scientific reports

Check for updates

OPEN

An efficient semantic segmentation method for road crack based on EGA-UNet

Li Yang[1,2]✉, Jingwei Deng[1], Hailong Duan[1,2] & Chenchen Yang[1]

Road cracks affect traffic safety. High-precision and real-time segmentation of cracks presents a challenging topic due to intricate backgrounds and complex topological configurations of road cracks. To address these issues, a road crack segmentation method named EGA-UNet is proposed to handle cracks of various sizes with complex backgrounds, based on efficient lightweight convolutional blocks. The network adopts an encoder-decoder structure and mainly consists of efficient lightweight convolutional modules with attention mechanisms, enabling rapid focusing on cracks. Furthermore, by introducing RepViT, the model's expressive ability is enhanced, enabling it to learn more complex feature representations. This is particularly important for dealing with diverse crack patterns and shape variations. Additionally, an efficient global token fusion operator based on Adaptive Fourier Filter is utilized as the token mixer, which not only makes the model lightweight but also better captures crack features. Finally, to demonstrate the method's effectiveness and accuracy, we compare the proposed approach with some existing methods on three public datasets. Experimental results demonstrate that the proposed method outperforms existing approaches in detecting cracks of diverse shapes and sizes within complex backgrounds, satisfying the requirements for both high precision and real-time segmentation.

Keywords Road defect, U-Net, Semantic segmentation

Cracks constitute the most prevalent road defects. As urbanization accelerates and infrastructure construction advances, the detection and segmentation of cracks exhibit extensive application value in road safety assessments and maintenance. Unrepaired cracks can pose severe threats to traffic safety[1]. The timely detection and repair of cracks represent a significant responsibility for transportation departments. Traditional crack detection methods rely heavily on manual inspections—a process that is cumbersome, labor-intensive, and inefficient, with accuracy further compromised by operator experience and subjective judgment. Consequently, there is an urgent need for high-precision and real-time segmentation methods. Segmenting cracks poses significant challenges due to their erratic intensity, erratic contrast, and densely packed backgrounds. Moreover, diverse road textures resembling cracks and their inherently low contrast complicate crack recognition. The advent of deep learning has facilitated the emergence of computer vision techniques, presenting researchers with novel solutions to these problems. Researchers have increasingly directed their focus on automated crack segmentation methods grounded in computer vision techniques.

Crack segmentation serves as a critical preprocessing step for quantitative structural health assessment, aiming to precisely extract crack morphologies from complex backgrounds. Crack segmentation methodologies are primarily categorized into traditional image processing methods[2-7] and deep learning methods[9-13]. Traditional frameworks typically employ edge detection algorithms, morphological operations, and threshold segmentation strategies. Ron et al.[2] introduced the concept of gravitational field intensity as a substitute for image gradient and proposed an adaptive threshold selection approach based on the mean and standard deviation of image gradient magnitude. Yan-Yin et al.[3] presented an image segmentation algorithm leveraging an adaptive weighted mathematical morphology edge detector. Cuevas et al.[4] introduced an image segmentation algorithm based on an automated threshold determination technique. These methods exhibit efficacy in handling images with simple backgrounds and distinct cracks but falter in images with complex backgrounds, blurred and microscale crack, and cannot fulfill real-time segmentation demands. Recently, advancements in deep learning technology, particularly the pervasive application of convolutional neural networks[8] in image processing, have significantly propelled crack segmentation methods based on deep learning[14-20]. Deep learning-based crack segmentation

[1]School of Automation and Electrical Engineering, Tianjin University of Technology and Education, Tianjin, China. [2]Tianjin Key Laboratory of Information Sensing and Intelligent Control, Tianjin University of Technology and Education, Tianjin, China. ✉email: yangli@tute.edu.cn

demonstrates superior accuracy and computational efficiency through three fundamental advantages. Firstly, deep learning models exhibit robust feature learning capabilities, enabling automatic learning of crack feature representations, thereby enhancing segmentation accuracy. Secondly, deep learning models possess excellent generalization ability, achieving effective segmentation across different scenarios and crack types. Finally, deep learning models can be trained and optimized end-to-end, simplifying the segmentation process and improving segmentation efficiency. Liu et al.[9] proposed a deep level of convolutional neural network (CNN) called DeepCrack which predicted pixel-wise crack segmentation in an end-to-end approach, achieving high performance. Choi et al.[10] proposed a semantic damage detection network (SDDNet), which was trained on a manually created crack dataset, achieving a high mean Intersection over Union (mIoU) on the test set and demonstrating a certain degree of real-time performance. Qu et al.[11] combined a novel multi-scale convolutional feature fusion module and proposed a crack detection method based on a deeply supervised convolutional neural network, which achieved better performance in terms of F1 score and mean IoU. These crack segmentation methods demonstrate satisfactory performance but struggle with accurately segmenting topological cracks and lack real-time segmentation capabilities. Real-time segmentation is crucial for enhancing efficiency and ensuring segmentation integrity. Real-time denotes an algorithm's ability to complete a task within a specified time frame without significantly increasing the number of parameters to improve accuracy. Our focus is on real-time segmentation methods that preserve semantic information during transmission, as this information tends to degrade over time. Employing lightweight convolutions, residual connections, and the Spatial Pyramid Pooling Fast (SPPF) module in deep network layers can address these challenges, enhancing multi-scale information representation and overcoming limitations of single spatial hierarchy in feature maps. This paper mainly has the following contributions:

1. An Efficient Ghost Sparse Convolution(GSConv) Block (EG-Block) is proposed, consisting of three lightweight convolutions with kernel sizes of 3×3, incorporating residual connections and an efficient multi-scale attention (EMA) module. This module enhances the extraction and transmission of crack features and reduces the loss of crack information.
2. The A-RepViT Block module is proposed, wherein the original token mixer is substituted by Adaptive Frequency Filtering (AFF). AFF facilitates the model in capturing crack features more effectively, enhances crack segmentation accuracy, and diminishes a certain quantity of parameters, thereby achieving an improved equilibrium between accuracy and efficiency.
3. The EGA-UNet network is proposed, which embeds the proposed Efficient GSConv Block into UNet and incorporates the SPPF module and A-RepViT Block into the deeper layers of the encoder. EGA-UNet improves the effectiveness of crack segmentation while being a lightweight network, meeting the requirements for real-time and high-precision crack segmentation.

The rest of the paper is organized as follows: Sect. 2 presents the related work; Sect. 3 presents the proposed algorithm and designed methodology; Sect. 4 explores the experimental results and discussion. Finally, conclusions are drawn in Sect. 5.

Related works

This section mainly explains the practical application of semantic segmentation technology in road crack detection, as well as the historical development of transformers and attention mechanisms and their practical application in crack detection.

Road crack detection based on semantic segmentation

Semantic segmentation of road cracks is predominantly implemented through deep convolutional neural networks (CNNs). In this paper, it is possible to learn the complex relationships between crack pixels in an image and generate a crack prediction map of the same size as the input image by training the network, ensuring that each pixel is assigned a specific category label. Leveraging the advantages of CNN in feature representation, various semantic segmentation techniques for crack image detection have emerged. Cheng et al.[22] proposed an automatic crack detection method based on the U-Net deep convolutional neural network, achieving excellent results. A supervised method[23] based on deep learning was proposed, which had the ability to handle different road surface conditions. However, the encoder of the traditional UNet network had relatively weak feature extraction capabilities and a large number of parameters. This had prompted efforts to enhance the encoder's feature extraction capabilities and to make the entire network more lightweight. Yu et al.[24] proposed a U-shaped encoder-decoder semantic segmentation network that combined UNet and ResNet for pixel-level road surface crack image segmentation, improved crack detection performance. Zhang et al.[25] proposed a customized deep learning model architecture named Efficient Crack Segmentation Network to accelerate real-time pavement crack detection and segmentation without compromising performance. Building upon the work of previous researchers, the algorithm proposed in this paper, named EGA-UNet, integrates efficient lightweight convolutional blocks to quickly and precisely segment topological cracks of varying shapes with complex backgrounds.

The application of transformers in semantic segmentation

Since Transformer architecture was introduced by Vaswani et al. in 2017[26], it was initially primarily used for natural language processing tasks. However, over time, the powerful capabilities of Transformer have also been applied to the field of computer vision, including object detection[27] and image segmentation[28]. Conventional convolutional neural networks (CNNs) predominantly utilize local receptive fields for image processing, whereas Transformers leverage self-attention mechanisms to capture global dependencies. This enables

effective integration of features across various levels, enhancing the model's ability to comprehend global image information and inter-regional relationships. At the same time, the self-attention mechanism of Transformers[29] also facilitates self-supervised learning, which is particularly useful in image segmentation since labeling a large amount of image data is both time-consuming and labor-intensive. Through self-supervised learning, the model can be trained with little to no labeled data. Therefore, scholars have begun applying the Transformers architecture to image segmentation. Gao Y et al.[30] proposed UTNet, which integrated the self-attention mechanism into convolutional neural networks to enhance medical image segmentation. Cheng B et al.[31] proposed a new architecture capable of solving any image segmentation task (panoptic, instance, or semantic), which achieved excellent results on public datasets like COCO. Xie E et al.[32] proposed a semantic segmentation framework called SegFormer, which combined transformers with a lightweight multi-layer perceptron (MLP) decoder, improving segmentation performance. While Transformers have demonstrated strong performance, they have also incurred significant computational costs and inference speed trade-offs. The advancement of Transformers has traditionally been credited to the attention-based mechanism for token mixing. Yu T et al.[33] completely replaced the attention-based module with Spatial MLPs as the token mixer and discovered that the derived MLP-like model could easily achieve competitive performance on image classification benchmarks. Yu W et al.[34] further abstracted the entire Transformer as a general architecture MetaFormer by viewing the attention module as a specific token mixer. They did not specify a token mixer, as they believe that MetaFormer is the key to achieving superior results in visual task models similar to Transformers and MLPs. Huang Z et al.[35] developed an efficient global token fusion operator named Adaptive Fourier Filter to replace the self-attention module in Transformers. To address the computational complexity of Transformers, Wang et al.[36] proposed a lightweight Transformer architecture with a slight increase in the number of parameters and inference speed, which accuracy has been improved. In the task of crack segmentation, Liu et al.[37] proposed a Transformer-based crack detection network called CrackFormer, and conducted experiments on three public datasets, all of which achieved good results. Building upon prior research, this study enhances the RepViT Block, introduces the A-RepViT Block, and incorporates it into the EGA-UNet al.gorithm. This integration aims to deepen image comprehension and enhance the precision of crack segmentation.

Attention mechanisms

In recent years, attention mechanisms have been widely applied in various fields of deep learning, including object detection[38,39], semantic segmentation[40,41], object tracking[42], and face recognition[43], among others. The core principle of the attention mechanism is to emulate human information processing by selectively prioritizing important information based on task requirements, while disregarding irrelevant information, rather than treating all information equally. Hu et al.[44] proposed the Squeeze-and-Excitation Network (SENet) to model the interdependencies between feature channels. Wang et al.[45] introduced an Efficient Channel Attention (ECA) module, overcoming the trade-off between performance and complexity. Xu et al.[46] presented an Efficient Local Attention (ELA) method that effectively utilized spatial information without reducing channel dimensions or increasing the complexity of the neural network, meeting the demands of various vision tasks. Ouyang et al.[47] proposed an Efficient Multi-Scale Attention(EMA) module that focused on retaining information of each channel while reducing computational overhead. Additionally, the method further aggregates the output features of two parallel branches through cross-dimensional interaction. The results show that EMA significantly outperforms several recent attention mechanisms without changing the network depth. In the task of crack detection, Qiao et al.[48] and Liang et al.[49] respectively proposed a pavement crack detection method using the attention mechanism, both having good performances. This illustrates the capability of the attention mechanism to integrate multi-scale feature data and identify distinctive features. Building upon prior work, the present study introduces the EG-block, incorporating an attention mechanism to facilitate rapid identification of cracks within the network.

Methods

In this section, the proposed EGA-UNet network is introduced firstly, and then the EG-block, A-RepViT Block, SPPF module and finally the loss function in detail.

The structure of EGA-UNet

EGA-UNet consists mainly of EG-block, A- RepViT Block and SPPF module. Input a road crack image into the segmentation model and divide all the pixels in the image into crack pixels and background pixels based on the corresponding probability values. Crack pixels have higher probability values, while background pixels have lower probability values. Therefore, crack segmentation can be seen as a pixel-by-pixel binary classification problem. Figure 1 shows the architecture of the EGA-UNet. Based on the encoder-decoder structure, the EGA-UNet network model is proposed. It utilizes a combination of residual connections and attention mechanism modules to form the EG-block, replacing the convolutional parts in the encoder and decoder to extract feature information from different layers. Integrating the SPPF module into the encoder's lower layers enables the network to better capture spatial hierarchies in images. This enhancement improves multi-scale crack feature perception while boosting flexibility and robustness. Following the SPPF module, the A-RepViT Block is included to help the model better capture crack features and handle multi-scale features. The decoder integrates multi-layer feature information in the form of feature pyramid to achieve accurate crack segmentation.

The structure of EG-Block

The EG-Block comprises three 3×3 lightweight convolutions and a single 1×1 standard convolution. It also incorporates a residual connection mechanism and an EMA module, as shown in the Fig. 2. This process can be represented as Eq. (1). Increasing the network depth can enhance crack recognition in models. The EG-Block structure employs three concatenated lightweight convolutional layers to extract features efficiently. Stacking

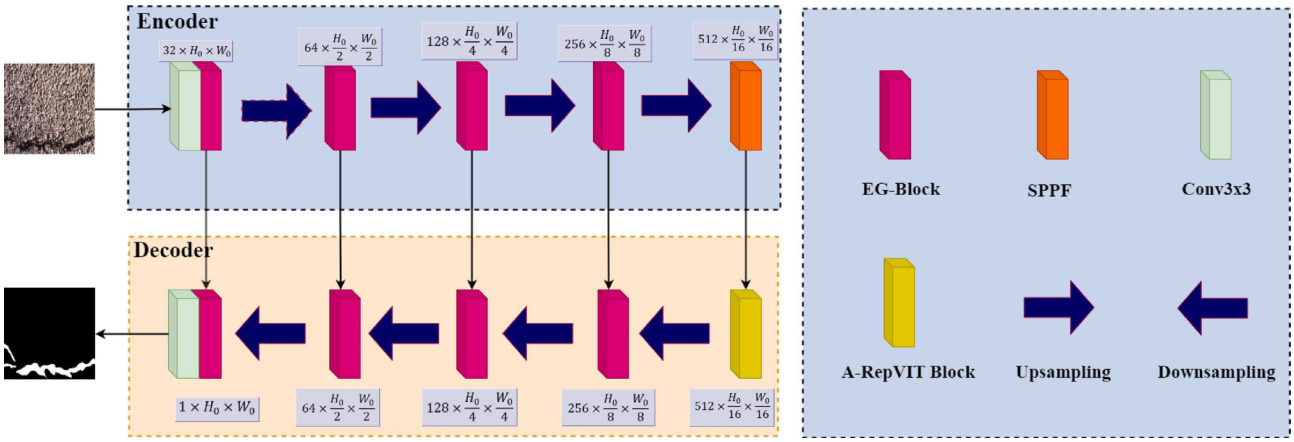


Fig. 1. The structure of EGA-UNet.

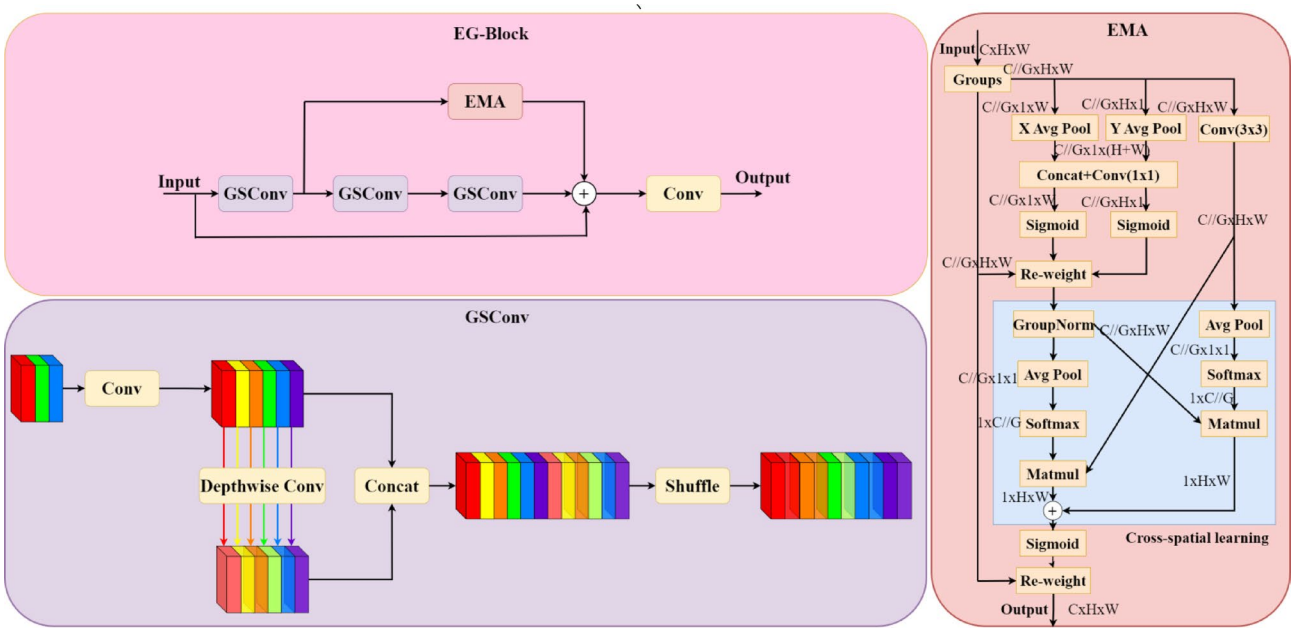


Fig. 2. The structure of Efficient GSConv Block(EG-Block).

convolutional layers augments the number of parameters, hence lightweight convolutions were utilized. Each convolutional layer captures distinct levels of features from input data, enabling the network to amalgamate simple features into complex representations. Moreover, stacking convolutional layers enhances the network's robustness against translations and minor deformations in the input data, thereby enhancing generalization. However, excessive convolutional layer stacking can degrade network performance. To address this issue, the residual method is employed after three lightweight convolutional layers, directly incorporating input features into the feature map. This approach facilitates feature reuse, minimizes information loss, and sustains robust feature extraction capabilities with increased network depth. Furthermore, an EMA module is integrated after the initial lightweight convolutional layer to integrate multi-scale feature information and select discriminative features. This module reconstructs some channels into batch dimensions and groups the channel dimensions into multiple sub-features to ensure that spatial semantic features are evenly distributed within each feature group, which can be represented by Eq. (2)[47]. The purpose of this is not only to encode global information to recalibrate the channel weights in each parallel branch but also to further aggregate the output features of the two parallel branches through cross-dimensional interaction. To capture pixel-level paired relationships, it can retain more crack pixels.

$$f_{out} = f_C(f_{GS}(f_{GS}(f_{GS}(x))) + f_{EMA}(f_{GS}(x))) \tag{1}$$

where x represents the input, f_{out} represents the output, f_{EMA} represents the channel through the EMA module, f_{GS} represents the operation of GSConv with a kernel size of 3× 3, f_C represents the result of a standard convolution operation with a kernel size of 1 × 1.

$$X = [X_0, X_i, \cdots, X_{G-1}], X_i \in \mathbb{R}^{C//G \times W \times H} \tag{2}$$

where C//G represents division into G groups along the channel direction, and W and H represent the width and height of the feature map.

The structure of A-RepViT block

The RepViT Block is the main structure in the lightweight convolutional neural network RepViT[36], and it uses structural re-parameterization techniques to separate the Token mixer and Channel mixer. This preserves the original expressive capacity by decreasing parameters and computational complexity, thereby enhancing the model's learning and inference efficiency. Moreover, the expansion ratio in the Channel mixer is adjusted to 2, augmenting network width, improving model performance, and decreasing computational burden. A 3 ×3 kernel convolution is employed alongside the SE layer arranged in a cross-block configuration akin to the Token mixer to decrease latency without compromising model performance. Despite the reduction in delay, the SE layer exhibits lower precision for target segmentation compared to the self-attention mechanism in the conventional Transformer architecture. Therefore, the efficient global token fusion operator based on AFF proposed in [35] is adopted in this paper instead of the SE layer as the Token mixer in RepViT Block, and the A-RepViT Block is proposed, as shown in Fig. 3(a). The adaptive frequency filtering can help the model better capture the crack characteristics and deal with multi-scale features better. The AFF(as shown in the Fig. 3(b)) based efficient global Token fusion operator is mathematically equivalent to using a dynamic convolution kernel with the same spatial resolution as the Token set to perform Token fusion in the original domain (as shown in the Fig. 3(c)), and has the function of content adaptive Token fusion in the global scope. Thus, the depth of the image understanding and the accuracy of crack segmentation can be improved. Moreover, AFF optimizes computational resource utilization by prioritizing crucial information, leading to improved model performance.

The structure of SPPF

SPPF is improved on the basis of Spatial Pyramid Pooling(SPP), as shown in Fig. 4. SPP uses three maximum pooling layers with different kernel sizes to maximize feature maps, thus obtaining multi-scale feature maps. Finally, these feature maps are joined together to enrich the spatial hierarchy of the feature map. SPPF is modified from SPP by changing it into three max pooling layers of the same kernel size and transforming it into a serial structure for computation. This reduces redundant calculations and enhances the model's speed while better handling multi-scale objects and improving robustness to scale variations.

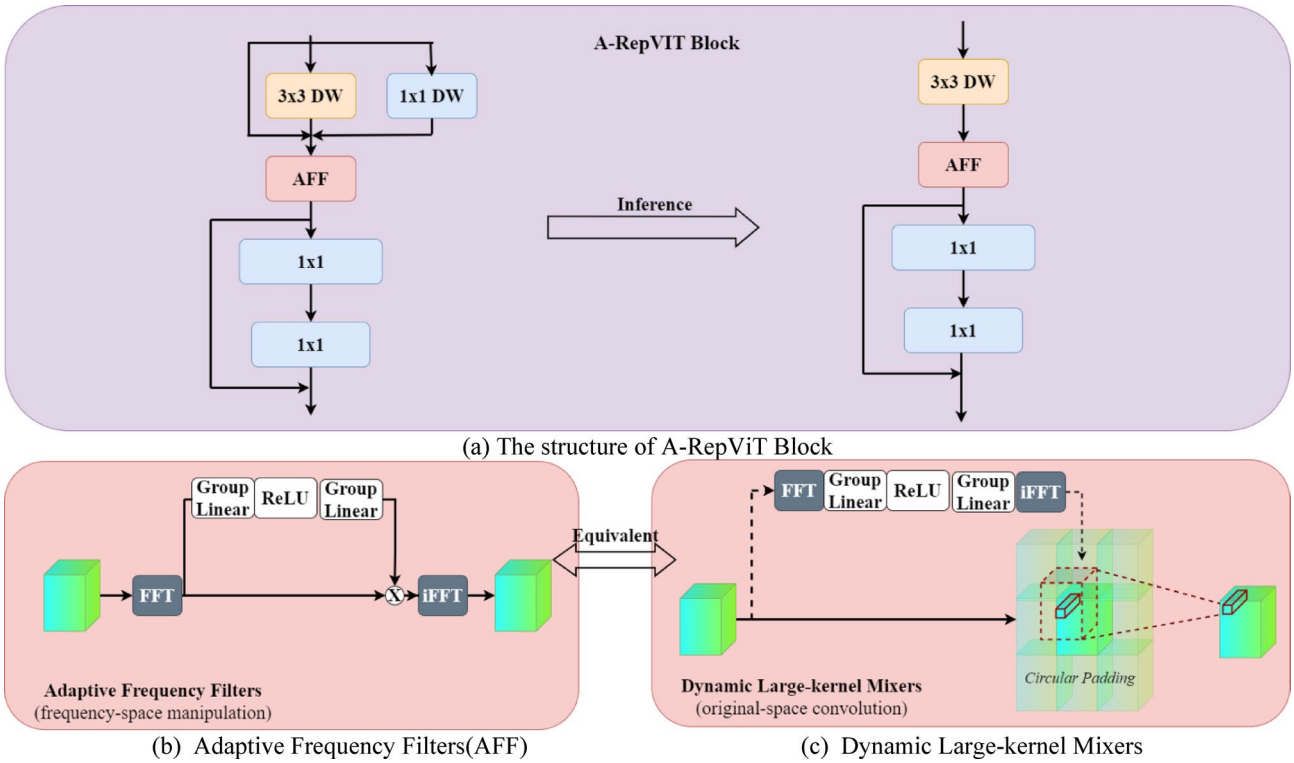


Fig. 3. The structures of A-RepViT Block, Adaptive Frequency Filters and Dynamic Large-kernel Mixers.

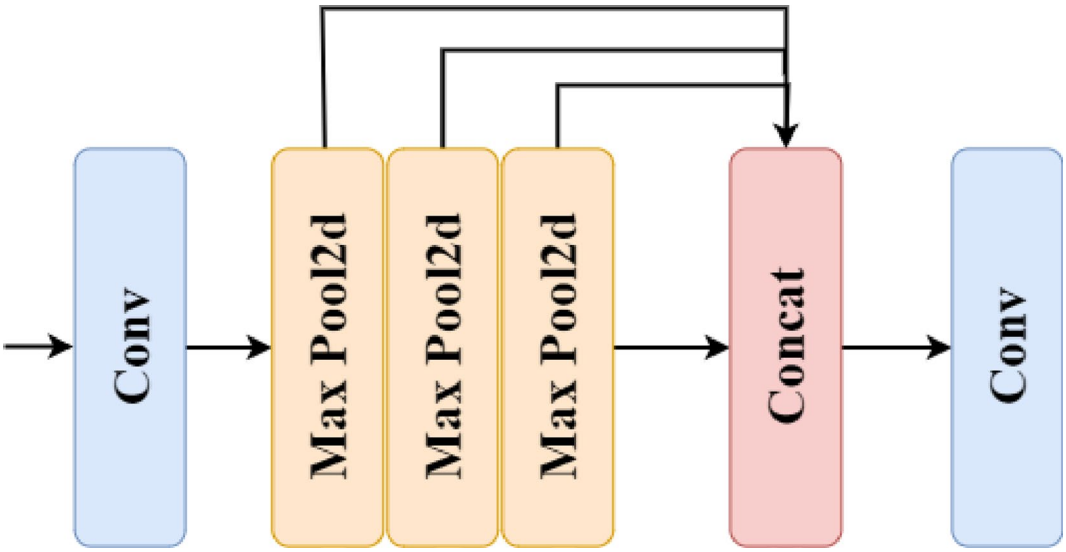


Fig. 4. The structure of SPPF

Loss function

The loss function is the most fundamental and crucial element in deep learning, as it measures the quality of predictions. EGA-UNet is regarded as a per-pixel classifier and generates a prediction map. Firstly, binary cross-entropy loss is introduced for training the EGA-UNet. The predicted probability for the pixel is denoted by $P = \{P_i, i = 1, 2, \cdots, N\}$, where N represents the number of pixels in the image. $Y = \{y_i, i = 1, 2, \cdots, N\}, y_i \in \{0, 1\}$ is the true label. The binary cross-entropy loss function is defined as:

$$L_{bce} = -\sum_{i=1}^{N} y_i \log(p_i) - (1 - y_i) \log(1 - p_i) \tag{3}$$

Dice loss can handle severely imbalanced classes, improving the balance between accuracy and recall during training. Dice loss is defined as:

$$L_{dice} = 1 - \frac{2\sum_{i=1}^{N} p_i y_i + 1 \times 10^{-6}}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} p_i^2 + 1 \times 10^{-6}} \tag{4}$$

Add 1×10^{-6} to both the numerator and denominator to smooth out fluctuations across epochs and enhance stability. Considering that most of the pixels in the image are non-crack (background), binary cross-entropy loss and Dice loss are combined for network training. The overall loss is the weighted sum of binary cross-entropy loss and Dice loss.

$$L = rL_{bce} + (1 - r)L_{dice} \tag{5}$$

where r is the weighting factor that controls the contribution of the two losses to the total loss. In this paper, r is set to 0.5.

Results and discussion

In order to verify the performance of the proposed EGA-UNet on crack segmentation, three public data sets are used to train and test the algorithm, and the EGA-UNet is compared with other mainstream detection algorithms. It is verified that the proposed algorithm shows good performance on topological cracks with different shapes and complex backgrounds. Finally, the ablation experiments are carried out to verify the effectiveness of the improved modules.

Experimental detail

The proposed method is implemented by PyTorch, and CUDA is used to accelerate the algorithm. The algorithm adopts batch normalization after each convolutional layer to accelerate the convergence speed during training. After batch normalization, SiLU is used to add the nonlinear relationship between the layers of the network to overcome the problems of overfitting and gradient disappearance. The following Table 1 lists the specific information.

experimental facilities	configuration
CPU	Intel(R) Xeon(R) Gold 6248R
GPU	NVIDIA RTX3090
PyTorch	1.13.0
CUDA	11.7
CUDNN	8.6.0
Optimizer	SGD
learning rate	0.01
weight decay	0.0001
momentum	0.9
Python	3.8

Table 1. Experimental configurations.

Dataset

Crack500[50]: This dataset consists of 500 images of road surface cracks, each with a size of 2000 ×1500 pixels. These images often blend with similar backgrounds with their complex background noise and foreground, which make detection difficult. Due to the small dataset, the images in the dataset are divided into 3368 road surface crack images, and each size of image is 640 ×360 pixels, as shown in Fig. 5(a). These images were then split into 1896 training images, 348 validation images, and 1124 test images. The cracks in this dataset are diverse, including single and topological cracks, as well as thick and thin cracks.

DeepCrack[9]: This dataset consists of 537 complex crack images with a resolution of 544 ×384, as shown in Fig. 5(b), which 300 images are for the training set and 237 images for the test set. The image samples contain cracks with various textures, different scenes and varying scales with crack widths ranging from 1 to 180 pixels. This allows for testing the network's performance in handling multi-scale cracks in different scenarios.

CrackTree206[51]: This dataset contains 206 images of road cracks with a resolution of 800 ×600 pixels, as shown in Fig. 5(c), which 126 images are for training and 80 images are for validation. The cracks in the dataset are relatively thin and there are obstructions and shadow interferences. It can test the generalization performance of network models.

CFD[52]: This dataset contains 118 images of size 480 ×320 pixels, as shown in Fig. 5(d). And these images contain background noises such as shadows and stains. All the images have manually labeled ground truth contours. The CFD dataset was used to test model generalization performance.

Evaluation metrics

Pixel-level crack detection can be regarded as a form of binary semantic segmentation, aiming to divide the image into foreground (cracks) and background (non-crack features). Therefore, Dice coefficient and mean intersection over union (mIoU) are used as evaluation metrics for segmentation performance. Dice coefficient is a statistic method used for comparing the similarity of two samples, while mIoU measures the ratio of the intersection to the union of the predicted results and the ground truth labels to assess model performance. Model Parameters and FLOPs are used simultaneously as standards for measuring model complexity. The equations are defined as follows.

$$Dice = \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{6}$$

$$mIoU = \frac{1}{2} \times (IoU_b + IoU_f) \tag{7}$$

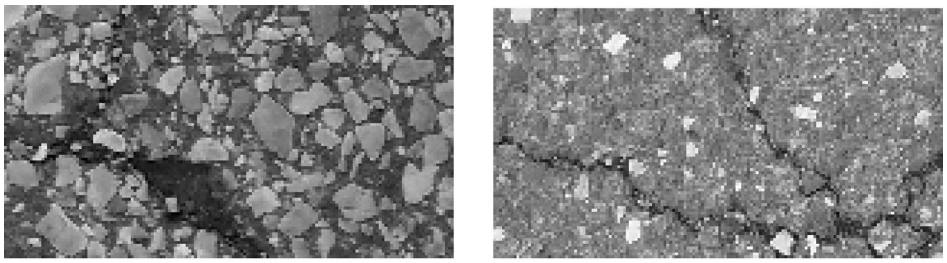
where p_i and g_i represent the predicted and ground truth values for pixel i respectively, and N is the total number of pixels. IoU_b represents Intersection over Union (IoU) of the true background and the predicted background. IoU_f represents Intersection over Union of actual cracks and predicted cracks.

The effectiveness of the attention mechanism

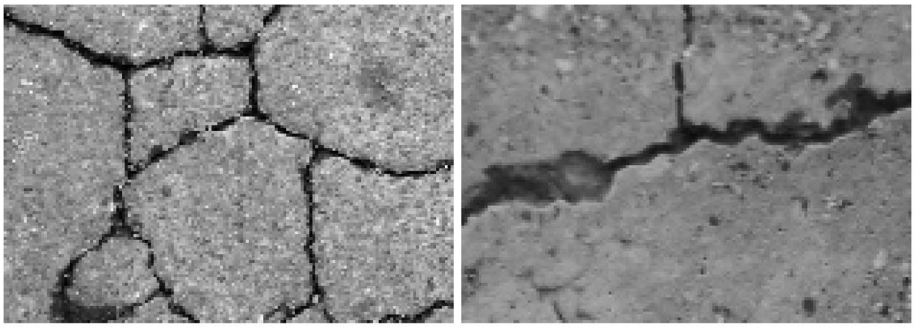
To verify the effectiveness of the attention mechanism in the EG-Block, we conducted experiments on EG-Block(without attention mechanism), EG-Block(with SE), EG-Block(with ECA), EG-Block(with CBAM) and EG-Block(with EMA) respectively in Crack500. Then it can be seen from Table 2 that the segmentation performance of EG-Block(with EMA) is the best.

Evaluation

To evaluate the effectiveness of the proposed method, this paper conducts comparative experiments between the proposed method and five mainstream methods—FCN, U-Net, SegNet, PSPNet, and DeepLabV3—on three public datasets. To ensure the comparability of the experiments, all compared methods were applied with their default parameter settings and underwent the same training process. A total of 100 rounds of training were conducted, followed by qualitative and quantitative analysis of the resulting models to evaluate the performance

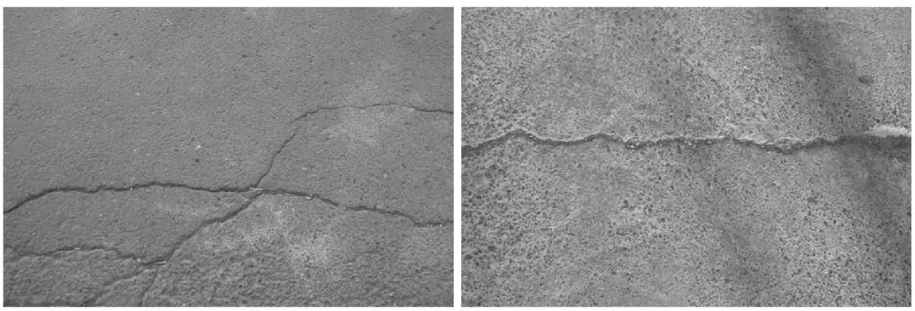


(a) Crack500 dataset samples



(b) DeepCrack dataset samples

(c) CrackTree206 dataset samples



(d) CFD dataset samples

Fig. 5. Samples of four datasets.

of each method. Firstly, we display the loss curve to determine whether the model overfits the current data as shown in Fig. 6.

Then, this paper qualitatively analyzes method performance through the detection results of different models. The detection results of FCN, UNet, SegNet, PSPNet, and DeepLabV3, compared to the method proposed in this paper on the Crack500, DeepCrack, and CrackTree206 public datasets are shown in the Fig. 7. The experiment selected images with topological, coarse, normal and fine cracks. For the images with topological and coarse cracks, it can be seen from the first row that FCN, UNet, SegNet, PSPNet, and DeepLabV3 all missed some

Method	Dice(%)	IoU_b(%)	IoU_f(%)	mIoU(%)	Params(M)	FLOPs(G)
EG-Block(without attention) mechanism)	76.4	97.6	68.4	83.0	4.68	33.09
EG-Block(with SE)	76.8	97.6	68.8	83.2	4.71	33.16
EG-Block(with ECA)	77.0	97.7	69.2	83.5	4.69	33.13
EG-Block(with CBAM)	76.5	97.7	68.1	82.9	4.71	33.16
EG-Block(with EMA)	77.8	97.8	70.0	83.9	4.73	34.28

Table 2. Test the effectiveness of the attention mechanism in the EG-Block on Crack500.

detections; from the sixth row, it can be observed that the FCN and UNet networks had incomplete detections, while the other networks performed well. For topological and normal cracks, it can be seen from the second row that in the presence of edge stone influence, all comparison networks show missed detections; from the third row, it can be seen that the comparison networks are less effective in segmenting cracks; from the fourth row, it can be seen that the comparison networks are poor at identifying scattered cracks; from the fifth row, it can be seen that FCN, UNet, PSPNet, and DeepLabv3 have incomplete segmentation of fine cracks at the edges. The last three rows show images of topological and fine cracks. It is evident from these images that FCN exhibits insensitivity towards fine cracks, rendering it incapable of their identification. Conversely, UNet, SegNet, PSPNet, and DeepLabV3 demonstrate a tendency to overlook numerous fine cracks. In contrast, the approach introduced in this study successfully detects nearly all fine cracks. In summary, the method proposed in this paper surpasses these methods in segmenting topological cracks of diverse shapes against intricate backgrounds, thereby meeting the practical demands for crack segmentation.

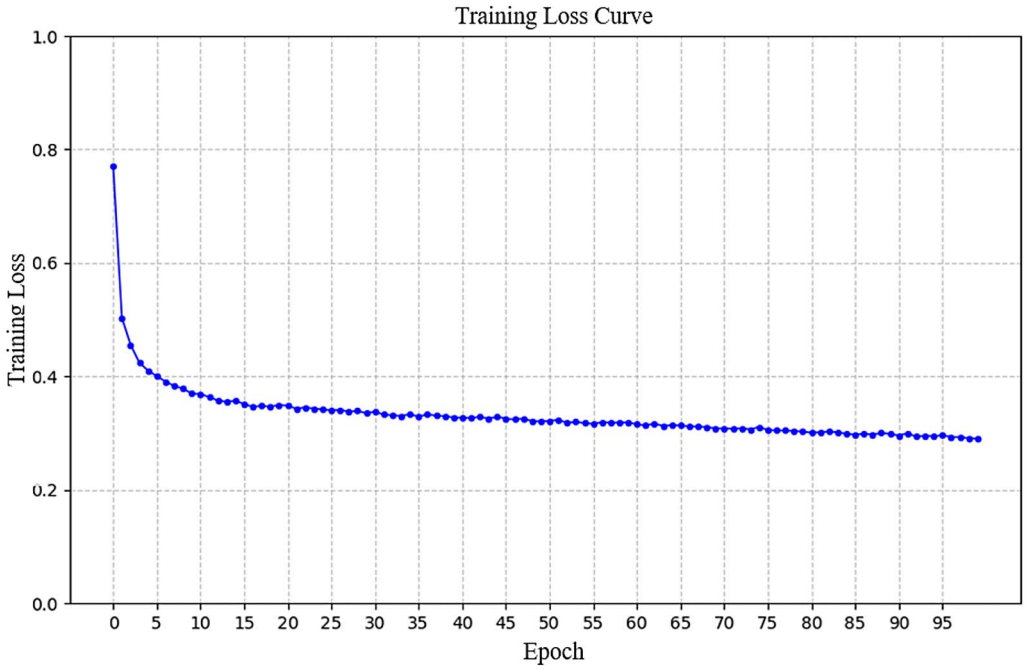
To better verify the effectiveness of the method proposed in this paper, quantitative analysis is conducted using three datasets. The experimental results for different methods are shown in the Tables 3, 4 and 5, which include Dice coefficient, background intersection over union (IoU_b), crack intersection over union (IoU_f), mean intersection over union (mIoU), parameters, FLOPs and FPS. On the Crack500 dataset, the proposed EGA-UNet achieves a Dice coefficient of 77.8%, which is 5.2%, 2.3%, 3.0%, 1.8% and 2.5% higher than FCN, U-Net, SegNet, PSPNet and DeepLabV3 respectively. The mIoU of EGA-UNet reaches 83.9%, which is higher than FCN, U-Net, SegNet, PSPNet and DeepLabV3 by 3.1%, 1.3%, 1.3%, 1.1% and 2.0% respectively. On the DeepCrack dataset, the proposed EGA-UNet achieves a Dice coefficient of 82.6%, which is higher than FCN, U-Net, SegNet, PSPNet and DeepLabV3 by 14.9%, 3.4%, 2.8%, 5.1% and 3.7% respectively. The mIoU of EGA-UNet reaches 86.4%, which exceeds FCN, U-Net, SegNet, PSPNet and DeepLabV3 by 8.8%, 1.5%, 1.9%, 3.0% and 2.2% respectively. On the CrackTree206 dataset, because the FCN network is too deep and does not make good use of the shallow information, the fine crack information is not extracted. The proposed EGA-UNet achieves a Dice coefficient of 73.1%, which is higher than U-Net, SegNet, PSPNet and DeepLabV3 by 3.1%, 11.9%, 44.9% and 52.3% respectively. The mIoU of EGA-UNet reaches 78.9%, surpassing U-Net, SegNet, PSPNet and DeepLabV3 by 1.3%, 4.7%, 21.1% and 23.4% respectively. Moreover, the number of parameters and FLOPs of EGA-UNet are significantly lower than those of models like FCN, SegNet, PSPNet, and DeepLabV3, thus EGA-UNet meets the requirements for real-time and high-precision crack segmentation. Additionally, it can be seen that EGA-UNet has a faster detection speed than other mainstream methods from FPS, which proves the high efficiency of this method. Finally, we compared EGA-UNet with the latest methods on three datasets in Table 6, and EGA-UNet has a highest mIoU that it meets the requirement of real-time segmentation of road cracks.

Furthermore, the model's generalization performance will be evaluated by testing on the CFD dataset. We performed quantitative and qualitative analyses on this dataset. And the experimental setting is exactly the same as the training to ensure the fairness of the experiment. Firstly, quantitative analysis as shown in Table 7, we can see that EGA-UNet has higher accuracy and speed than the compared network. And as shown in Fig. 8, it can be seen that EGA-UNet can basically segment pavement cracks, which reflects the good generalization performance of the model.

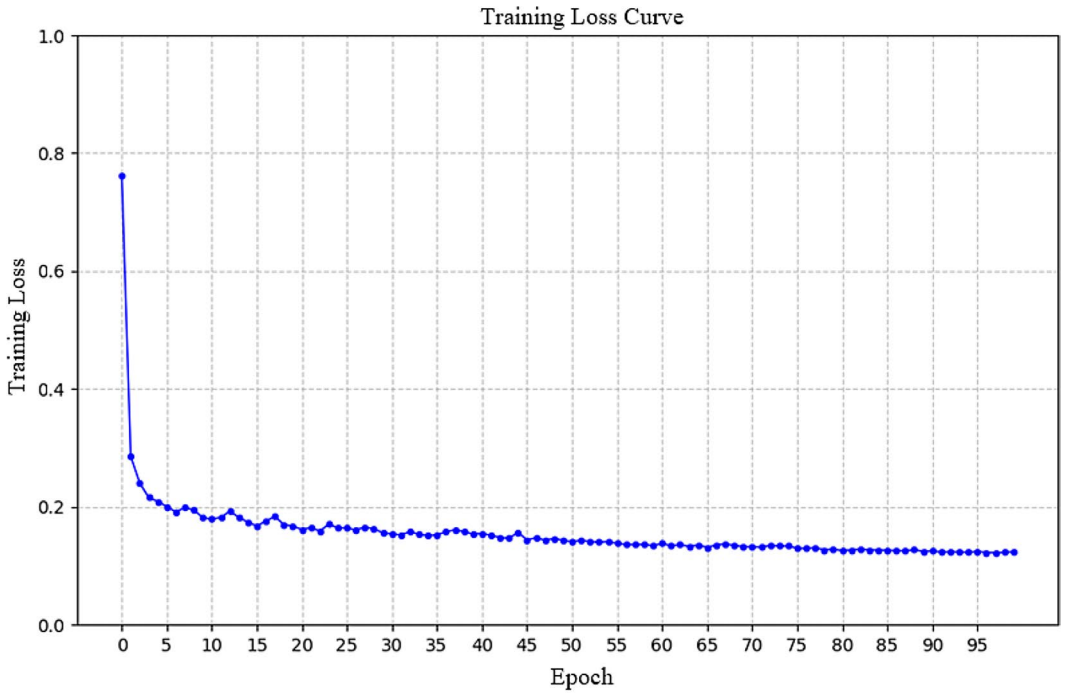
Finally, the superiority of the proposed method is demonstrated through comparisons on various complex background images, as illustrated in Fig. 9, showcasing its capability to accurately segment cracks amidst intricate backgrounds.

Ablation

In this section, ablation experiments are designed on the Crack500 dataset to examine the roles of different mechanisms in the U-Net architecture. The contribution of various improved modules to the final model's performance is evaluated, and the results are shown in Table 8. Among them, U-Net (c = 32) indicates that the channel parameters are 32, 64, 128, 256 respectively; U-Net (c = 64) indicates that the channel parameters are 64, 128, 256, 512 respectively. By comparing the first row and the second row, it can be seen that U-Net(c = 32) achieves higher metrics with fewer parameters. Comparing the last four lines with the first line, it can be seen that the added EG-Block module, SPPF module and A-RepViT Block module all effectively improve the model's performance in segmenting images. According to the results of the ablation experiments, compared to the basic U-Net network, the Dice coefficient increased by 2.3% and the mIoU increased by 1.3%, with almost no change in the number of parameters, which satisfies the requirements for detecting cracks of various sizes in complex backgrounds.



(a)Loss curve of the training process of the FCN network

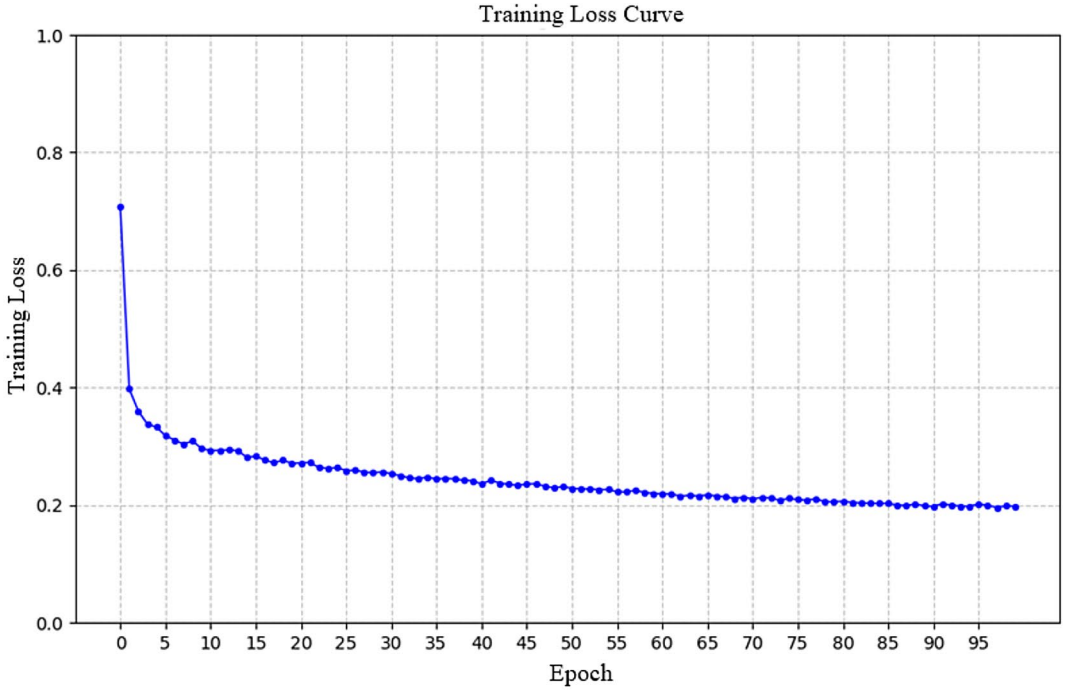


(b) Loss curve of the training process of the UNet network

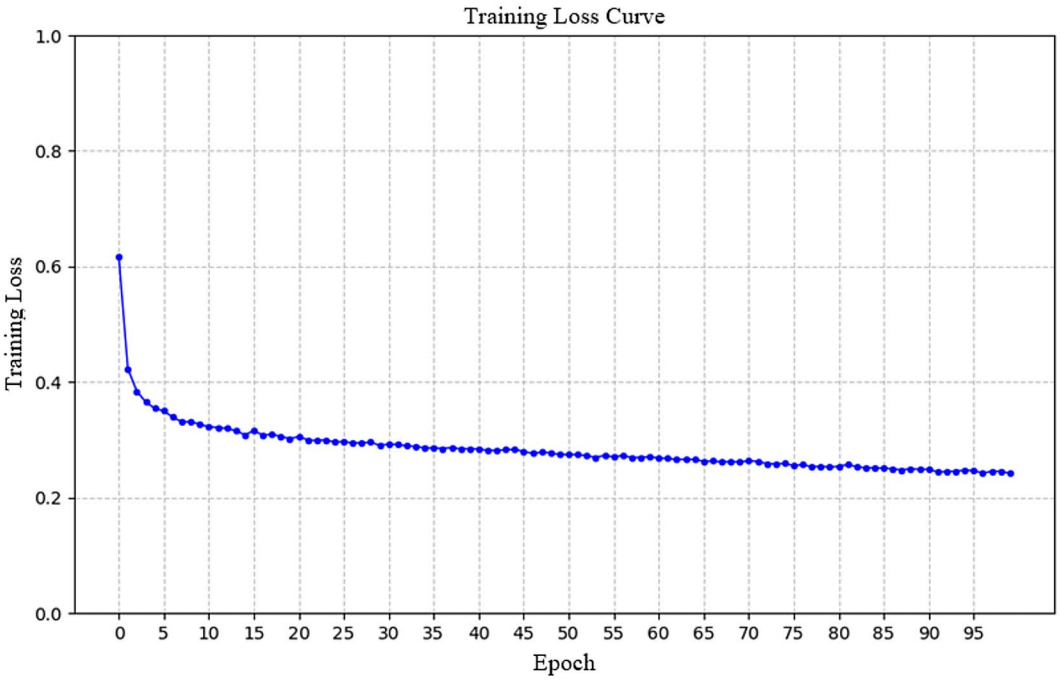
Fig. 6. Loss curve of the training process of the different network.

Conclusion

This research presented an innovative and efficient crack segmentation method. Specifically, lightweight modules are proposed, EG-Block and A-RepViT Block, which address the issue of feature loss in transmission and retains crack features better, while also improving the segmentation performance and efficiency of the network. This method addresses critical limitations of traditional crack segmentation methods, such as relying on slow and error-prone manual segmentation. Extensive experimentation on three publicly available datasets has confirmed that EGA-UNet outperforms in terms of higher mIoU and reduced computational demands, thereby affirming



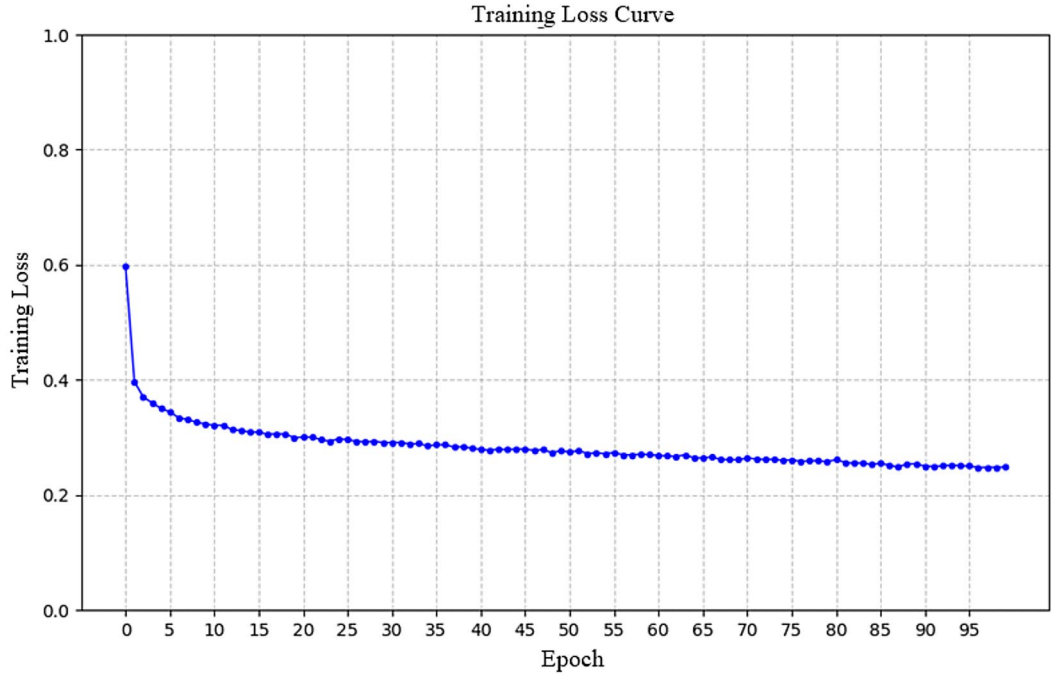
(c)Loss curve of the training process of the SegNet network



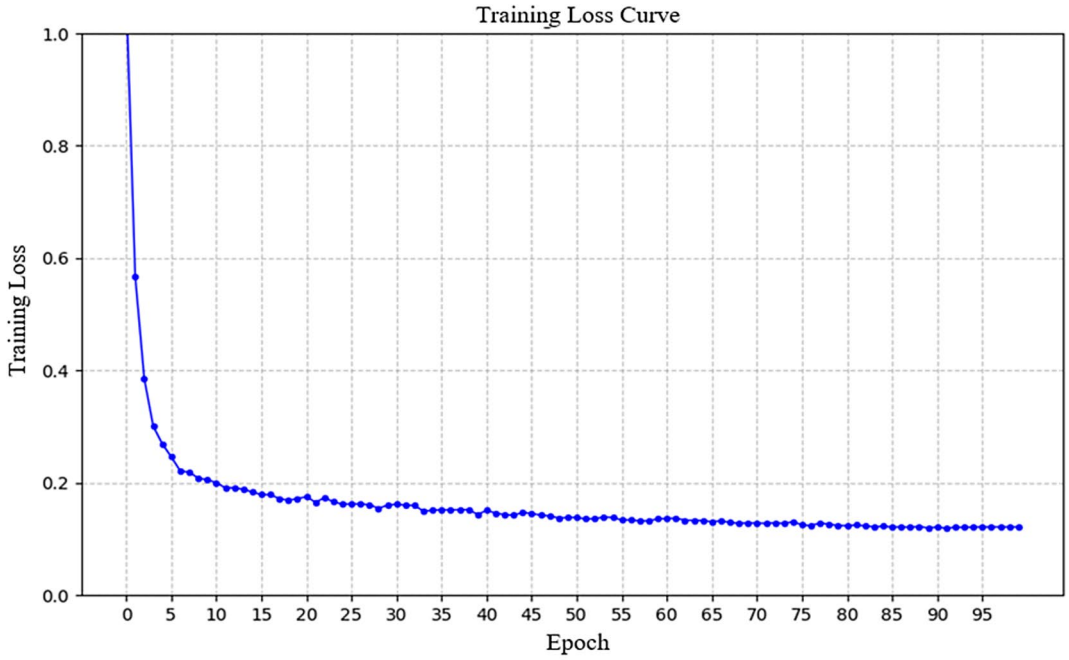
(d) Loss curve of the training process of the PSPNet network

Fig. 6. (continued)

the efficacy and versatility of the approach. This research holds significant implications by offering a streamlined approach for road maintenance personnel to expedite and enhance the accuracy of crack detection, thereby optimizing road maintenance operations. Future endeavors will focus on further refining crack segmentation techniques for roads and enhancing algorithmic efficiency. Moreover, efforts will be directed towards enhancing the model's generalizability across various segmentation domains.



(e)Loss curve of the training process of the DeepLabV3 network



(f) Loss curve of the training process of the EGA-UNet network

Fig. 6. (continued)

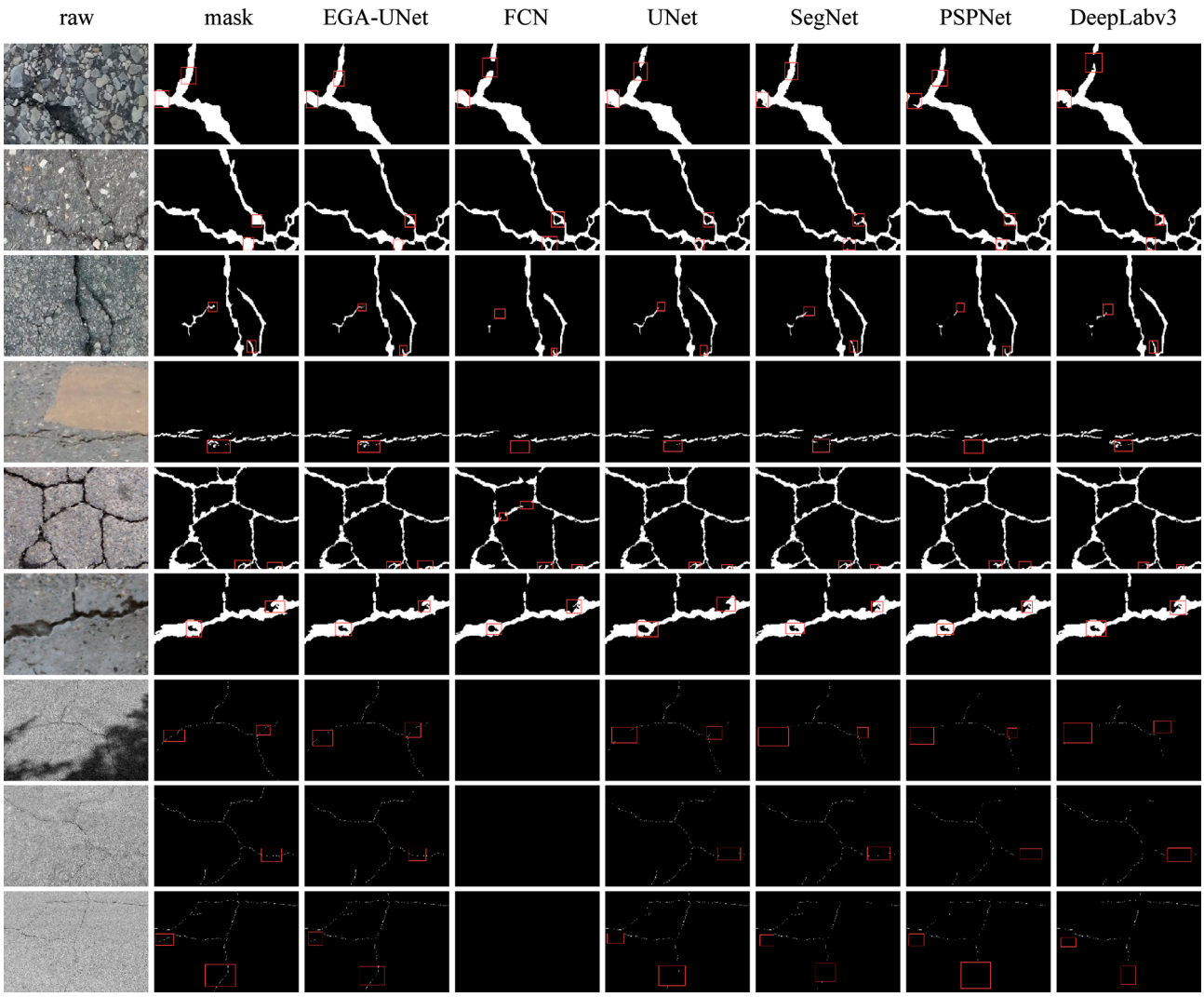


Fig. 7. Comparison pictures of various methods test results (The first three rows are for the crack500, the middle three rows are for the DeepCrack, and the last three rows are for the CrackTree206).

Method	Dice(%)	IoU_b(%)	IoU_f(%)	mIoU(%)	Params(M)	FLOPs(G)	FPS
FCN	72.6	97.2	64.6	80.8	134.27	213.27	37
U-Net	75.5	97.5	67.8	82.6	4.32	35.59	88
SegNet	74.8	97.6	67.6	82.6	29.48	149.49	43
PSPNet	76.0	97.5	68.1	82.8	46.59	156.10	42
DeepLabV3	75.3	97.4	66.4	81.9	73.22	144.28	45
EGA-UNet	77.8	97.8	70.0	83.9	4.73	34.28	91

Table 3. Comparison of the performance of different models on the Crack500.

Method	Dice(%)	IoU_b(%)	IoU_f(%)	mIoU(%)	Params(M)	FLOPs(G)	FPS
FCN	67.7	97.4	57.7	77.6	134.27	197.38	41
U-Net	79.2	98.5	71.4	84.9	4.37	32.29	91
SegNet	79.8	98.5	69.9	84.5	29.48	135.79	46
PSPNet	77.5	98.3	68.6	83.4	46.59	141.54	43
DeepLabV3	78.9	98.4	70.1	84.2	73.22	130.82	49
EGA-UNet	82.6	98.7	74.0	86.4	4.73	31.12	95

Table 4. Comparison of the performance of different models on the deepcrack.

Method	Dice(%)	IoU_b(%)	IoU_f(%)	mIoU(%)	Params(M)	FLOPs(G)	FPS
FCN	-	-	-	-	134.27	370.64	26
U-Net	70.1	99.8	55.5	77.6	4.32	74.16	61
SegNet	61.2	99.7	45.0	74.2	29.48	311.67	28
PSPNet	28.2	99.2	16.4	57.8	46.59	325.16	27
DeepLabV3	20.8	98.8	12.2	55.5	73.22	300.54	30
EGA-UNet	73.1	99.8	58.1	78.9	4.73	71.46	64

Table 5. Comparison of the performance of different models on the CrackTree206.

Dataset Method	Crack500	DeepCrack	CrackTree206
Okran et al[53].	78.30	-	-
Jia et al[54].	50.54	-	-
Saberironaghi et al[55].	77.00	83.90	-
Xie et al[56].	-	-	78.10
Wang et al. [57]		77.19	76.59
ours	83.90	86.40	78.90

Table 6. mIoU comparison of the latest models on three datasets.

Method	Dice(%)	IoU_b(%)	IoU_f(%)	mIoU(%)	Params(M)	FLOPs(G)	FPS
FCN	-	-	-	-	134.27	144.92	45
U-Net	72.4	99.8	63.8	81.8	4.32	25.86	96
SegNet	65.1	99.7	52.3	76.0	29.48	121.26	50
PSPNet	42.5	99.4	32.6	66.0	46.59	126.75	49
DeepLabV3	33.8	99.1	25.5	62.3	73.22	114.74	56
EGA-UNet	75.2	99.8	71.1	85.5	4.73	23.97	101

Table 7. Comparison of the performance of different models on the CFD. Raw UNet EGA-UNet.

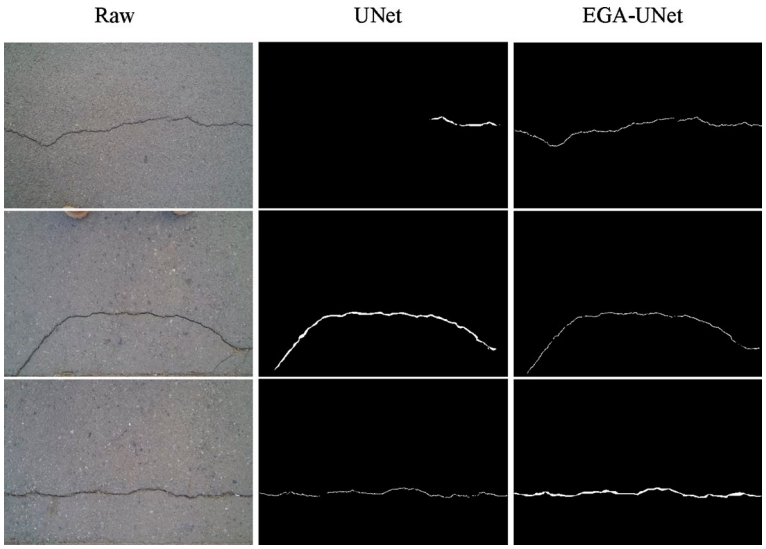
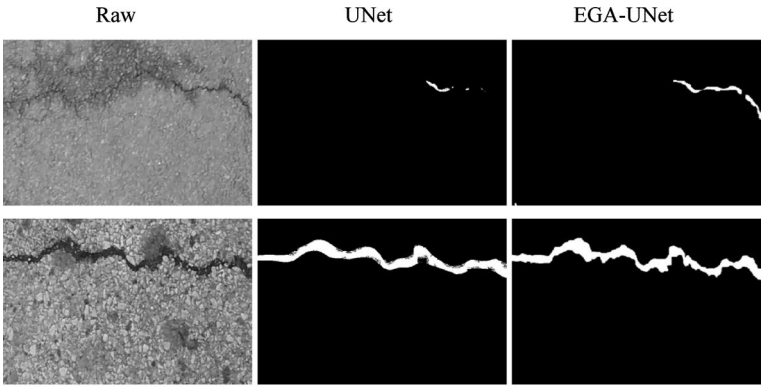
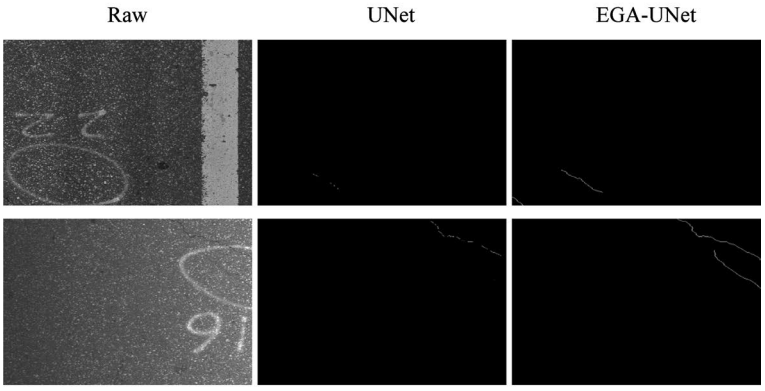


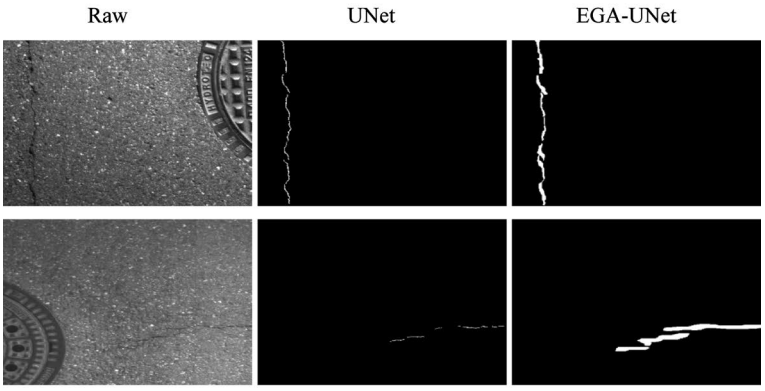
Fig. 8. Test on the CFD dataset



(a) Crack with oil or water stain on the road surface

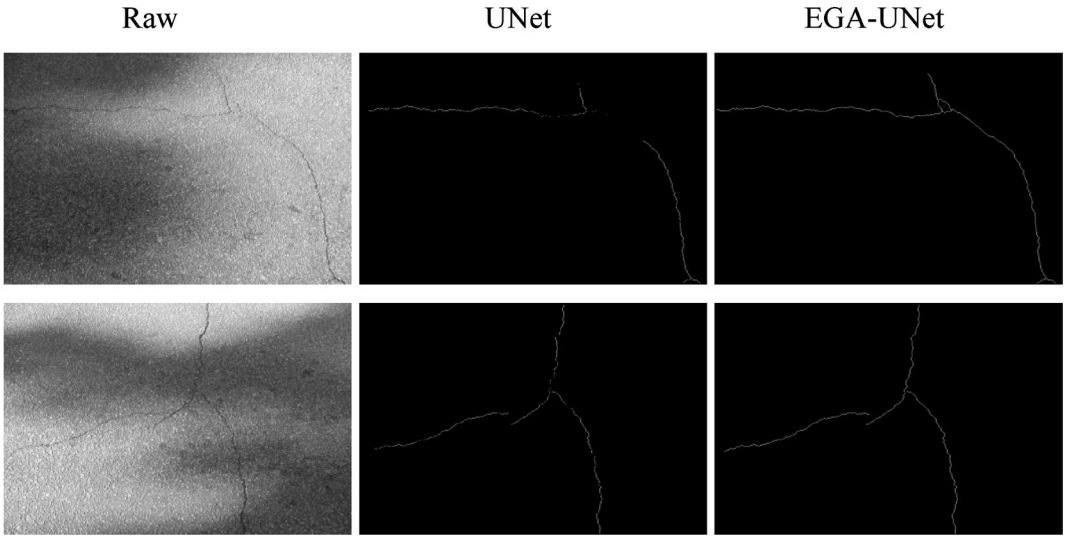


(b) Crack with font on the road surface

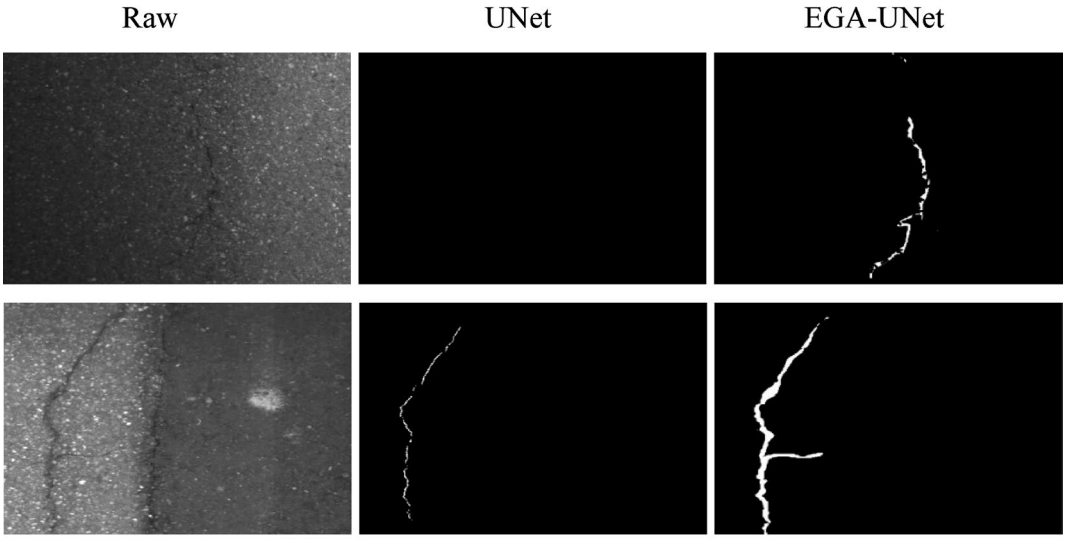


(c) Crack with manhole cover on the road surface

Fig. 9. Compare the segmentation of various complex background images.



(d) Crack with shadows on the road surface



(e) Crack with different light intensities on the road surface

Fig. 9. (continued)

Method	Dice(%)	IoU_b(%)	IoU_f(%)	mIoU(%)	Params(M)	FLOPs(G)
U-Net(c = 32)	75.5	97.5	67.8	82.6	4.32	35.59
U-Net(c = 64)	74.8	97.4	66.3	81.9	17.26	141.20
+EG-Block	76.7	97.7	68.9	83.3	4.38	35.68
+EG-Block(GSConv)& RepViT Block	76.9	97.7	69.3	83.5	4.65	35.90
+ EG-Block(GSConv)& A-RepViT Block	77.3	97.7	69.5	83.6	4.53	31.80
EGA-UNet	77.8	97.8	70.0	83.9	4.73	31.93

Table 8. Effectiveness of each improved module.

Data availability

The data that support the findings of this study are available upon reasonable request from the corresponding author.

Received: 17 December 2024; Accepted: 9 May 2025

Published online: 30 September 2025

References

1. Adlinge, S. S. & Gupta, A. Pavement deterioration and its causes. *Int. J. Innovative Res. Dev.* **2**, 437–450 (2013).
2. Rong, W., Li, Z., Zhang, W. & Sun, L. An improved canny edge detection algorithm. In. *IEEE international conference on mechatronics and automation*, 577–582 (IEEE, 2014). (2014).
3. Yan-Ying, G., Guo-Qing, Y. & Ii-hui, J. Adaptive weighted morphology detection algorithm of plane object in Docking guidance system. *Int. J. Adv. Robotic Syst.* **7**, 16 (2010).
4. Cuevas, E., Zaldivar, D. & Pérez-Cisneros, M. A novel multi-threshold segmentation approach based on differential evolution optimization. *Expert Syst. Appl.* **37**, 5265–5271 (2010).
5. Dollár, P. & Zitnick, C. L. Fast edge detection using structured forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 1558–1570 (2014).
6. Gao, W., Zhang, X., Yang, L. & Liu, H. An improved sobel edge detection. In. *3rd International conference on computer science and information technology*, vol. 5, 67–71 (IEEE, 2010). (2010).
7. Yuan, L. & Xu, X. Adaptive image edge detection algorithm based on canny operator. In *4th International Conference on Advanced Information Technology and Sensor Application (AITS)*, 28–31 (IEEE, 2015). (2015).
8. Gu, J. et al. Recent advances in convolutional neural networks. *Pattern Recogn.* **77**, 354–377 (2018).
9. Liu, Y. et al. A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing* **338**, 139–153 (2019).
10. Choi, W., Cha, Y. J. & Sddnet Real-time crack segmentation. *IEEE Trans. Ind. Electron.* **67**, 8016–8025 (2019).
11. Qu, Z., Cao, C., Liu, L. & Zhou, D. Y. A deeply supervised convolutional neural network for pavement crack detection with multiscale feature fusion. *IEEE Trans. Neural Networks Learn. Syst.* **33**, 4890–4899 (2021).
12. Kang, D., Benipal, S. S., Gopal, D. L. & Cha, Y. J. Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning. *Autom. Constr.* **118**, 103291 (2020).
13. Di Benedetto, A., Fiani, M. & Gujski, L. M. U-net-based Cnn architecture for road crack segmentation. *Infrastructures* **8**, 90 (2023).
14. Yao, H. et al. Encoder–decoder with pyramid region attention for pixel-level pavement crack recognition. *Comput. Civ. Infrastructure Eng.* **39**, 1490–1506 (2024).
15. Zhu, G. et al. A lightweight encoder–decoder network for automatic pavement crack detection. *Comput. Civ. Infrastructure Eng.* **39**, 1743–1765 (2024).
16. Chen, J. & He, Y. A novel u-shaped encoder–decoder network with attention mechanism for detection and evaluation of road cracks at pixel level. *Comput. Civ. Infrastructure Eng.* **37**, 1721–1736 (2022).
17. Li, P. et al. Our-net: a multi-frequency network with octave max unpooling and octave Convolution residual block for pavement crack segmentation. *IEEE Trans. Intell. Transp. Syst.* (2024).
18. Huang, Y., Liu, Y., Liu, F. & Liu, W. A lightweight feature attention fusion network for pavement crack segmentation. *Comput. Civ. Infrastructure Eng.* **39**, 2811–2825 (2024).
19. Fan, Z. et al. Ensemble of deep convolutional neural networks for automatic pavement crack detection and measurement. *Coatings* **10**, 152 (2020).
20. Chu, H., Wang, W. & Deng, L. Tiny-crack-net: A multiscale feature fusion network with attention mechanisms for segmentation of tiny cracks. *Comput. Civ. Infrastructure Eng.* **37**, 1914–1931 (2022).
21. Targ, S., Almeida, D. & Lyman, K. Resnet in resnet: Generalizing residual architectures. *arxiv 2016. arXiv preprint arXiv:1603.08029* (2016).
22. Cheng, J., Xiong, W., Chen, W., Gu, Y. & Li, Y. Pixel-level crack detection using u-net. In *TENCON 2018–2018 IEEE region 10 conference*, 0462–0466IEEE, (2018).
23. Fan, Z., Wu, Y., Lu, J. & Li, W. Automatic pavement crack detection based on structured prediction with the convolutional neural network. *arXiv preprint arXiv:1802.02208* (2018).
24. Yu, G., Dong, J., Wang, Y. & Zhou, X. Ruc-net: A residual-unet-based convolutional neural network for pixel-level pavement crack segmentation. *Sensors* **23**, 53 (2022).
25. Zhang, T., Wang, D., Lu, Y. & Ecsnet An accelerated real-time image segmentation Cnn architecture for pavement crack detection. *IEEE Trans. Intell. Transp. Syst.* **24**, 15105–15112 (2023).
26. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inform. Process. Syst.* 30 (2017).
27. Zhang, Z. et al. Vit-yolo: Transformer-based yolo for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2799–2808 (2021).
28. Chen, J. et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021).
29. Berg, A., O'Connor, M. & Cruz, M. T. Keyword transformer: A self-attention model for keyword spotting. *arXiv preprint arXiv:2104.00769* (2021).
30. Gao, Y., Zhou, M. & Metaxas, D. N. Utnet: a hybrid transformer architecture for medical image segmentation. In Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, Part III 24, 61–71Springer, (2021).
31. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A. & Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299 (2022).
32. Xie, E. et al. Segformer: simple and efficient design for semantic segmentation with Transformers. *Adv. Neural Inform. Process. Syst.* **34**, 12077–12090 (2021).
33. Yu, T., Li, X., Cai, Y., Sun, M. & Li, P. S2-mlp: Spatial-shift mlp architecture for vision. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 297–306 (2022).
34. Yu, W. et al. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10819–10829 (2022).
35. Huang, Z. et al. Adaptive frequency filters as efficient global token mixers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6049–6059 (2023).
36. Wang, A. et al. Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15909–15920 (2024).
37. Liu, H. et al. Transformer network for fine-grained crack detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3783–3792 (2021).
38. Lv, Z., Zhong, P., Wang, W., You, Z. & Falco, N. Multiscale attention network guided with change gradient image for land cover change detection using remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **20**, 1–5 (2023).
39. Zhu, W. et al. Concrete crack detection using lightweight attention feature fusion single shot multibox detector. *Knowledge-Based Syst.* **261**, 110216 (2023).
40. Yang, L., Bai, S., Liu, Y. & Yu, H. Multi-scale triple-attention network for pixelwise crack segmentation. *Autom. Constr.* **150**, 104853 (2023).
41. Sun, X., Xie, Y., Jiang, L., Cao, Y. & Liu, B. Dma-net: deeplab with multi-scale attention for pavement crack segmentation. *IEEE Trans. Intell. Transp. Syst.* **23**, 18392–18403 (2022).

42. Chen, X. et al. Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14572–14581 (2023).
43. Huang, R. & Wang, X. Face anti-spoofing using feature distilling and global attention learning. *Pattern Recognit.* **135**, 109147 (2023).
44. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141 (2018).
45. Wang, Q. et al. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11534–11542 (2020).
46. Xu, W., Wan, Y. & Ela Efficient local attention for deep convolutional neural networks. *arxiv 2024. arXiv preprint arXiv:2403.01123* (2024).
47. Ouyang, D. et al. IEEE,. Efficient multi-scale attention module with cross-spatial learning. In *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5 (2023).
48. Qiao, W., Liu, Q., Wu, X., Ma, B. & Li, G. Automatic pixel-level pavement crack recognition using a deep feature aggregation segmentation network with a Scse attention mechanism module. *Sensors* **21**, 2902 (2021).
49. Liang, J., Gu, X., Jiang, D. & Zhang, Q. Cnn-based network with multi-scale context feature and attention mechanism for automatic pavement crack segmentation. *Autom. Constr.* **164**, 105482 (2024).
50. Amhaz, R., Chambon, S., Idier, J. & Baltazart, V. Automatic crack detection on two-dimensional pavement images: an algorithm based on minimal path selection. *IEEE Trans. Intell. Transp. Syst.* **17**, 2718–2729 (2016).
51. Zou, Q., Cao, Y., Li, Q., Mao, Q. & Wang, S. Cracktree: automatic crack detection from pavement images. *Pattern Recognit. Lett.* **33**, 227–238 (2012).
52. Shi, Y., Cui, L., Qi, Z., Meng, F. & Chen, Z. Automatic road crack detection using random structured forests. *IEEE Trans. Intell. Transp. Syst.* **17**, 3434–3445 (2016).
53. Okran, A. M., Rashwan, H. A. & Puig, D. Enhanced crack segmentation network: Leveraging multi-dimensional attention. In *Artificial Intelligence Research and Development*, 94–96IOS Press, (2024).
54. Jia, G., Lu, Y., Song, W., He, H. & Ma, R. A method for superfine pavement crack continuity detection based on topological loss. *Electron. Lett.* **59**, e12963 (2023).
55. Saberironaghi, A., Ren, J. & Depthcracknet A deep learning model for automatic pavement crack detection. *J. Imaging.* **10**, 100 (2024).
56. Xie, Y., Zhuo, A., Chen, Y. & Bai, Y. Nffnet: An encoder-decoder net for crack detection based on noise filtering fusion. (2023).
57. Wang, Y. et al. Ggmnet: Pavement-crack detection based on global context awareness and multi-scale fusion. *Remote Sens.* **16**, 1797 (2024).

Author contributions
All the authors contributed extensively to the manuscript. L.Y. contributed to the experimental equipment, research directions and opinions. J.D. performed the experiments and analyzed the results. L.Y., J.D., H.D., and C.Y. wrote and edited the article. All authors have read and agreed to the publication of the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Conflict of interest
The authors report no conflict of interest.

Additional information
Correspondence and requests for materials should be addressed to L.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.