Computer Science and Information Systems 21(4):1435–1455 https://doi.org/10.2298/CSIS230725042W

Advancing Crack Segmentation Detection:Introducing AAMC-Net Algorithm for Image Crack Analysis

Wang Xiaofang[1], Liu Chenfang[2], Hou Junliang[1], and Zhou Liang[1]

[1] Geely University of China,
Chengdu Sichuan, 641423
939549393@qq.com
houjunliang@guc.edu.cn
zhouliang@guc.edu.cn
[2] Chengdu College of University of Electronic Science and Technology of China,
Chengdu Sichuan, 611731
2975431869@qq.com

Abstract. This research paper presents an innovative solution to address the challenges of poor detail detection effectiveness and prolonged training time in image segmentation. The proposed approach leverages the Adaptive Attention Multiscale Convolution Network (AAMC-Net), incorporating a multi-scale dilated convolution VGG_L network for feature extraction and a deconvolution method for image segmentation. Extensive experiments demonstrate the superior performance of the proposed algorithm concerning intersection over Union (IOU), accuracy, precision, recall, F1, average training efficiency, and segmentation efficiency when compared to several traditional algorithms. On average, the proposed algorithm achieves remarkable improvements of 3.9%, 3.1%, 1.7%, 4.9%, 17.9%, 14.8% ,and 20.2% in these metrics. Moreover, the enhanced algorithm exhibits notable advantages in detail processing and real-time image segmentation detection.

Keywords: Image crack segmentation, Convolutional neural network,VGG_L, Attention mechanism.

1. Introduction

As a major infrastructure construction center, China has built a road network spanning 6 million kilometers by 2022 [2]. Effective road damage detection is crucial for ensuring road safety. Road cracks are early indicators of pavement deterioration [9]. Manually extracting cracks based solely on workers' experience and subjective judgments is inefficient and costly. Rapid and precise road crack detection is vital for reducing road maintenance expenses, enhancing driving safety, lowering fuel consumption, and prolonging road lifespan [27]. As a key technique for crack detection, effective crack segmentation can advance crack detection technology to some extent.

With the fast evolution of computer processors, image processing and computer vision have been extensively applied in crack segmentation, including threshold-based road crack extraction, morphological detection, region segmentation, and edge detection methods [12] [16]. Thresholding techniques like Otsu's method and histogram thresholding, despite being simple and fast, are prone to environment and illumination effects, resulting in poor performance on images with blurry crack edges. Morphological methods such

as erosion and dilation, multistructural and multiscale mathematical morphology, though effectively suppressing noise interference and achieving good detection with fast computation, struggle with detailed segmentation. Region-based crack segmentation methods like watershed and region growing enable automatic segmentation but perform poorly on faint and tiny cracks, requiring enhancement algorithms to suppress noise. Common edge detectors like Sobel, Roberts, gradient operators, Laplacian operators, Marr and Canny, though easy to implement and fast, are sensitive to noise due to template size and direction limitations.

With the development of artificial intelligence, deep learning and convolutional neural network (CNN) technologies have become a research trend for applying image crack segmentation. Models including FCN [34] [10], UNet [35], and UNet++ [33]have improved crack segmentation accuracy. However, these models have drawbacks like large computational loads and many network parameter settings, making them less ideal for real-time detection.

The shortcomings in fine-grain crack segmentation and the absence of real-time detection capabilities are pressing issues that directly impede advancements in intelligent crack detection systems. The VGG neural network is notable for its minimal computational demands and swift processing speeds, making it a comparatively lightweight model apt for rapid image analysis [23]. Attention mechanisms facilitate end-to-end learning, thereby aiding in the interpretation and understanding of data. When amalgamated with neural network models, these mechanisms markedly improve the precision of segmentation networks, demonstrating exceptional efficacy in image detection tasks [18]. Moreover, the Convolutional Block Attention Module (CBAM) enhances feature extraction in convolutional layers, accentuating key image information and excelling in local texture extraction. Dilated convolutions maintain spatial resolution whilst effectively capturing a broader contextual scope, all without augmenting the number of parameters or computational complexity. Consequently, this study aims to refine the VGG neural network by leveraging the benefits of CBAM and dilated convolutions to develop the AACM-Net model. The objective is to optimize the segmentation of crack details whilst concurrently reducing the time required for crack detection, thereby elevating the overall accuracy of segmentation.

2. Related Work

As the country places greater emphasis on intelligent transportation systems, the focus has shifted towards intelligent detection of road cracks. Image processing, particularly digital image processing, has witnessed significant advancements in crack detection [17]. Notably, researchers like Xiu et al. [31]proposed using the Sobel operator and CV model for image edge detection to address uneven gray image segmentation. Similarly, Jiang et al. [11] employed adaptive thresholding to extract crack edges and effectively mitigate noise interference. However, challenges persist due to complex lighting conditions, background variations, and other external factors, leading to suboptimal results.

With the rapid advancement of processors, artificial intelligence and computer vision have increasingly come to the fore in the digital economy era. Deep learning and convolutional neural networks point the way forward for road crack detection. Notably, like J.Long et al [15]utilized a semantic segmentation technique using a Fully Convolutional

Network(FCN8s). The FCN8s approach is quite adaptable, capable of handling images of any dimension. It streamlines the training regimen by producing classification labels in just one forward pass, enabling seamless training and prediction from start to finish. However, it's worth noting that this method demands a lot of computational power and falls short when it comes to detailing the textures of smaller objects. To address this, Dung et al. [4] proposed an FCN-based automatic crack identification algorithm, employing convolutional layers instead of fully connected layers to classify images of any scale. However, FCN's inefficiency and limited receptive field hinder it from capturing global and detailed information effectively. Further, Sun et al. [24] used a hybrid approach combining multi-resolution features with Transformers for crack semantic segmentation(Multi-Transformer). The method use a global receptive field for scene understanding and incorporates Transformer networks for semantic segmentation, overcoming the limitations of CNNs' small receptive fields. However, the use of Transformers and multi-resolution features increases the model's computational complexity.

At the same time, Wang et al [29] proposed an efficient road surface crack segmentation algorithm using a deep learning encoder-decoder structure. They used a pre-trained DenseNet121 as the encoder to extract road features and a global attention upsampling module as the decoder for crack segmentation. Although the method effectively improves training efficiency, it is highly dependent on the quality and quantity of the training data. The use of multiple attention modules and a pre-trained DenseNet121 adds to the model's complexity. Lin et al [13] used a deep semantic segmentation method combining Conditional Random Fields (CRF) and CNN (CRF-CNN) to capture contextual area information and background information for crack segmentation. The inclusion of contextual information enhances the model's image understanding capabilities but also increases its complexity and computational costs. Building on this, they introduced multi-scale inputs and sliding in pyramid pooling layers the following year to capture context for patch-background [14], improving semantic segmentation accuracy. While the introduction of efficient training methods somewhat optimizes the CRF and CNN architecture, the model remains complex and resource-intensive.

Furthermore, Fan [6] and his team introduced a road crack detection algorithm based on deep learning and adaptive image segmentation. The algorithm trains a deep convolutional network to determine whether an image contains cracks and uses adaptive thresholds to extract road surface cracks. Although the method performs well in extracting various types of road cracks, it requires substantial computational resources and has a long training time.

To overcome the limitations of existing methods and enhance crack detection accuracy, we propose a novel algorithm based on AAMC-Net image. The algorithm leverages the CBAM hybrid domain attention mechanism to strengthen global and local image features. Furthermore, we employ multi-scale dilated convolutions VGG_L for feature extraction and utilize deconvolution for fracture detection. The entire model training is based on MSELoss Adam for adaptive learning, optimizing performance and efficiency. Through extensive experiments, we demonstrate the superior capabilities of our AAMC-Net based algorithm in intelligent road crack detection, providing an innovative solution for infrastructure maintenance and safety enhancement.

3. Research Method

The CBAM attention mechanism is a lightweight, general-purpose module that can be seamlessly integrated into convolutional neural network architectures for end-to-end training and the capture of useful information. VGG achieves learning by deepening the network with multiple nonlinear layers, albeit with fewer network parameters. In this study, we use the VGG network as the foundational structure and propose a multi-scale dilated convolution variant, termed VGG_L. This is combined with the CBAM attention mechanism to construct the AACM-Net network model specifically designed for road crack segmentation. The architecture of this model is depicted in Figure 1.

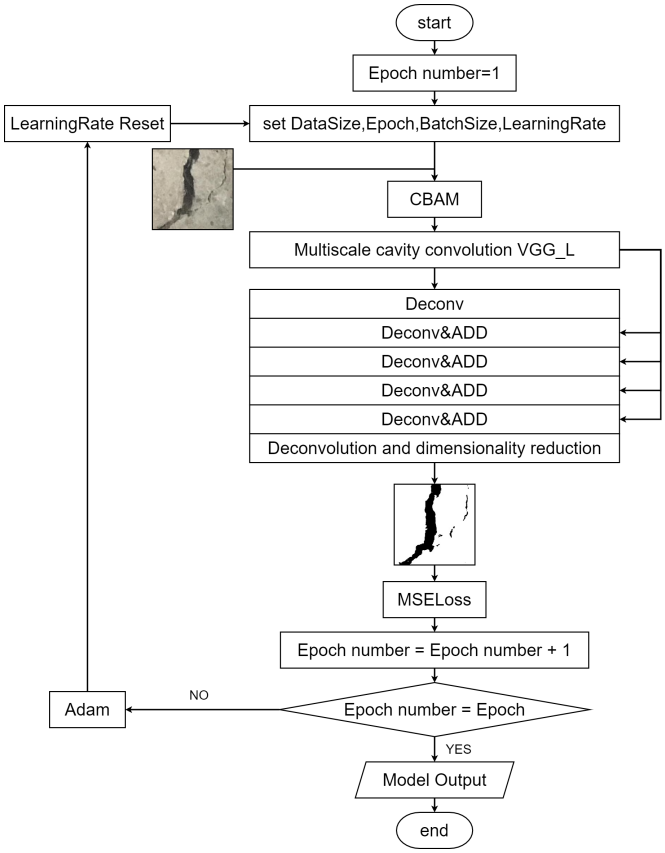


Fig. 1. Algorithm flow chart

A deep learning-based road crack image segmentation model is designed, with its training process illustrated in Figure 1. The initial step involves setting model parameters, including the size of the dataset, the number of training iterations, batch size, and the initialization of the learning rate. The model's input, crack images, first pass through the CBAM module, which enhances key features of the image using channel and spatial

attention mechanisms. Subsequently, feature extraction is performed using the VGG_L model, which employs multi-scale dilated convolutions to capture image details across different scales. Following feature extraction, the model reconstructs and segments features through a series of deconvolution and feature fusion steps (Deconv&ADD). At the end of each training cycle (Epoch), the model evaluates the segmentation effectiveness by calculating the Mean Squared Error (MSELoss) and adjusts parameters based on the loss using the Adam optimizer. This iterative process continues until the predetermined number of training iterations is reached, at which point the model outputs the final crack segmentation results, completing the training process. This training approach ensures that the model effectively learns to segment cracks from complex backgrounds, enhancing the accuracy and robustness of segmentation through the attention mechanism and multi-scale feature fusion. In this way, the model is equipped to handle crack detection challenges across various road surfaces and lighting conditions.

3.1. Image Processing

CBAM based on hybrid attention mechanism is used to preprocess the input crack feature image to enhance the global and local texture feature information. The hybrid attention mechanism based CBAM [30]model implements the image feature enhancement process by processing global and local texture features in channel features and spatial features. It consists of convolution of channel and spatial attention mechanisms, the overall process is shown in Figure 2.

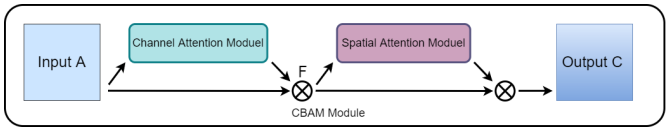


Fig. 2. Overall CBAM flow chart

The input A is initially processed through a channel attention module, resulting in a feature map that is element-wise multiplied (denoted as $\otimes$) with the input A to produce a feature map F. This process enhances features associated with specific channels. Subsequently, the feature map F undergoes further processing through a spatial attention module, and the resulting feature map is again element-wise multiplied with F, leading to the final output C. This illustration demonstrates how channel and spatial attention modules are sequentially employed to enhance the network's capacity to represent input images effectively.

This study integrates a Combined Channel and Spatial Attention Mechanism (CBAM) to enhance the precision of crack segmentation in images. As shown in Figure 2, the input image A initially undergoes processing through a channel attention module. This module focuses on the importance of different channels within the image, generating a feature map. The primary goal of this step is to highlight feature channels that are most relevant for crack detection.After processing by the channel attention mechanism, the feature map is element-wise multiplied with the original input A, resulting in the feature map F. This

approach applies channel weights to the original features, intensifying the focus on significant features. Subsequently, the feature map F is directed to a spatial attention module, concentrating on the crack regions within the image. The feature map, post spatial attention processing, is element-wise multiplied with the input F, further enhancing spatial focus. Ultimately, the input A, after undergoing channel and spatial attention weighting, transforms into the output C. This output contains a richer and more detailed representation of crack features, crucial for the following steps of crack segmentation. The model is capable of more precise crack detection and segmentation, particularly in complex crack images. Notably, throughout this process, the size of the output features aligns with that of the input features, maintaining dimensions of 214×214. The computational method for the CBAM cross-domain attention mechanism is presented in formula 1.

$$C = CBAM(A) = SAM(CAM(A) \otimes A) \otimes (CAM(A) \otimes A) \tag{1}$$

Among them, both C and $CBAM(A)$ are the results of CBAM mixed domain attention mechanism operation, A is the original crack image, CAM is the channel domain attention operation mechanism for crack image processing, SAM is the spatial domain attention mechanism operation for crack image processing, $\otimes$ is the matrix multiplication operation, and the matrix multiplication operation formula is shown in formula 2.

$$F_{ik \times jl} = A_{i \times j} \bullet A\prime_{k \times l} \tag{2}$$

Among them,the symbol $\bullet$ signifies element-wise matrix multiplication, $A_{i \times j}$ is the matrix A whose pixel size is $i \times j$, $A\prime_{k \times l}$is the matrix$A\prime$ whose pixel size is $k \times l$, and $F_{ik \times jl}$ is the result matrix F whose pixel size is $ik \times jl$.

Channel Domain Attention Mechanism The Channel Attention Mechanism (CAM) enhances the capability for global feature extraction in crack images by evaluating the importance of each channel feature map. It achieves this by compressing and reducing dimensions in the spatial domain to obtain the channel attention map. The computational process is illustrated in Figure 3.

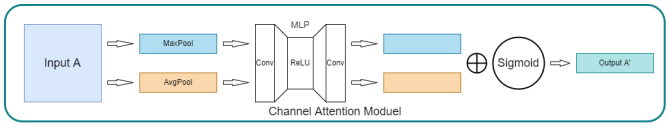


Fig. 3. CAM calculation flow chart

As can be seen from Figure 3, the original fissure image A is fed into the channel attention mechanism model, processed by the maximum pooled layer and the average pooled layer respectively to obtain two different feature graphs, which are then fed into the MLP neural network [25] for feature vector weight assignment. The whole process adopts the consistency principle, and finally obtains the global texture enhancement feature map through Sigmoid activation function [28]. The calculation is shown in formula 3.

$$A' = CAM\,(A) = Sigmoid\,(MLP\,(AvgPool\,(A)) + MLP\,(MaxPool\,(A))) \quad (3)$$

$Sigmoid$ represents Sigmoid activation function, $AvgPool$ represents average pooling operation, $MaxPool$ represents maximum pooling calculation, and MLP represents MLP neural network, which is used to extract attention map in CAM operation, essentially convolves two convolution layers with 1 convolution core with a ReLU activation layer [28], as shown in formula 4.

$$MLP\,(pool) = Conv(ReLU(Conv(pool)) \quad (4)$$

Where is $pool$ stands for pooled input, $Conv$ for convolution, and $ReLU$ for ReLU activation. The global texture feature image processed by channel attention is convolved with the original feature image to get the feature map F.

Spatial Attention Mechanism The spatial attention mechanism (SAM) enhances the ability of the model to extract local texture features by obtaining the importance of different regions in the feature map. SAM compresses the channel dimension of the feature graph to obtain the attention graph in the spatial domain. The calculation flow is shown in Figure 4. As shown in Figure 4, the maximum and average pooled layers are still used

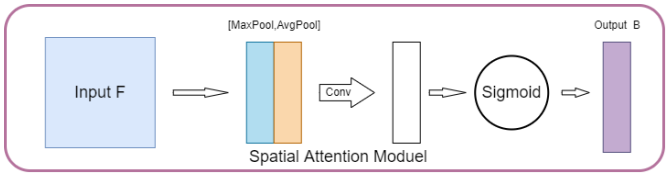


Fig. 4. SAM calculation process

in the input of Feature Diagram F into the spatial attention mechanism, resulting in two features with the same size and number of channels. Then, the two feature images are stitched together and processed by Sigmoid activation function after convolution operation of 7×7 to obtain the local texture enhanced crack image. As shown in formula 5.

$$B = SAM\,(F) = Sigmoid\left(Conv^{7\times7}\left([AvgPool\,(F) + MaxPool\,(F)]\right)\right) \quad (5)$$

Of which, $Conv^{7\times7}$ represents the convolution operation with a convolution nucleus size of 7×7.

3.2. Crack Image Feature Extraction

The study uses the VGG_16 [23] deep neural network as a base model, improves it, and proposes a multiscale null convolution VGG_L neural network for crack image feature

extraction. The network retains the first 10 convolutional layers and 4 pooling layers of VGG_16, and connects 1 layer of CBAM mixed-domain attention mechanism and 1 layer of dual-scale null convolution. The architecture lightens the VGG_16 model while enhancing the feature focusing ability of the network, the contextual information capturing ability, which can better extract the crack detail information and improve the recognition accuracy. The architecture is shown in Figure 5.

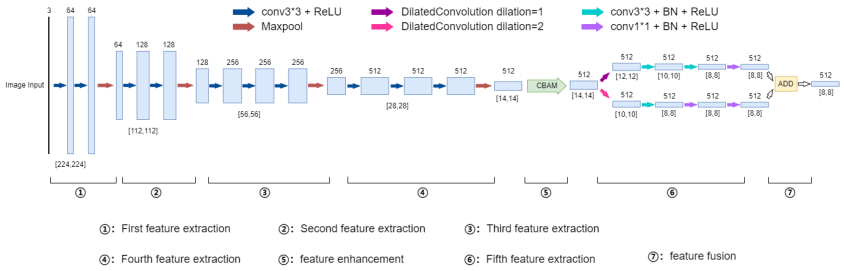


Fig. 5. VGG_L network architecture for multiscale void convolution

As shown in Figure 5, the network sequentially processes the input image through convolutional layers, max pooling layers, and dilated convolutional layers. The mid-section of the network includes convolutional layers with Batch Normalization (BN) and the ReLU activation function, along with an attention module (CBAM), for further refinement of features. After a series of convolutional and dilated convolutional layer processing, features are ultimately fused through an ADD operation, resulting in the generation of the output feature map. In the figure, the content within brackets [] indicates the size of the feature map, numbers denote the quantity of feature layers, and arrows of different colors represent various operations within the network.

The study inputs images of 214×214 dimensions into the VGG_L model, where they first undergo four feature extractions and one feature enhancement. The feature map, post-enhancement, is then subjected to a fifth feature extraction and one feature fusion. The first and second feature extraction processes are identical, involving two rounds of conv3*3+ReLU operations followed by one Maxpool operation. The third and fourth feature extractions are similar, consisting of three conv3*3+ReLU operations and one Maxpool operation each. The fifth feature extraction involves dual-scale dilated convolution, where the feature map enhanced by CBAM undergoes convolution operations at dilation rates of 1 and 2 simultaneously. After the dilation rate 1 convolution, the feature map undergoes two conv3*3+BN+ReLU operations, followed by one conv1*1+BN+ReLU operation to generate the feature map. Conversely, the feature map post dilation rate 2 convolution first undergoes one conv3*3+BN+ReLU operation, then two conv1*1+BN+ReLU operations to generate its feature map.

For feature fusion, the two feature maps obtained from the fifth feature extraction, being of equal size, are fused by adding corresponding width and height dimensions. This results in the final feature map from the feature extraction process.

The first feature extraction is the same as the second feature extraction. After two convolution operations with a 3×3 kernel and padding of 1, followed by a ReLU activation operation [28], the convolution kernel is pooled with a 3×3 size, a stride of 2, and a kernel of 2. The computation is as shown in formula 6.

$$C_i = MaxPool(ReLU\ \ (Conv\,(ReLU\,(Conv\,(C_{i-1}))))\quad i \in (1,2) \tag{6}$$

Among them, C_{i-1}represents the input fracture data set image, i represents the number of feature extraction processing times, C_i represents the feature extraction result, and when $i\ =\ 1$, C_0 is C, represents the fracture image after CBAM feature enhancement.

The third and the fourth feature extraction processes are the same, consisting of three layers of convolution core size is 3×3, padding 1 and ReLU activation, and the feature extraction results are obtained by the maximum pool size is 3×3, step is 2 and kernel is 2, as shown in formula 7.

$$C_i = MaxPool(ReLU\ (Conv(ReLU(Conv\,(ReLU\,(Conv\,(C_{i-1})))\ \ (i \in (3,4)) \tag{7}$$

After processing, the feature extraction results of crack map were obtained C_1, C_2, C_3 and C_4. The fourth feature extraction result C_4 is further enhanced by the CBAM mixed domain attention mechanism module to solve the problem of fine-grained information loss.

Dilated convolutions [20] keeps the image size constant while enhancing the receptive field of the convolution neural network. Double-scale dilated convolutions is used to achieve the fifth feature extraction to reduce the loss of context information and improve the accuracy of crack recognition. The processing involves a convolution operation with a void ratio of 1 and a void ratio of 2 and three ordinary convolution operations, as shown in Formula 8.

$$DC_\gamma(x,y,n,\gamma) = Conv(Conv(Conv(DConv\,(x,y,n,\gamma)))) \tag{8}$$

Among them, $DC_\gamma(x,y,n,\gamma)$represents the improved multi-scale convolution function, γ represents the void ratio, determines the sampling interval of the convolution kernel, x and y represent the length and width of the input crack image respectively, and n represents the edge length of the convolution core. $DConv$ represents the convolution operation of holes, the calculation is shown in formula 9.

$$DConv\,(x,y,n,\gamma) = \sum_{n}^{x}\sum_{n}^{y} I\,[x+\gamma n, y+\gamma n]\,w\,[n,n]) \tag{9}$$

Among them, $DConv$ represents the dilated convolutions operation, w is the void filter, I is the input image, and γ-1 represents the number of holes.

Dilated convolutions enhances the effect of image extraction by enlarging the receptive field of the convolution process without any additional parameters. The receptive field computation is shown in formula 10.

$$G = (m+1)(m-1)+m \tag{10}$$

Among them, m represents convolution nucleus size, G represents receptive field size.

In this paper, the dilated convolutions conv+BN+ReLU is composed of convolution core = 3×3, padding = 1, stride = 2. The image is processed by convolution with voidage 1 and voidage 2, respectively, and two features are extracted after three ordinary convolution operations. The convolution layer is composed of conv+BN+ReLU. Among them, the void ratio is 1, and the convolution parameter of the first two convolution treatments is 3×3, padding = 0, stride = 1. The convolution parameter of the third convolution is convolution kernel =1×1, padding = 0, stride = 1. When the void ratio is 1, the first convolution parameter is convolution kernel =3×3, padding = 0, stride = 1. The convolution parameters of the second and third convolution are convolution kernel =1×1, padding = 0, stride = 1. Because the sizes of the two feature maps are the same, the weighted operation is used to fuse the two feature maps. The result is the fifth feature extraction graph. , as described in formula 11.

$$MDC = \sum_{r=1}^{2} DC_{\gamma} \tag{11}$$

Among them, MDC is the fused crack feature extraction result.

3.3. Image Segmentation and Dimension Reduction

In order to reduce the channel dimension and improve the resolution of the feature map, the fused deconvolution and skip connection methods are used for five deconvolution operations [7] [8]. In deconvolution, which consists of ConvTanspose+BN+ReLU, the deconvolution layer parameter is set to convolution kernel =3×3, padding = 1, stride = 2, out_padding = 1. The output size of the deconvolution is calculated as Formula 12.

$$o = s\,(i-1) + 2p - k + 2 \tag{12}$$

Here, i represents the input size, k the convolution size, p the boundary extension, s the convolution size, and o is the output size.

The feature fusion feature map MDC is subjected to the deconvolution process. The processed feature map is skip connect with the first pooled processed feature map ,to achieve feature fusion. Then skip connect the output feature map with the second pooled feature map to realize feature fusion. According to this idea, sequential deconvolution processing and skip connect with the first four pooling results C_1,C_2,C_3,$C_4$1 to complete the image feature fusion. Because the dimension and channel dimension of the deconvolution feature map are the same as those of feature extraction. The method of wide-high correspondence addition is used to fuse the image features, and the MDC of fracture extraction is obtained. Although MDC has the same resolution as the input image, the channel dimension is much higher than the input image. Convolution reduction is used to reduce the high dimensional feature map to 3 dimensions. The number of convolution kernels is equal to the number of low-dimensional images, that is, the width and height of convolution kernels are 1. Thus obtains the final image segmentation result Y.

3.4. Merging MSELoss and Adam Parameter Optimization

SGD gradient descent learning rate optimization algorithm [19] has a slow convergence rate and local optimal solution. The model learning process was optimized using a fusion

of MSELoss and Adam methods. First, the loss value is calculated using MSELoss, and the loss function is used to calculate the mean square deviation, as shown in Formula 13.

$$MSELoss\,(x,y) = (x-y)^2 \tag{13}$$

Where x is the forecast and y is the label. During the model training process, MSELoss loss values are calculated every iteration and the results are iterated using the Adam adaptive learning rate algorithm [1] [26]. The handling steps are as follows:

1. Initialization. The initial chemical accessibility is θ ,the step length is α ,and the initialization matrix estimates the exponential rate of decay as β_1,β_2. β_1,β_2 as a value of [0,1]. Initialize the numeric constant δ, initialize the number of updates $t = 0$, and compute the MSELoss.

2. Gradient calculation. MSELoss is gradient computed to get the current loss gradient, as shown in Formula 14.

$$g = \nabla_{MSELoss} \tag{14}$$

Where g represents the MSELoss gradient, ∇ is the gradient calculation symbol.

3. Calculation of biased moment estimation. The second moment estimate is obtained by using the partial first moment estimate to calculate the partial first moment estimate of the gradient as shown in formula 15.

$$s = \beta_1 s + (1-\beta_1)\, g \tag{15}$$

Where s represents the partial first moment estimate of the gradient, and β_1 represents the exponential decay rate of the gradient partial first moment estimate.The partial first moment estimation of the gradient is introduced into the formula of partial second moment estimation to correct the influence factor. The calculation is shown in formula 16.

$$r = \beta_2 r + (1-\beta_2)\, g \odot g \tag{16}$$

Among them, r represents gradient second moment estimation, β_2 represents gradient second moment estimation exponential decay rate.

4. Deviation correction. The first moment estimation is modified to obtain the deviation coefficient. The calculation is shown in formula 17.

$$\hat{r} = \frac{r}{1-\beta_2^t} \tag{17}$$

The $\hat{s}$ represents the correction of the partial first moment deviation of the gradient, and t represents the number of iterations of the learning rate.The second-order deflection coefficients are obtained by revising the second-order moment estimation. The calculation is shown in formula 18.

$$\hat{r} = \frac{r}{1-\beta_2^t} \tag{18}$$

Among them, $\hat{r}$ represents the correction of deviation of gradient partial second moment.

5. Update learning rates. The first-order moment estimation correction coefficient and second-order moment estimation correction coefficient are incorporated into the learning

rate updating formula to calculate and update the learning rate.The calculation is shown in formula 19.

$$\theta_i = \theta_{i-1} - \alpha \frac{\hat{s}}{\sqrt{\hat{r}} + \delta} \tag{19}$$

Where θ_i is the updated learning rate, θ_{i-1} is the previous learning rate. After processing, the corresponding learning rate is obtained, which is used for the next training, and the whole process is repeated until the end of the training.

4. Experimental Results and Discussion

Research using AMD EPYC 7302 16-Core Processor CPU, NVIDIA GeForce RTX 3090GPU computing power resources. The development uses the Pytorch deep learning framework and the Numpy framework.

4.1. Data Source

The study used self-collected data, CFD datasets [21] [3],CRACK500 [32] and GAPs384 [5] datasets, a total of 5800, including different lighting, different scenes. In order to reduce the difference caused by the external environment, the scale normalization is carried out. Labelme is used to annotate the data, and the annotation information includes background and crack. Because of the deficiency of the public dataset and the self-mining dataset, the existing dataset is expanded by image augmentation. By means of rotation, mirroring, salt and pepper noise, Gaussian noise, random noise, etc. [22], the original data set is enlarged to 12800 images, and all image sizes are set to 224×224pixel.

Salt and pepper noise and random noise were added to the data augmentation, respectively 1.5% and 1%. Random Gaussian noise is used for noise amplification, as shown in Formula 20.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{20}$$

Among them, μ represents mean, σ represents standard deviation, $\mu = 0, \sigma = 0.05$. After processing, the experimental dataset is obtained. The dataset sample is shown in Figure 6.

Figure 6shows that A is the original figure, B is the random noise image, C is the salt and pepper noise image, and D is the Gaussian noise image.

4.2. Analysis of Results

Loss rate analysis of different algorithms The training parameters of UNet networks [35], UNet++ networks [33], FCN8s networks [15], CRF-CNN [13], Multi-Transformer [24] are shown in Table 1.

Based on the parameters set in Table 1, the model training of UNet networks [35], UNet++ networks [33], FCN8s networks [15], CRF-CNN [13],Multi-Transformer [24],

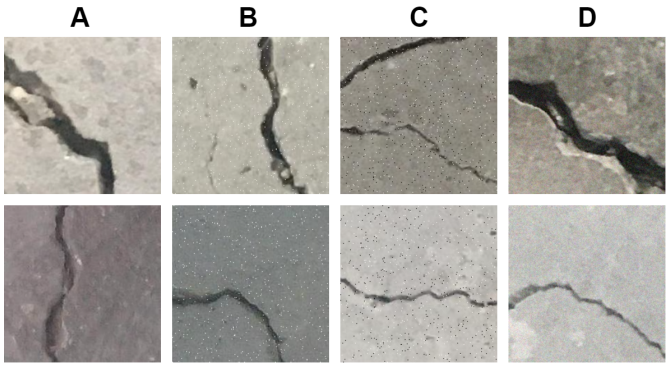


Fig. 6. Partial data set

Table 1. Training Parameters Configuration Table

Parameter Name	Value
InputSize	[3,224,224]
OutputSize	[3,224,224]
DataSize	12800
BatchSize	2
Epoch	16
LearningRate	10^{-6}
Max iteration	102400

and the training loss process of several algorithms is analyzed. The loss curve is shown in Figure 7.

As Figure 7 shows,the image illustrates the diminishing trend of iteration losses during the training process for six models: AAMCNet, UNet, Multi-Transformer, U-Net++, FCN8s, and CRF-CNN. It uses the horizontal axis to denote the number of iterations and the vertical axis to represent the loss rate. This effectively showcases each model's efficiency in reducing losses over successive training iterations. In the training process of the six models, with the increase of the number of iterations, the loss value decreases rapidly at first, and then decreases gradually, and finally approaches convergence. The initial loss of AAMC-Net model is the lowest, only 0.29. This is due to the use of CBAM mixed domain attention mechanism twice, which can more effectively focus the important features of crack image and enhance the network feature extraction ability. In addition, the AAMC-Net model converges nearly 10000 iterations, and AAMC-Net converges fastest. This is due to the use of lightweight VGG_L networks and the use of Adma's MSELoss adaptive function in model training, which adaptively adjusts the learning rate and helps the model find the minimum value of the loss function faster.

Subjective Evaluation Analysis In the same experimental environment, the image was segmented using improved Sobel [31], adaptive threshold algorithm [11], UNet [35], UNet++ [33], FCN8s [15], CRF-CNN [13],Multi-Transformer [24] and AAMC-Net network respectively. The segmentation effect was shown in Figure 8 and Figure 9.

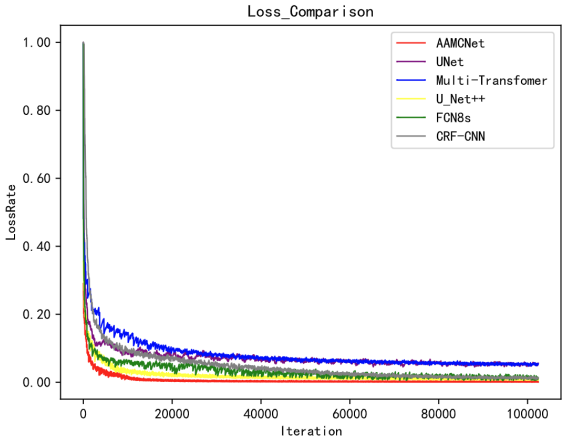
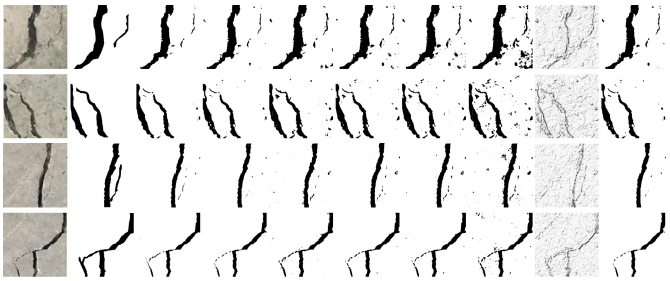


Fig. 7. Loss ratio chart

An analysis of algorithm performance was conducted using subjective human evaluation. As shown in Figures 8 and Figures 9, two digital image processing methods outperform the CFD dataset in image segmentation. The CFD dataset contains numerous road surface textures. However, the improved Sobel operator and adaptive thresholding are highly sensitive, effortlessly capturing subtle variations within the images. In road crack detection, identifying road surface textures as granular cracks in the CFD dataset adversely affects the effectiveness of image segmentation.

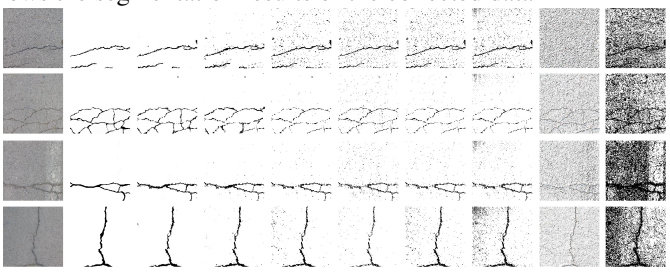
Within the neural network models, particularly in the segmentation results of UNet, stemming from the CFD dataset, significant noise is present, which to some extent hampers the segmentation performance. Even though the upgraded version, UNet++, demonstrates clear crack segmentation results, it struggles in handling image details. This is attributed to UNet++ filtering redundant information during model training, inadvertently filtering out crack details. Consequently, UNet++ performs poorly in applications involving small crack details. While FCN8s progressively integrates Pool3 and Pool4 layers, thereby effectively retaining some crack details, it still retains a considerable amount of noise. Although CRF-CNN and Multi-Transformer excel in segmenting image details, residual noise still impacts the segmentation performance.

In contrast, AAMC-Net effectively mitigates noise interference and simultaneously captures intricate crack segmentation details. This achievement is owed to the integration of CBAM during preprocessing and feature extraction stages, enabling the capture and enhancement of both global and local information. Moreover, the fusion of CBAM and multi-scale atrous convolution modules for feature extraction facilitates multi-layer perception of crack features at varying scales, effectively mitigating information loss resulting from pooling. Additionally, this approach mitigates interference from non-crack information during crack image segmentation processing.



1.Original drawing 2.Label 3. Improved algorithm 4.UNet++ 5.Muti-Transfomer 6.CRF-CNN 7.FCN8s 8.UNet 9.Improved Sobel operator 10.Adaptive threshold

Fig. 8. shows the segmentation results of the collected data



1.Original drawing 2.Label 3. Improved algorithm 4.UNet++ 5.Muti-Transfomer 6.CRF-CNN 7.FCN8s 8.UNet 9.Improved Sobel operator 10.Adaptive threshold

Fig. 9. shows the segmentation results of the CFD data

4.3. Objective Performance Analysis

The study used improved Sobel [31], adaptive threshold algorithm [11], UNet [35], UNet++ [33], FCN8s [15], CRF-CNN [13],Multi-Transformer [24] and AAMC-Net network to segment 10 road crack images randomly selected in the validation concentration. After processing, the processing images shall be evaluated by using cross and merge ratio, accuracy ratio (Acc), precision ratio, recall ratio and F1 value [36], the processed average results will be retained to 3 decimal places, as shown in Table 2.

From the table 3, we can see that among the eight comparison algorithm models, the combination ratio, accuracy ratio, precision ratio, recall ratio and F1 value of the improved algorithm are the best, 96.310%, 96.379%, 99.550% and 97.008% respectively. Multitransformer takes second place, and the two digital image processing algorithms are the worst. The worst of the six models is the Unet network. Compared with the five network models, the improved algorithm has an average increase of 2.650%, 1.860%, 0.626%, 3.092% and 1.731% respectively.

The AAMC-Net model exhibits superior performance in image crack segmentation, attributed to the utilization of dual CBAM processes that enhance the crack feature maps through image strengthening, thereby augmenting the network's feature capturing capabilities. Additionally, the model benefits from dual-scale atrous convolutions for feature extraction, where atrous convolutions expand the image's receptive field. The dual-scale atrous convolutions effectively capture crack feature information at different scales, en-

Table 2. Training Parameters Configuration Table

Algorithms	IOU	Acc	Precision	Recall	F1
AAMC-Net	96.310	96.321	96.379	99.550	97.008
UNet	90.118	91.205	96.134	92.668	94.112
FCN8s	94.014	94.623	95.501	95.486	94.418
UNet++	94.808	95.812	95.806	97.912	95.542
CRF-CNN	95.163	95.513	95.328	98.116	96.231
Multi-Transformer	95.217	95.826	96.134	98.915	96.532
Adaptive Threshold	93.450	94.685	95.215	98.695	83.115
Improved Sobel	86.650	86.693	86.690	97.195	70.126

compassing local details and global structures, while maintaining robustness against various interferences. This enhancement amplifies the model's expressive capacity. Through analysis of five objective evaluation metrics, the algorithm's proficiency in crack detail handling and segmentation performance is once again affirmed.

4.4. Reasoning Speed

In order to verify the real-time detection performance of the algorithm, the real-time performance of road crack processing is evaluated by using a single image crack segmentation time, the calculation is shown in formula 21.

$$T = \frac{1}{h}\sum_{i=1}^{h} Q_h \tag{21}$$

Among them, T represents the average time of single image segmentation, h represents the number of images processed in batches, and Q_h represents the time used to segment the nth image.

The study used UNet [35], UNet++ [33], FCN8s [15], CRF-CNN [13], Multi-Transformer [24] and AAMC-Net network to train in the same experimental environment. The training time of six network models is analyzed and the average segmentation time of single crack image after training is compared. The result is 3 decimal places, as shown in Table 3.

Table 3. Training Parameters Configuration Table

Model	AAMC-Net	UNet	FCN8s	UNet++	CRF-CNN	Multiscale CRF
Average split time(s)	0.240	0.242	0.256	0.276	0.332	0.344
Training time(min)	241	245	278	306	332	345

As can be seen from Table 4, among the six algorithms, AAMC-Net has the shortest model training time and average detection time per image, followed by UNet, while Multi-Transformer takes the most time. When analyzed in conjunction with Table 3, although the training and average detection time of the UNet model are not significantly

different from those of AAMC-Net, the average accuracy rate of crack segmentation increased by 5.609%. Compared to the most effective Multi-Transformer, the average accuracy of crack segmentation increased by 0.517%, but the time for crack detection and model training decreased by 30.233% and 30.145% respectively. In real-time scenarios like road monitoring or instant image analysis, the response speed and accuracy of algorithms are paramount. The efficiency advantage of the improved algorithm is confirmed through actual model training and random crack image segmentation detection. The time advantage of the AAMC-Net architecture stems from reducing convolution and pooling layers in the VGG_16 model and integrating CBAM to optimize the output of traditional convolutions.

By decreasing the number of convolution and pooling layers in the VGG_16 model, AAMC-Net significantly reduces computational complexity, leading to a marked decrease in model training time. This is especially valuable in real-time scenarios, enabling quicker retraining of the model to adapt to environmental changes like weather or lighting conditions.

The integration of CBAM in the algorithm optimizes feature extraction, focusing more on key image areas and enhancing crack detection capabilities. The inclusion of this attention mechanism heightens sensitivity to crack features, maintaining high accuracy even in complex backgrounds or various crack types.

AAMC-Net, using multi-scale dilated convolutions, captures a broader context without adding extra parameters. This means comprehensive image content understanding without increasing computational burden, crucial for analyzing images at different resolutions and aiding real-time systems to adapt to diverse input conditions.

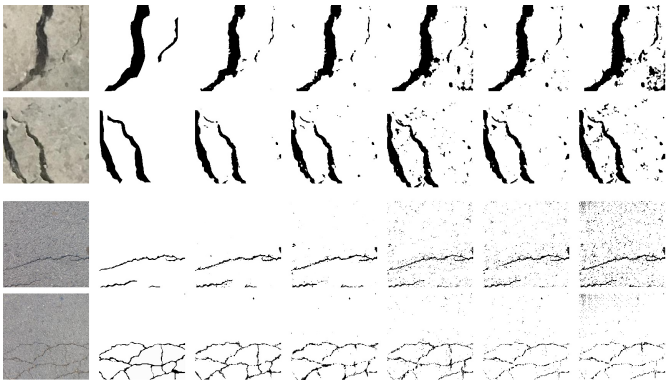
The algorithm also simplifies the training process and speeds up model convergence by automatically adjusting the learning rate with an adaptive MSELoss learning function. In real-time applications, this translates to rapid transfer training across different task scenarios, quicker convergence, and reduced model training time.

In experiments, AAMC-Net, compared with other advanced algorithms, displayed shorter average detection times and higher accuracy in crack segmentation. These results demonstrate its potential in real-time applications, as it quickly identifies and locates cracks while maintaining high accuracy. This is critical for applications requiring immediate response, such as monitoring road conditions in intelligent traffic systems or rapidly identifying and assessing road damage in disaster management systems for swift deployment of repair work and emergency measures.

4.5. Ablation Experiment

In order to study the effectiveness of the algorithm, the image segmentation of VGG_16, VGG_10, VGG_L, VGG_10 + CBAM and the improved algorithm model were studied under the same experimental environment. The segmentation results are shown in Figure 10, and the ablation analysis results are shown in Table 4.

The results of ablation experiments in Table 4 show that the average segmentation time of IOU, ACC, Precision, Recall and fracture images of AAMC-Net is the best. Compared with the ablation experiment, the VGG_10 fused with CBAM is better than VGG_10. The results show that CBAM can enhance the ability of image feature extraction and improve the performance of image segmentation. VGG_10 image fused with CBAM attention mechanism module has better performance of crack segmentation and



1.Original image 2.Label Image 3.AAMC-Net Algorithm 4.VGG_L 5.VGG_16 6.CBAM+VGG_10 7.VGG_10

Fig. 10. Ablation experimental segmentation

Table 4. Training Parameters Configuration Table

Algorithms	IOU	Acc	Precision	Recall	F1	Average split time(s)	Training time(min)
VGG_16	94.221	94.332	94.618	96.126	95.012	0.271	298
VGG_10	89.112	88.604	88.482	88.909	87.826	0.226	221
VGG_L	94.815	95.321	95.212	98.132	96.810	0.235	236
VGG_10+CBAM	94.151	94.214	94.371	95.825	93.712	0.229	230
AAMC-Net	96.25	96.351	96.382	99.532	97.008	0.240	241

less interference in image segmentation. The training time of the model and the average segmentation time of single image increase little. The results showed that multi-scale cavity convolution could enhance the receptive field in convolution process. The aim is to enhance the ability of different size feature extraction and improve the performance of image segmentation. Compared AAMC-Net with VGG_16, the performance of AAMC-Net is better than VGG_16. The model training time and detection time are better than VGG_16. Ablation experiments show that AAMC-Net is robust and optimized based on VGG_16.

5. Conclusion

In this paper, we have introduced the AAMC-Net convolutional neural network algorithm, which leverages the mixed domain attention mechanism to enhance feature extraction from the feature map. Through the integration of MSE loss and Adam during global learning and iterative training, the network model achieves improved performance for road crack segmentation and demonstrates its applicability in various image segmentation domains.

Our algorithm exhibits notable strengths in detail processing, model training, and real-time detection on single images. It successfully addresses the challenge of road crack segmentation with promising results. Furthermore, the flexibility of its application to other image segmentation tasks highlights its potential in broader contexts.

However, it is essential to acknowledge that the algorithm demands substantial computing resources, necessitating further optimization efforts. Additionally, the algorithm can only directly process images that conform to the input dimensions, which may not meet the general requirements of real-time applications. It's worth noting that there is still room for improvement in terms of multi-object real-time segmentation, in order to enhance segmentation efficiency. In the future, we plan to concentrate on optimizing the algorithm to enhance its computational performance, explore a universal detection architecture for multi-scale images, all while maintaining its effectiveness.

In conclusion, the AAMC-Net algorithm showcases significant advancements in image crack segmentation detection. As we continue to refine and optimize this approach, we anticipate its continued relevance and applicability to various image segmentation tasks, contributing to the advancement of the field.

References

1. Adam, K.D.B.J.: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 1412 (2014)
2. Cajas, Y.R.A., Guisado, Y.Z., Vergaray, A.D.: Identify faults in road structure zones with deep learning. Journal of System and Management Sciences 12(6), 163–191 (2022)
3. Cui, L., Qi, Z., Chen, Z., Meng, F., Shi, Y.: Pavement distress detection using random decision forests. In: International Conference on Data Science. pp. 95–102. Springer (2015)
4. Dung, C.V., et al.: Autonomous concrete crack detection using deep fully convolutional neural network. Automation in Construction 99, 52–58 (2019)
5. Eisenbach, M., Stricker, R., Seichter, D., Amende, K., Debes, K., Sesselmann, M., Ebersbach, D., Stoeckert, U., Gross, H.M.: How to get pavement distress detection ready for deep learning? a systematic approach (2017)
6. Fan, R., Bocus, M.J., Zhu, Y., Jiao, J., Wang, L., Ma, F., Cheng, S., Liu, M.: Road crack detection using deep convolutional neural network and adaptive thresholding. In: 2019 IEEE Intelligent Vehicles Symposium (IV). pp. 474–479. IEEE (2019)
7. Gao, H., Yuan, H., Wang, Z., Ji, S.: Pixel transposed convolutional networks. IEEE transactions on pattern analysis and machine intelligence 42(5), 1218–1227 (2019)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Hu, G.X., Hu, B.L., Yang, Z., Huang, L., Li, P.: Pavement crack detection method based on deep learning models. Wireless Communications and Mobile Computing 2021, 1–13 (2021)
10. Islam, M.M., Kim, J.M.: Vision-based autonomous crack detection of concrete structures using a fully convolutional encoder–decoder network. Sensors 19(19), 4251 (2019)
11. Jiang, X., Yang, X., Ding, X.: An efficient and reliable approach based on adaptive threshold for road defect detection. International Journal of Innovative Computing and Applications 12(5-6), 321–329 (2021)
12. Kheradmandi, N., Mehranfar, V.: A critical review and comparative study on image segmentation-based techniques for pavement crack detection. Construction and Building Materials 321, 126162 (2022)
13. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3194–3203 (2016)
14. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Exploring context with deep structured models for semantic segmentation. IEEE transactions on pattern analysis and machine intelligence 40(6), 1352–1366 (2017)

15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
16. Mohan, A., Poobal, S.: Crack detection using image processing: A critical review and analysis. alexandria engineering journal 57(2), 787–798 (2018)
17. Munawar, H.S., Hammad, A.W., Haddad, A., Soares, C.A.P., Waller, S.T.: Image-based crack detection methods: A review. Infrastructures 6(8), 115 (2021)
18. Niu, Z., Zhong, G., Yu, H.: A review on the attention mechanism of deep learning. Neurocomputing 452, 48–62 (2021)
19. Ruder, S.: An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747 (2016)
20. Schmidt, C., Athar, A., Mahadevan, S., Leibe, B.: D2conv3d: Dynamic dilated convolutions for object segmentation in videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1200–1209 (2022)
21. Shi, Y., Cui, L., Qi, Z., Meng, F., Chen, Z.: Automatic road crack detection using random structured forests. IEEE Transactions on Intelligent Transportation Systems 17(12), 3434–3445 (2016)
22. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data 6(1), 1–48 (2019)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
24. Sun, Z., Zhou, W., Ding, C., Xia, M.: Multi-resolution transformer network for building and road segmentation of remote sensing image. ISPRS International Journal of Geo-Information 11(3), 165 (2022)
25. Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J.: Mlp-mixer: An all-mlp architecture for vision. Advances in neural information processing systems 34, 24261–24272 (2021)
26. Tong, Q., Liang, G., Bi, J.: Calibrating the adaptive learning rate to improve convergence of adam. Neurocomputing 481, 333–356 (2022)
27. Tran, V.P., Tran, T.S., Lee, H.J., Kim, K.D., Baek, J., Nguyen, T.T.: One stage detector (retinanet)-based crack detection for asphalt pavements considering pavement distresses and surface objects. Journal of Civil Structural Health Monitoring 11, 205–222 (2021)
28. Villmann, T., Ravichandran, J., Villmann, A., Nebel, D., Kaden, M.: Activation functions for generalized learning vector quantization-a performance comparison. arXiv preprint arXiv:1901.05995 (2019)
29. Wang, W., Su, C.: Convolutional neural network-based pavement crack segmentation using pyramid attention network. IEEE Access 8, 206548–206558 (2020)
30. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
31. Xiu, C., Yin, H., Liu, Y.: Image segmentation of cv model combined with sobel operator. In: The 32nd China Control and Decision Making Conference. pp. 124–128 (2020)
32. Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X., Ling, H.: Feature pyramid and hierarchical boosting network for pavement crack detection. IEEE Transactions on Intelligent Transportation Systems 21(4), 1525–1535 (2019)
33. Yang, Q., Ji, X.: Automatic pixel-level crack detection for civil infrastructure using unet++ and deep transfer learning. IEEE Sensors Journal 21(17), 19165–19175 (2021)
34. Yang, X., Li, H., Yu, Y., Luo, X., Huang, T., Yang, X.: Automatic pixel-level crack detection and measurement using fully convolutional network. Computer-Aided Civil and Infrastructure Engineering 33(12), 1090–1109 (2018)
35. Zhang, L., Shen, J., Zhu, B.: A research on an improved unet-based concrete crack detection algorithm. Structural Health Monitoring 20(4), 1864–1879 (2021)

36. Zhao, F., Dong, B., Pan, H., Anqi, S.: A mining algorithm to improve lstm for predicting customer churn in railway freight traffic. Studies in Informatics and Control 32, 25–38 (2023)

Wang Xiaofang is with Geely University of China, located in Eastern New District, Chengdu, Sichuan Province (641423). Research interests include computer vision and data mining.

Liu Chenfang is with Chengdu College of University of Electronic Science and Technology of China, located in Gaoxin West District, Chengdu, Sichuan Province (611731). Research interests include convolutional neural network.

Hou Junniang is with Geely University of China, located in Eastern New District, Chengdu, Sichuan Province (641423).Research interests include data analysis and mining.

Zhou Liang is with Geely University of China, located in Eastern New District, Chengdu, Sichuan Province (641423).Research interests include image processing.

Received: July 25, 2023; Accepted: February 07, 2024.