applied sciences

MDPI

Article

A Multi-Scale Contextual Information Enhancement Network for Crack Segmentation

Lili Zhang [1], Yang Liao [1], Gaoxu Wang [2,*], Jun Chen [3] and Huibin Wang [1]

[1] College of Computer and Information Engineering, Hohai University, Nanjing 211100, China
[2] State Key Laboratory of Hydrology–Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing 210029, China
[3] College of Harbour, Coastal and Offshore Engineering, Hohai University, Nanjing 211100, China
* Correspondence: gxwang@nhri.cn

Abstract: In recent years, convolutional neural-network-based crack segmentation methods have performed excellently. However, existing crack segmentation methods still suffer from background noise interference, such as dirt patches and pitting, as well as the imprecise segmentation of fine-grained spatial structures. This is mainly due to the fact that convolutional neural networks dilute low-level spatial information in the process of extracting deep semantic features, and the network cannot obtain accurate context awareness because of the limitation of the actual receptive field size. To address these problems, an encoder–decoder crack segmentation network based on multi-scale contextual information enhancement is proposed. First, a new architecture of skip connection is proposed, enabling the network to obtain refined crack segmentation results; then, a contextual feature enhancement module is designed to make the network more effective at distinguishing between cracks and background noise; finally, the deformable convolution is introduced into the encoder network to further enhance its ability to extract the diverse morphological features of cracks by adaptively adjusting the sampling area and the receptive field size. Experiments show that the proposed method is effective in crack segmentation and outperforms mainstream segmentation networks such as DeepLab V3+ and UNet++.

Keywords: convolutional neural network; crack segmentation; skip connections; contextual features; deformable convolution

Citation: Zhang, L.; Liao, Y.; Wang, G.; Chen, J.; Wang, H. A Multi-Scale Contextual Information Enhancement Network for Crack Segmentation. *Appl. Sci.* **2022**, *12*, 11135. https://doi.org/10.3390/app122111135

Academic Editors: Christian W. Dawson

Received: 23 September 2022
Accepted: 31 October 2022
Published: 2 November 2022

1. Introduction

In recent years, as the urbanization rate of countries around the world increases, a large number of infrastructures, such as bridges, tunnels, and dams, are constructed, providing a solid guarantee for economic development and livelihood security. However, the supervision and maintenance of these facilities has also brought us new challenges. These infrastructures commonly use concrete as the construction material and the surface crack is one of the main symptoms of their damage and destruction [1,2]. Without timely maintenance, cracks will have a significant impact on the service life and safety of those infrastructures. Other facilities, such as asphalt roads, also need to be checked regularly to ensure that surface cracks can be maintained and repaired in a timely way. Therefore, the automatic identification of surface cracks from optical images of various scenes is of great research importance [3]. Due to the development of computer science and image processing technology, it is now possible to partially automate the process of surface crack inspection. However, it is still a difficult task to accurately separate the cracks from the complex image background, as there may be dirt patches, oil stains, pitting, or other noise interferences.

Most of the early crack segmentation techniques rely on traditional digital image processing methods, which often involve multiple pre-processing processes, such as morphological filtering [4,5], fuzzy theory methods [6,7], and wavelet transform [8,9], as well as various crack segmentation methods, such as methods based on the threshold algorithm [10,11] or the edge detection algorithm [12,13]. Traditional digital image processing methods are sensitive to interference from external factors, such as light changes and shadow occlusion, making them unusable in complex scenes. Meanwhile, digital image processing methods require manually designed feature operators, which are more difficult and less efficient to implement.

Recently, the application of deep-learning-based convolutional neural networks (CNNs) in the field of computer vision has developed rapidly and has even surpassed human performance in a variety of tasks, such as image classification [14,15], object detection [16,17], and semantic segmentation [18,19]. Compared with traditional digital image processing methods, CNNs are characterized by their high level of automation and strong feature extraction capability, as CNNs do not rely on manually designed feature operators. In terms of crack recognition applications, some studies localize cracks in images by classification [20,21] or object detection [22,23] methods. However, these methods cannot obtain detailed information about the cracks, making them less optimal. Segmentation-based crack recognition methods annotate cracks in images at the pixel level, providing a better level of detailed information, as part of the current mainstream research direction [24].

Due to the special morphological characteristics of cracks, the crack segmentation task faces two challenges: the accurate segmentation of fine-grained spatial structures and the ability to adapt to complex background environments. The former requires that the multi-level feature information extracted by the feature extraction network can be fully utilized, while the latter requires the network to possess accurate context awareness. It is shown in [25] that feature maps in different levels explore distinctive information, with shallow feature maps possessing fine spatial information and deep feature maps capturing rich semantic information, while the conversion process from shallow to deep feature maps leads to a loss of detailed spatial information. To recover the lost spatial information in the decoder network, SegNet [26] assists the decoder in up-sampling by means of maximum pooling indexing, while U-Net [19] feeds the shallow feature information generated in the encoder directly to the decoder network by means of a skip connection. Both of them are based on the symmetric encoder–decoder architecture, and there are some recent studies of crack segmentation which also use similar architectures [27,28]. However, it is demonstrated in [29] that the multi-scale feature information in the encoder cannot be fully utilized by delivering information between the same layers of the encoder and decoder networks. Meanwhile, due to the limitation of the empirical receptive field size [30], the plain convolutional neural networks cannot provide sufficient contextual feature information, which is necessary to adapt to complex scenarios. To address these problems, this paper proposes a multi-scale contextual information enhancement network (MCIE-Net), which redesigns the connection structure between the encoder and the decoder of the U-Net to capture multi-scale feature information and enhance the decoder's ability to restore fine-grained the spatial structure of cracks; meanwhile, a contextual feature enhancement module, which consists of the pyramid pooling network and channel attention mechanism, is designed to enhance the context awareness of the network.

Specifically, the contributions of this paper can be summarized as follows:

i. In order to obtain refined crack segmentation results, a multi-scale skip connection structure is designed to aggregate the multi-level feature information extracted from the decoder and improve the network's ability to capture the spatial features. The multi-scale skip connection also optimizes the feature aggregation method, which reduces the number of network parameters and lowers the computational cost.

ii. To accurately distinguish between crack and other interferences in the background, a contextual feature enhancement module is proposed to extract contextual information at multiple scales to improve the context awareness of the network. It consists of a pyramid pooling network and a channel attention mechanism which can recalibrate the channel weights of feature maps to guide the network to focus on important contextual information.
iii. Since the fixed sampling scale of plain convolution is not conducive to extracting the diverse morphological features of cracks, we improve the feature extraction network by introducing the deformable convolution into the deep layers of the encoder, changing the receptive field size of the network through adaptive sampling area adjustment, and enhancing its feature extraction capability.

2. Related Works

2.1. Traditional Image Processing Methods

Most of the traditional crack segmentation methods mainly rely on the color difference between cracks and background or the edge features of cracks to extract cracks from images [31]. Kirschke et al. [10] used a histogram-based threshold segmentation method to extract road cracks. Cheng et al. [11] proposed a threshold segmentation algorithm with reduced sample space and interpolation to optimize the efficiency of crack segmentation. Katakam [32] used the method of chunking the image first and then threshold-handling each sub-block separately to improve the accuracy of crack segmentation. Oliveira and Correia [33] firstly pre-processed the images using morphological filters and then used dynamic threshold segmentation to segment the cracks. Zhang et al. [34] integrated spatial clustering, threshold segmentation, and region-growing methods to obtain a coarse-to-fine segmentation of cracks. In [9,35], wavelet transform was used for crack segmentation, while in [12], the Canny operator was used to detect the contours of cracks. In addition, there are some studies that identify cracks with the help of machine learning methods. Considering the connectivity of cracks, Fernandes et al. [36] used a graph-based (graph-based) approach to extract crack features, and then support vector machines were used to classify the features to obtain a classification of crack types. In [37], crack structure features were extracted and learned from annotation data, and, based on this, a crack recognition framework was generated using random structure forest to achieve pixel-level crack segmentation.

2.2. Deep-Learning-Based Methods

Deep-learning-based crack segmentation methods mostly use semantic segmentation models. In 2015, Long et al. [18] achieved the first end-to-end segmentation of natural images using fully convolutional neural (FCN) networks, which have thus become the most classical network model in the field of semantic segmentation. Liu et al. [25] used a FCN backbone and a deeply supervised approach to upscale and fuse the feature maps from all levels of the backbone, and then applied a guided filter to fuse all feature maps as well as the side outputs to create a segmentation output. Ren et al. [38] used dilated convolution with a different dilation rate in the last four layers of the FCN to expand the receptive field without changing the feature map scale, and used skip connections to deliver shallow feature information, assisting the decoder in generating segmentation results. However, the methods based on FCN networks still suffer from information loss when up-sampling low-resolution feature maps generated in the deep layer of the feature extraction network. To solve the problem, symmetric encoder–decoder-based network structures, such as SegNet [26] and U-Net [19], have been proposed. In particular, U-Net has had a profound impact on many subsequent studies due to its pioneering concept and excellent performance, and a series of semantic segmentation models such as UNet++ [39] and Unet 3+ [29] have been derived on its basis. Since the detailed spatial information of cracks can be more effectively restored, many recent studies of crack segmentation are

based on the SegNet and U-Net structures. Ran et al. [40] introduced a spatial attention mechanism and a channel attention mechanism in SegNet and used spatial pyramidal pooling to capture crack features from different scales. Zou et al. [3] pair-wisely fused the feature maps generated in the encoder and decoder network at the same scale, and generated segmentation results by extracting features from the fused feature maps at multiple scales using a multi-scale fusion component. Lau et al. [27] replaced the plain convolutional neural network of the encoder of U-Net with a residual network and added spatial and channel compression excitation modules to the decoder. Based on U-Net, Han et al. [28] designed a skip-level round-trip sampling structure, in which the deep feature maps of the encoder network were up-sampled and aggregated with some shallow feature maps, and then down-sampled and fed into the decoder network. These up- and down-sampling actions enhanced the network's memory of transmitting low-level features in the shallow layer, helping the network to pay attention to the distinction between the cracks and the background. Zhao et al. [30] proposed PSPNet, which applies special pyramid pooling to the semantic segmentation task and extracts multi-scale contextual information. Some other studies also explored spatial pyramid pooling, such as the DeepLab series [41–43], although the difference is that DeepLabs use a dilated convolution rather than pooling to obtain contextual information at multiple scales. Sun et al. [44] adopted and enhanced DeepLabv3+, in which a multi-attention module was introduced to dynamically adjust the weights of different feature maps for pavement crack image segmentation. Yuan et al. [45] proposed OCR-Net, which uses object contextual feature representation for contextual information extraction based on object regions, thus explicitly enhancing object information and achieving good results on several mainstream semantic segmentation databases. Zhou et al. [46] explored an exemplar-based regime which provides a nonparametric segmentation framework based on non-learnable prototypes, where several typical points in the embedding space are selected for class prototypical representation, and distance to the prototypes determines how a pixel sample is classified. For deep learning models, there has been a bottleneck over the years to acquire sufficient ground-truth supervision, especially for segmentation tasks that require pixel-level annotations. Zhou et al. [47] proposed a group-wise learning framework for weakly supervised semantic segmentation that explicitly encodes semantic dependencies in a group of images to discover a rich semantic context for estimating more reliable pseudo ground truths, which are subsequently employed to train more effective segmentation models. König et al. [48] proposed a weakly supervised approach for crack segmentation that leverages a CNN classifier to create a rough crack localization map. The map was fused with a thresholding-based approach to segment the mostly darker crack pixels, and the pseudo labels were used to train the standard CNN for surface crack segmentation.

3. Proposed Method

In this paper, we propose a multi-scale contextual information enhancement network (MCIE-Net) for crack segmentation, and its structure is shown in Figure 1. In crack segmentation, there are strict requirements to accurately localize fine-grained spatial information, such as crack edges, which means that the network must extract sufficient figurative spatial information, as well as abstract semantic information. Since the plain skip connection of U-Net cannot fully utilize the multi-level features information generated in the encoder, MCIE-Net designs a new skip connection structure, so that the deep layer of the decoder network can aggregate feature information from multiple shallower layers. By fusing the multi-scale information, the decoder can obtain a richer representation of semantic information, and, more importantly, the decoder can make full use of the low-level features to improve the network's ability to restore detailed spatial information. Due to the limitation of fixed sampling scales, the plain convolutional neural networks cannot provide enough contextual scene information, which leads to the weak ability of the network to distinguish the segmented object from the interference information in the background. Therefore, we propose the contextual feature enhancement module, through

which the network can extract contextual features at multiple scales and obtain a richer representation of contextual scene information. The contextual feature enhancement module is placed in the deep layers of the decoder network instead of all layers, because the shallow layers do not explore the contextual features and the feature enhancement methods may dilute the fine-grained spatial information in them, affecting the network's segmentation capability. Our experiments in Section 4.5.2 demonstrate that when both performance and efficiency are taken into account, it is the best choice to place the module in the fourth and fifth layers of the decoder network. Additionally, in order to better extract the diverse morphological features of cracks, we improve the feature extraction network by introducing the deformable convolution into the encoder network to change the size of the receptive field through adaptive sampling area adjustment. Similar to [49], we do not replace all the layers of the encoder network with deformable convolution layers because there will be a limited improvement in network performance while using the deformable convolution in the shallow layers, but a significant increase in computational cost. Our experiments in Section 4.5.3 show that replacing the plain convolution of the fourth and fifth layers of the encoder network with deformable convolution is the optimal choice to achieve both performance and efficiency.

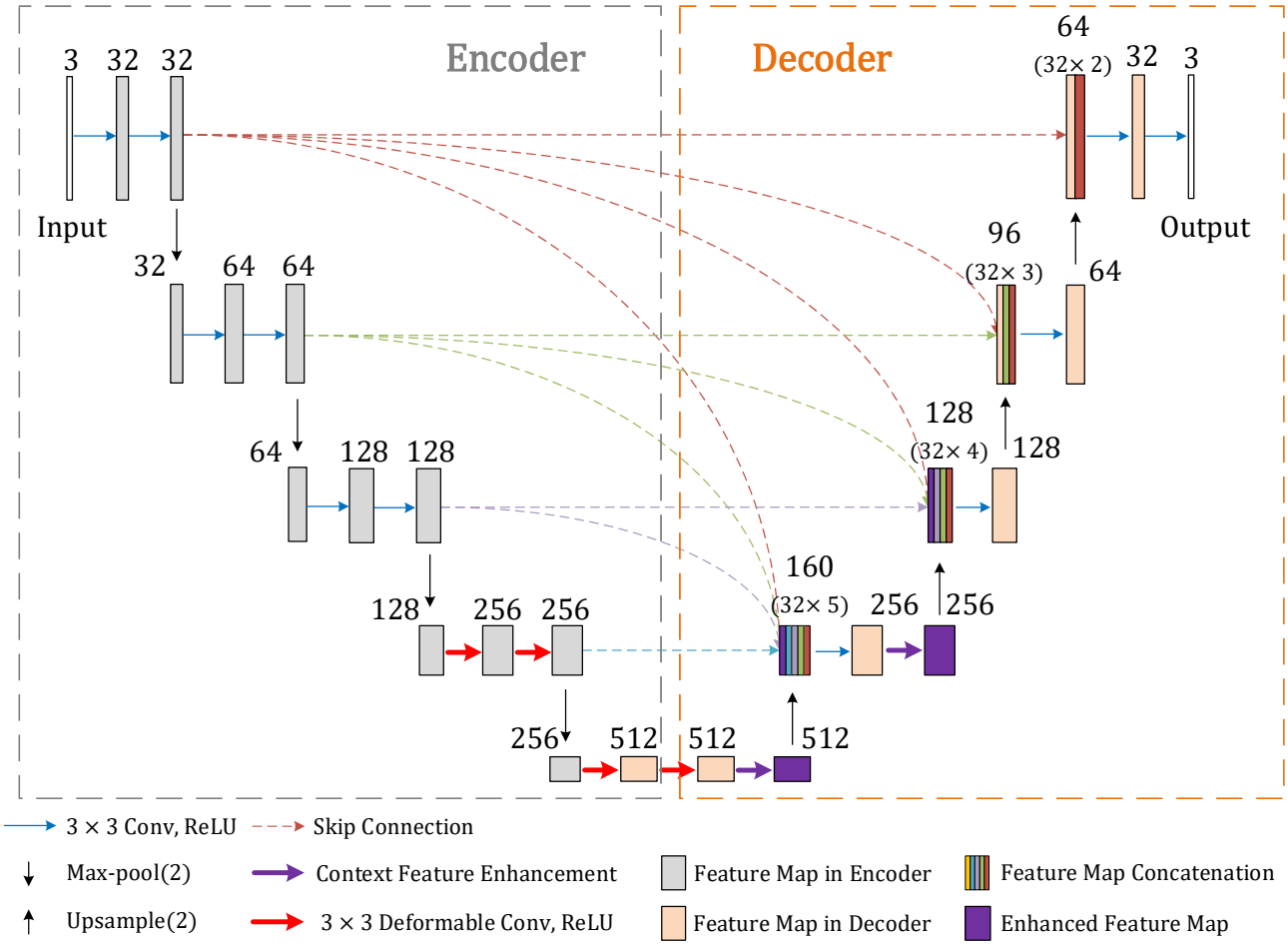

Figure 1. An illustration of the MCIE-Net architecture.

3.1. Multi-Scale Skip Connection Structure

The encoder network of U-Net has a distinct hierarchical character, and different stages of its convolutional layers obtain diverse meaningful features. The shallow layers keep abundant structure information, while the deep layers obtain more abstract features which play a crucial role in object recognition. Due to the loss of spatial information caused by the down-sampling operation in the coding process, it is not enough to only

use deep coarse feature maps to obtain fine segmentation results. Specifically, in the application of crack segmentation, the slender and tortuous structural characteristics of the cracks place high demands on the edge prediction capability of the model. In order to better restore the spatial information of cracks, we propose a multi-scale skip connection method to connect the encoder and decoder of U-Net. By feeding multi-scale feature maps to the decoder through the multi-scale skip connection, the network can better capture fine-grained spatial features.

Figure 2 shows four different skip connection structures. Among them, the plain skip connection and the densely connected skip connection can only deliver information at the same level; thus, the utilization of multi-level feature information generated in the encoder is relatively limited. In contrast, our multi-scale skip connection can provide the decoder with multiple feature maps generated in different layers of encoder network, enabling each layer of the decoder to learn rich fine-grained structural features. Compared with the full-scale skip connection, our multi-scale skip connection has the following two improvements. First, the multi-scale skip connection does not build the intra-connection between the decoder layers. The deep feature maps in the decoder do not contain spatial information, and the semantic information they contain can be effectively utilized through the backbone network; thus, there is no need to build additional connection paths. In addition, after concatenating the feature maps from different sources, we adjust the number of channels of the feature maps to be consistent with the same layer of feature maps in the encoder. The experiments in Section 4.5.1 demonstrate that the two improvements in this paper can significantly reduce the number of network parameters and lower the computational cost compared with full-scale skip connection.

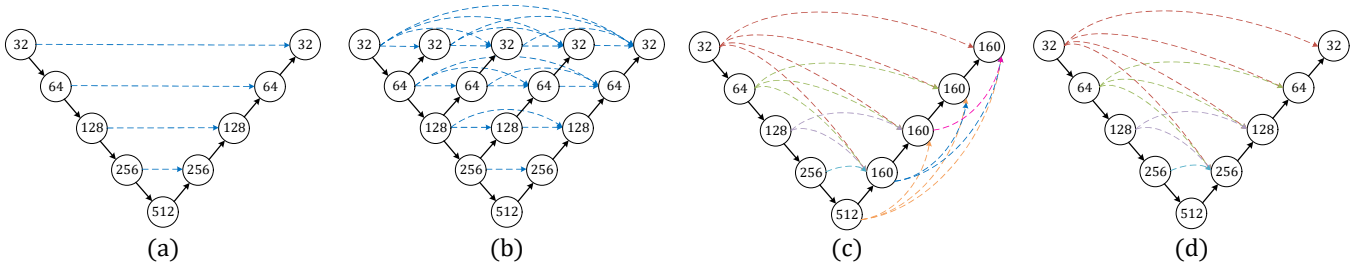

Figure 2. Comparison of different architectures of skip connections. (**a**) Plain skip connection, (**b**) densely connected skip connection, (**c**) full-scale skip connection, and (**d**) multi-scale skip connection.

As an example, Figure 3 illustrates how to construct the feature map X_{De}^{4} (the fourth layer of the decoder). The decoder receives the feature map X_{En}^{4} from the same encoder layer and X_{De}^{5} from a deeper decoder layer; meanwhile, the feature maps of X_{En}^{1}, X_{En}^{2}, and X_{En}^{3}, generated in several shallower layers of the encoder, are also transmitted through the skip connection paths. Furthermore, while the plain skip connection directly concatenates feature maps from different sources, we unify the number of channels of each feature map before concatenating, so that the superfluous information can be reduced [29]. Lastly, a feature aggregation mechanism, which consists of a convolution operation, a batch normalization, and a ReLU activation function, is used to seamlessly merge the abundant information owned by the concatenated feature map.

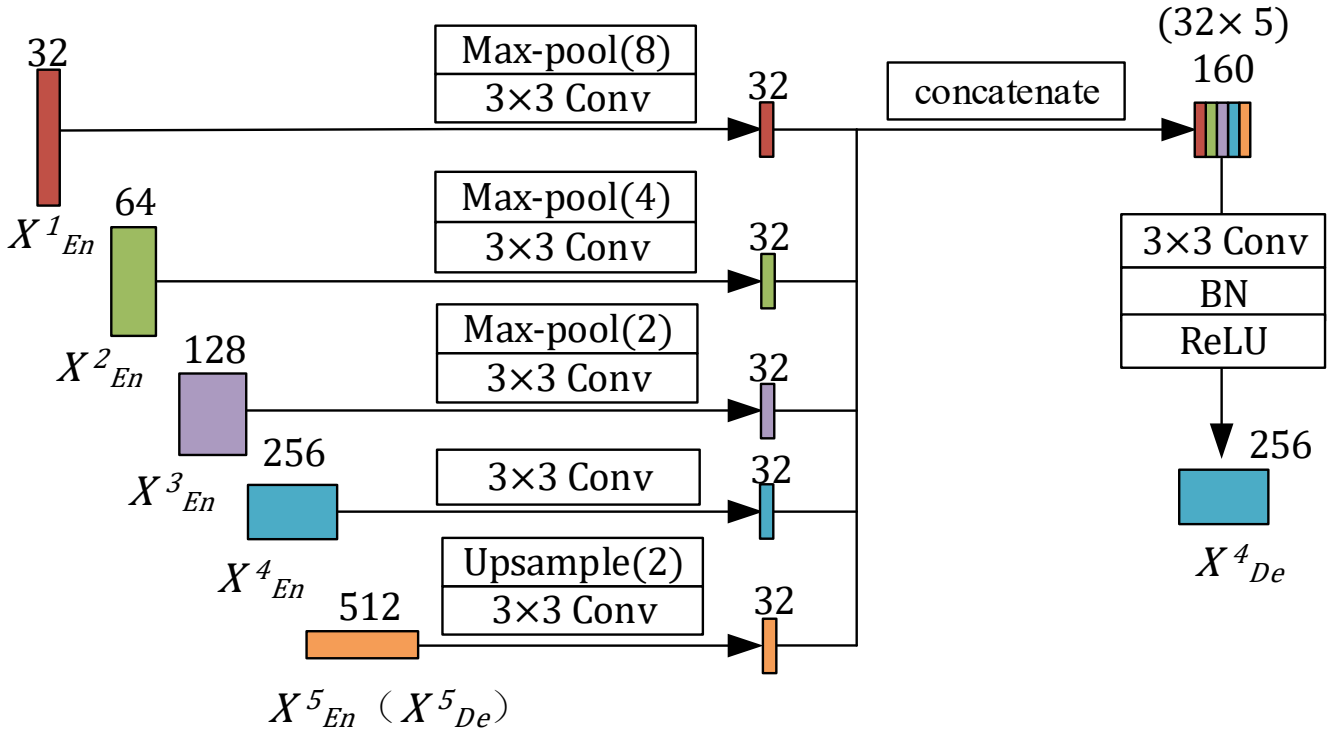

Figure 3. An illustration of how to construct the feature map of the fourth decoder layer.

3.2. Contextual Feature Enhancement Module

In the practical application of crack segmentation, some complex and variable environmental factors, such as shadow and pitting, are very likely to interfere with the recognition of cracks and lead to the occurrence of segmentation error. This is because the pixel characteristics of these interfering factors are very similar to the cracks and are not easily discernible. In order to accurately distinguish cracks from other interfering factors, it is necessary to obtain accurate scene perception with the help of sufficient contextual information [30]. Based on this, this paper introduces a contextual feature enhancement module (CFEM) to enhance the deep layers of the decoder network to obtain sufficient contextual information.

Figure 4 shows the structure of the contextual feature enhancement module, which mainly includes two sub-modules of a pyramidal pooling module and a squeeze-and-excitation block. In the convolutional neural network, the size of the receptive field can roughly indicate the extraction of contextual information. Theoretically, convolutional neural networks can expand the receptive field by increasing the network depth, and thus obtain a larger range of contextual information, but it is shown in [30] that the actual receptive field of deep neural networks is much smaller than the theoretical value. To address this problem, global average pooling is often utilized as a typical global contextual prior model in many tasks such as image classification [50,51]. However, in semantic segmentation tasks, this strategy is not enough to cover the necessary contextual information of the complex scene images. Therefore, the pyramid pooling module down-samples the feature map X_1 to be enhanced by averaging pooling to four scales, and the sizes are 1×1, 2×2, 3×3, and 6×6, respectively. The small feature map of each size contains the contextual feature information extracted at that scale. After adjusting the number of channels and bilinear interpolation up-sampling, the four new feature maps are restored to their original size. Then, the four feature maps are concatenated with X_1 as the enhanced feature map X_2. Note that the number of channels of the feature map X_2 becomes

twice that of the original feature map X_1, which is not adapted to the symmetrical structure of our U-shaped network. So, we reduce it to half via a convolution operation and then obtain feature map X_3.

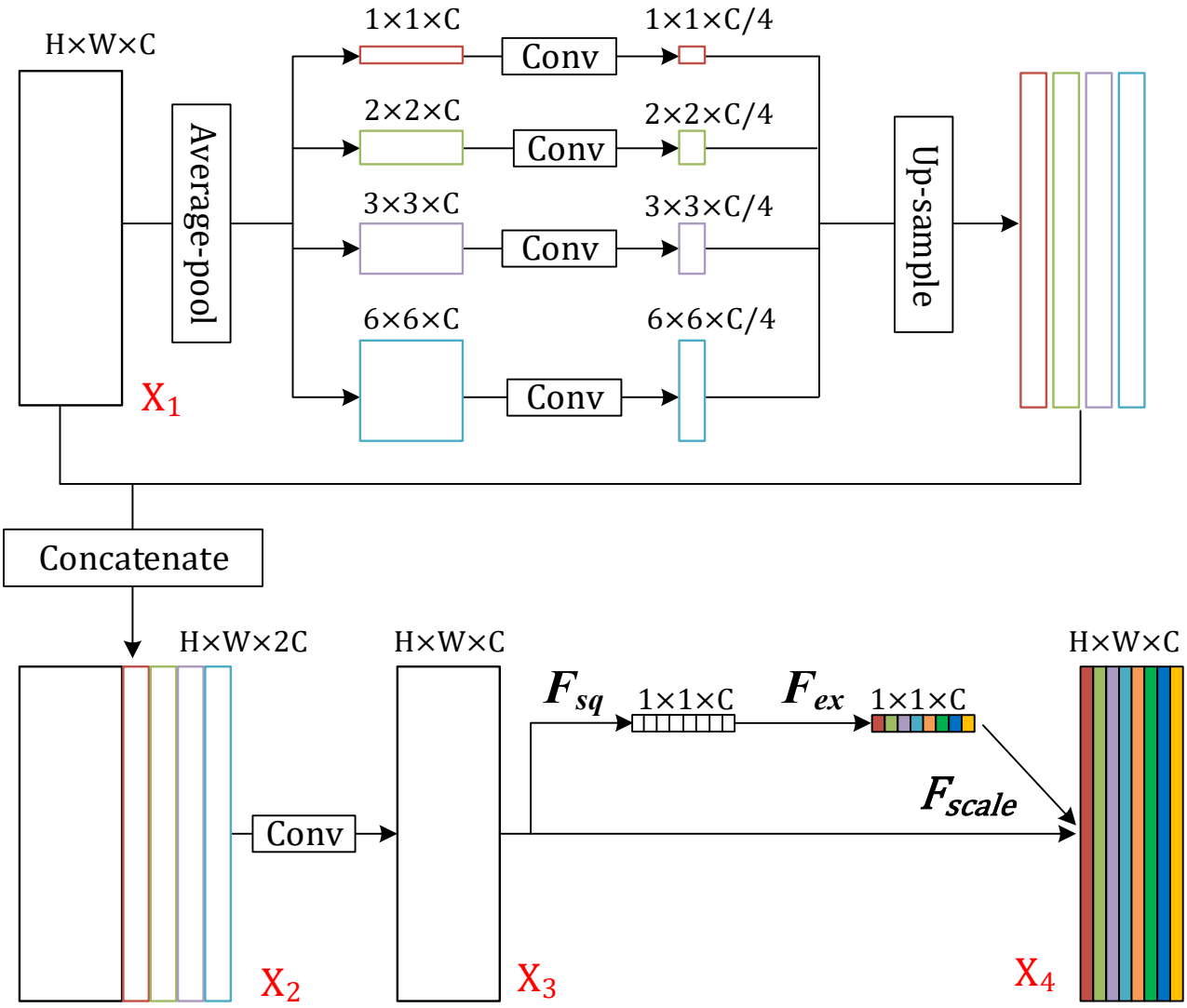

Figure 4. An illustration of the context feature enhancement module.

In order to make effective use of the rich semantic information owned by the enhanced feature map X_3, a channel attention mechanism is performed to recalibrate the channel weights, which selectively emphasize the important features and suppress unimportant features by learning the global information. Specifically, the channel attention mechanism is a squeeze-and-excitation block consisting of three operations: squeeze F_{sq}, excitation F_{ex}, and the channel-wise multiplication F_{scale} [52]. The operation F_{sq} squeezes the feature map to $1 \times 1 \times C$ with global average pooling to achieve global information embedding. The operation F_{ex} transmits the squeezed feature map through a fully connected layer, a RELU activation layer, a fully connected layer, and a sigmoid activation layer in turn to capture the dependency information on the channel dimension. Then, the output $1 \times 1 \times C$ feature map will be used as the weight matrix to recalibrate the channel weights of feature map X_3 in the operation F_{scale}, and finally we obtain the contextual feature enhanced feature map X_4. The channel attention mechanism can guide the network to focus on important channel information, making it more sensitive to critical contextual feature information.

3.3. Introduction of Deformable Convolution

The high-level CNN layers encode the semantic information over spatial locations [49], and different locations may correspond to objects with different scales or deformation. However, the receptive field size of the same CNN layer is invariable, which can affect the network to extract features. Specifically, in the application of crack segmentation, the cracks have diverse morphologies and different orientations. Therefore, it is difficult for the traditional CNNs limited by fixed geometric structures to learn all the morphological features of cracks.

To address this problem, we replace some of the standard convolutions in the encoder network with deformable convolutions, which can better capture various morphological information of cracks by adaptively adjusting the receptive field sizes. In the standard convolution, given the input feature map X, for each location p_0 on the output feature map Y, we have

$$Y(p_0) = \sum_{p_n \in R} \omega(p_n) \cdot X(p_0 + p_n) \quad (1)$$

where $\omega(p_n)$ denotes the weight of the convolution kernel at p_n, p_n enumerates the locations in R, and R is the full set of sampling points. The size of the sampling region of standard convolution is fixed and determined by the convolution kernel size. In deformable convolution, R is unchanged, but the sampling region is augmented by sampling offsets. As illustrated in Figure 5, the sampling offsets are obtained by a convolutional layer over the same input feature map. The spatial resolution of the output offset is the same as the input feature map, while its channel dimension is $2N$, corresponding to N 2D offsets. Therefore, in deformable convolution, Equation (1) becomes

$$Y(p_0) = \sum_{p_n \in R} \omega(p_n) \cdot X(p_0 + p_n + \Delta p_n) \quad (2)$$

where Δp_n is the sampling offset. The introduction of deformable convolution allows the network to adaptively adjust the sampling area and receptive field size to improve its ability to extract crack features.

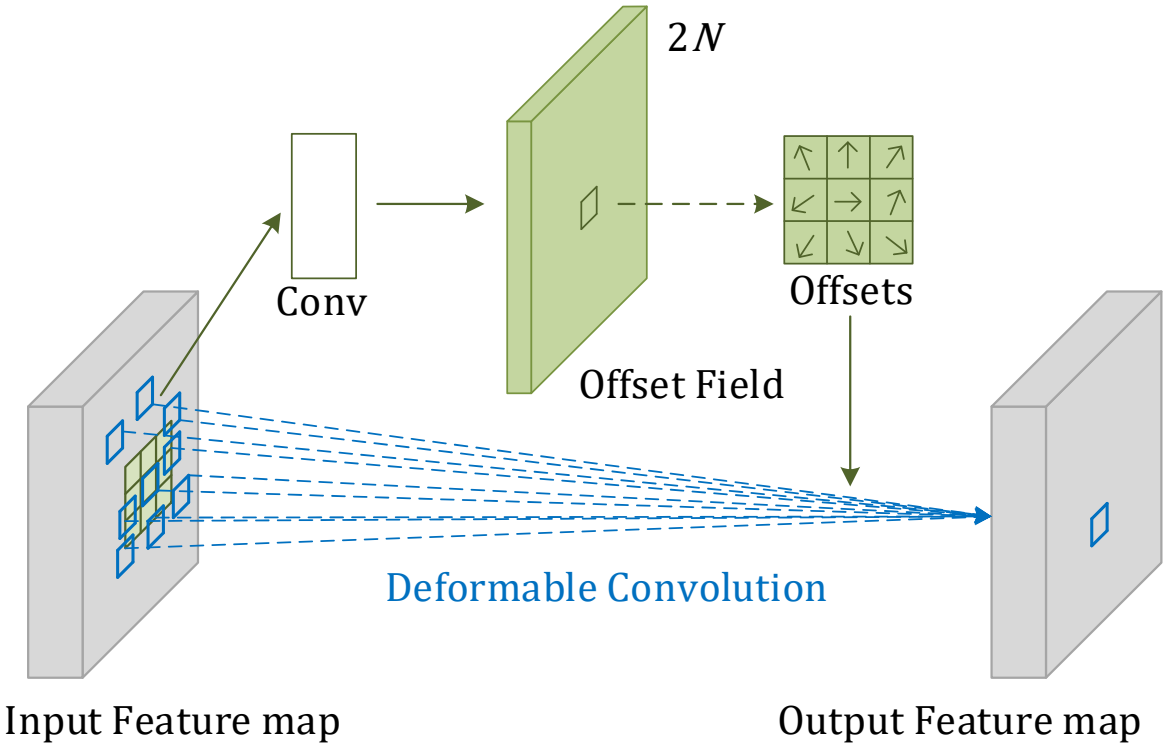

Figure 5. An illustration of the 3 × 3 deformable convolution.

4. Experiments

4.1. Training Configuration

In the crack segmentation task, since pixels can only be assigned a probability of being or not being part of a crack, it can be viewed as a binary classification problem. The binary cross-entropy loss (BCELoss) is commonly used for binary segmentation tasks. However, this loss function is not suitable for crack segmentation, because in the crack images, there are far more non-crack pixels than crack pixels, which may cause the network to prefer to segment the background pixels. We use the sum of binary cross-entropy and Dice loss as the loss function, called BCEDiceLoss. The BCELoss can be calculated as:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left(t_i log^{p_i} + (1-t_i) log^{(1-p_i)}\right) \quad (3)$$

where N is the number of pixels in the image, t_i is the truth value of the pixel, and p_i is the predicted value of the pixel. The Dice Loss can be expressed as:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} t_i p_i + \varepsilon}{\sum_{i=1}^{N} t_i + \sum_{i=1}^{N} p_i + \varepsilon} \quad (4)$$

where ε is the smoothing factor for preventing the denominator from being 0. Therefore, the BCEDiceLoss we use can be calculated as:

$$L_{BCEDice} = 1 - \frac{2\sum_{i=1}^{N} t_i p_i + \varepsilon}{\sum_{i=1}^{N} t_i + \sum_{i=1}^{N} p_i + \varepsilon} - \frac{1}{N}\sum_{i=1}^{N}\left(t_i log^{p_i} + (1-t_i) log^{(1-p_i)}\right) \quad (5)$$

where we set the smoothing factor ε to $1e^{-5}$.

We use the stochastic gradient descent with momentum as the optimizer to minimize the loss function, and the momentum is set as 0.9. The initial learning rate is $1e^{-3}$, and the poly learning rate strategy is used to adjust the learning rate while training.

Our experiments are conducted on a system with an NVIDIA RTX A4000 GPU(Manufactured by Nvidia in Santa Clara, California, the United States) and an Intel Xeon Gold 5320 CPU(Manufactured by Intel in Santa Clara, California, the United States). The software environment of the system is Python 3.6 and Pytorch 1.7.

4.2. Datasets and Metrics

We verify the effectiveness of our method on three publicly available crack datasets: DeepCrack-DB (proposed in [25]), CFD (proposed in [37]), and CCSD [53]. There are 537 images in the dataset DeepCrack-DB with a resolution of 544×384. These images contain various cracks of roads, walls, bridges, and so on. The images are relatively clear, but the background environment is complex with many kinds of disturbing factors. We split the dataset DeepCrack-DB randomly in a ratio of 8:2, resulting in 429 images in the training dataset and 108 images in the test dataset. The dataset CFD has 118 images taken from road surfaces with a resolution of 480×320. The images are blurred and the cracks are very thin, making segmentation more difficult. We remove some images from the dataset because they have obvious errors in the annotation. The remaining images are randomly partitioned in a ratio of 8:2, resulting in a training dataset with 89 images and a test dataset with 23 images. There are 458 high-resolution images in the dataset CCSD and we resize them to the resolution of 512×384. These images are taken approximately 1 m away from the surfaces and the concrete surfaces have variation in terms of surface finishes (exposed, plastering, and paint). We split the dataset CCSD randomly in a ratio of 8:2, resulting in 366 images in the training dataset and 92 images in the test dataset.

Additionally, random image augmentation methods are adopted to each image during the network training, including methods such as rotations, flips, and brightness shifts. Some of the images of the three datasets are shown in Figure 6.

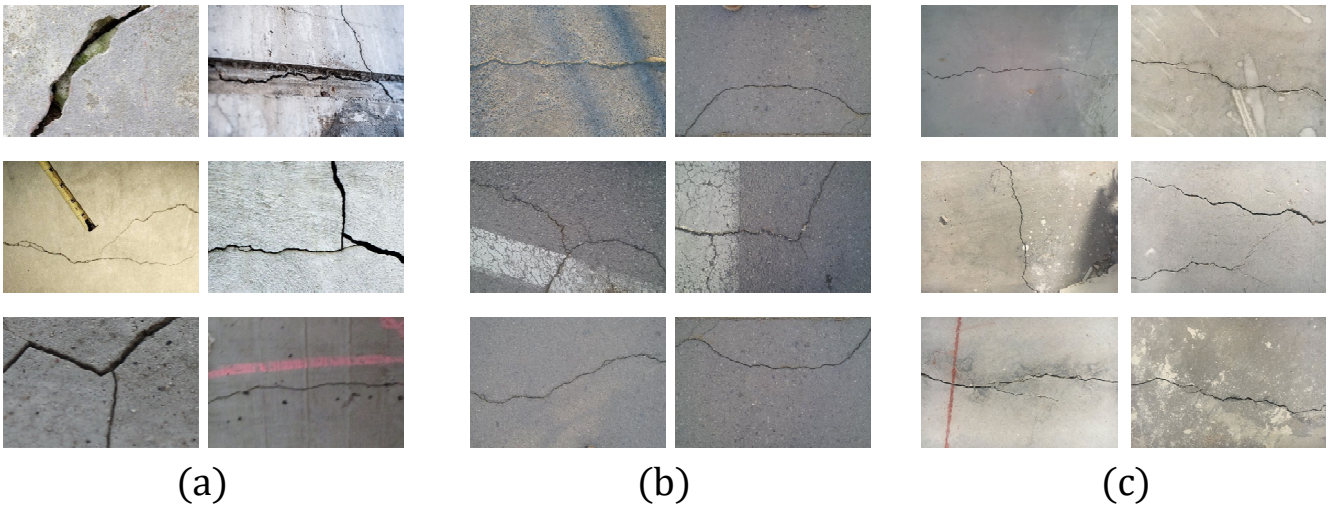

Figure 6. Crack images of the two datasets. (**a**) DeepCrack-DB, (**b**) CFD, (**c**) CCSD.

We select intersection over union (IoU), precision, recall, and F1 score as the evaluation metrics. Their precise definitions are:

$$IoU = \frac{TP}{TP + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

where TP denotes the number of true positives, FP denotes the number of false positives, and FN denotes the number of false negatives. Considering that the images in dataset CFD are blurred and there are transition regions between the crack pixels and the non-crack pixels in the subjectively labeled ground truth, the two pixel points around the labeled crack pixels are also considered as TP. It should be noted that this evaluation method has been commonly used in other studies [24,27,37,54].

4.3. Comparison with Other Methods

Our method is compared with five other deep-learning-based methods which are: DeepCrack [25], DeepLab v3+ [43], U-Net, UNet++, and Unet 3+. DeepCrack is one of the classical networks of crack segmentation. It consists of a FCN backbone and a deep supervision mechanism, and its prediction results are refined by the guided filtering and conditional random field (CRF) methods. DeepLabV3+ is the latest and optimal network framework of Google's DeepLab series [41,42,55]. It is based on null convolution, spatial pyramid pooling modules, and coding region-decoding region structures, and its further fusion of the underlying features with the higher-level features improves the segmentation accuracy. UNet++ and Unet 3+ are evolved from U-Net, and all three of them have a U-shaped network structure. They have very good performance and are widely used in various fields.

4.3.1. Results on DeepCrack-DB

Figure 7 shows the segmentation results of the six methods on some images in the DeepCrack-DB dataset, and Table 1 shows a quantitative comparison of the IoU, precision, recall, and F1 score metrics. As can be seen in Figure 7, all methods, except our MCIE-Net, exhibit varying degrees of false detection when there are interference factors in the image background region. As shown in Table 1, U-shaped networks such as U-Net and MCIE-Net achieve better results in crack segmentation, because their skip connection structures play an important role in capturing the fine structural features of the cracks, which avoids the problems of detail loss and segmentation discontinuities. MCIE-Net further enhances the contextual semantic information extraction and improves its adaptability to complex background conditions; thus, it achieves the best results in all four metrics, which are 91.28%, 95.67%, 94.59%, and 94.82%, respectively.

Table 1. Performance comparison of different methods on the DeepCrack-DB dataset.

Models	IoU (%)	Precision (%)	Recall (%)	F1 Score (%)
DeepCrack	66.49	66.67	85.28	70.01
DeepLab v3+	75.95	86.18	85.04	84.52
U-Net	85.97	94.45	89.26	90.31
UNet++	87.68	94.82	91.89	92.36
Unet 3+	88.05	94.64	92.07	93.27
MCIE-Net	91.28	95.67	94.59	94.82

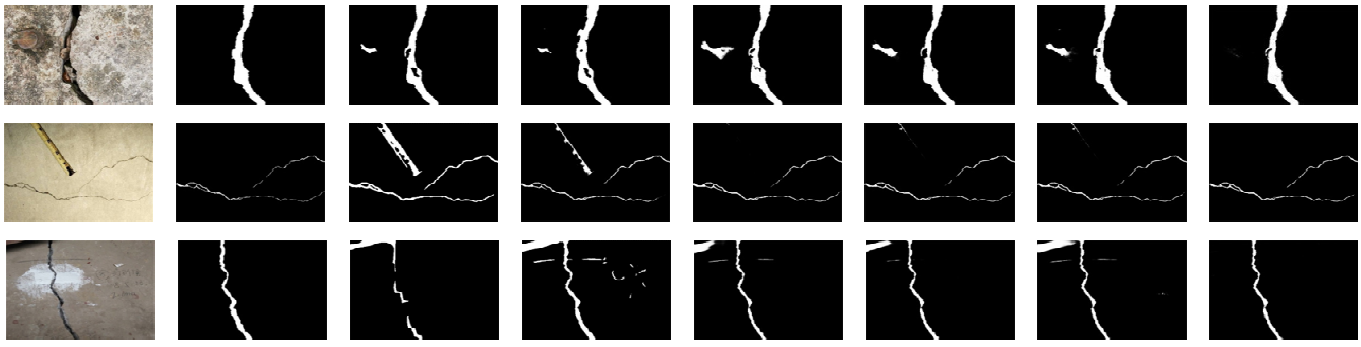

Figure 7. Sample results of using different methods on the DeepCrack-DB dataset. (**a**) Images, (**b**) ground truth, (**c**) DeepCrack, (**d**) DeepLab v3+, (**e**) U-Net, (**f**) UNet++, (**g**) Unet 3+, (**h**) MCIE-Net.

4.3.2. Results on CFD

Figure 8 shows the segmentation results of the six methods on some images in the CFD dataset, and Table 2 shows the quantitative comparison of the IoU, precision, recall, and F1 score metrics. As can be seen in Figure 8, DeepCrack and DeepLab v3+ perform poorly for predicting the crack boundary, making the crack region in the segmentation results wider than the ground truth. Unet 3+ and our MCIE-Net show a stronger ability to capture details and ensure the continuity in the crack area. Meanwhile, compared with other methods, our MCIE-Net can distinguish cracks and interference factors in background more effectively. As shown in Table 2, U-Net and Unet 3+ achieve the highest precision of 98.65% but a lower recall, while DeepCrack achieves higher recall of 88.62% but a lower precision. The F1 score takes into account both precision and recall, and the best result in this metric is achieved by our MCIE-Net, which is 89.47%. In addition, MCIE-Net also achieves the best result in the IoU metric, which is 81.32%.

Table 2. Performance comparison of different methods on the CFD dataset.

Models	IoU (%)	Precision (%)	Recall (%)	F1 Score (%)
DeepCrack	75.52	83.83	88.62	84.71
DeepLab v3+	68.9	79.02	84.32	81.31
U-Net	73.73	98.65	74.5	83.84
UNet++	76.42	98.45	77.32	85.39
Unet 3+	77.53	98.65	78.43	86.29
MCIE-Net	91.32	96.32	84.02	89.47

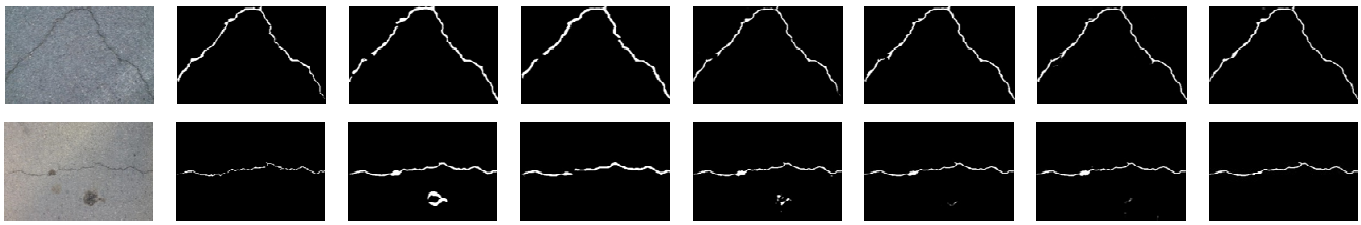

Figure 8. Sample results of using different methods on the CFD dataset. (**a**) Images, (**b**) ground truth, (**c**) DeepCrack, (**d**) DeepLab v3+, (**e**) U-Net, (**f**) UNet++, (**g**) Unet 3+, (**h**) MCIE-Net.

4.3.3. Results on CCSD

Figure 9 shows the segmentation results of the six methods on some images in the CCSD dataset, and Table 3 shows q quantitative comparison of the IoU, precision, re-call, and F1 score metrics. As can be seen in Figure 9, DeepCrack and DeepLab v3+ fail to obtain accurate crack edge segmentation results; thus, the crack region is rougher than that in the ground-truth images. The segmentation results show that the skip connection structure plays a very critical role in capturing the detailed information of the cracks, and our MCIE-Net has the best performance in identifying thin cracks which are hard to distinguish from the background. Meanwhile, as shown in Table 3, the MCIE-Net achieves the best results in all four metrics, which are 91.28%, 95.67%, 94.59%, and 94.82%, respectively.

Table 3. Performance comparison of different methods on the CCSD dataset.

Models	IoU (%)	Precision (%)	Recall (%)	F1 Score (%)
DeepCrack	78.77	83.76	90.86	87.37
DeepLab v3+	76.03	96	77.22	84.91
U-Net	85.59	98.22	86.08	91.31
UNet++	86.31	98.43	86.67	91.77
Unet 3+	89.61	98.32	90.1	93.75
MCIE-Net	90.45	98.46	92.43	94.22

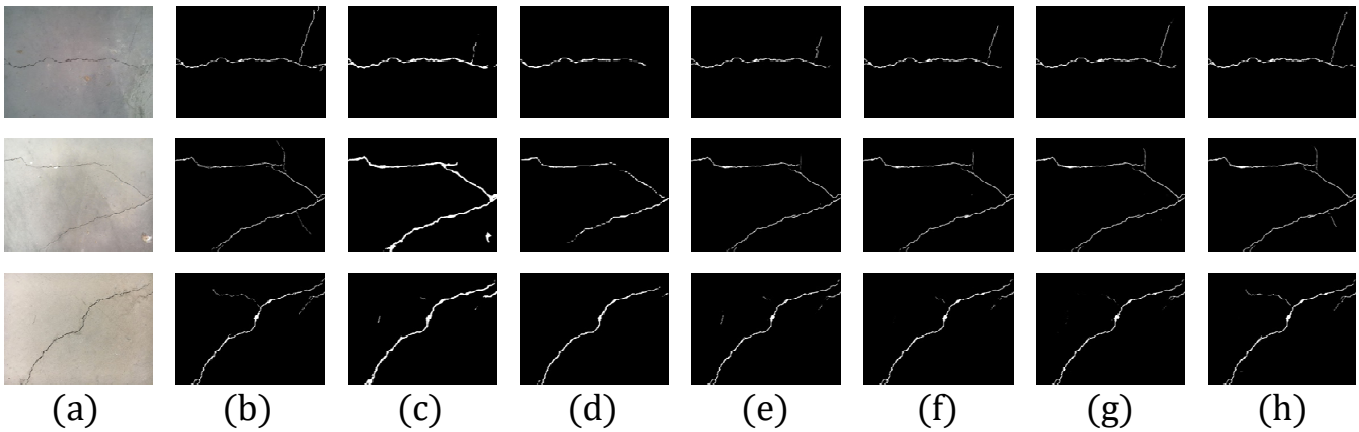

Figure 9. Sample results of using different methods on the CCSD dataset. (**a**) Images, (**b**) ground truth, (**c**) DeepCrack, (**d**) DeepLab v3+, (**e**) U-Net, (**f**) UNet++, (**g**) Unet 3+, (**h**) MCIE-Net.

4.3.4. Complexity and Efficiency Comparison

We further calculate the number of model parameters and computational complexity of the six models, and the results are shown in Table 4. The used metrics are parameters, FLOPs, and FPS, which refer to the number of parameters the network needs to learn, the number of floating point operations for a single forward propagation during training, and the number of images that can be predicted by the model per second, respectively. While calculating the FLOPs and FPS, we use images in dataset DeepCrack-DB with a resolution of 480×320, and the batch sizes are unified to be 4. As shown in Table 4, the parameters of the proposed model is relatively small and the FLOPs is the lowest. In terms of FPS, U-Net, which has the simplest network structure, performs best, processing more than 50 images per second. The UNet++ and Unet 3+ can process about 30 images per second, and the segmentation speed of our method is slightly reduced, about 24 images per second. Considering that the need for real-time crack segmentation is not very necessary, a slight decrease in segmentation efficiency in exchange for an increase in segmentation performance is desirable.

Table 4. Complexity and efficiency comparison of different models.

Models	Parameters (10^6)	FLOPs (10^9)	FPS (Frames/s)
DeepCrack	14.72	181.11	26.28
DeepLab v3+	16,385	187.69	12.4
U-Net	7.85	126.9	54.32
UNet++	9.16	314.13	30.48
Unet 3+	6.75	454.8	28.16
MCIE-Net	7.87	107.93	24.36

4.4. Ablation Experiments

To verify the effectiveness of improved schemes, such as multi-scale skip connection (MSSC), the contextual feature enhancement module (CFEM), and deformable convolution (DConv), an ablation study is carried out on the DeepCrack-DB dataset. IoU and F1 score metrics are used as the metrics and the results are shown in Table 5. From experiment 1 and experiment 2, we can learn that the IoU is improved by 1.36% and the F1 score is improved by 1.85% after using the multi-scale skip connection structure. This indicates that our multi-scale skip connection structure can help the network to better capture fine-grained spatial information, such as crack edges. Comparing experiments 2 and 3, it can be seen that the IoU and the F1 score improve by 2.48% and 1.58%, respectively, which indicates that the network can improve its ability to adapt to complex backgrounds and

distinguish between interference information and segmented objects. Comparing experiment 3 and experiment 4, the introduction of deformable convolution improves the IoU by 1.47% and the F1 score by 1.08%, which shows that the use of deformable convolution is also an effective method to improve the segmentation performance.

Table 5. Ablation study of the MCIE-Net.

Group	MSSC	CFEM	DConv	IoU (%)	F1 Score (%)
1				85.97	90.31
2	✓			87.33	92.16
3	✓	✓		89.81	93.74
4	✓	✓	✓	91.28	94.82

To highlight the advantages of our multi-scale skip connection in extracting spatial feature information and obtaining refined crack segmentation results, we select some images with tiny cracks from the test sets of DeepCrack-DB, CFD, and CCSD datasets to form the mixed test set A (MTSA), and evaluate the segmentation results of the images. The MTSA has a total of 51 images, of which 20 are selected from DeepCrack-DB, 7 are selected from CFD, and 24 are selected from CCSD. Our network is compared to Network A, whose structure is mostly the same as MCIE-Net, but without the multi-scale skip connection. Some representative samples are shown in Figure 10, and Table 6 shows the quantitative comparison of the IoU, precision, recall, and F1 score. As can be seen in Figure 10, Network A fails to extract refined spatial information and recognize the tiny cracks. Meanwhile, as shown in Table 6, our MCIE-Net achieves the best results for IoU, recall, and F1 score metrics, and the result for the precision metric is very close to that of Network A. It can be concluded that our multi-scale skip connection plays an important role in obtaining refined crack segmentation results.

Table 6. Performance comparison of our networks with or without MSSC on MTSA.

Models	IoU (%)	Precision (%)	Recall (%)	F1 Score (%)
Network A	74.83	90.09	78.79	81.76
MCIE-Net	83.38	89.86	91.35	90.48

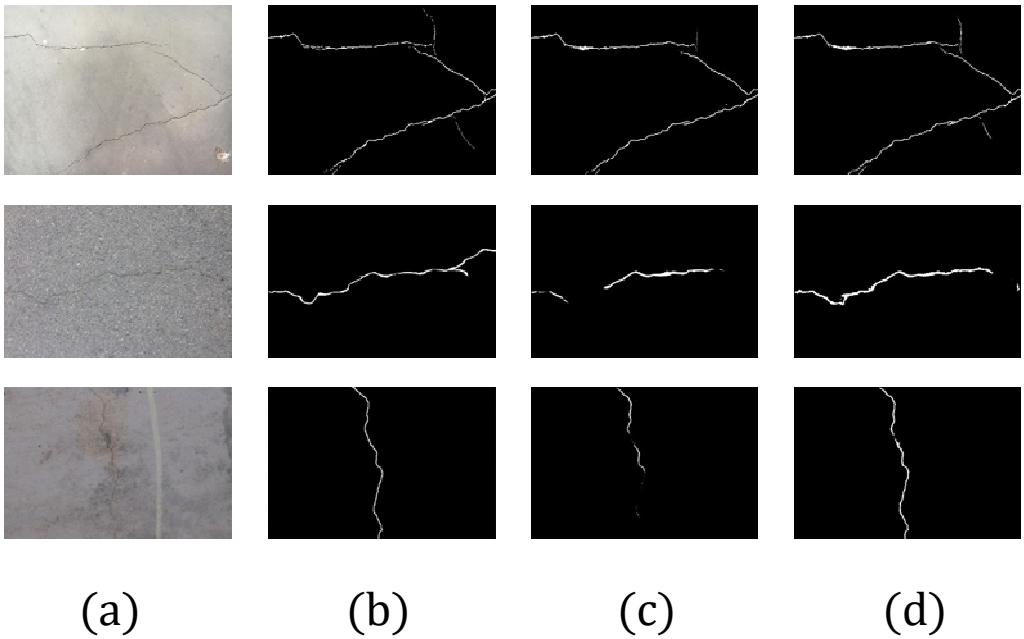

Figure 10. Sample results of using our networks with or without multi-scale skip connection on the mixed test set A. (**a**) Images, (**b**) ground truth, (**c**) Network A, (**d**) MCIE-Net.

We also select some images with background noise from the test sets of the three data sets mentioned above to form the mixed test set B (MTSB). There are 50 images in the MTSB, of which 31 are selected from A, 6 from B, and 13 from C. We evaluate the segmentation results of images in MTSB and our MCIE-Net is compared to Network B, whose structure is mostly the same as the propose network, but without the contextual feature enhancement module. Table 7 shows a quantitative comparison of the IoU, precision, recall, and F1 score metrics, in which our MCIE-Net achieves the best results in all four metrics. The segmentation results in Figure 11 highlight the advantage of our MCIE-Net in distinguishing between cracks and noise interference in the background, which is attributed to the contextual feature enhancement module.

Table 7. Performance comparison of our networks with or without CFEM on MTSB.

Models	IoU (%)	Precision (%)	Recall (%)	F1 Score (%)
Network B	73.82	82.07	90.32	84.82
MCIE-Net	86.5	95.05	90.74	92.41

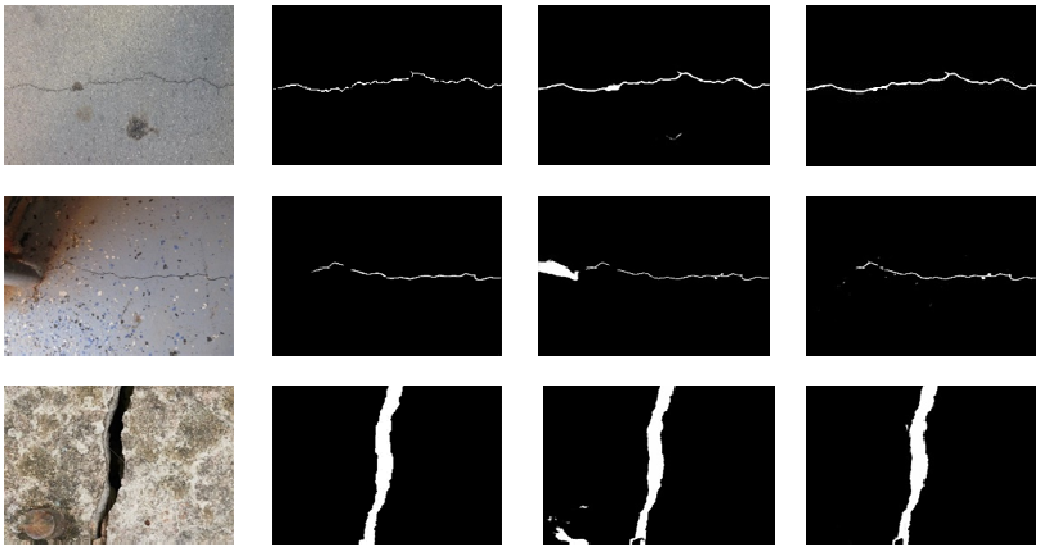

Figure 11. Sample results of using our networks with or without contextual feature enhancement module on the mixed test set B. (**a**) Images, (**b**) ground truth, (**c**) Network B, (**d**) MCIE-Net.

4.5. Effect of Different Settings of the Network Components

To verify that the parameter settings of our network structure are optimal, we perform several side-studies, further analyzing the different components of our network. All experiments are conducted on the DeepCrack-DB dataset, and the batch sizes are unified to be 4.

4.5.1. Comparison of Different Skip Connection Structures

We compare the segmentation performance and model complexity of our proposed multi-scale skip connection (MSSC) with the other three skip connection structures, namely the plain skip connection (PSC), the densely connected skip connection (DCSC), and the full-scale skip connection (FSSC), as mentioned in Section 3.1. Experiments are conducted using the four network models shown in Figure 12 with no other changes. IoU, F1 score, parameters, and FLOPs are used as the metrics and the results are shown in Table 8. It can be seen in the table that the segmentation performance of the multi-scale skip connection is better than that of the plain skip connection, and is very close to that of the

densely connected skip connection. While the full-scale skip connection achieves the best results in the IoU and F1 score metrics, its result in model complexity is the worst, with the highest FLOPs of 454.80×10^9. Our multi-scale skip connection achieves the best results in parameters and FLOPs, which are 5.82×10^6 and 101.58×10^9, respectively. Obviously, compared to the densely connected skip connection and full-scale skip connection, our multi-scale skip connection structure significantly reduces the computational cost at the cost of only a slight degradation in segmentation performance. Considering all the data in the Table 8, it can be concluded that the proposed multi-scale skip connection structure achieves a good balance between performance and model complexity.

Table 8. Segmentation performance and model complexity of the networks with different skip connection structures.

Structures	IoU (%)	F1 Score (%)	Parameters (10^6)	FLOPs (10^9)
PSC	85.97	90.31	7.85	126.90
DCSC	87.68	92.36	9.16	314.13
FSSC	88.05	93.27	6.75	454.80
MSSC	87.33	92.16	5.82	101.58

4.5.2. Analysis of the Contextual Feature Enhancement Module

The proposed contextual feature enhancement module includes two sub-modules: a pyramid poling module (PPM) and a squeeze-and-excitation block (SEB). In order to verify the effectiveness of the two sub-modules on model performance improvement, experiments are conducted. IoU and F1 score are used as the metrics and the results are shown in Table 9. The data in the table show that the introduction of the pyramid pooling module or the squeeze-and-excitation block alone can only improve the performance of the model to a small extent. The proposed contextual feature enhancement module can better improve the performance of the model with only a slight increase in FLOPs and number of model parameters, achieving the best results in both IoU and F1 scores.

Table 9. Ablation study of the contextual feature enhancement module.

Group	PPM	SEB	Parameters (10^6)	FLOPs (10^9)	IoU (%)	F1 Score (%)
1			5.95	102.15	88.62	92.42
2	✓		7.83	107.92	89.89	93.44
3		✓	5.99	102.16	89.24	93.05
4	✓	✓	7.88	107.93	91.28	94.82

Our contextual feature enhancement module (CFEM) can be plug-and-played into any layer of the decoder network. To further explore the best optimization scheme, we place CFEM at different layers of the decoder network while keeping the encoder network structure unchanged. F1 score, FPS, and FLOPs are used as the metrics, and the results are shown in Figure 12. It can be seen in the figure that as the number of network layers to which CFEM is applied increases, the network efficiency gradually decreases; however, the segmentation performance does not improve with it. This is because contextual features are semantic information, which is usually obtained by deep layers of the network. The shallow layers mainly explore structural features, and the multi-scale pooling and channel weight recalibration mechanism in CFEM will dilute the structural information in low-level feature maps, so that the network performance suffers. Taking both performance and efficiency into account, we apply the contextual feature enhancement module to the fourth and fifth layers of the decoder network.

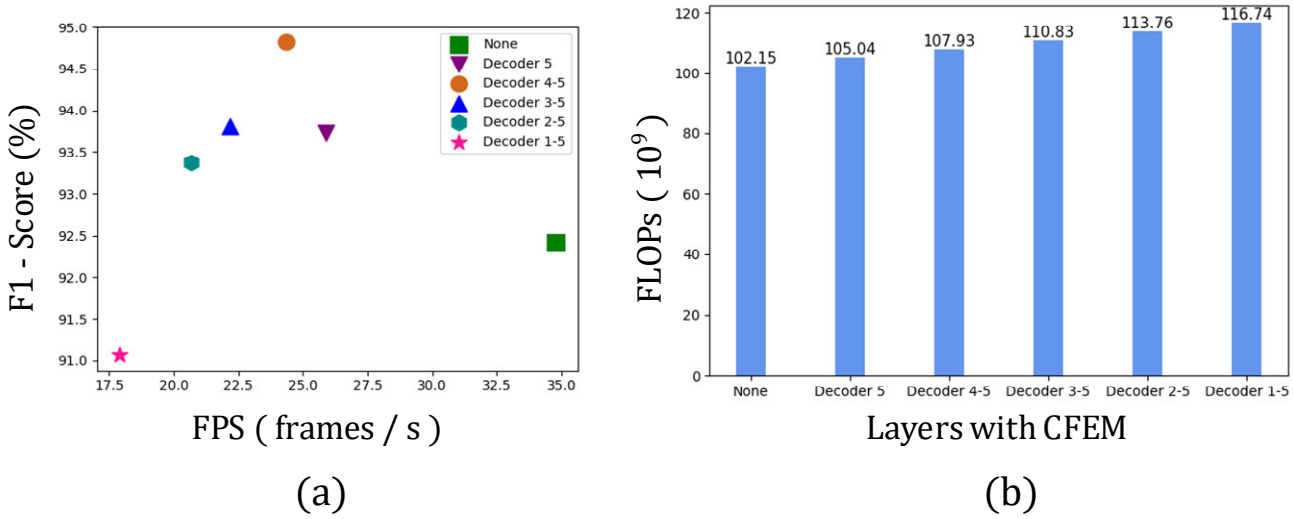

Figure 12. Segmentation performance and efficiency of the networks with the CFEM in different decoder layers. (**a**) Comparison of the F1 score and FPS of the six networks; (**b**) FLOPs of the six networks.

4.5.3. Deformable Convolution in Different Encoder Layers

The deformable convolution (DConv) adds 2D offsets, learned from the preceding feature maps via additional convolutional layers, to the regular grid sampling locations in the standard convolution, enabling free-form deformation of the sampling area. This module can readily replace their plain counterparts in any of the CNN layers. In order to explore the best optimization scheme for the encoder network while taking into account the operational efficiency, deformable convolutions are used to replace the standard ones in different layers of the encoder network, while the decoder network structure is kept unchanged. We evaluate the segmentation results with three metrics, including F1 score, FPS, and FLOPs, and the results are shown in Figure 13. It can be concluded from the figure that the introduction of deformable convolution has improved the segmentation performance of the network. However, as the number of network layers to which deformable convolution is applied continues to increase, the segmentation performance does not improve significantly. On the contrary, there is a significant decrease in FPS. This is because the feature maps in shallow layers of the network are large in size and it takes lots of time to generate the offset fields, leading to a decrease in the prediction efficiency. Considering both performance and efficiency, we apply deformable convolution to the fourth and fifth layers of the encoder network.

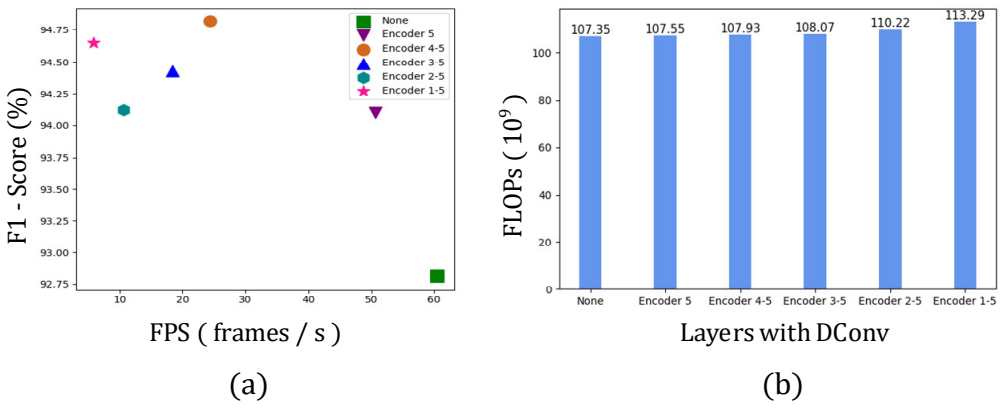

Figure 13. Segmentation performance and efficiency of the networks with deformable convolution in different encoder layers. (**a**) Comparison of the F1 score and FPS of the six networks; (**b**) FLOPs of the six networks.

5. Conclusions and Future Work

In this paper, a multi-scale contextual information enhancement network for crack segmentation is proposed. The redesigned skip connection structure enables the decoder to fuse feature information from multiple scales generated in the encoder, which improves the network's ability to capture the fine-grained spatial structure of cracks. The proposed contextual feature enhancement module allows our network to adapt to complex scenarios. In addition, we introduce deformable convolution in the encoder, which further improves the network's ability to extract crack features. The experiments on two public datasets demonstrate that the proposed MCIE-Net shows competitive performance on the crack segmentation task.

Regarding the direction of future work, on the one hand, we will continue to optimize the crack segmentation model to balance prediction accuracy and operational efficiency to further enhance its usefulness in engineering applications. On the other hand, the effectiveness of a deep learning model depends heavily on the number of training samples; however, it is very difficult to label a large number of crack images in detail. Therefore, ways of performing crack segmentation under small-dataset conditions represent another research direction.

Author Contributions: Conceptualization, L.Z. and Y.L.; methodology, L.Z. and Y.L.; validation, J.C. and G.W.; resources, G.W. and H.W.; data curation, L.Z. and H.W.; writing—original draft preparation, L.Z. and Y.L.; writing—review and editing, L.Z. and Y.L.; supervision, L.Z. and J.C.; project administration, G.W.; funding acquisition, G.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Nation Nature Science Foundation of China (grant nos. 42075191, 91847301, 92047203, and 52009080).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Salam, M.; Mathavan, S.; Kamal, K.; Rahman, M. Pavement crack detection using the Gabor filter. In Proceedings of the 16th international IEEE conference on intelligent transportation systems, The Hague, Netherlands, 6–9 October 2013. https://doi.org/10.1109/ITSC.2013.6728529.
2. 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. In Proceedings of the 2017 International Joint Conference on Neural Networks, Anchorage, Alaska, 14–19 May 2017. https://doi.org/10.1109/IJCNN.2017.7966101.
3. Zou, Q.; Zhang, Z.; Li, Q.; Qi, X.; Wang, Q.; Wang, S. Deepcrack: Learning hierarchical convolutional features for crack detection. *IEEE Trans. Image Process.* **2019**, *28*, 1498–1512. https://doi.org/10.1109/TIP.2018.2878966.
4. Li, G. Improved pavement distress detection based on contourlet transform and multi-direction morphological structuring elements. *Adv. Mater. Res.* **2012**, *466*, 371–375. https://doi.org/10.4028/www.scientific.net/amr.466-467.371.
5. Su, Z.; Guo, Y. Algorithm on Contourlet Domain in Detection of Road Cracks for Pavement Images. In Proceedings of the 2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science, Hong Kong, China, 10–12 August 2010. https://doi.org/10.1109/dcabes.2010.111.
6. Das, H.C.; Parhi, D.R. Detection of the Crack in Cantilever Structures Using Fuzzy Gaussian Inference Technique. *AIAA J.* **2009**, *47*, 105–115. https://doi.org/10.2514/1.35927.
7. Zhang, D.; Qu, S.; He, L.; Shi, S. Automatic ridgelet image enhancement algorithm for road crack image based on fuzzy entropy and fuzzy divergence. *Opt. Lasers Eng.* **2009**, *47*, 1216–1225. https://doi.org/10.1016/j.optlaseng.2009.05.014.
8. Zuo, Y.; Wang, G.; Zuo, C. Wavelet Packet Denoising for Pavement Surface Cracks Detection. In Proceedings of the 2008 International Conference on Computational Intelligence and Security, Suzhou, China, 13–17 December 2008. https://doi.org/10.1109/cis.2008.208.
9. Zhou, J.; Huang, P.; Chiang, F.P. Wavelet-Based Pavement Distress Classification. *Transp. Res. Rec. J. Transp. Res. Board* **2005**, *1940*, 89–98. https://doi.org/10.1177/0361198105194000111.
10. Kirschke, K.R.; Velinsky, S.A. Histogram-Based Approach for Automated Pavement-Crack Sensing. *J. Transp. Eng.* **1992**, *118*, 700–710. https://doi.org/10.1061/(asce)0733-947x(1992)118:5(700).

11. Cheng, H.D.; Shi, X.J.; Glazier, C. Real-Time Image Thresholding Based on Sample Space Reduction and Interpolation Approach. *J. Comput. Civ. Eng.* **2003**, *17*, 264–272. https://doi.org/10.1061/(asce)0887-3801(2003)17:4(264).
12. Zhao, H.; Qin, G.; Wang, X. Improvement of canny algorithm based on pavement edge detection. In Proceedings of the 2010 3rd International Congress on Image and Signal Processing, Yantai, China, 16–18 October 2010. https://doi.org/10.1109/cisp.2010.5646923.
13. Abdel-Qader, I.; Abudayyeh, O.; Kelly, M.E. Analysis of Edge-Detection Techniques for Crack Identification in Bridges. *J. Comput. Civ. Eng.* **2003**, *17*, 255–263. https://doi.org/10.1061/(asce)0887-3801(2003)17:4(255).
14. Jmour, N.; Zayen, S.; Abdelkrim, A. Convolutional neural networks for image classification. In Proceedings of the 2018 International Conference on Advanced Systems and Electric Technologies, Hammamet, Tunisia, 25 March 2018; pp. 397–402. https://doi.org/10.1109/ASET.2018.8379889.
15. He, K., Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015. https://doi.org/10.1109/iccv.2015.123.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. https://doi.org/10.1109/tpami.2016.2577031.
17. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
18. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015. https://doi.org/10.1109/CVPR.2015.7298965.
19. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
20. Cha, Y.J.; Choi, W.; Büyüköztürk, O. Deep learning-based crack damage detection using convolutional neural networks. *Comput. Aided Civ. Infrastruct. Eng.* **2017**, *32*, 361–378. https://doi.org/10.1111/mice.12263.
21. Pauly, L.; Hogg, D.; Fuentes, R.; Peel, H. Deeper networks for pavement crack detection. In Proceedings of the 34th The International Association for Automation and Robotics in Construction, Taipei, Taiwan, 28 June 2017. https://doi.org/10.22260/isarc2017/0066.
22. Tang, J.; Mao, Y.; Wang, J.; Wang, L. Multi-task Enhanced Dam Crack Image Detection Based on Faster R-CNN. In Proceedings of the 2019 IEEE 4th International Conference on Image, Vision and Computing, Xiamen, China, 5–7 July 2019. https://doi.org/10.1109/icivc47709.2019.8981093.
23. Suh, G.; Cha, Y.J. Deep faster R-CNN-based automated detection and localization of multiple types of damage. In Proceedings of the Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2018, Denver, CO, USA, 5–8 March 2018. https://doi.org/10.1117/12.2295954.
24. König, J.; Jenkins, M.D.; Mannion, M.; Barrie, P.; Morison, G. Optimized deep encoder-decoder methods for crack segmentation. *Digit. Signal Process.* **2021**, *108*, 102907. https://doi.org/10.1016/j.dsp.2020.102907.
25. Liu, Y.; Yao, J.; Lu, X.; Xie, R.; Li, L. DeepCrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing* **2019**, *338*, 139–153. https://doi.org/10.1016/j.neucom.2019.01.036.
26. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. https://doi.org/10.1109/tpami.2016.2644615.
27. Lau, S.L.H.; Chong, E.K.P.; Yang, X.; Wang, X. Automated Pavement Crack Segmentation Using U-Net-Based Convolutional Neural Network. *IEEE Access* **2020**, *8*, 114892–114899. https://doi.org/10.1109/access.2020.3003638.
28. Han, C.; Ma, T.; Huyan, J.; Huang, X.; Zhang, Y. CrackW-Net: A novel pavement crack image segmentation convolutional neural network. *IEEE Trans. Intell. Transp. Syst.* **2021**. https://doi.org/10.1109/TITS.2021.3095507.
29. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.W.; Wu, J. Unet 3+: A full-scale connected unet for medical image segmentation. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020. https://doi.org/10.1109/icassp40776.2020.9053405.
30. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017. https://doi.org/10.1109/cvpr.2017.660.
31. Chambon, S.; Moliard, J.M. Automatic Road Pavement Assessment with Image Processing: Review and Comparison. *Int. J. Geophys.* **2011**, *2011*, 989354. https://doi.org/10.1155/2011/989354.
32. Katakam, N. Pavement Crack Detection System Through Localized Thresholding. Doctoral dissertation, University of Toledo, Toledo, OH, USA, 2009.
33. Oliveira, H.; Correia, P.L. Automatic road crack segmentation using entropy and image dynamic thresholding. In Proceedings of the 2009 17th European Signal Processing Conference, Glasgow, Scotland, 25 August 2009.
34. Zhang, D.; Li, Q.; Chen, Y.; Cao, M.; He, L.; Zhang, B. An efficient and reliable coarse-to-fine approach for asphalt pavement crack detection. *Image Vis. Comput.* **2017**, *57*, 130–146. https://doi.org/10.1016/j.imavis.2016.11.018.
35. Wang, K.C.P.; Li, Q.; Gong, W. Wavelet-Based Pavement Distress Image Edge Detection with À Trous Algorithm. *Transp. Res. Rec. J. Transp. Res. Board* **2007**, *2024*, 73–81. https://doi.org/10.3141/2024-09.
36. Fernandes, K.; Ciobanu, L. Pavement pathologies classification using graph-based features. In Proceedings of the 2014 IEEE International Conference on Image Processing, Paris, France, 27–30 October 2014. https://doi.org/10.1109/icip.2014.7025159.

37. Shi, Y.; Cui, L.; Qi, Z.; Meng, F.; Chen, Z. Automatic Road Crack Detection Using Random Structured Forests. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 3434–3445. https://doi.org/10.1109/tits.2016.2552248.
38. Ren, Y.; Huang, J.; Hong, Z.; Lu, W.; Yin, J.; Zou, L.; Shen, X. Image-based concrete crack detection in tunnels using deep fully convolutional networks. *Constr. Build. Mater.* **2020**, *234*, 117367. https://doi.org/10.1016/j.conbuildmat.2019.117367.
39. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. https://doi.org/10.1109/tmi.2019.2959609.
40. Ran, R.; Xu, X.; Qiu, S.; Cui, X.; Wu, F. Crack-SegNet: Surface Crack Detection in Complex Background Using Encoder-Decoder Architecture. In Proceedings of the 2021 4th International Conference on Sensors, Signal and Image Processing, Nanjing, China, 15–17 October 2021. https://doi.org/10.1145/3502814.3502817.
41. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
42. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. https://doi.org/10.1109/tpami.2017.2699184.
43. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Lect. Notes Comput. Sci.* **2018**, 833–851. https://doi.org/10.1007/978-3-030-01234-2_49.
44. 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.* **2022**, *23*, 18392–18403. https://doi.org/10.1109/tits.2022.3158670.
45. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In Proceedings of the 16th European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020. https://doi.org/10.1007/978-3-030-58539-6_11.
46. Zhou, T.; Wang, W.; Konukoglu, E.; Goo, L.V. Rethinking Semantic Segmentation: A Prototype View. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022. https://doi.org/10.1109/cvpr52688.2022.00261.
47. Zhou, T.; Li, L.; Li, X.; Feng, C.M.; Li, J.; Shao, L. Group-Wise Learning for Weakly Supervised Semantic Segmentation. *IEEE Trans. Image Process.* **2022**, *31*, 799–811. https://doi.org/10.1109/tip.2021.3132834.
48. König, J.; Jenkins, M.D.; Mannion, M.; Barrie, P.; Morison, G. Weakly-Supervised Surface Crack Segmentation by Generating Pseudo-Labels Using Localization With a Classifier and Thresholding. *IEEE Trans. Intell. Transp. Syst.* **2022**, 1–12. https://doi.org/10.1109/TITS.2022.3204853.
49. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017. https://doi.org/10.1109/iccv.2017.89.
50. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santiago, Chile, 7–13 December 2015. https://doi.org/10.1109/CVPR.2015.7298594.
51. 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, Las Vegas, NV, USA, 26–30 June 2016. https://doi.org/10.1109/CVPR.2016.90.
52. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. https://doi.org/10.1109/CVPR.2018.00745.
53. Özgenel, Ç.F. *Concrete Crack Segmentation Dataset*, version 1; Mendeley Data. https://doi.org/10.17632/jwsn7tfbrp.1.
54. Fan, Z.; Li, C.; Chen, Y.; Di Mascio, P.; Chen, X.; Zhu, G.; Loprencipe, G. Ensemble of deep convolutional neural networks for automatic pavement crack detection and measurement. *Coatings* **2020**, *10*, 152. https://doi.org/10.3390/coatings10020152.
55. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.