

Chapman University

Chapman University Digital Commons

Engineering Faculty Articles and Research

Fowler School of Engineering

7-24-2023

Multi-Scale Attention Networks for Pavement Defect Detection

Junde Chen

Yuxin Wen

Yaser Ahangari Nanekaran

Defu Zhang

Adan Zeb

Follow this and additional works at: https://digitalcommons.chapman.edu/engineering_articles



Part of the [Computer and Systems Architecture Commons](#), [Other Computer Sciences Commons](#), [Software Engineering Commons](#), and the [Systems Architecture Commons](#)

Multi-Scale Attention Networks for Pavement Defect Detection

Comments

This is a pre-copy-editing, author-produced PDF of an article accepted for publication in *IEEE Transactions on Instrumentation and Measurement* in 2023 following peer review. This article may not exactly replicate the final published version. The definitive publisher-authenticated version is available online at <https://doi.org/10.1109/TIM.2023.3298391>.

Copyright

© 2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Multi-scale attention networks for pavement defect detection

Junde Chen, Yuxin Wen, Yaser Ahangari Nanekaran, Defu Zhang, *Member, IEEE*, and Adan Zeb

Abstract—Pavement defects such as cracks, net cracks, and pit slots can cause potential traffic safety problems. The timely detection and identification play a key role in reducing the harm of various pavement defects. Particularly, the recent development in deep learning-based CNNs has shown competitive performance in image detection and classification. To detect pavement defects automatically and improve effects, a multi-scale mobile attention-based network, which we termed MANet, is proposed to perform the detection of pavement defects. The architecture of the encoder-decoder is used in MANet, where the encoder adopts the MobileNet as the backbone network to extract pavement defect features. Instead of the original 3×3 convolution, the multi-scale convolution kernels are utilized in depth-wise separable convolution layers of the network. Further, the hybrid attention mechanism is separately incorporated into the encoder and decoder modules to infer the significance of spatial points and inter-channel relationship features for the input intermediate feature maps. The proposed approach achieves state-of-the-art performance on two publicly-available benchmark datasets, i.e., the Crack500 (500 crack images with 2,000×1,500 pixels) and CFD (118 crack images with 480×320 pixels) datasets. The mean intersection over union (*MIoU*) of the proposed approach on these two datasets reaches 0.7219 and 0.7788, respectively. Ablation experiments show that the multi-scale convolution and hybrid attention modules can effectively help the model extract high-level feature representations and generate more accurate pavement crack segmentation results. We further test the model on locally collected pavement crack images (131 images with 1024×768 pixels) and it achieves a satisfactory result. The proposed approach realizes the *MIoU* of 0.6514 on the local dataset and outperforms other compared baseline methods. Experimental findings demonstrate the validity and feasibility of the proposed approach and it provides a viable solution for pavement crack detection in practical application scenarios. Our code is available at <https://github.com/xtu502/pavement-defects>.

Index Terms—Pavement defect detection, deep neural network, multi-scale convolution, attention module, image identification.

I. INTRODUCTION

PAVEMENT defect detection is a challenging task in traffic transportation maintenance. Defects on the surface of the

Corresponding author: Yuxin Wen

J. Chen and Y. Wen are with the Dale E. and Sarah Ann Fowler School of Engineering, Chapman University, CA 92866, USA. (e-mail: jundchen@chapman.edu; yuwen@chapman.edu).

J. Chen is with the School of Informatics, Xiamen University, Xiamen 361005, China, and also with the Department of Electronic Commerce, Xiangtan University, Xiangtan 411100, China.

Y.A. Nanekaran is with the School of Information Engineering, Yancheng Teachers University, Yancheng 224000, China (e-mail: yaser@yctu.edu.cn).

D. Zhang is with the School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: dfzhang@xmu.edu.cn).

A. Zeb is with the southern university of science and Technology, Shenzhen 518000, China. (e-mail: adnanzeb@sustech.edu.cn).

Manuscript received Nov 2, 2022.

road may distress people's traveling and cause economic loss if these defects are not properly checked and maintained in time [1]. Therefore, the detection of pavement defects is one of the vital tasks in road maintenance operations and has attained increasing attention in recent years. Nevertheless, the conventional manual approaches relying on visual observations of experienced specialists or inspectors are dangerous, inefficient, error-prone, labor-intensive, costly, and cannot be extended in large areas [2]-[5]. To promote the advancement of pavement checks and alleviate the workload of experts, it is necessary to realize the automation of defect detection. Hence, there are a great demand and important realistic significance to develop an efficient, fast, and accurate tool to automatically detect various pavement defects.

The new era of pavement defect detection is being presented with the rapid advancement of digital cameras and image processing techniques, which delivers a scientific basis for the automatic detection of pavement defects using road defect images. Automatic crack detection based on image processing techniques can be generally divided into three types: classification [1], [4], object detection [6], [23], and pixel-level detection [7], [20]. Pixel-level crack detection is the most critical task as it can provide more information about cracks, including the area, shape, and orientation, which is useful for identifying the severity level of infrastructure defects [20]. In the early days, threshold-based methods were usually employed to detect defect regions depending on the assumption that the real defect pixel is consistently darker than its surroundings. Kaseko et al. [8] used an auto-thresholding technique to perform pavement crack detection and their research results demonstrated the potential and feasibility of the proposed approach. Using the threshold-based method, Liu et al. [9] developed an automated pavement distress inspection system to recognize pavement cracks, and their findings showed that the pavement crack was identified accurately. Nonetheless, the thresholding methods can only generate discontinuous defect fragments because the intensity along the defect may not always be lower than that of the backgrounds. Additionally, road shadows usually produce non-uniform illuminance in pavement images, which can further weaken the effects of the thresholding methods. Edge detection-based methods have also been proposed for pavement defect detection [10], [11], but they are susceptible to low contrast between the defect regions and the backgrounds, thereby misclassifying some speckle noises as defect fragments. Apart from that, some wavelet-transform based approaches have been applied to detect pavement defects as well [12], [13]. These methods detect defects efficiently, however, they cannot handle cracks

with low continuity or high curvature well owing to the anisotropic characteristic of the wavelet. In the recent decade, major machine learning (ML) algorithms, such as support vector machine (SVM) [14], artificial neural networks (ANN) [8], [16], random forests (RF) [15], and Adaboost [17], have attained popularity in the field of image recognition and are commonly employed in pavement defect detect applications. In [14], Quintana et al. utilized the SVM model to identify pavement defects by splitting up the image into small patches as the input. Chen et al. [16] adopted a block-based ANN method for the detection of road defects and achieved promising performance. In [17], the AdaBoost method was used to identify pavement defects from road surface images based on textural information, etc. Although impressive results have been reported in the literature, the ML-based methods are greatly dependent upon hand-crafted features. Due to the complex pavement conditions, it is hard to exploit effective characteristics for all pavements.

Most recently, deep learning (DL), particularly convolution neural networks (CNN), has been developed to address most computer vision tasks due to its outstanding representation capability. Some works have also been dedicated to leveraging the property of DL for the detection of pavement defects [18]-[27]. For example, Choi and Young-Jin [18] proposed an original CNN architecture named SDDNet, which consists of multiple modules, such as standard convolutions and densely connected separable convolution modules. Although promising results are obtained, this model has high computational complexity. The dense connections increase the amount of calculation and consume more memories. Using a deep hierarchical CNN, Liu et al. [19] built an end-to-end network called DeepCrack for pixel-wise crack segmentation. Many downsampling layers are used in their network, resulting in lost spatial resolution on the feature maps, which is difficult to recover in upsampling layers and sacrifices much performance in thin crack detection. Chen and Huiping [20] recommended a hybrid atrous CNN named HACNet for crack detection. Despite reasonable good results reported, it is hard to determine various dilation rate settings and the gridding issue is also a challenge for atrous convolutions. Based on a Holistically-nested edge detector, Yang et al. [21] proposed a Feature Pyramid and Hierarchical Boosting Network (FPHBN) to detect pavement cracks. Multi-scale feature extraction is performed by their method, however, it is developed on standard convolutions with a relatively large number of parameters. Besides, the edge detection-based method is easily affected by the possible low contrast between the defect regions and the backgrounds, as mentioned previously. In another work, reference [22] introduced a method of splitting the image into different blocks and CNN to detect whether the block had defects or not. However, this approach is inconvenient due to image splitting and it is sensitivity to patch scale. In [23], Schmugge et al. applied a CNN-based object detection method to implement crack detection from multiple overlapping frames in a video. But their method neglects the spatial relations between pixels and also overvalues crack width. Similarly, using a Faster Region-based Convolutional Neural Network (Faster R-CNN) architecture, Cha et al. [24] reported a visual

inspection method for detecting multiple damage types. An impressive performance is obtained by the model, however, their method relies on more manual annotation information like the bounding box, key points, and coordinate information of target objects. In practice, it is time-consuming and labor-intensive to obtain a large amount of annotation information for model training. To address these critical challenges, Ali and Young-Jin [25] developed an attention-based generative adversarial network (GAN) to generate new synthetic images for training the damage segmentation model. Nevertheless, the attention-based GAN is just used for data augmentation rather than damage segmentation tasks. Moreover, the attention mechanism presented in their study only considers pixel relationships regardless of their spatial features. In another research, Zhang et al. [26] trained a CNN model to identify the category for each pixel of images, while they still relied on manually designed feature extractors and just utilized CNN as a classifier. Besides, the fixed feature extractor also prevented the popularization of their method.

Despite the limitation, the latest studies have demonstrated the effectiveness of CNN-based methods. In this study, we propose an end-to-end network architecture, namely MANet, to perform the detection of pavement defects. Precisely, the MobileNet is chosen as the backbone extractor, and to enlarge the convolutional receptive field, we modify the architecture of classical MobileNet. The multi-scale convolution kernels are utilized instead of the existing 3×3 convolution kernels in depth-wise separable convolution layers. Then, to infer the significance of spatial points and channel interdependency features, a hybrid attention mechanism is incorporated into our network, where the attention modules are embedded into both the encoder and the decoder modules. Experimental findings indicate the effectiveness and feasibility of the proposed approach. To summarize, the major contributions of this paper can be recapitulated as follows.

- We have collected a pavement defect image dataset from real-life pavement scenarios. 4 types of pavement defects including crack, net crack, map crack, and pit slot along with one normal category were collected in this dataset. This dataset is expected to facilitate further research on pavement defect detection.
- The MobileNet is chosen as the backbone extractor, and the multi-scale depthwise separable convolutions are substituted for the original ones to enlarge the convolution receptive fields and improve the richness of modeling feature information.
- The hybrid attention mechanism which consists of spatial and channel-wise attention modules is embedded into the encoder and decoder compositions of the network to separately learn the significance of spatial points and achieve the maximum reuse of inter-channel relation features.
- In addition to detecting whether there are defects in the pavement, the method proposed in this paper also identifies the specific types of pavement defects.

The remainder of this paper is structured as follows. Section II displays the dataset of pavement defect images used and

primarily discusses the methodology. Section III dedicates to the algorithm experiments, with a comparative analysis of the experimental results. Finally, Section IV summarizes the research and points out the direction of future work.

II. MATERIALS AND METHODS

A. Image Datasets

In this work, two publicly available datasets and one locally collected dataset are used for our experiments. The first publicly available dataset is the CRACK500 [21], which primarily includes cracking defect images comprised of 500 pavement defect samples with a uniform size of 2,000×1,500 pixels for each image. These crack images have been annotated by a pixel-level binary map (ground truth), and each sample image is cropped into 16 non-overlapped patches that each patch contains more than 1,000 pixels of crack kept. As a consequence, 1,896 training images, 348 validation images, and 1,124 test images are included in this dataset, which is currently the largest open-access pavement crack dataset with pixel-wise annotation [21]. Another dataset is the CFD dataset [2], which consists of 118 images collected from urban pavement conditions in Beijing, China. The size of each image in the CFD dataset is 480×320 pixels and all the images have been manually labeled with ground truth contours. The device applied to acquire the photographs is an iPhone5 with a 4mm focus, f/2.4 aperture, and 1/135s exposure time. Also, the CFD dataset is used for evaluating the model. Fig. 1 displays the partial crack sample images on the publicly available datasets.

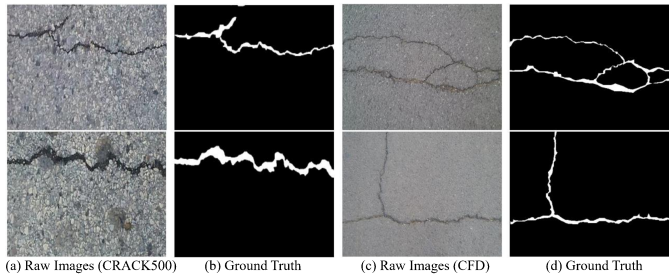


Fig. 1: Sample images of CRACK500 and CFD datasets.

To take more crack types into consideration, we have performed a widespread collection of pavement defect images photographed under practical pavement scenarios with heterogeneous background conditions and varied illumination intensities. Most pavement defect images were captured using a consumer-level color digital camera with Nikon S3100, which was used to photograph without digital or optical zoom and with flash always off. Some other images were obtained from publicly available sources through popular search engines like Google, Yahoo, Baidu, and Bing. As a consequence, a total of 131 pavement defect images with 4 types including crack, net crack, map crack, and pit slot are captured in our experiments. All the images are uniformly processed into the RGB model using Photoshop software, and then the sizes of images are adjusted to 1024 × 768 pixels. Fig. 2 presents the sample images and specification information of pavement

defects, including the crack width, average seam width, block size, and defect area.


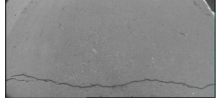


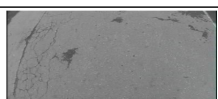



Defect types	Items	Slight defects	Severe defects
Crack	Defect images		
	Symptoms	The crack wall is not scattered (average crack width < 5mm).	Serious crack, more branch crack (average crack width > 5mm)
Net crack (crack width: 0.2~0.3m, block area 0.1~10m ²)	Defect images		
	Symptoms	Little crack, block size > 100cm	large crack, block size locates in 50cm~100cm
Map crack (crack width: < 0.3m)	Defect images		
	Symptoms	Small cracks, no scattering, block size : 20cm~50cm	Block broken, wide cracks, block size < 20cm
Pit slot	Defect images		
	Symptoms	Depth <= 25mm, area (< 1m ²)	Depth > 50mm, area (> 1m ²)

Fig. 2: The typical pavement defect types.

B. Related Work

1) *MobileNet*: MobileNet is a mobile-first convolution neural network designed to efficiently maximize accuracy while considering the restriction of computational resources for deploying deep learning applications [28]. Based on a streamlined structure, MobileNet adopts depth-wise separable convolutions (DSConv) to construct lightweight CNNs and it decomposes a regular convolution into two compositions including a depth-wise convolution (DConv) and a 1×1 convolution named point-wise convolution (PConv). For the input feature map, the DConv with one filter is first performed on each channel, and thus the results of DConv are implemented with the PConv to obtain the final output results of DSConv. The formulas of DConv and PConv are presented in Eqs. (1,2), respectively.

$$DConv(\theta_d, y)_{(i,j)} = \sum_{w=0}^W \sum_{h=0}^H \theta_{d(w,h)} \odot y_{(i+w, j+h)}, \quad (1)$$

$$PConv(\theta_p, y)_{(i,j)} = \sum_{l=0}^L \theta_l \times y_{(i,j,k)}, \quad (2)$$

where θ signifies the weights of convolutional kernels, H and W imply the height and width of the images, respectively. (i, j) index the position of images, \odot symbolizes the dot products of elements, L denotes the number of channels, and y represents the input images. Further, the calculation of DSConv can be written as

$$DSConv(\theta_p, \theta_d, y)_{(i,j)} = PConv_{(i,j)}(\theta_p, DConv_{(i,j)}(\theta_d, y)). \quad (3)$$

MobileNet explicitly incorporates the DSConv as its core component of network architecture, which consists of a 3×3 DConv layer followed by a Batch Normalization (BN) and

Rectified Linear Unit (ReLU) layer, 1×1 PConv layer followed by a BN and ReLU layer, Global Average Pooling (GAP) layer, Dropout layer, and fully-connected (FC) layer. Fig. 3 portrays the key difference between the MobileNet and other conventional CNNs.

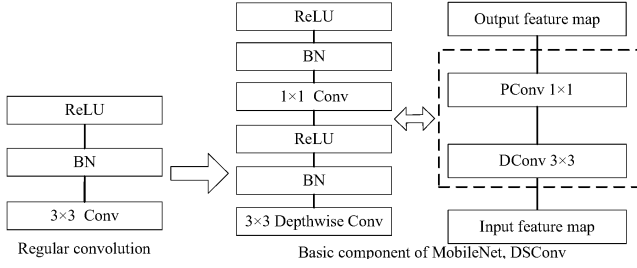


Fig. 3: Basic blocks of MobileNet.

2) *Attention Mechanism*: Similar to human visual attention, the attention mechanism in deep CNNs can make the model keep attention on useful information while ignoring unwanted noises. Attention mechanisms [29], [30] have proven helpful in numerous computer vision tasks including image segmentation [31]-[33] and classification [34]-[36]. They have also shown state-of-the-art performance in the internal damage image segmentation [37], [38]. One of the most typical representatives is SENet [39], which simply squeezes each 2D feature map to recalibrate the weights between channels. Although it can efficiently build interdependencies among channels, the SENet only involves channel attention while ignoring spatial attention. Woo et al. [36] further advanced this idea by introducing a spatial attention mechanism, called convolutional block attention module (CBAM). However, it is slightly inferior to the SE attention module in mobile neural networks compared with the SE block [29]. For these reasons, the SE block paired with the spatial attention of CBAM is introduced in our lightweight network architecture. Combining the merits of channel-wise attention (CA) and spatial attention (SA), a hybrid attention mechanism is employed in the networks, where the CA is well in finding the desired object in multiple feature maps while the SA is particularly prominent when probing the target regions in feature maps. The specific details are described as follows.

Suppose an intermediate feature map $f \in R^{W \times H \times C}$ is input into the CA (SE block) and the SA modules, the CA module will rescale the original feature f by channel-wise multiplication, written as

$$CA(f) = F_c(f) = f_c * s_c, \quad (4)$$

where $f_c \in R^{W \times H}$, c means the c -th channel, $*$ symbolizes the dot products of elements, and s_c (scalar) is the weight obtained by the Squeeze and Excitation operations of the SE block. Then, the SA module concatenates the results output from the CA and performs the convolution operation using a normal convolution layer, thereby obtaining the spatial attention map. The formula can be presented as

$$SA(f) = F_s(f) = \sigma(c^{7 \times 7}([GMP(f); GAP(f)])), \quad (5)$$

where σ symbolizes the sigmoid function, $c^{7 \times 7}$ implies the 7×7 convolution, GMP and GAP signify the global maximum pooling (GMP) and global average pooling (GAP)

operations, respectively. Consequently, the overall calculation result of the hybrid attention module is written as

$$F_{att} = CA(f) + SA(f) = f_c * s_c + F_s(f). \quad (6)$$

C. Proposed Approach

1) *MANet Model*: To achieve effective detection of pavement defects, the MANet model that adopts an encoder-decoder framework is proposed in the paper. As mentioned earlier, MobileNet is a type of lightweight CNNs depending upon DSCConv and has shown outstanding capability in dealing with both large-scale and small-scale problems of image recognition. Motivated by the competitive performance, the MobileNet is chosen as the backbone extractor of the proposed MANet to extract the features of pavement defect images. For the task of defect region segmentation, we need to preserve the spatial information of the target images. Thereupon, the original completely associated layers are removed from the classical MobileNet, and the convolutional layers coupled with the downsampling and the upsampling layers are used in the network architecture, where the former composition is named as the encoder while the latter one is as the decoder. In other words, this is an encoder-decoder network structure for the MANet, where the former layers downsampling the input are the encoder part and the latter layers which upsample the feature maps are the decoder part. Furthermore, the hybrid attention mechanism can make full use of spatial and channel-wise attention to infer the significance of spatial points and channel interdependency features for the input intermediate feature maps. Thence, the hybrid attention module is incorporated into our network and it includes two aspects of characteristics: (1) The hybrid attention module is embedded into the extractor part of the network to make it focus on more information related to cracks, and can better extract the characteristics of pavement defects. (2) The hybrid attention module is also introduced into the decoder module, which makes the decoder position more accurate and obtains richer detail features when recovering cracks or other defect types. In addition, the size of the convolution kernel is uniformly adopted as 3×3 in the depthwise separable convolution of the MobileNet, which makes the extracted information relatively unitary owing to the limited receptive field of this single-scale convolution kernel. Therefore, to enlarge the convolutional receptive field and enhance the richness of the convolutional feature channels, we adopt the multi-scale convolution by replacing the original 3×3 convolution kernel with the 1×1 , 3×3 , and 5×5 multi-scale convolution kernels, respectively. In brief, the main modules of the MANet model are described below.

(1) *Encoder module*. The encoder module of the MANet, or the feature extractor of the network, is based on the MobileNet which includes an input convolution layer and other 12 depthwise separable convolution blocks DSConv1~DSConv12. The dimensions of input images are assigned as the fixed size of $224 \times 224 \times 3$, and the convolutional kernel size of the input convolution layer is $3 \times 3 \times 32$ with a stride of 2. DSConv1~DSConv8 use a 3×3 convolution kernel, and the number of channels is 64, 128, 256, and 512, respectively.

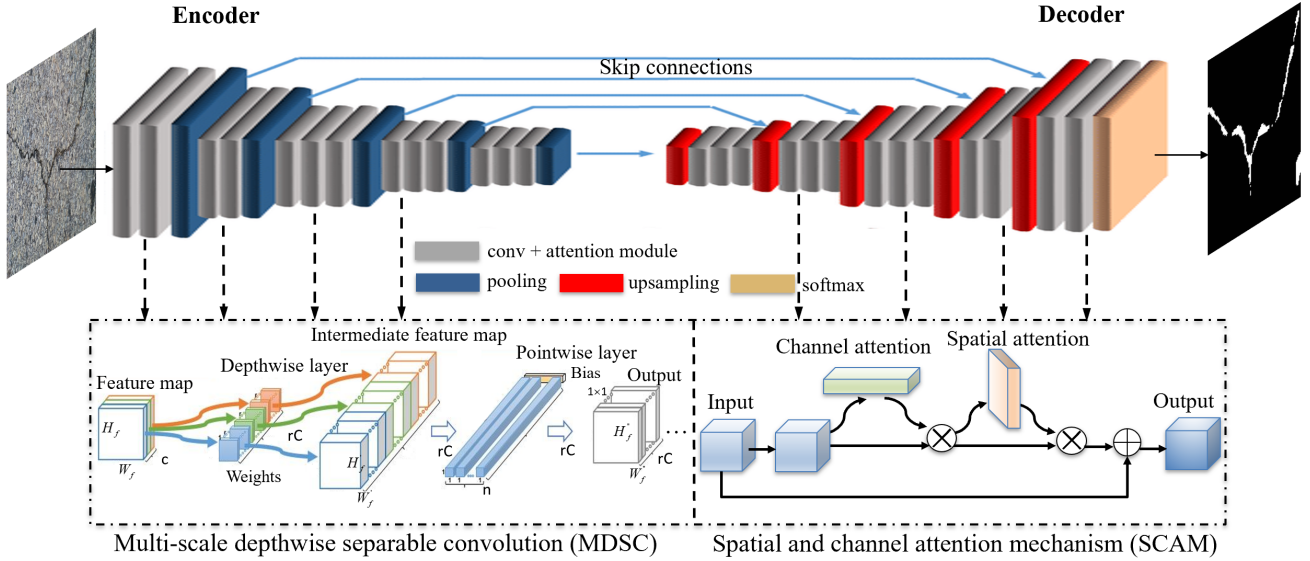


Fig. 4: The architecture of the proposed MANet model.

TABLE I: The main parameters of the MANet.

Module type	Input shape	Convolution kernel	Output shape	Repeated times	Stride
Input layer	224×224×3	-	(None, 224, 224, 3)	1	-
Conv1	226×226×3	3×3, 32	(None, 112, 112, 32)	1	2
DSConv block	112×112×32	3×3, 64	(None, 112, 112, 64)	2	2
CAM+SAM	112×112×32	-	(None, 112, 112, 32)	1	-
DSConv block	114×114×64	3×3, 128	(None, 56, 56, 128)	2	2
CAM+SAM	56×56×128	-	(None, 56, 56, 128)	1	-
DSConv block	56×56×128	3×3, 256	(None, 28, 28, 256)	2	2
CAM+SAM	28×28×256	-	(None, 28, 28, 256)	1	-
DSConv block	28×28×256	3×3, 1×1, 5×5, 512	(None, 14, 14, 512)	6	2
ZeroPadding (decoder)	14×14×512	-	(None, 16, 16, 512)	1	-
Conv2	16×16×512	3×3, 512	(None, 14, 14, 512)	1	-
CAM+SAM	14×14×512	-	(None, 14, 14, 512)	1	-
Upsampling	14×14×512	-	(None, 28, 28, 512)	3	2
Conv block	16×16×512	3×3, 256, 128, 64	(None, 112, 112, 64)	3	-
Conv3	112×112×64	3×3, 2	(None, 112, 112, 2)	1	-
CAM+SAM	112×112×2	-	(None, 112, 112, 2)	1	-
Output	112×112×2	-	(None, 12544, 2)	Reshape	-

The multi-scale convolution kernels are conducted in the DSConv blocks of DSConv8~DSConv12, and the 1×1, 3×3, and 5×5 convolution kernels are utilized alternatively. After each depthwise separable convolution block, the size of the output feature map is reduced to 1/2 of the input feature map, and the number of channels is doubled. The size of the feature map obtained after the encoder is 14×14×512.

(2) Decoder module. The decoder module restores the target details in the image layer by layer, and the most common one is the U-Net framework, where the notion is to upsample the high-level features and merge the corresponding low-level features for gaining the gradual restoration of the target details. Nevertheless, the degree of target detail recovery is limited to the simple fusion of the high-level and low-level features. Therefore, the hybrid attention is also embedded into the decoder module to remove redundant information, locate

cracks accurately, and restore defect details, thereby improving the quality of output. In short, the attention mechanism is incorporated into both the encoder module and decoder module to enhance the model performance. The input feature map is the output of the encoder part, and the 4 convolution blocks comprised of a 3×3 convolution layer, ReLU activation function, and batch normalization are separately conducted with the number of channels as 512, 256, 128, and 64. In particular, the attention module including the cascaded channel-wise attention module (CAM) and spatial attention module (SAM) is embedded after the convolution layers of each convolution block. In this manner, the features obtained by CAM and SAM are fused to generate the output of the decoder part, which merges both the low-level and high-level features and effectively restores the detailed information of the pavement defect images. Fig. 4 depicts the architecture of the

proposed MANet and the main parameters are summarized in Table I.

2) *Model Training*: For modeling training, the Cross-Entropy (CE) Loss Function is the commonly-used loss function in deep CNNs for pixel-wise image segmentation, and the formula can be defined by

$$L = - \sum_{k=1}^K y_k \log(p_k), \quad (7)$$

where K denotes the number of classes, $y_k \in \{0, 1\}$ (if the class k is consistent with the type of the sample, y_k is equal to 1; otherwise, it is 0), p_k is the predicted distribution of a specific sample belonging to class k . Due to the demerit that the prediction loss weights are regarded as the same for the negative and positive instances in CE Loss Function, Lin et al. [40] recommended the Focal Loss (FL) function instead of the classical CE function. The FL function is calculated as

$$FL(p_k) = - \omega_k (1 - p_k)^\gamma \log(p_k), \quad (8)$$

In Eq. (8), ω is the weighting factor, and γ is a hyperparameter of modulating factor. It is essential to emphasize that the Focal Loss function is primarily developed to solve the problems of imbalanced and indistinguishable samples for the target detection tasks that need binary classification. However, multi-classification tasks are more required in practical applications, such as multilabel image classification and multiscale segmentation, etc. Thereupon, we enhanced the Focal Loss function and substituted it for the traditional CE Loss Function. The formula of the enhanced Focal Loss function is presented using Eqs. (9-11).

$$FL_{mult}(p_k) = - \sum_{k=1}^K \omega_k (1 - p_k)^\gamma y_k \log(p_k), \quad (9)$$

$$\omega_k = \text{count}(x_i) / \text{count}(x_i \in k), \quad (10)$$

$$y_k = \begin{cases} 1, & k = \text{actual class}, \\ 0, & k \neq \text{actual class}, \end{cases} \quad (11)$$

where K signifies the total number of categories, and x_i represents the sample. On the basis of this, the detailed training procedure of the proposed method is displayed in Algorithm 1.

III. EXPERIMENTS AND ANALYSIS

In this section, the experiments are conducted to investigate the performance of the proposed approach. Except that some image pre-processing work was implemented by Photoshop, we primarily conducted the experiments using Python 3.6, where Keras, Tensorflow, and OpenCV3 libraries were used for algorithm running. The hardware configuration to execute the pavement defect detection algorithm contains the Intel® Xeon(R) E5-2620V4 processor, RTX 2080 TI graphics card (GPU), and 64 GB memory.

Algorithm 1: The detailed training procedure of the model

Input: The training sample $T = \{x_1, x_2, \dots, x_n\}$ where $i = 1, 2, \dots, n, x \in R^n$

- 1 **Begin**
- 2 Randomly initialize the parameters of the models $w \in R^d$;
- 3 **while not done do**
- 4 The network parameters are trained using the target dataset, where the Adam solver is used to update the weights.
- 5 $w_{c+1} = w_c - \eta * \hat{m}_c / (\sqrt{\hat{s}_c} + \varepsilon)$
- 6 where w is the weight matrix, c indexes the classes, η means the learning rate, \hat{m}_c and \hat{s}_c denote the bias-corrected first and second moments.
- 7 The optimized Focal Loss function is used instead of the classical Cross-Entropy loss function, and the model is evaluated by $\nabla_w L_{train}(w) // L_{train}$ refers to Eq.(9).
- 8 Continuously update the weight parameters $w \leftarrow w - \eta \nabla_w L_{train}(w)$ using T sample set.
- 9 **End while**
- Output:**
- 10 Obtain the best model parameter w_{best} .
- 11 **End**

A. Experiments on the Public Datasets

1) *Experimental results*: A series of experiments were conducted on the publicly available CRACK500 and CFD datasets to evaluate the performance of the proposed approach. Using the method proposed in Section II C, we implemented both the model training and testing on the pavement defect images. The well-known algorithms including U-Net [41], PSPNet [42], FCNet [43] and CrackForest [2], were chosen as the baseline models for comparative analysis. As described in Section II A, the training and validation sets of CRACK500 were utilized to train and determine if the models were overfitted, while the test dataset was used to evaluate the models. In addition, to ensure a fair comparison, the parameters of all the models were kept the same, including the hyper-parameters of batch size, training epochs, optimizer, and others. Each experiment ran for 5 epochs with the steps_per_epoch of 512, the batch size of 2, the true shuffle, and the optimizer of Adam. The Cross-Entropy loss function was used for the other compared methods except that the enhanced Focal Loss function was employed for our proposed approach.

Taking the statistics of accurate and false detections into account, we measure the model performance using the metrics like *Precision (Pr)*, *Recall (Re)*, *F1 - Score (F1)*, *Overlapping Rate (OR)*, and *Mean Intersection over Union (MIoU)*. Where, the higher the values of metrics are, the better the segmentation performance is. The formulas of these measurement metrics are defined as follows:

$$Pr = \frac{TP}{TP + FP}, \quad (12)$$

TABLE II: Performance on the CRACK500 dataset.

No.	Models	P_r	R_e	$F1$	OR	$MIoU$	Time (s)
1	U-Net	0.7498	0.7212	0.7352	0.6381	0.4721	00:45
2	PSPNet	0.8098	0.6077	0.6943	0.5792	0.6586	00:33
3	FCNet	0.8011	0.7810	0.7909	0.6942	0.6491	00:32
4	CrackForest [2]	0.4825	0.6293	0.5462	-	-	02:95 (CPU)
5	FPHBN [21]	0.7012	0.6993	0.7002	-	0.5600	00:20
6	DMANet [44]	0.6950	0.8000	0.7440	-	0.5590	-
7	MANet	0.8663	0.8454	0.8557	0.7337	0.7219	00:52

$$Re = \frac{TP}{TP + FN}, \quad (13)$$

$$F1 = \frac{2P_r \times R_e}{P_r + R_e}, \quad (14)$$

$$OR = \frac{|P_{result} \cap GT|}{|P_{result} \cup GT|}, \quad (15)$$

$$MIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{TP}{FN + FP + TP} \quad (16)$$

where TP (true positive) means the number of pixels that belong to the defect regions and are correctly detected. FP (false positive) denotes the number of pixels that do not belong to the defect regions but are wrong detected as the defects. FN (false negative) is the number of pixels that belong to the defect regions incorrectly detected. P_{result} denotes the predicted crack regions and GT means the real crack regions. k indicates the number of classifications. Fig. 5 displays the examples of the results detected by the different methods and relevant metrics measurements are presented in Table II. Also, some results reported by the latest literature are summarized in this table for comparative analysis.

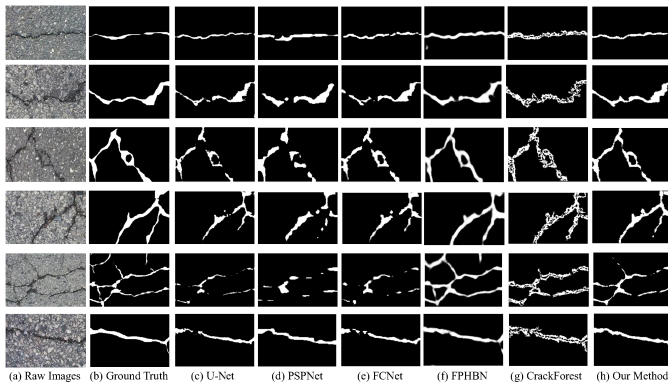


Fig. 5: The test results on the CRACK500 dataset.

As can be observed from Fig. 5, when the interferences in the images are relatively small and the crack is clear, these comparative models can better detect cracks, e.g., the crack samples in the first and the last rows have been detected well by these methods. Conversely, when there are many cracks in the images or the image definition is not high or there are other interferences, the test results of each model are quite different. As seen in the fifth row of Fig. 5, the upper cracks are even difficult to be distinguished by the naked eye in some areas, and each model has missed detection to a certain extent. U-Net, PSPNet, and FCNet missed seriously, while MANet

probed more details of these cracks. From the second, third, and fourth rows of Fig. 5, it can be seen that the U-Net, FCNet, and CrackForest have some missed detections, and the detected cracks are not continuous too. The FPHBN has over-segmentation issue since the segmented crack area is generally thicker than the ground truth, as shown in the first, second, third, and fifth rows. In addition, the missed detection areas also exist, such as the second row of Fig. 5 (f). Particularly, the CrackForest misses most details of the cracks and PSPNet has more false detections. Whilst, U-Net and FCNet have some under-segmentation areas since some crack images are not correctly segmented out, as shown in Fig. 5 (c,e). Although MANet also has some misdetection to a certain extent, the overall detection effect is the best and the cracks are detected more accurately. In brief, it can be seen from Fig. 5 that the MANet detects cracks more correctly and has preserved more crucial information compared with other state-of-the-art methods. Qualitative analysis reveals the effectiveness of the proposed MANet in pavement crack detection. Further, as seen in Table II, the comprehensive indicators $F1$, OR , and $MIoU$ of the proposed MANet reach 0.8557, 0.7337, and 0.7219, respectively. They are the highest among all the algorithms, so the MANet has the best crack detection effect, which is consistent with the results of the quantitative analysis. Additionally, in terms of time consumption of crack detection, the MANet takes slightly more time than other compared methods. The maximum time-consuming difference between the proposed approach and other comparison methods is not greater than the 20s, which does not pose a challenge for the current hardware level.

In like manner, the experiments were further conducted on the CFD dataset. The ratio of the samples randomly assigned to the training set to those in the test set was 8:2, except that 10% of the sample images were drawn to verify the validity of the models. The partial detection samples of different methods are displayed in Fig. 6. Besides, apart from the metrics measurements of the detection samples for different methods, a performance investigation of our approach compared with the latest methods in existing literature is accomplished, as listed in Table III. The experimental findings on the CFD dataset also show the promising performance of the proposed approach compared with other state-of-the-art methods.

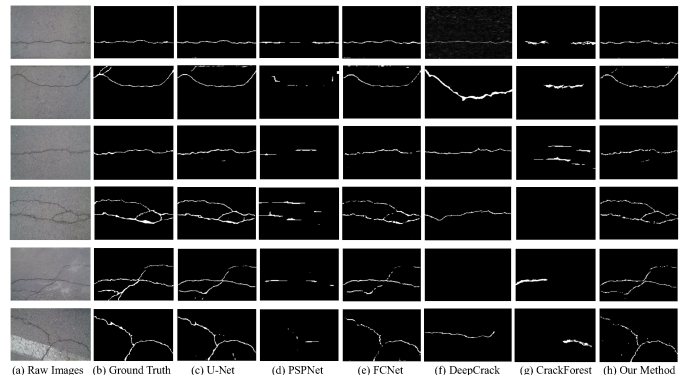


Fig. 6: Detection results on CFD dataset.

TABLE III: Performance on the CFD dataset.

No. Models	Pr	Re	$F1$	OR	$MIOU$	Time (s)
1 U-Net	0.7833	0.7729	0.7781	0.6892	0.7089	01:48
2 PSPNet	0.7062	0.5269	0.6035	0.5177	0.5188	00:49
3 FCNet	0.7905	0.6755	0.7284	0.6356	0.4915	00:45
4 Canny [45]	0.4337	0.7307	0.4570	-	-	-
5 CrackForest [2]	0.7466	0.9514	0.8318	-	-	03:74 (CPU)
6 TuFF [46]	0.5521	0.4177	0.4465	0.2987	-	59:35
7 DeepCrack [19]	0.6550	0.7600	0.7040	-	0.7660	00:18
8 HACNet [20]	0.6780	0.7510	0.7100	-	0.7700	00:42
9 MANet	0.7634	0.8908	0.8221	0.7153	0.7788	02:11

The following still analyzes the effectiveness of MANet from two aspects: quantitative and qualitative. When the crack locations are clear and the interference is less, the U-Net, FCNet, and MANet can all detect the cracks well, as shown in rows 1-3 of Fig. 6. The cracks of these samples are properly segmented by the U-Net, FCNet, and MANet methods basically. However, PSPNet and CrackForest have more missed detections on these samples. When the cracks in the image are more complex or the background interference is large, each model has different degrees of discontinuity or error detection of pavement cracks, as shown in rows 4-6 in Fig. 6. PSPNet, DeepCrack, and CrackForest have serious under-segmentation defects, and FCNet has a similar problem. Especially, DeepCrack has the problem of mis-segmentation, which may be caused by the overfitting of this model with its deep network structure. Besides, the output of multiple side maps increases the computational complexity and may incur some possible errors. U-Net has the over-segmentation issue since some residues are left in the segmented images of Fig. 6 (c), which implies the noise of the segmented results. By comparison, the MANet model proposed in this paper can basically maintain the integrity of the cracks and performs better in detail. Further, Table III presents the quantitative comparison of the detection results of different methods on the CFD test dataset. The comprehensive metrics OR and $MIOU$ of the proposed approach are 0.7153 and 0.7788, which are the best performance of all the algorithms. For another comprehensive index $F1$, it is the best for the proposed approach except for the CrackForest, while there are many missed detections for the CrackForest method. Moreover, the MANet also indicates superior effectiveness compared to the results of related literature, as shown in Table III. The main reason behind the solid performance of the proposed approach is that the MANet adopts the multi-scale convolution kernels in the DSConv layers instead of the single 3×3 convolution kernels, which expands the convolutional receptive field and enhances the richness of the convolutional feature channels. Besides, the hybrid attention mechanism incorporated in the network realizes the maximum reuse of inter-channel relations and infers the significance of spatial point features. As a consequence, the promising experimental results are obtained by the proposed approach, which indicates the model has obtained an increasing performance gain relative to other influential methods and can be used to detect pavement crack defects.

2) *Ablation study*: We perform the ablation experiments

TABLE IV: The performance of ablation experiments.

Ablation approach	Pr	Re	$F1$	$MIOU$
Delete multi-scale conv	0.8443	0.7192	0.7767	0.6791
Delete attention	0.8237	0.7997	0.8115	0.7231
Replace EFL with CE	0.8573	0.7047	0.7735	0.6718
This study	0.7634	0.8908	0.8221	0.7788

on our model. Specifically, we analyze the efficacy of multi-scale convolution and hybrid attention modules on the experimental dataset of the CFD crack images. First, we separately remove the modules of multi-scale convolutions and hybrid attention in the network to investigate the performance of the model training. Then, we evaluate the effect of the optimized loss function by substituting the enhanced Focal Loss (EFL) function with the traditional Cross-Entropy (CE) loss function (see Eq. (7)). Table IV summarizes the comparison results of ablation experiments. From Table IV we notice a significantly decreased performance in the results of the ablated models. The $MIOU$ of removing multi-scale convolution and hybrid attention modules drop to 0.6791 (decrease by 0.0997) and 0.7231 (decrease by 0.0557), respectively. It is worthy to note that the ablated models still perform better than some benchmark methods, as shown in Table III. The ablation experiment indicates that both the multi-scale convolution and hybrid attention modules contribute to the performance gain of the proposed approach. For the effect of the optimized loss function, we also notice an obvious decrease, where the $MIOU$ drops to 0.6718 (decrease by 0.107). The key explanation for this is that the crack region only occupies a small part of the whole image, and the CE function does not consider the sample imbalance problem, resulting in decreased accuracy. This ablation experiment demonstrates that the EFL function delivers better results than that of the CE loss function used in our model for pavement crack detection.

B. Experiments on the Local Dataset

1) *Image segmentation*: Similar to the above experiments performed on the publicly accessible datasets, the proposed approach is further tested on our collected local pavement defect images. To improve the detection ability of the system, the data augmentation scheme is utilized to produce new synthetic images for enhancing the diversity and variety of samples. Except for 131 original samples, the 416 synthetic samples and their corresponding Ground Truth (GT) are generated by random scaling and rotation, horizontal or vertical flipping, and shifting to synthesize new images for enriching the dataset. Among these sample images, 472 images are used for the training set, 52 images are for the validation set, and 23 images are for the test set. The detailed parameter assignments of the data augmentation methods are presented below: the sample images were zoomed in or out with a scale transformation from 0.9 to 1.1, the random rotation was implemented on the images with the angle of ($90^\circ, 180^\circ$), and the raw images were flipped 90° along the horizontal or vertical axes. On this basis, the experiments were conducted on the local pavement images. Fig. 7 displays the partial sample images detected by different methods, and the corresponding measurement metrics are calculated in Table V.

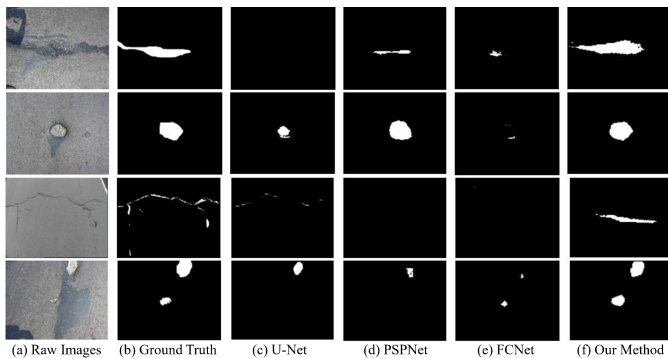


Fig. 7: Detection results on the local dataset.

TABLE V: Metrics measurements on the local dataset.

No.	Models	Pr	Re	$F1$	OR	$MIOU$	Time (s)
1	U-Net	0.5775	0.5379	0.5570	0.5185	0.4928	01:05
2	PSPNet	0.7773	0.5475	0.5425	0.5368	0.5568	00:30
3	FCNet	0.6327	0.5250	0.6288	0.5141	0.4917	01:01
4	MANet	0.8833	0.6268	0.7332	0.6121	0.6514	00:56

As seen in Fig. 7, the test results of the algorithms on the local dataset are different, and each algorithm has some missing detection more or less. Such as columns 3, 4, and 5, the missing detection of U-Net, PSPNet, and FCNet is relatively serious, while the performance of the proposed method is superior to that of the compared methods, as displayed in the column 6 of Fig. 7. Besides, although the performance of all algorithms on the local dataset is inferior to that on the open-source datasets because of the more complicated background and interference conditions, the proposed approach outperforms the other compared methods in terms of quantitative analysis metrics like *Precision*, *Recall*, *F1 – Score*, *Overlapping Rate*, and *MIOU* as shown in Table V. Moreover, to further evaluate the generalization ability of the proposed network, the leave-one-out-cross-validation approach is implemented to obtain the model performance evaluation results. A total of 5-fold cross-validation experiments are performed in our work. Each time, the 472 images are used as the training set to learn the model and the other 52 images are used as the validation set to evaluate the model. The segmentation performance on the test set excluded at the modeling stage is reported as the results of each cross-validation experiment. Consequently, using the leave-one-out-cross-validation approach, the 5 different experimental results are obtained respectively. Table VI summarizes the cross-validation results on the test dataset. As seen in Table VI, the average *Precision* and *MIOU* reach 0.8464 and 0.6661 in multiple cross-validation experiments, and the corresponding standard deviations (*Std*) are 0.0496

TABLE VI: Comparative results of cross-validation experiments.

Cross validation	Pr	Re	$F1$	$MIOU$
Fold-1	0.8154	0.7183	0.7637	0.6698
Fold-2	0.8912	0.6288	0.7373	0.6145
Fold-3	0.8696	0.6798	0.7630	0.6549
Fold-4	0.7747	0.8845	0.8259	0.7319
Fold-5	0.8812	0.6824	0.7691	0.6592
<i>Mean</i>	0.8464	0.7187	0.7718	0.6661
<i>Std</i>	± 0.0496	± 0.0979	± 0.0326	± 0.0423

and 0.0423, respectively, which demonstrates that the proposed method is robust and effective across a range of different data sources and conditions. Thereafter, according to the experimental results, it can be assumed that the proposed method has a certain capability to detect pavement structural defects, and can also be transplanted in other fields.

2) *Image classification*: In practice, the aim of pavement defect detection needs to address the problems of whether there are defects in the pavements and what the pavement defects are. Thus, in addition to extracting the defect regions of pavements, we have to know the specific types of pavement defects too. As illustrated in Section II C, we added the fully connected layer in the extractor module of MANet and utilized it to generate a new classification network for the identification of pavement defect types. That is, the multi-scale attention mobile network, where the multi-scale convolution kernels were substituted for the traditional 3×3 convolution kernel in DSCov layers and the attention module was incorporated into the network to highlight the useful features while suppressing the needless information, was used to perform the identification of pavement defect types. The ratio of the samples randomly assigned to the training set to those in the test set was 3:1, and the data enhancement scheme was also utilized in model training. Approximately 1,000 sample images were guaranteed for the augmented samples, in which 10% of the images were drawn from the training set as the validation set of the model. The hyper-parameter training optimizer was Adam, with epochs of 100, a minibatch size of 64, and a learning rate of 1×10^{-3} . After 100 epochs of training, the training accuracy of the proposed method achieves 97.08%, the validation accuracy attains 92.31%, the training loss is 0.0605, and the validation loss is 0.1182, as depicted in Fig. 8. On the ground of this, the trained optimum model can be used for the test of pavement defect detection, and Fig. 9 depicts the confusion matrix of test results.

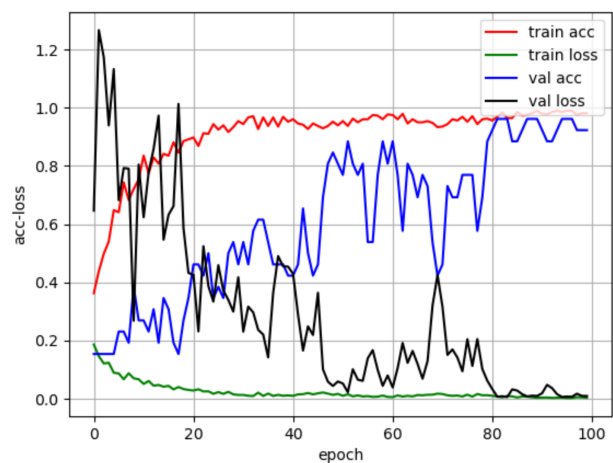


Fig. 8: The performance of model training.

It can be visualized from Fig. 8 that the curves of training and validation accuracy all tend to be stable and achieve a higher value after around 80 epochs of training, which demonstrates the effectiveness of the proposed approach. Thereupon, the obtained model was used for the prediction

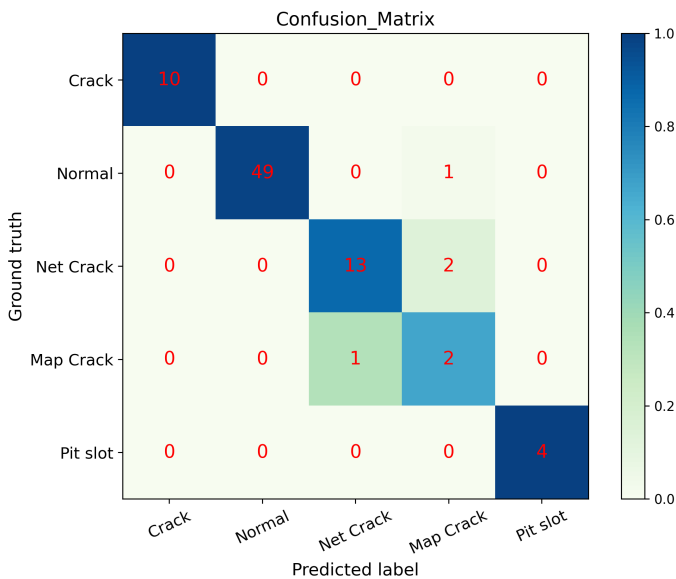


Fig. 9: The prediction result of the proposed approach.

TABLE VII: Identification results compared with literature [16].

ID	Pavement images	This study	EANN [16]	Params (m/cm/m ²)		
				width	diameter	area
1	normal1	normal	normal	/	/	/
2	normal2	normal	normal	/	/	/
3	normal3	normal	normal	/	/	/
4	normal4	normal	normal	/	/	/
5	pit_slot1	pit slot	pit_slot	/	/	0.75
6	pit_slot2	pit slot	pit_slot	/	/	1.33
7	pit_slot3	pit slot	pit_slot	/	/	1.36
8	pit_slot4	pit slot	pit_slot	/	/	1.32
9	slight map crack1	net crack	map crack	/	35	/
10	slight crack1	crack	crack	1.35	/	/
11	slight crack2	crack	crack	2.20	/	/
12	slight crack3	crack	crack	1.68	/	/
13	slight crack4	crack	crack	3.34	/	/
14	slight crack5	crack	crack	2.75	/	/
15	slight crack6	crack	crack	3.29	/	/
16	slight crack7	crack	crack	1.88	/	/
17	slight crack8	crack	crack	3.42	/	/
18	slight net crack1	net crack	crack	/	123	4.2
19	slight net crack2	map crack	crack	/	154	5.7
20	slight net crack3	net crack	crack	/	118	6.4
21	severe map crack	map crack	map crack	/	18	/
22	severe map crack	map crack	map crack	/	15	/
23	severe crack1	crack	crack	14.6	/	/
24	severe crack2	crack	crack	13.8	/	/
25	severe net crack1	net crack	crack	/	55.3	5.2
26	severe net crack2	net crack	crack	/	60.9	6.7

of pavement defect types. The proposed approach shows a good ability to identify crack damage images, even if some oil marks or road marks interfere with images. A promising performance can be reflected by the confusion matrix of Fig. 9, where most of the sample images in each category have been successfully identified by the MANet. For example, 10 crack samples have all been accurately recognized by the proposed approach. Except that 1 instance is misclassified into the types of map crack, 50 normal samples have been correctly identified. 15 net-crack samples have been accurately recognized by the proposed approach except for 2 samples incorrectly identified as map-crack type. Similarly, the 3 map-crack defect samples have been properly recognized by the

proposed approach apart from 1 sample mistakenly classified into the net crack type. The 4 pit slot samples have been accurately identified by the proposed method too. Furthermore, a performance investigation of our method compared to the results of the existing literature has also been implemented, and Table VII summarizes the representative samples identified by our method compared with the results in the literature [16]. From this table, we can see that the number of samples correctly identified by the proposed method is more than that of the approach reported in [16]. Thereupon, through the comparative analysis, it can be known that the proposed method outperforms the existing state-of-the-art and exhibits a competitive performance for pavement defect detection.

IV. CONCLUSIONS

With the rapid economic development, great achievements have been made in road construction, and to improve the service life of the road, the maintenance of the pavement has become more and more important along with the advancement of the road network system. The traditional methods that rely on manual detection of road damage can no longer meet the needs of road development. Therefore, the research and application of automatic pavement defect detection has great need and realistic significance, which is also a hot and difficult research topic in the field of intelligent recognition systems. Aiming at the problem that various defects are difficult to detect, this paper built a pavement defect detection network, in which the MobileNet was selected as the backbone extractor, and a hybrid attention module was separately introduced in the encoder and decoder modules. Moreover, the multi-scale convolution kernels were substituted for the original 3×3 convolution kernels in depth-wise separable convolution layers of the network, which enlarged the convolution receptive fields and improved the capability of feature extraction. In the experiments, the publicly available CRACK500 and CFD datasets along with our collected local pavement defect images are utilized to verify the relevant performance of the proposed approach. The experimental results indicate that the proposed approach is completely better than other reference methods in terms of metrics like *OR* and *MIoU*. The experimental findings also prove that it is a very effective approach by using the extractor module of MANet and adding the fully connected layer to perform the identification of pavement defect types.

In our experiments, the proposed MANet has proven to be quite promising. However, it does have some limitations: First, the model has high segmentation performance and identification accuracy, but at the expense of a higher computational complexity. It consumes slightly more computational time. Model pruning algorithms can be added to simplify the model in the future work. Besides, since most publicly-available datasets in pavement crack detection are still static image datasets, our method has only been performed on image data so far. In the future, we will test our method on video datasets. Furthermore, we would like to transplant the model on more real-world applications.

ACKNOWLEDGMENT

The authors would like to thank Fundamental Research Funds for the Central Universities (No. 20720181004), and the authors also want to thank editors and unknown reviewers for the constructive advice.

REFERENCES

- [1] Zhang, Lei, et al. "Road crack detection using deep convolutional neural network." *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016.
- [2] Shi, Yong, et al. "Automatic road crack detection using random structured forests." *IEEE Transactions on Intelligent Transportation Systems* 17.12 (2016): 3434-3445.
- [3] Lin, Dongyun, et al. "CAM-guided Multi-Path Decoding U-Net with Triplet Feature Regularization for Defect Detection and Segmentation." *Knowledge-Based Systems* 228 (2021): 107272.
- [4] Nguyen, Tien Sy, Manuel Avila, and Stéphane Begot. "Automatic detection and classification of defect on road pavement using anisotropy measure." *2009 17th European Signal Processing Conference*. IEEE, 2009.
- [5] Chen, Qi, et al. "Pavement crack detection using hessian structure propagation." *Advanced Engineering Informatics* 49 (2021): 101303.
- [6] Suh, Gahyun, and Young-Jin Cha. "Deep faster R-CNN-based automated detection and localization of multiple types of damage." *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems* 2018. Vol. 10598. SPIE, 2018.
- [7] Doğan, Gürkan, and Burhan Ergen. "A new mobile convolutional neural network-based approach for pixel-wise road surface crack detection." *Measurement* 195 (2022): 111119.
- [8] Kaseko, Mohamed S., and Stephen G. Ritchie. "A neural network-based methodology for pavement crack detection and classification." *Transportation Research Part C: Emerging Technologies* 1.4 (1993): 275-291.
- [9] Liu, Fanfan, et al. "Novel approach to pavement cracking automatic detection based on segment extending." *2008 International Symposium on Knowledge Acquisition and Modeling*. IEEE, 2008.
- [10] Ayenu-Prah, Albert, and Nii Attoh-Okine. "Evaluating pavement cracks with bidimensional empirical mode decomposition." *EURASIP Journal on Advances in Signal Processing* 2008 (2008): 1-7.
- [11] Chanda, Sukalpa, et al. "Automatic bridge crack detection—a texture analysis-based approach." *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer, Cham, 2014.
- [12] Zhou, Jian, Peisen S. Huang, and Fu-Pen Chiang. "Wavelet-based pavement distress detection and evaluation." *Optical Engineering* 45.2 (2006): 027007.
- [13] Subirats, Peggy, et al. "Automation of pavement surface crack detection using the continuous wavelet transform." *2006 International Conference on Image Processing*. IEEE, 2006.
- [14] Quintana, Marcos, Juan Torres, and José Manuel Menéndez. "A simplified computer vision system for road surface inspection and maintenance." *IEEE Transactions on Intelligent Transportation Systems* 17.3 (2015): 608-619.
- [15] Kodovsky, Jan, Jessica Fridrich, and Vojtěch Holub. "Ensemble classifiers for steganalysis of digital media." *IEEE Transactions on Information Forensics and Security* 7.2 (2011): 432-444.
- [16] Chen, Junde, Anwar Ul Haq, and Defu Zhang. "Block-based automatic road defect recognition approach." *Journal of Electronic Imaging* 28.5 (2019): 053023.
- [17] Cord, Aurélien, and Sylvie Chambon. "Automatic road defect detection by textural pattern recognition based on AdaBoost." *Computer-Aided Civil and Infrastructure Engineering* 27.4 (2012): 244-259.
- [18] Choi, Wooram, and Young-Jin Cha. "SDDNet: Real-time crack segmentation." *IEEE Transactions on Industrial Electronics* 67.9 (2019): 8016-8025.
- [19] Liu, Yahui, et al. "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation." *Neurocomputing* 338 (2019): 139-153.
- [20] Chen, Hanshen, and Huiping Lin. "An effective hybrid atrous convolutional network for pixel-level crack detection." *IEEE Transactions on Instrumentation and Measurement* 70 (2021): 1-12.
- [21] Yang, Fan, et al. "Feature pyramid and hierarchical boosting network for pavement crack detection." *IEEE Transactions on Intelligent Transportation Systems* 21.4 (2019): 1525-1535.
- [22] Cha, Young-Jin, Wooram Choi, and Oral Büyüköztürk. "Deep learning-based crack damage detection using convolutional neural networks." *Computer-Aided Civil and Infrastructure Engineering* 32.5 (2017): 361-378.
- [23] Schmutge, Stephen J., et al. "Crack segmentation by leveraging multiple frames of varying illumination." *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017.
- [24] Cha, Young-Jin, et al. "Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types." *Computer-Aided Civil and Infrastructure Engineering* 33.9 (2018): 731-747.
- [25] Ali, Rahmat, and Young-Jin Cha. "Attention-based generative adversarial network with internal damage segmentation using thermography." *Automation in Construction* 141 (2022): 104412.
- [26] Zhang, Allen, et al. "Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network." *Computer-Aided Civil and Infrastructure Engineering* 32.10 (2017): 805-819.
- [27] Kang, Dongho, and Young-Jin Cha. "Autonomous UAVs for structural health monitoring using deep learning and an ultrasonic beacon system with geo-tagging." *Computer-Aided Civil and Infrastructure Engineering* 33.10 (2018): 885-902.
- [28] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017).
- [29] Hou, Qibin, Daquan Zhou, and Jiashi Feng. "Coordinate attention for efficient mobile network design." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
- [30] Tsotsos, John K. A computational perspective on visual attention. MIT Press, 2011.
- [31] Huang, Zilong, et al. "Ccnet: Criss-cross attention for semantic segmentation." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
- [32] Fu, Jun, et al. "Dual attention network for scene segmentation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
- [33] Hou, Qibin, et al. "Strip pooling: Rethinking spatial pooling for scene parsing." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
- [34] Hu, Jie, et al. "Gather-excite: Exploiting feature context in convolutional neural networks." *arXiv preprint arXiv:1810.12348* (2018).
- [35] Bello, Irwan, et al. "Attention augmented convolutional networks." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.
- [36] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
- [37] Kang, Dong H., and Young-Jin Cha. "Efficient attention-based deep encoder and decoder for automatic crack segmentation." *Structural Health Monitoring* 21.5 (2022): 2190-2205.
- [38] Lewis, John, Young-Jin Cha, and Jongho Kim. "Dual encoder–decoder-based deep polyp segmentation network for colonoscopy images." *Scientific Reports* 13.1 (2023): 1183.
- [39] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
- [40] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." *Proceedings of the IEEE international conference on computer vision*. 2017.
- [41] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.
- [42] Zhao, Hengshuang, et al. "Pyramid scene parsing network." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- [43] Sheikhjafari, Ameneh, et al. "Unsupervised deformable image registration with fully connected generative neural network." (2018).
- [44] Sun, Xinzi, et al. "Dma-net: Deeplab with multi-scale attention for pavement crack segmentation." *IEEE Transactions on Intelligent Transportation Systems* 23.10 (2022): 18392-18403.
- [45] Fan, Zhun, et al. "Automatic pavement crack detection based on structured prediction with the convolutional neural network." *arXiv preprint arXiv:1802.02208* (2018).
- [46] Mukherjee, Suvadip, Barry Condron, and Scott T. Acton. "Tubularity flow field—A technique for automatic neuron segmentation." *IEEE Transactions on Image Processing* 24.1 (2014): 374-389.



Junde Chen receives his master's degree in Sichuan University and the PHD degree in school of informatics, Xiamen University, China. Currently, he is doing PostDoctor research in Dale E. and Sarah Ann Fowler School of Engineering at Chapman University, Orange, CA, USA. His research interests include the aspects of Data Mining, Image Processing, Big data and Decision Support System etc.



Yuxin Wen works as an assistant professor in the Dale E. and Sarah Ann Fowler School of Engineering at Chapman University, Orange, CA, USA currently. Her research interests focus on data mining, data analysis, statistics, and mathematical modeling for quality improvement, prognostics in complex systems with applications in manufacturing, healthcare, and traffic, etc.



Yaser Ahangari Nanekaran received the B.E. degree from IAU of Ardabil Branch, Ardabil, Iran, in Power Electrical Engineering, M.Sc. degree in IT from Cankaya University, Ankara, Turkey, and PHD degree in school of informatics, Xiamen University, Chian. He is currently working in School of Information Engineering, Yancheng Teachers University, Yancheng, China. His research area mainly includes data mining, big data and deep learning techniques.



Defu Zhang works in School of Informatics at Xiamen University currently. His research interests include all aspects of computational intelligence, image analysis and data mining, etc. He has published papers in the following Journals: INFORMS Journal on Computing, Computers & Operations Research, European Journal of Operational Research, Expert System with Applications, etc.



Adnan Zeb received the master's degree in computer science from the department of Computer Science, COMSATS University, Islamabad, Pakistan, in 2017, and his PHD degree in school of informatics, Xiamen University. He is currently pursuing Post-Doctor research in the southern university of science and Technology, China. His research interests include Machine learning, Knowledge representation learning, Image Processing, etc.