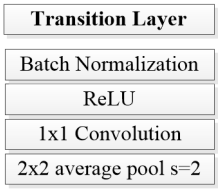


FIGURE 4. The architecture of transition layer.

C. FEATURE PYRAMID ATTENTION (FPA)

Recent models mainly use SPP or ASPP modules, but when performing image feature extraction, it may cause the lack of spatial local information and loss of pixel positioning, which will do harm to the consistency of the feature map and the precision of image segmentation. We use FPA module to solve this problem. The design of FPA module is inspired by Attention Mechanism, which can realize pixel-level attention of high-level feature maps. Fig. 5 shows the implementation of FPA. The FPA module has a U-shaped network structure like Feature Pyramid Network, which fuses the features of the input feature map under three different scales. The convolution kernel in pyramid structure has three sizes, 3×3, 5×5, and 7×7, which can extract effective semantic information in a larger range. The feature information obtained after the input feature map passes through the pyramid structure is multiplied by the feature information after the 1×1 convolution operation, and finally concatenate with the feature information after the global pooling and 1×1 convolution in order to achieve multi-scale pixel-level information feature extraction, improve model performance.

D. GLOBAL ATTENTION UPSAMPLE(GAU)

The function of the decoder module is to map the features of the low-resolution encoder to the full input resolution to

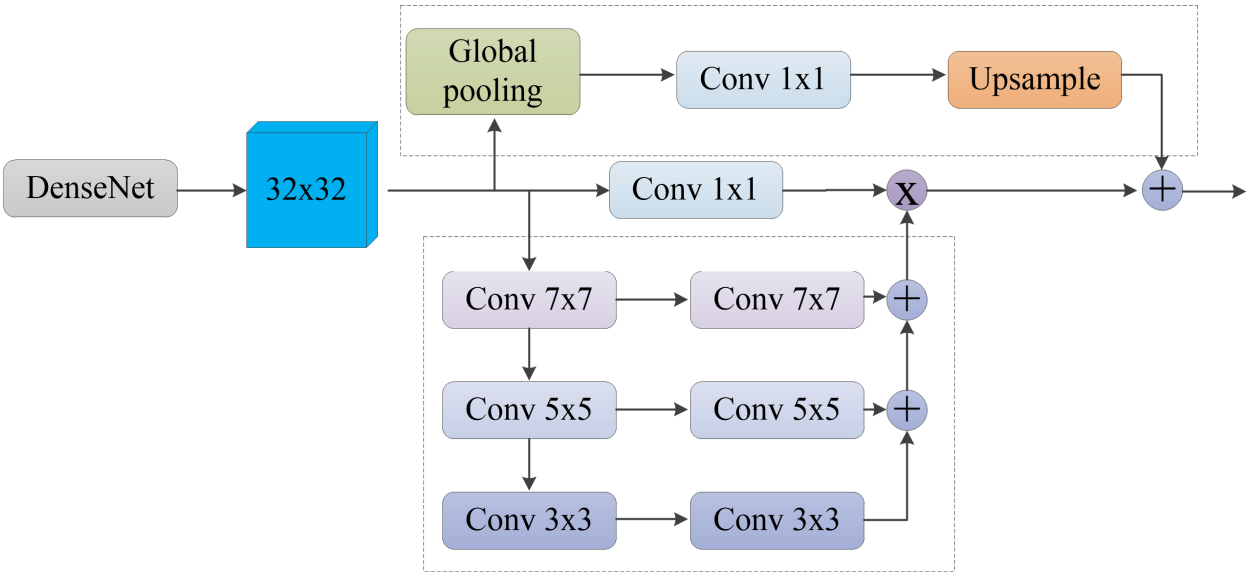


FIGURE 5. Feature pyramid attention representation.

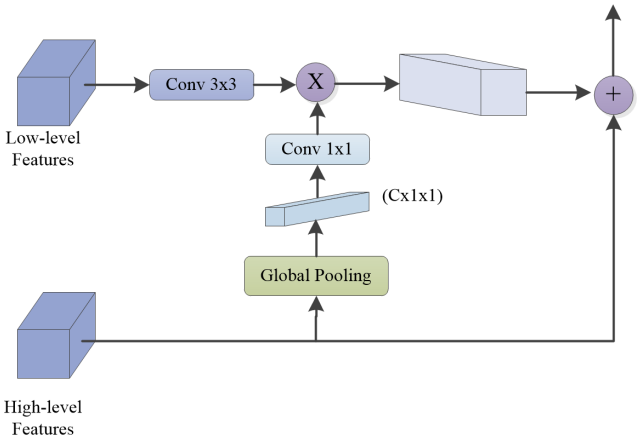


FIGURE 6. The component of global attention Upsample.

achieve the classification of pixel. The GAU module provides global information for the category positioning of the low-level features through the global pooling operation. The component of the module is illustrated in Fig. 6. Firstly, we perform a 3 × 3 convolution operation on the low-level features, perform global pooling and 1 × 1 convolution operations on the high-level features, then multiply the two features, and finally concatenated with the high-level features to guide the classification of low-level features.

E. LOSS FUNCTION

The loss function is a representative of the optimization objective, which can be described as the "baton" of the entire network model. It is mainly used to measure the performance of the model's prediction. The learning of network parameter is guided through the back propagation of the error between predicted value and true value. Therefore, how to choose a loss function to make it closer to the optimization goal is extremely important. For semantic segmentation tasks, the commonly used loss functions mainly include cross-entropy loss function, Dice loss function, IOU loss function, Tversky loss function and others.

In our research, loss function consists of two parts: cross-entropy loss and Dice loss [32].

The cross-entropy loss function is distribution-based loss. When the semantic segmentation platform uses Softmax to classify pixels, it is used and can be calculated by the following formula:

$$L_{cross-entropyloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{i=c}^{C} g_i^c log p_i^c \tag{2}$$

Dice loss is Region-based loss. Dice coefficient is an ensemble similarity measurement function. It is usually used to calculate the similarity of two samples. The value range is [0,1]. The Dice loss function can be expressed as:

$$L_{Diceloss} = 1 - \frac{2\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^c p_i^c}{\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^c + \sum_{i=1}^{N}\sum_{c=1}^{C} p_i^c} \tag{3}$$

where i is a single pixel, N is the total number of pixels, c is the classification, C is the total number of categories, g_i^c indicates whether the classification is correct; p_i^c is the probability of belonging to a certain category.

The following equation determines the loss function we use:

$$Loss = 1 - \frac{2\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^c p_i^c}{\sum_{i=1}^{N}\sum_{c=1}^{C} g_i^c + \sum_{i=1}^{N}\sum_{c=1}^{C} p_i^c} - \frac{1}{N}\sum_{i=1}^{N}\sum_{i=c}^{C} g_i^c log p_i^c \tag{4}$$

F. OPTIMIZATION

Most deep learning algorithms involve optimization. Researchers have proposed a series of optimization algorithms to update and calculate network parameters that affect model training and output. The optimizers are helpful to approximate or reach the optimal value, thereby minimizing the loss function. We choose an adaptive learning rate Adam optimizer [33] to minimize the loss function, which utilizes the first moment estimate and second raw moment estimate of the gradient to adjust the learning rate of parameters dynamically. The implementation of the Adam optimizer is as follows:

For the initial vector θ, we initialize the first moment vector s, second moment vector r, and time step t to 0, and then take out m samples $\left\{x^{(1)}, \ldots, x^{(m)}\right\}$ with the target of $y^{(i)}$ from the training dataset. Gradient g is defined as:

$$g_t = \frac{1}{m}\nabla_\theta \sum_i L\left(f\left(x^{(i)}; \theta\right), y^{(i)}\right) \tag{5}$$

For t+1, update biased first moment estimate and biased second raw moment estimate:

$$s_{t+1} = \rho_1 s_t + (1-\rho_1)\, g_t \tag{6}$$

$$r_{t+1} = \rho_2 r_t + (1-\rho_2)\, g_t \odot g_t \tag{7}$$

Then bias-corrected first moment estimate and second raw moment estimate are computed:

$$\widehat{s_{t+1}} = \frac{s_{t+1}}{1-\rho_1^{t+1}} \tag{8}$$

$$\widehat{r_{t+1}} = \frac{r_{t+1}}{1-\rho_2^{t+1}} \tag{9}$$

Finally, we update parameters with the following equation:

$$\theta_{t+1} = \theta_t - \epsilon \frac{\widehat{s_{t+1}}}{\sqrt{\widehat{r_{t+1}}} + \delta} \tag{10}$$

where ρ_1, ρ_2 is the decay rate of moment estimate (setting to 0.9, 0.999 respectively), ϵ is step size (setting to 1e-3), δ is a constant to ensure that we don't divide by zero (setting to 10e-8).

G. TUNING THE LEARNING RATES

When training the neural network, the learning rate is the most significant hyperparameter that we need to set. The learning rate controls the speed of updating model weights and has a great influence on the effective capacity of model. If the learning rate is too small, it will cause the neural network to converge slowly or get stuck in the local minimum; if the learning rate is too large, it will cause the neural network to fail to reach the global minimum and the model cannot converge.

At the beginning of training, a larger learning rate can speed up the convergence of the model. As the training epoch increases, the learning rate should gradually decay to avoid skipping the optimal value and to improve training stability. In our research, we set the initial learning rate to 3e-5, and use poly learning rate strategy to adjust the learning rate. The learning rate of each step can be calculated by the following formula:

$$LearningRate = Initial_{lr} \times \left(1 - \frac{epoch}{\max\_epoch}\right)^{0.9} \tag{11}$$

The learning rate of each epoch during training can be seen from Fig. 7.

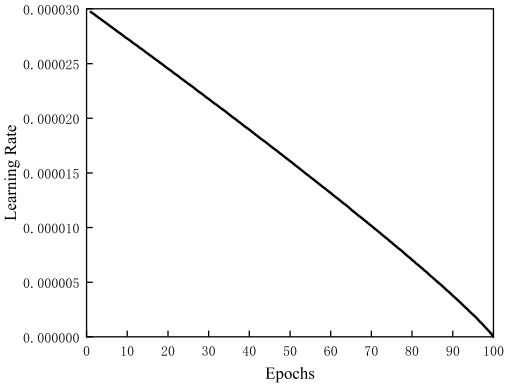


FIGURE 7. Learning rate of each epoch.

H. TRANSFER LEARNING

Transfer learning refers to the transfer of a pre-trained model to other tasks, so that the training of model has a higher starting point, without needing to train from scratch. The pre-trained model is more effective for feature extraction, which can greatly shorten model training time and accelerate model convergence [34]. In the crack segmentation model training stage, we use the transfer learning method to transfer the pretrained DenseNet121 (pre-trained on ImageNet, its last two layers removed) weights to the crack segmentation task. Using the pre-trained weights to initialize the decoder part of our model, and initialize the rest of the parameters in the neural network using he normal initialization. Finally, we train the entire neural network.

III. DATASETS

A. CRACK500

The Crack 500 [35] dataset contains 500 images with a resolution around 2000 × 1500. These images are obtained on main campus of Temple University via a phone. Due to the limitation of computing resources, each image is cropped to 16 non-overlapped image regions, and only the regions containing more than 1000 crack pixels are saved. Each crack image is annotated at the pixel level. Therefore, Crack 500 contains 3368 crack images, of which the training data contains 1896 images, the test data contains 1124 images, and the validation data contains 348 images. Some data samples can be seen in Fig. 8.

B. DEEPCRACK

The DeepCrack [36] dataset contains 537 crack images, with complex background and various crack scales, which can

FIGURE 8. Crack500 data sample. (a) image; (b) ground truth.

better reflect crack characteristics. The DeepCrack dataset contains three textures (bare, dirty, rough) and two scenes (concrete, asphalt) and the crack width ranges from 1 pixel to 180 pixels. The crack area in each image only accounts for a small percentage, which is similar to the actual situation. The dataset is divided into a training data (300 images) and a test data (237 images). All crack images have been manually annotated and presented as a binary image. Some representative samples in DeepCrack are shown in Fig. 9.

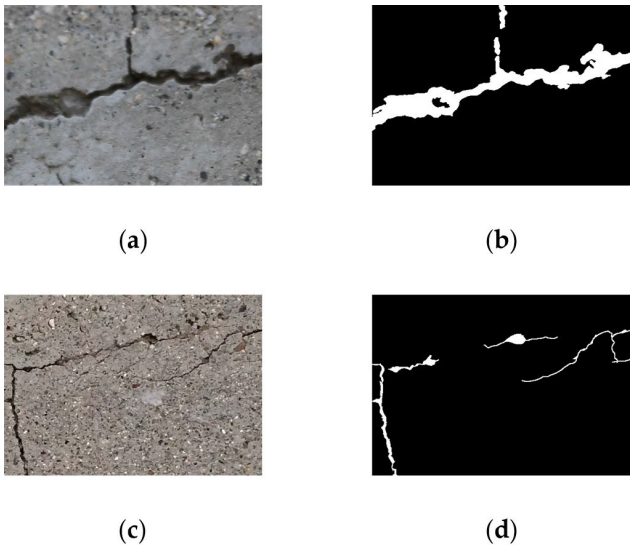


FIGURE 9. DeepCrack data sample. (a)(c) image; (b)(d) ground truth.

C. GAPS384

GAPS384 [35] is obtained via manually annotating 384 crack images selected from the GAPS dataset [37] at pixel level by Yang *et al.* Due to the limitation of GPU memory, each image is then cropped to 6 non-overlapped images. Therefore, GAPS384 includes 509 raw images and 509 annotated images. The images in the GAPS dataset were captured in summer 2015 by a measuring vehicle with mobile mapping system S.T.I.E.R, which is manufactured by LEHMANN + PARTNER GmbH. Specifically, the measuring vehicle using the surface camera system which is composed of two JAI Pulnix TM2030 cameras equipped with Kodak KAI-2093 1'' progressive scan CCD imager. The GAPS dataset provides high-quality crack images with a resolution of 1920 × 1080 as a neural network training dataset. Some samples are shown in Fig. 10.

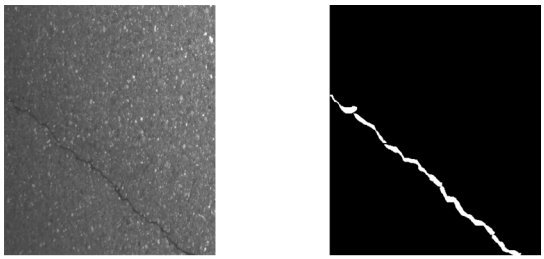


FIGURE 10. GAPS384 data sample. (a)image; (b)ground truth.

D. MIXED CRACK DATASET (MCD)

Crack segmentation model needs to be sufficiently robust to accommodate surface crack in different environment and materials. To improve the robustness of our model, we merge the images in Crack500, DeepCrack, and GAPS384 datasets to form a mixed crack dataset (MCD), which contains concrete cracks and asphalt cracks. All the images are resized to the size of 512 × 512. The MCD contains a total of 4414 raw images and 4414 annotated images. The detailed information is illustrated in Table1.

TABLE 1. The detail of mixed crack dataset.

Dataset	train	valid	Test
Crack500	1896	348	1124
DeepCrack	290	10	237
GAPS384	465	5	39
Total number of images	2651	363	1400

IV. CRACK SEGMENTATION EXPERIMENTAL RESULTS AND DISCUSSIONS

A. TRAINING SETTINGS

All the experiments in this paper are carried out on Tensorflow in Windows system.

Hardware settings of the computer are as follows:
CPU: Intel(R)Core (TM)i7CPU@3.20GHz
RAM: two 8GB DDR4 memories
GPU: NVIDIA GTX1080Ti

The input image size of our model is 512 × 512. When the input image does not meet the requirement, the image is then resized. During model training process, we use data augmentation to create fake data and add them to the training data to increase the number of training data, so that the crack segmentation model can obtain better generalization capabilities. The data augmentation methods we use include: flipping, rotation, and brightness change.

B. EVALUATION CRITERIA

For the crack segmentation task in this research, we introduce four essential metrics, Precision (Pr), Recall (Re), Dice Coefficient (Dice), and Intersection over Union (IoU) to evaluate the performance of segmentation model. Precision

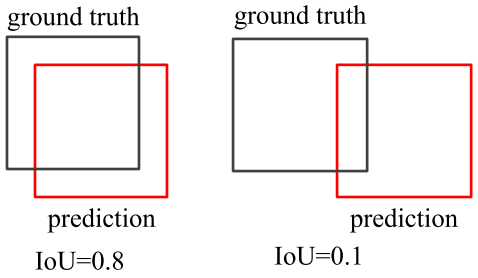


FIGURE 11. Examples of positive instance(left) and negative instance(right).

describes the purity of crack pixel detection, which means the ratio of all pixels predicted to be cracks that are actually positive. Recall is a measure of predictive completeness, that is, the proportion of identified crack pixels. Dice Coefficient considers the precision and recall comprehensively, which is the harmonic mean of them. IoU is also known as Jaccard Index, which is the ratio of the intersection and union between prediction and ground truth. MIoU can be calculated via taking the IoU of each class and averaging them. We make the following definitions: the crack pixels are seen as positive instances. As shown in Fig. 11, if the value of p_i^c (refer to (2)) is over 0.5, the pixel is a positive instance, otherwise it is a negative instance. The precise definition of Pr, Re, Dice, IoU are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$Dice = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{14}$$

$$IoU = \frac{GroundTruth \cap Prediction}{GroundTruth \cup Prediction} \tag{15}$$

where TP is true positive (the number of crack pixels that are correctly detected), FP is false positive (the number of pixels that are wrongly detected as cracks), FN is false negative (the number of crack pixels that are wrongly detected).

Considering the subjectivity of manual labeling and the existence of crack transition region, we consider all predictions to be true positive pixels, if crack pixels are two pixels near the manually labeled crack pixel. The method has been used in [38].

C. RESULTS

In order to evaluate the performance of our model on unseen images, we evaluate our model on Crack500 dataset firstly. Fig. 12 shows the output of crack segmentation at different epochs. We can see that as the training epoch increases, the crack segmentation results become more detailed and closer to the ground truth. The comparison of test results among our proposed crack segmentation model and other four recently proposed crack segmentation models on Crack500 dataset (the first three models are reimplemented by Lau *et al.*) can be seen from Table 2. The results are

TABLE 2. Comparison of test results using different methods on Crack500 dataset.

Method	Pr	Re	Dice	IoU
U-Net by Nguyen et al. [38]	0.6954	0.6744	0.6895	0.5261
CNN by Fan et al. [35]	0.7123	0.6955	0.7056	0.5451
Split-Attention Network [39]	0.7368	0.7165	0.7295	0.5742
U-Net by Lau et al. [40]	0.7426	0.7285	0.7327	0.5782
Our method	**0.8163**	**0.7654**	**0.7681**	**0.6235**

TABLE 3. Performance comparison when tested on MCD and Crack500 datasets.

Dataset	Pr	Re	Dice	IoU
MCD	0.8131	0.6633	0.6909	0.5278
Crack500	0.8163	0.7654	0.7681	0.6235

quantitatively represented using four metrics: Pr, Re, Dice, and IOU (the optimal results have been highlighted in bold). From Table 2, we can see that compared with the other four methods, our method has a great improvement in terms of Pr, Re, Dice, IoU evaluation metrics, which are 0.8163, 0.7654, 0.7681, 0.6235 respectively. From the results, we can conclude that our model has a good generalization capability. It is worth noting that our model sacrifices some recall in exchange for higher precision. Fig. 13 shows some segmentation results of our best model.

To further verify the performance of our model, we re-train and test it on the MCD dataset. The results are shown in Table3, the values of Pr, Re, Dice, IoU are all decreased compared with that on the Crack500 dataset. This illustrates that complex background, multi-surface materials, and multi-scale crack conditions have great influence on crack segmentation performance.

TABLE 4. Performance comparison when training with different loss functions.

Loss Function	Pr	Re	Dice	IoU
cross-entropy	0.8207	0.7459	0.7603	0.6133
ours	0.8163	0.7654	0.7681	0.6235

V. ABLATION STUDY

A. ABLATION FOR LOSS FUNCTION

In order to show the effect of loss function we proposed on improving the performance of segmentation model, we perform ablation study on Crack500 dataset. In this experiment, we train two neural networks, one with the loss function of cross-entropy (refer to (2)), and the other with the loss function of the sum of cross-entropy and Dice loss (refer to (4)), other conditions remain unchanged. Table4 shows the test results of the two methods. From Table4 we can see that, when the sum of the cross-entropy and Dice loss is used

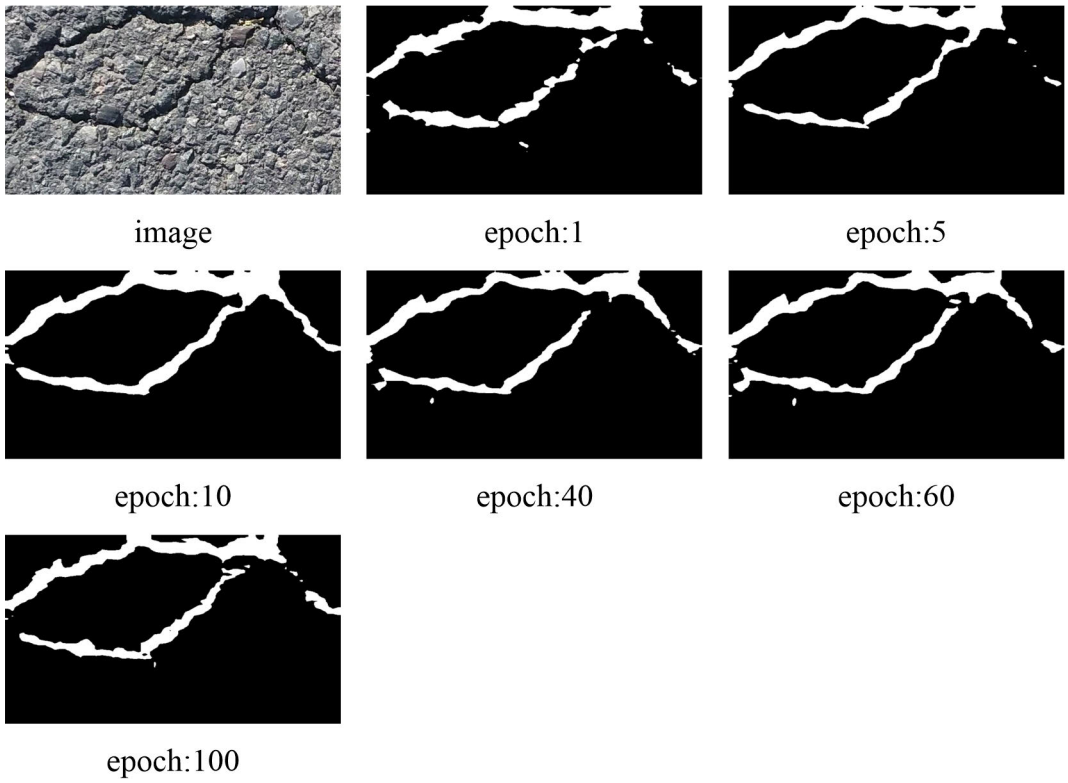


FIGURE 12. Output at different epochs.

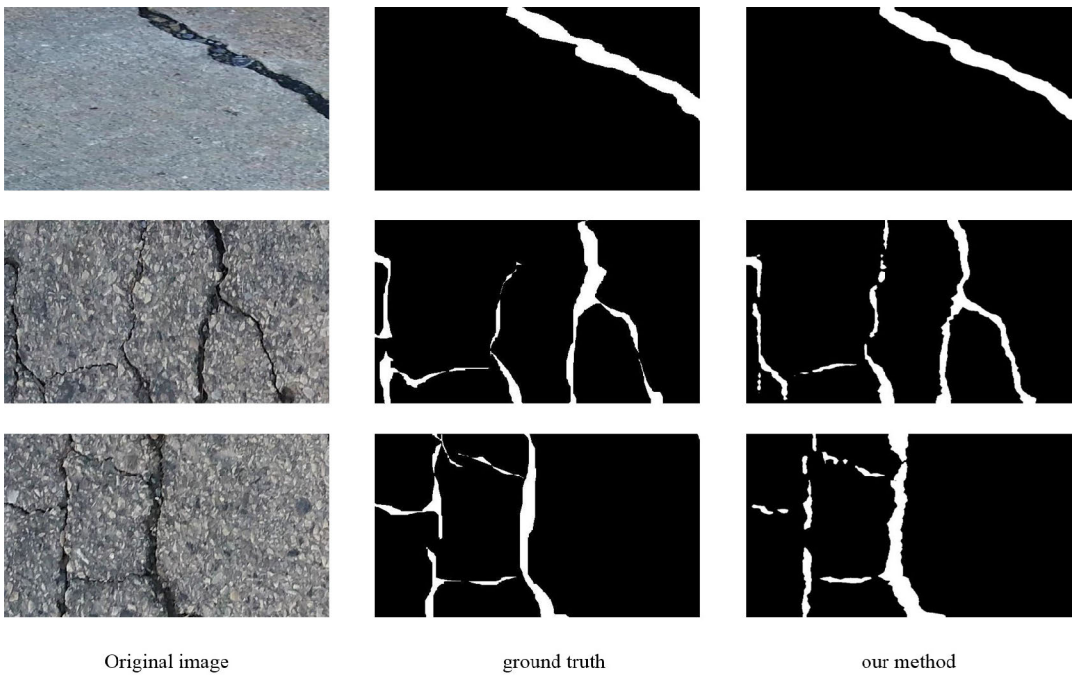


FIGURE 13. Segmentation results.

as loss function, the values of Re, Dice and IoU have been improved by a small margin at the cost of precision. The loss function we designed can produce better model parameters and better model performance during neural network optimization.

B. ABLATION FOR ENCODER

The semantic segmentation architecture can be seen as an encoder-decoder network. After giving input images, the encoder generates feature images with semantic information through neural network learning. The decoder gradually

TABLE 5. Performance comparison of different encoders.

Encoder	Pr	Re	Dice	IoU
MobileNetV2[41]	0.7841	0.7248	0.7286	0.5732
DenseNet121	0.8163	0.7654	0.7681	0.6235
ResNet50[42]	0.7942	0.7253	0.7294	0.5741
Xception[43]	0.7929	0.7474	0.7552	0.6067

implements the category labeling of each pixel after the encoder provides feature maps. The choice of encoder is very important to the performance of segmentation models. We choose four classic convolutional neural networks as the encoder, keeping other conditions unchanged, and conducte training and testing on the Crack500 dataset. The results are shown in Table 5. We can see that, for pavement crack segmentation tasks, Pyramid Attention Network using DenseNet121 as the encoder has the best performance, the Dice coefficient and IoU are 0.7681 and 0.6235 respectively.

VI. CONCLUSION

This study concentrates on the method of applying convolutional neural network to detect pavement cracks. Our network structure is a Pyramid Attention Network with an encoder of pretrained DenseNet121. The crack features in the image are well extracted in the encoder part. Compared with the commonly used methods such as dilated convolution which may cause local information missing, our model combines the attention mechanism with the spatial pyramid network to extract precise crack features information. In addition, through fusion of contextual information at different scales, better pixel-level attention is obtained. In the decoding network part, our model makes full use of feature information at different scales and combines CNN and pyramid modules to reduce computational cost and realize the classification guidance of high-level semantic information to low-level feature maps. We use the Adam optimizer which is an adaptive learning rate algorithm to guide the training of the neural network. Cross-entropy focuses on a fitting situation of the overall pixel and describes the difference between two probability distributions. When the semantic segmentation model uses softmax to classify the pixels, it is used. The Dice loss focuses on the overlap between the ground truth and the prediction. In order to have a comprehensive measure of the error between the predicted value and the true value from a global and local perspective, we use the sum of the Dice loss and the Cross-entropy loss as the final loss function to help the model converge faster and perform better. The performance of our method is tested and compared with the other four models on the Crack500 dataset. The Dice and IoU values achieved by our approach are 0.7681 and 0.6235 on the Crack500 dataset which outperforms other four recent models. In order to test the performance of our model under different environmental and material conditions, we retrained and tested the model on the MCD dataset. The Dice and IoU values achieved by our approach are 0.6909 and 0.5278 on the MCD dataset. Compared with other machine learning methods, our model requires less feature engineering and has better performance.

The proposed pavement crack segmentation model achieves good performance on the Crack500 dataset and the MCD dataset. For areas lacking professionals in pavement cracks analyzing, our model is a good choice. Although the method in this paper shows good performance, it still has a long way to go for automated detection of pavement cracks. One limitation of our proposed method is that we need to input lots of manual annotated pixel-level crack images to train efficient and accurate models. The manual annotation method is time-consuming and subjective, in addition, the performance of the model is closely related to the dataset. With the development of unsupervised learning, this problem may be solved in the near future. Another limitation is that our model only realizes the texture representation of pavement cracks, but cannot characterize the extent and severity of distresses. Combining the power of computer vision with natural language processing provides a good solution to this problem. The actual road conditions are complex and changeable. How to achieve high-precision and real-time crack detection is a difficult problem. For multi-task detection of different materials and different damages, how to build a robust model is still an important research direction in the future.

REFERENCES

[1] *American Society of Civil Engineers (ASCE) 2017 Infrastructure Report Card: Roads; American Society of Civil Engineers (ASCE)*, ASCE, Reston, VA, USA, 2017.

[2] M. S. Kaseko and S. G. Ritchie, "A neural network-based methodology for pavement crack detection and classification," *Transp. Res. C, Emerg. Technol.*, vol. 1, no. 4, pp. 275–291, Dec. 1993.

[3] A. Akagic, E. Buza, S. Omanovic, and A. Karabegovic, "Pavement crack detection using otsu thresholding for image segmentation," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2018, pp. 1092–1097.

[4] A. Ayenu-Prah and N. Attoh-Okine, "Evaluating pavement cracks with bidimensional empirical mode decomposition," *EURASIP J. Adv. Signal Process.*, vol. 2008, no. 1, pp. 1–7, Dec. 2008.

[5] P. Subirats, J. Dumoulin, V. Legeay, and D. Barba, "Automation of pavement surface crack detection using the continuous wavelet transform," in *Proc. Int. Conf. Image Process.*, Oct. 2006, pp. 3037–3040.

[6] Y. Hu, C.-X. Zhao, and H.-N. Wang, "Automatic pavement crack detection using texture and shape descriptors," *IETE Tech. Rev.*, vol. 27, no. 5, pp. 398–405, 2010.

[7] A. Hizukuri and T. Nagata, "Development of a classification method for a crack on a pavement surface images using machine learning," in *Proc. 13th Int. Conf. Qual. Control Artif. Vis.*, vol. 10338, Mar. 2017, Art. no. 103380.

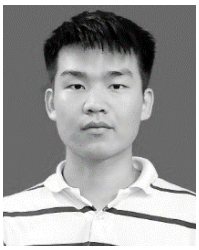
[8] N.-D. Hoang and Q.-L. Nguyen, "A novel method for asphalt pavement crack classification based on image processing and machine learning," *Eng. with Comput.*, vol. 35, no. 2, pp. 487–498, Apr. 2019.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[10] J. Cheng, W. Xiong, W. Chen, Y. Gu, and Y. Li, "Pixel-level crack detection using U-Net," in *Proc. IEEE Region 10 Conf. (TENCON)*, Oct. 2018, pp. 0462–0466.

[11] M. David Jenkins, T. A. Carr, M. I. Iglesias, T. Buggy, and G. Morison, "A deep convolutional neural network for semantic pixel-wise segmentation of road and pavement surface cracks," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2120–2124.

[12] A.-R. Kim, D. Kim, Y.-S. Byun, and S.-W. Lee, "Crack detection of concrete structure using deep learning and image processing method in geotechnical engineering," *J. Korean Geotechnical Soc.*, vol. 34, no. 12, pp. 145–154, 2018.
[13] D. Mazzini, P. Napoletano, F. Piccoli, and R. Schettini, "A novel approach to data augmentation for pavement distress segmentation," *Comput. Ind.*, vol. 121, Oct. 2020, Art. no. 103225.
[14] N. Wang, S. Ma, J. Li, Y. Zhang, and L. Zhang, "Multistage attention network for image inpainting," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107448.
[15] J. S. Lee, S. H. Hwang, I. Y. Choi, and Y. Choi, "Estimation of crack width based on shape-sensitive kernels and semantic segmentation," *Struct. Control Health Monitor.*, vol. 27, no. 4, p. e2504, Apr. 2020.
[16] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geosci. Remote Sens. Lett.*, early access, Oct. 1, 2020, doi: 10.1109/LGRS.2020.3026587.
[17] Z. Wang, C. Zou, and W. Cai, "Small sample classification of hyperspectral remote sensing images based on sequential joint deeping learning model," *IEEE Access*, vol. 8, pp. 71353–71363, 2020.
[18] J. J. Rubio, T. Kashiwa, T. Laiteerapong, W. Deng, K. Nagai, S. Escalera, K. Nakayama, Y. Matsuo, and H. Prendinger, "Multi-class structural damage segmentation using fully convolutional networks," *Comput. Ind.*, vol. 112, Nov. 2019, Art. no. 103121.
[19] S. Bang, S. Park, H. Kim, and H. Kim, "Encoder–decoder network for pixel-level road crack detection in black-box images," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 34, no. 8, pp. 713–727, Aug. 2019.
[20] Y. Huang, H. Zhang, H. Li, and S. Wu, "Recovering compressed images for automatic crack segmentation using generative models," *Mech. Syst. Signal Process.*, vol. 146, Jan. 2021, Art. no. 107061.
[21] G. Li, J. Wan, S. He, Q. Liu, and B. Ma, "Semi-supervised semantic segmentation using adversarial learning for pavement crack detection," *IEEE Access*, vol. 8, pp. 51446–51459, 2020.
[22] H. You, S. Tian, L. Yu, and Y. Lv, "Pixel-level remote sensing image recognition based on bidirectional word vectors," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1281–1293, Feb. 2020.
[23] W. Choi and Y.-J. Cha, "SDDNet: Real-time crack segmentation," *IEEE Trans. Ind. Electron.*, vol. 67, no. 9, pp. 8016–8025, Sep. 2020.
[24] V. Hoskere, Y. Narazaki, T. A. Hoang, and B. Spencer, Jr., "MaDnet: Multi-task semantic segmentation of multiple types of structural materials and damage in images of civil infrastructure," *J. Civil Struct. Health Monit.*, vol. 10, pp. 757–773, Jun. 2020.
[25] Z. Dong, J. Wang, B. Cui, D. Wang, and X. Wang, "Patch-based weakly supervised semantic segmentation network for crack detection," *Construct. Building Mater.*, vol. 258, Oct. 2020, Art. no. 120291.
[26] S. Wang, X. Wu, Y. Zhang, X. Liu, and L. Zhao, "A neural network ensemble method for effective crack segmentation using fully convolutional networks and multi-scale structured forests," *Mach. Vis. Appl.*, vol. 31, nos. 7–8, pp. 1–18, Nov. 2020.
[27] B. Kim and S. Cho, "Automated vision-based detection of cracks on concrete surfaces using a deep learning technique," *Sensors*, vol. 18, no. 10, p. 3452, Oct. 2018.
[28] C. V. Dung and L. D. Anh, "Autonomous concrete crack detection using deep fully convolutional neural network," *Autom. Construct.*, vol. 99, pp. 52–58, Mar. 2019.
[29] B. Oliveira Santos, J. Valença, and E. Júlio, "Automatic mapping of cracking patterns on concrete surfaces with biological stains using hyperspectral images processing," *Struct. Control Health Monitor.*, vol. 26, no. 3, p. e2320, Mar. 2019.
[30] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*. [Online]. Available: http://arxiv.org/abs/1805.10180
[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[32] R. R. Shamir, Y. Duchin, J. Kim, G. Sapiro, and N. Harel, "Continuous dice coefficient: A method for evaluating probabilistic segmentations," 2019, *arXiv:1906.11031*. [Online]. Available: http://arxiv.org/abs/1906.11031
[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[34] S. Kornblith, J. Shlens, and Q. V. Le, "Do better ImageNet models transfer better?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2661–2671.
[35] F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, and H. Ling, "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1525–1535, Apr. 2020.
[36] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation," *Neurocomputing*, vol. 338, pp. 139–153, Apr. 2019.
[37] M. Eisenbach, R. Stricker, D. Seichter, K. Amende, K. Debes, M. Sesselmann, D. Ebersbach, U. Stoeckert, and H.-M. Gross, "How to get pavement distress detection ready for deep learning? A systematic approach," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2039–2047.
[38] A. Sheta, H. Turabieh, S. Aljahdali, and A. Alangari, "Pavement crack detection using convolutional neural network," in *Proc. 9th Int. Symp. Inf. Commun. Technol.*, 2018, pp. 251–256.
[39] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, Z. Zhang, H. Lin, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*. [Online]. Available: http://arxiv.org/abs/2004.08955
[40] S. L. H. Lau, E. K. P. Chong, X. Yang, and X. Wang, "Automated pavement crack segmentation using U-Net-Based convolutional neural network," *IEEE Access*, vol. 8, pp. 114892–114899, 2020.
[41] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[43] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.



WENJUN WANG received the B.S. degree in agricultural water conservancy engineering from Hohai University, Nanjing, China, in 2019, where he is currently pursuing the Ph.D. degree in hydraulic structure engineering. His research interests include computer vision and deep learning.

CHAO SU received the Ph.D. degree in hydraulic structure engineering from Hohai University, Nanjing, China, in 2005. He is currently a Professor with the College of Water Conservancy and Hydropower Engineering, Hohai University. His research interests include numerical calculation method, experimental study of hydraulic structure, stability analysis of large underground cavern group, and deep learning.