machines

MDPI

Article

A Crack Defect Detection and Segmentation Method That Incorporates Attention Mechanism and Dimensional Decoupling

Lixin He [1,†,*], Wangwei Liu [1,†], Yiming Li [1], Handong Wang [1], Shenjie Cao [1] and Chengying Zhou [2]

[1] School of Artificial Intelligence and Big Data, Hefei University, Hefei 230601, China
[2] School of Electronic and Information Engineering, Anhui Jianzhu University, Hefei 230009, China
* Correspondence: hlxiniim@mail.ustc.edu.cn
† These authors contributed equally to this work.

Abstract: In this work, we propose a new crack image detection and segmentation method for addressing the issues regarding the poor detection of crack structures in certain complex background conditions, such as the light and shadow, and the easy-to-lose details in segmentation. This method can be categorized into two phases, where the first one is the coding phase. In this phase, the channel attention mechanism and crack characteristics, using the correlation channel with different scales increasing the network robustness and ability of feature extraction, have been introduced to decouple the channel dimension and space dimension. It also avoids underfitting caused by information redundancy during the jumping connection. In the second stage, i.e., the decoding stage, the spatial attention mechanism has been introduced to capture the crack edge information through the global maximum pooling and global average pooling of the high-dimensional features. Then, the correlation between the space and channel has been recovered through multiscale image information fusion to achieve accurate crack positioning. Furthermore, the Dice loss function has been employed to solve the problem of pixel imbalance between the categories. Finally, the proposed method has been tested and compared with existing methods. The experimental results illustrate that our method has a higher crack segmentation accuracy than existing methods. Furthermore, the mean intersection over the union ratio reaches 87.2% on the public dataset and 83.9% on the self-built dataset, and it has a better segmentation effect and richer details. It can solve the problem of crack image detection and segmentation under a complex background.

Keywords: crack detection; image segmentation; dimension decoupling; attention; multiscale

check for updates

Citation: He, L.; Liu, W.; Li, Y.; Wang, H.; Cao, S.; Zhou, C. A Crack Defect Detection and Segmentation Method That Incorporates Attention Mechanism and Dimensional Decoupling. *Machines* **2023**, *11*, 169. https://doi.org/10.3390/machines11020169

Academic Editor: Xiang Li

Received: 1 January 2023
Revised: 19 January 2023
Accepted: 21 January 2023
Published: 26 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cracks affect the beauty of concrete structures, reduce the durability of structures, and affect the structural safety. According to current practice, cracks are primarily detected by the manual inspection by professionals, which is time-consuming, subjective, and costly. Recently, the automatic crack detection has become a highly active research area [1]. Early researchers have employed the traditional machine learning algorithms to extract and identify crack features, and these methods require manually designed features that entails low detection accuracy.

Recently, CNNs have made significant breakthroughs in several fields such as target detection and image segmentation [2,3], which eliminate the need for manual feature extraction, thus achieving a higher accuracy and better robustness compared to traditional machine learning algorithms. Consequently, researchers started to apply it for the crack image classification task. Silva et al. [4] have used Visual Geometry Group (VGG16) as the base model and tuned the learning rate and the number of nodes in the fully connected layer to achieve good results in the crack classification task. Owing to the increasing demand, the classification task alone cannot meet the practical needs, and therefore, the crack

image detection and segmentation task have become the focus of research. However, the extraction of the semantic information of crack images in complex backgrounds is crucial, which affects the accuracy of detection and segmentation. Therefore, the fusion of semantic information under different modalities becomes the focus of research. For example, Tang et al. [5] have employed the migration learning for assisting in the network training and data enhancement, for the case of insufficient crack image samples that is combined with the front-end fusion approach. This is for the multimodal fusion to fuse the learned multiple feature data into a single data input to a multi-task enhanced RPN (Region Proposal Network) model to generate the candidate frames. The application of a multimodal fusion mechanism in the crack detection task effectively complements the semantic information extracted from different modalities, and it facilitates the removal of confusing candidate frames from images. The Fully Convolutional Network (FCN) network proposed by Long et al. [6] has shown great potential in semantic segmentation networks. Thereafter, several researchers have proposed certain improved semantic segmentation network models based on this approach, such as Jenkins et al. [7], who have used deep fully Convolutional Neural Networks (CNNs) to achieve pixel-level classification of cracked images. Furthermore, owing to the increasing difficulty of crack image segmentation arising from the variable application scenarios, the research is no longer limited to semantic information fusion in multimodality and has delved into the incorporation of multiscale information into multimodality. Therefore, Ronneberger et al. [8] have incorporated the multiscale information into the decoder stage of the feature extraction and restored the original image information through progressive up-sampling, which has achieved tremendous success in the field of medical image segmentation. Henceforth, Lau et al. [9] have replaced the encoder of U-Net with ResNet-34 network and added the channel attention mechanism to each module to obtain the crack feature extraction capability. To further obtain semantic information at different scales, Choi et al. [10] have incorporated a spatial pyramid pooling module in the decoder stage to fuse the semantic information at different scales, which is a medium-term fusion approach.

To achieve an accurate segmentation of the cracks under complex background, this work proposes a crack image detection and segmentation method that combines the attention and dimension decoupling. The following new design has been proposed to rectify the shortcomings of the existing methods.

(1) We consider the problems neglected by previous studies with respect to the semantic correlation between the high-level features of different modalities while performing multimodal fusion. Hence, this paper proposes a channel semantic information perception (CSIP) module that employs the channel attention mechanism [11,12] to learn the semantic information of the crack images on different channels, and it optimizes the combination of the semantic information between different channels through the weight parameters to improve the cross-channel semantic information feature extraction capabilities.

(2) We consider the problem where the previous studies have fused the multimodal features by equal-weight stitching or summation without attributing separate operations for the spatial and channel semantic information, thus leading to a bias in locating the location of cracks. Hence, this work proposes a dimensional decoupling module [13,14] that considers the channels and spaces separately, and it learns the encoding of semantic information of a specific category by different scale spaces and channels to solve the issue of deviation in crack location.

(3) This paper proposes a spatial semantic information perception (SSIP) module and a multimodal feature fusion (MMF) module with respect to the problem of detail information loss during up-sampling, where the spatial semantic information perception module employs the difference in the spatial position of the feature information at different scales. Furthermore, it employs the spatial attention mechanism [15] to extract the rich spatial position feature information at different scales. The multimodal feature fusion module employs the channel and spatial semantic information learned at different scales to fuse with the corresponding level of the encoder [16] and restores the loss and resolution

of the context information during up-sampling with respect to the coupled channel and spatial correlation.

(4) The Dice loss function, which tends to focus on the difference between pixels, has been employed to balance the prediction bias caused by unbalanced categories. This is to rectify the extremely unbalanced background and crack pixels in the dataset in this paper so that the model can better locate the segmentation position and achieve better segmentation results.

The remainder of this paper is organized as follows. Section 2 introduces the literature review of crack image segmentation. In Section 3, the methods proposed in this paper are introduced in detail. Section 4 discusses the results of the contrast test and the ablation test. Finally, Section 5 summarizes and clarifies the future research direction.

2. Related Work

Traditional machine learning algorithms. Recently, certain researchers have tried to use computer vision instead of manual vision for crack detection and obtained a series of results. Prasanna et al. [17] have proposed the automatic crack detection algorithm named spatially tuned robust multi-feature (STRUM) to achieve the collection and classification of bridge crack data. Shi et al. [18] have introduced random forest to generate a high-performance crack detector that can identify the cracks of arbitrary complexity. Amhaz et al. [19] have proposed an algorithm for the automatic detection of cracks in the 2D pavement images based on the minimum path selection in a natural way by considering the photometric and geometric features of the pavement images. Cord et al. [20] have performed an AdaBoost-based supervised learning method to capture the fluctuations in the texture information for the road defect detection using different texture information presented by the defective and non-defective pavement surfaces. These traditional machine learning methods, however, require the manual design of the features to be extracted. Alternatively, it is difficult for these algorithms to achieve the expected detection results while detecting road cracks under complex scenarios with changing illumination and background.

Deep learning methods. Since the deep learning has received increasing attention, the CNN models with automatic feature extraction have been employed for the classification and detection of crack images. Zhang et al. [21] have first applied a deep learning-based approach to the road crack detection problem for obtaining a better crack detection performance than that of the traditional machine learning algorithms. The work [22] has employed a faster R-CNN combined with Deep Convolutional Neural Network (DCNN) for crack detection. The method proposed by Fan et al. [23] can be divided into two steps as in the following. First, CNN is used for the classification of positive and negative samples, and then, an adaptive thresholding approach is used to extract the cracks for the positive sample crack images. The work [24] has employed DCNN and domain-specific heuristic post-processing techniques for crack detection in different scenarios. Zou et al. [25] have applied Long Short-Term Memory (LSTM) combined with residual networks to DCNN. Kim et al. [26] have proposed a classification algorithm to ascertain the presence of cracks from the surface images by extracting the crack candidate regions through an image binarization method, which is followed by an acceleration based on the robust features and CNN to construct a classification model.

In the crack segmentation task, the image segmentation method based on the full convolutional network (FCN) structure is a hotspot. While the classical CNN classifies the images through the fully connected layers after feature extraction by convolutional kernels, the FCNs classify the images at the pixel level by replacing all the final fully connected layers with the convolutional kernels, which is followed by the recovery of the extracted feature matrix to the size of the input image by deconvolution and up-sampling.

FCN, U-Net, and SegNet [6,8,27] are the representative models of the FCN structure. The U-Net network model is one type of FCN, and the model can be divided into two parts, viz., encoder and decoder. The encoder primarily takes the input image, through multiple convolutions and down-sampling, to obtain the high-dimensional abstract features and

sends them to the decoder. In the decoding stage, the feature maps at the same level as the encoding stage have been fused through multiple convolutions and up-sampling operations to recover the spatial information lost during the up-sampling process. Finally, the high-dimensional abstract features have been reduced to the same size as the original image size to realize the segmentation of the original image. Yang et al. [28] have integrated the contextual information with the low-level features of crack detection through the form of feature pyramids. Furthermore, they have employed weighting to rebalance the loss, thus ensuring better results. The work [29] has combined the attention mechanism with full CNN to achieve an end-to-end pixel-level crack segmentation. Ren et al. [30] have combined the spatial pyramid and jump connection module, based on SegNet, to improve the network with higher accuracy and generalization ability. Apart from the above methods, this paper has significantly improved the crack image segmentation accuracy by employing the spatial differences and channel semantic dependencies of multimodality, after dimensional decoupling, which is based on the full consideration of the semantic information at different scales.

3. Construction of the Network Model

3.1. Network Model Structure

To achieve an accurate detection and segmentation of cracks, we propose a crack image detection and segmentation method that combines the attention mechanism and dimensional decoupling for solutions of the problems of crack image feature extraction. This method is categorized into the encoding and decoding phases. In the encoding stage, a channel attention mechanism has been introduced in the channel semantic information perception (CSIP) module to enhance the crack information from the channel dimension and suppress irrelevant ones such as noise. Furthermore, the channel correlation and spatial correlation have been resolved in the dimension decoupling module. To avoid certain unnecessary channel and spatial correlation information from interfering with the decoding stage, for the coupling in the decoding stage, a multimodal feature fusion (MFF) module has been added to fuse with the corresponding level of the encoder to restore the lost and decoupled channels of context information during the up-sampling correlation with space. Then, in the spatial semantic information perception (SSIP) module, the crack edge information has been obtained through the global average pooling and global maximum pooling. Furthermore, it is combined with the spatial attention mechanism to strengthen the association of the context information in the decoding process and improve the segmentation accuracy. The Dice loss function, which focuses on the difference between pixels, has been employed to balance the prediction bias caused by the unbalanced categories so that the model can better locate the segmentation position and achieve better segmentation results. The model of this method is termed SEAD-U-Net. The overall structure is displayed in Figure 1. It has been divided into two parts, viz., the encoder and the decoder. The encoder and the decoder are each composed of a five-layer structure, and each layer structure is realized by different operations.

The output of each layer in the five-layer structure of the encoder is defined as $Z_E = \left[Z_E^1, \cdots, Z_E^5\right]$. To enhance the relationship between correlations of different channels, the recalibrated weight information can further improve the model's ability to extract crack features in the first layer of the encoder, after convolution extraction features, by using the channel attention mechanism to recalibrate the weight information of different channels. Furthermore, the output result is Z_E^1. The following three-layer structure further extracts the feature information through convolution and pooling, and the output results are Z_E^2, Z_E^3, and Z_E^4. To alleviate the disappearance of the gradient, skip connections have been used several times in this model, which generates certain unnecessary channel and spatial correlation information, thus making the feature analysis difficult and resulting in a reduced segmentation accuracy. Therefore, with respect to the fifth layer of the encoder in the structure, this work decouples the channel correlation and spatial correlation to ensure

that the subsequent decoding process focuses on analyzing the spatial information, and the output result is Z_E^5.

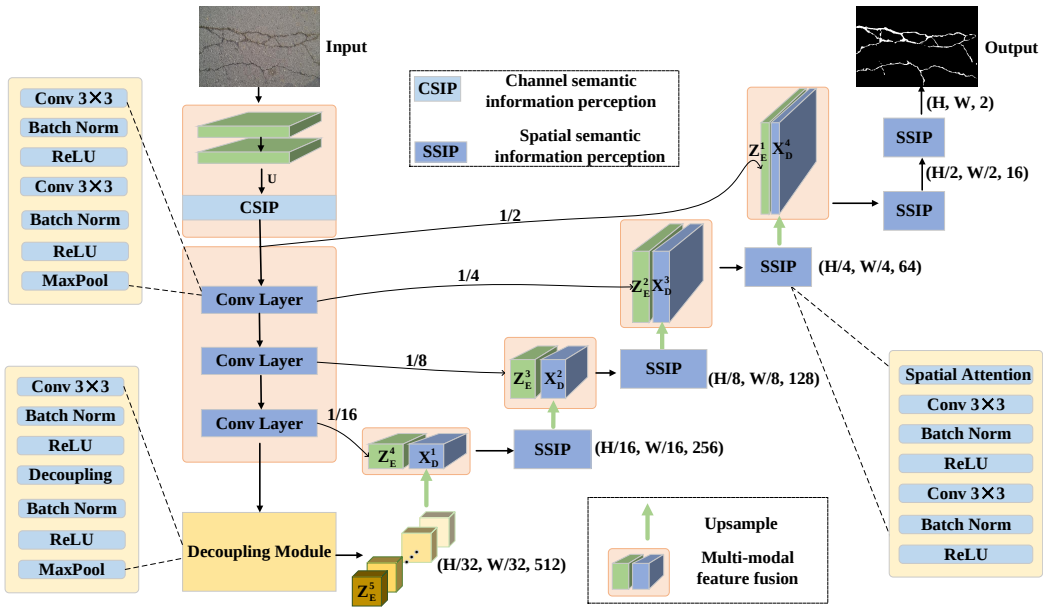


Figure 1. The structure of the proposed network for crack segmentation.

The decoder is also divided into five layers. The output of each layer is defined as $X_D = [X_D^1, \cdots, X_D^5]$. The first four layers continuously up-sample the deep features and fuse the shallow features to the image information through skip connections and make enhancements. In the crack image, owing to the highly uneven distribution of the number of pixels of each category, it is difficult for the model to capture the crack detail information while predicting the original image information. Therefore, a Spatial Attention Mechanism (SAM) has been added to each layer of the decoder that is used to capture the spatial correlation information during the up-sampling process. With respect to the specific operations, the global average pooling and global maximum pooling have been cascaded. Global average pooling can capture and distinguish between different categories in the feature map to avoid wrong segmentation. Global maximum pooling can sensitively capture the pixel change process to make the local detail features in the final segmentation map more complete. Then, the nonlinear relationship has been learned through the sigmoid function, and a non-mutually exclusive feature matrix is obtained, i.e., the weight matrix of the spatial attention mechanism through which the feature map is calibrated in the spatial dimension. When the model operates on spatial information, the channel and spatial correlations are supplemented by skip connections, which are then used as the input for the next layer. In the last layer, the final fracture segmentation map has been obtained through pixel-by-pixel prediction.

3.2. Encoder

Figure 2 shows the backbone network of the encoder, which is primarily divided into two parts. The first part adds an attention mechanism to the channel, re-calibrates the weights of different channels, and then enhances the crack information. The second part decouples the spatial dimension from the channel dimension and processes the channel and spatial information separately.

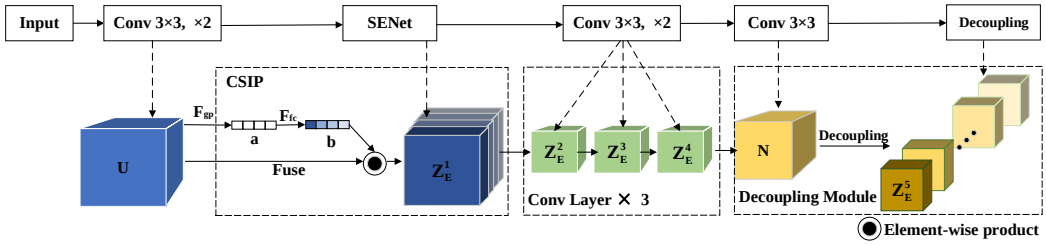


Figure 2. Encoder with CSIP and decoupling module.

Using the attention mechanism for the channel: In channel attention, we employ the feature map $U \in \mathbb{R}^{H \times W \times C}$ obtained after feature extraction, where H, W, and C represent the length and width of the feature map and the channel. Each channel of the number performs global average pooling to obtain an $a \in \mathbb{R}^{C}$:

$$a_c = F_{gp}(U_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i,j) \tag{1}$$

Among them, a_c is the c-th value of a, i and j are the positions of the current channel in U, and the feature map U is compressed into a set of feature vectors by the above formula.

Furthermore, we perform a fully connected operation on the feature vector, and learn the feature feedback on each channel through the ReLU activation function, to prepare for the subsequent feature weight re-calibration. This part transforms a to $b \in \mathbb{R}^{C}$:

$$b = F_{\mathrm{fc}}(a) = \delta(\beta(Wa)) \tag{2}$$

where δ is the ReLU activation function and β represents Batch Normalization, $W \in \mathbb{R}^{C \times C}$. Then, the feature map U is re-corrected at the channel level to form the feature map $Z_E^1 \in \mathbb{R}^{H \times W \times C}$, which is finally added to the attention mechanism.

$$Z_E^1 = F_{scale}(U, b) = bU \tag{3}$$

Finally, the decoupling operation has been performed on the channel, for the feature map $Z_E^1 \in \mathbb{R}^{H \times W \times C}$, after three Conv Layer convolution and down-sampling stacking operations. Furthermore, the feature extraction is completed in the decoupling layer to obtain $N \in \mathbb{R}^{H' \times W' \times C'}$, and we use the decoupling operation to perform a separate convolution on each channel of to obtain the output of the last layer $Z_E^5 \in \mathbb{R}^{H' \times W' \times C'}$.

3.3. Decoder

The decoder decodes the high-dimensional abstract features extracted during the encoding stage. Through the up-sampling, the continuously decoded features are restored to the original image size. Atrous Spatial Pyramid Pooling (ASPP) [31] employs the dilated convolutions with different sampling rates to enrich information. In the U-Net method, we connect the decoder with the corresponding hierarchical feature map of the encoding stage to enrich information. We have employed the combination of spatial attention mechanism and skip connection to capture the crack edge information during the up-sampling process. This is completed for accurately distinguishing between the background and crack information. The structure of the decoder is displayed in Figure 3.

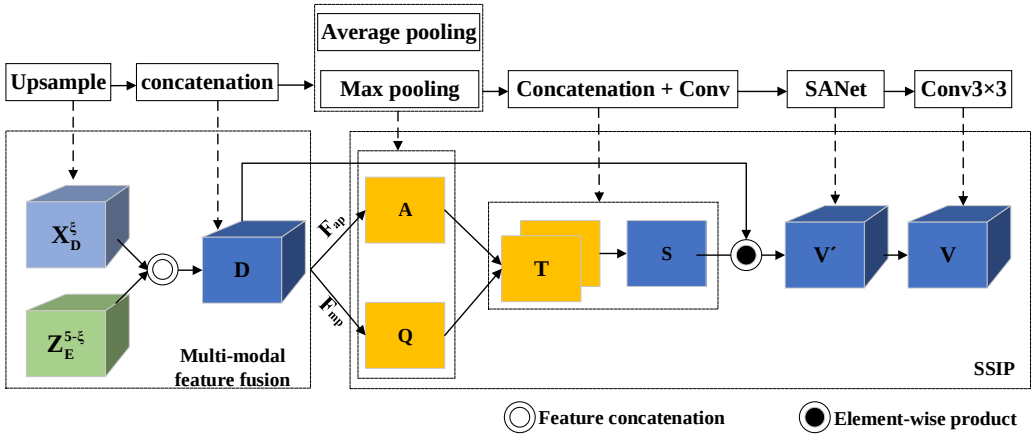


Figure 3. Decoder with multimodal feature fusion and SSIP.

We splice the output $Z_E^{5-\xi}(\xi \in [1,4])$ of each layer in the encoder with the input X_D^{ξ} of each layer of the decoder through a skip connection to obtain a high-dimensional feature map $D \in \mathbb{R}^{H^l \times W^l \times C^l}$, where H^l, W^l, and C^l represent the size of H, W, and C, respectively, in the first layer, and then follow the channel for the high-dimensional feature map D. The global average pooling of dimensions gives $A \in \mathbb{R}^{H^l \times W^l}$:

$$A_{ij} = F_{ap}(D_{ij}) = \frac{1}{C^l} \sum_{k=1}^{C^l} D_{ij}(k) \tag{4}$$

where $D_{ij}(k)$ is the feature of D with the spatial position i, j on the k-th channel, and A_{ij} is the feature of A with i, j in the spatial dimension. Concomitantly, we ascertain A that has been processed by the global average pooling.

Similarly, the global maximum pooling is performed on D along the channel dimension to obtain $Q \in \mathbb{R}^{H^l \times W^l}$:

$$Q_{ij} = D_{ij}(\arg\max_k D_{ij}(k) := \{k | \forall y : D_{ij}(y) \leqslant D_{ij}(k)\}) \tag{5}$$

where arg max is the maximum value operation, and Q_{ij} is the feature of the spatial position i and j in Q.

Subsequently, we have spliced A and Q along the channel dimension to obtain $T \in \mathbb{R}^{H^l \times W^l \times 2}$. Thereafter, T was a fully connected and sigmoid activation function, and we learned each spatial dimension in a targeted manner. According to the feature feedback to prepare for the re-calibration of the subsequent features in the spatial dimension, this part transforms T to $S \in \mathbb{R}^{H^l \times W^l}$:

$$S = \sigma F_{fs}(T) = \sigma(\otimes T) \tag{6}$$

where σ is the activation function, and sigmoid is used as the activation function. Furthermore, $\otimes$ implies convolution with T and aggregates T in the channel dimension.

Finally, we re-check the spatial information of the feature map D through S to obtain the feature map $V' \in \mathbb{R}^{H' \times W' \times C'}$, and we obtain V after the multi-layer convolution as the next layer of the decoder part of the input.

$$V' = F_{SA}(D, S) = SD \tag{7}$$

3.4. Loss Function

There are only two types of samples, viz., the foreground and background, in the crack image segmentation problem solved in this paper. The ratio of the crack pixels to the background pixels is highly unbalanced. Dice loss is a region-related loss function. The loss of the current pixel is related to the predicted value of the current pixel along with

the value of other points. It is usually employed to calculate the similarity between the two sample pixels and in the training process. The method focuses on the mining of the foreground area; hence, it has an appreciable performance in the scene where the positive and negative samples are seriously unbalanced. The formula is given as

$$DiceLoss = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \tag{8}$$

where X represents the ground truth segmented image, and Y represents the predicted segmented image. We have employed the Dice loss as the loss function to solve the problem of unbalanced ratio between the crack pixels and background pixels in the crack image to obtain accurate segmentation results.

4. Experiments

4.1. Experimental Environment

The hardware environment of this experiment is an Intel Core i7-8700 CPU processor, NVIDIA GeForce GTX 1060 GPU graphics card, and the software environment is WINDOWS 10 operating system by Microsoft, Pytorch deep learning framework, and Python libraries related to the neural networks.

4.2. Datasets

4.2.1. Concrete Crack Images for Classification

The dataset photos in this paper have been taken and processed by researchers at Middle East Technical University [21,32]. The dataset has a total of 40,000 images. It can be divided into two types, viz., the positive sample and negative sample sets. Each category has 20,000 images, 227 × 227 pixels, and RGB channels. There are differences in the surface finish and lighting conditions. We have selected 740 representative crack images for manual labeling.

We have selected the dataset from the aspects of surface finish to make the selected crack images represent the overall datasets of multiple, coarse, and fine cracks, and whether the junction between the cracks and the background is blurred, as follows.

- The crack samples in different backgrounds, including dry, wet, and crack samples with different textures.
- A sample with multiple cracks.
- Starting from the surface finish, select proofs with complex interference such as the large-area noise.
- Samples of cracks with different thicknesses.
- A sample where the edge of the crack is indistinguishable from the background.

The sample of the dataset is shown in Figure 4.

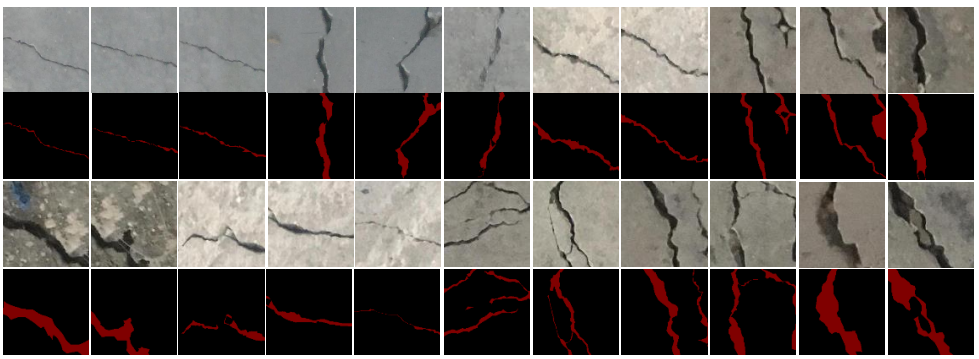


Figure 4. CCIC dataset: (**a**,**c**) original positive; (**b**,**d**) ground truth.

4.2.2. DeepCrack

The DeepCrack dataset includes a total of 537 RGB pavement crack images and the corresponding ground truth, 384 × 544 pixels, RGB channel characterized by small shadows, dense stains, and other noises. The sample of the dataset is displayed in Figure 5.

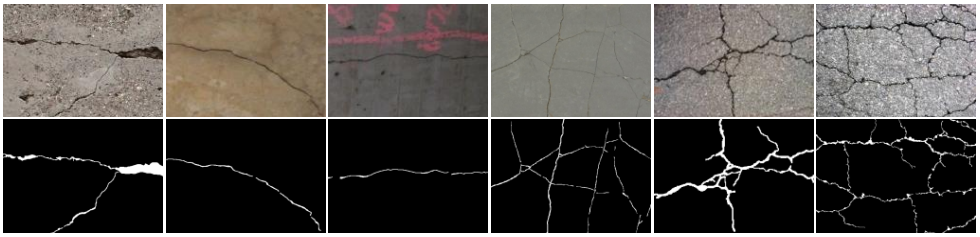


Figure 5. DeepCrack dataset: (**a**) original positive, (**b**) ground truth.

4.3. Evaluation Indicators

To evaluate the performance of the model quantitatively, this paper employs four indicators for the performance evaluation, viz., Precision (Precision, P), Recall (Recall, R), and F_1 score (F_1-score) [21]. The definition of each evaluation metric is as follows.

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{11}$$

where TP is a true positive example; i.e., the model correctly predicts the positive category sample as a positive category. TN is a true negative example; i.e., the model correctly predicts a negative category sample as a negative category. FP is a false positive example; i.e., the model correctly predicts a negative category sample as a negative category. It is predicted as a positive category, and FN is a false negative example; i.e., the model incorrectly predicts a positive category sample as a negative category. Concurrently, we have added mean intersection over union (MIOU) to evaluate the model segmentation effect. This is a standard evaluation metric for the segmentation problems, which calculates the overlap ratio of the intersection and union of two sets [33] as

$$MIOU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{12}$$

p_{ij} in the above formula implies predicting i as j, i indicates the real value, and j indicates the predicted value. It is used to calculate the intersection ratio of the real pixels and predicted pixels of cracks. Accordingly,

$$DSC = \frac{2|X \cap Y|}{|X| \cup |Y|} \tag{13}$$

The Dice Similarity Coefficient (DSC) is a set of similarity measurement functions, and X and Y are the prediction results and ground truth, respectively. In the image segmentation task, it is usually employed to calculate the similarity between the prediction and the label, and the value range is between [0, 1]. According to Equation (13), it should have the same value as F_1. However, the error value ε in the experimental results, owing to a protection parameter set in the experiment to prevent the output NaN, has been employed as an evaluation index.

The learning rate in our work employs the learning strategy proposed in [31] while performing the segmentation task training, and the number of iterations is 200 rounds. Compared with the fixed learning rate strategy, the learning rate changes with the number of iteration steps, as shown in Equation (14). The variable Iter represents the current number of iteration steps, max_iter represents the maximum number of iterations, and the power parameter is fixed at 0.9. The experiments have proved that setting a dynamic learning rate strategy is better for specific tasks than the fixed learning rate method.

$$lr \times \left(1 - \frac{iter}{\max_iter}\right)^{power} \tag{14}$$

4.4. Classification Experiment

To verify the performance of the encoding model proposed in this paper, the encoder has first been used as a classifier for the classification testing on the CCIC dataset, and 15,000 positive samples and 15,000 negative samples have been randomly selected from the dataset to train the CNN. The remaining 10,000 images have been employed to evaluate the performance of the designed network.

First, the encoder in this paper is compared with the traditional image processing method, boosting, and support vector machines (SVM). By comparing the Precision, Recall, and F_1 values, it is found that the network model proposed in this paper exceeds the boosting and SVM in terms of the three evaluation indicators. The network model proposed in this paper is compared with other deep learning methods to verify its performance. According to Table 1, our model has reached the best values for Recall and F_1. The effect is only 0.14% behind the best test method in Precision. The experiment proves that the network model proposed in this paper has a significant effect on the extraction of crack features.

Table 1. Classification result.

Method	Precision	Recall	F_1
SVM	0.8112	0.6734	0.7359
Boosting	0.7360	0.7587	0.7472
CNN by Zhang L et al. [21]	0.8696	0.9251	0.8965
CNN by Fan R et al. [23]	0.9967	0.9046	0.9484
Our method	0.9953	**0.9993**	**0.9973**

4.5. Segmentation Experiment

4.5.1. CCIC Dataset Experiment

Our model is compared with other related models, including ResU-Net34, U-Net by Woo et al., U-Net by Hu et al., VGG16-U-Net, and FCN, to verify the effectiveness of our method in the crack segmentation task. The quantitative comparison results are shown in Table 2, using the hand-labeled CCIC fracture dataset.

According to Table 2, in terms of quantitative comparison with the method proposed in this paper, other indicators are better than the compared methods, except that the R value is lower than ResU-Net34. Furthermore, the key indicators F_1 and MOIU are compared with the best algorithm. They have increased by 0.8% and 1.4%, respectively. Alternatively, the number of parameters required by the model in this paper is the least in the comparison model. It takes the least time for detecting each crack image, and each crack image only takes 0.040 s on an average. Figure 6 shows the MIOU value during the iteration process of each model, and Figure 7 shows the DSC value of each model. According to Figures 6 and 7, the MIOU value of the method proposed in this paper converges after two hundred iterations when compared with other models. This convergence is due to the stability, and the DSC value also achieves best results. The quantitative comparison results in the segmentation effect are better.

Table 2. Experimental results of segmentation on CCIC dataset.

Method	P	R	F_1	MIOU	#Params	Time/Image (s)
ResU-Net34 [16]	0.880	0.929	0.904	0.825	33.9M	0.042
U-Net by Woo S et al. [15]	0.879	0.899	0.889	0.800	17.2M	0.043
U-Net by Hu J et al. [11]	0.895	0.904	0.900	0.818	17.2M	0.041
VGG16-U-Net	0.888	0.911	0.899	0.817	28.1M	0.046
FCN [6]	0.898	0.889	0.893	0.807	23.9M	0.052
SEAD-U-Net	**0.906**	0.919	**0.912**	**0.839**	**15.1M**	**0.040**

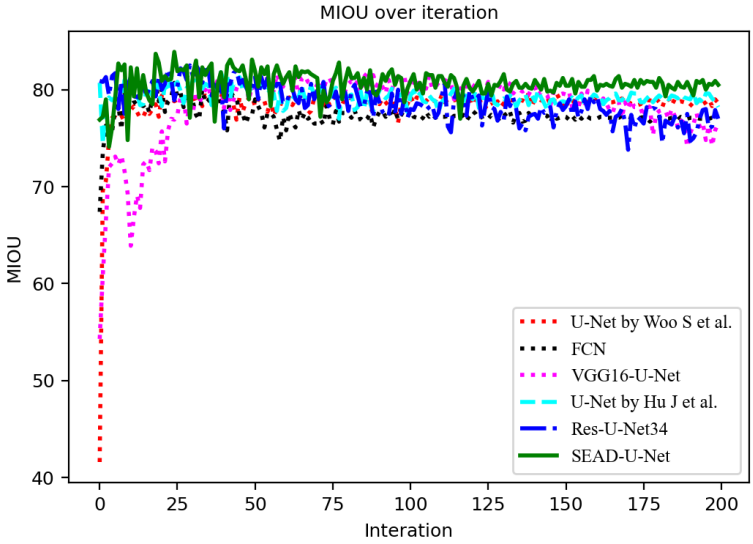


Figure 6. The MIOU comparison of the different methods.

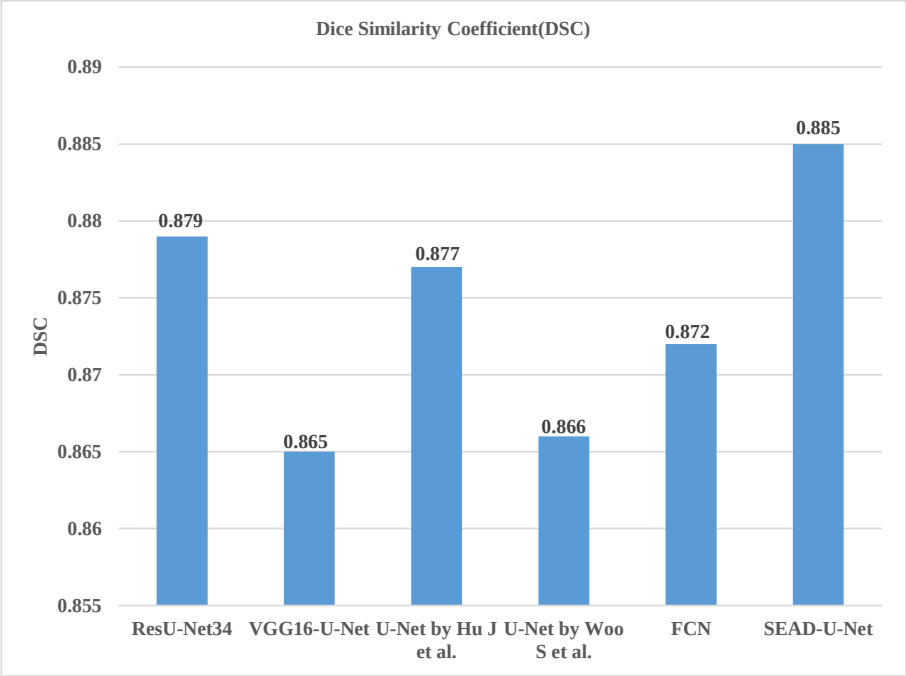


Figure 7. The DSC comparison of the different methods.

Figure 8 displays the partial segmentation results of the CCIC dataset. According to the seventh row in the figure, when the cracks in the picture are relatively clear and the interference is less, the five models can accurately segment the pixel positions of the cracks. According to the third row in the figure, when there are subtle invisible cracks in the picture,

all the models have different degrees of missed detection. The FCN segmentation effect is the worst in the fifth row, and the fine crack information is not captured. ResU-Net34 does not capture the crack information. However, the three models with the attention mechanism, compared with the above two methods, capture the crack information better. According to the sixth and second rows, the method proposed in this paper has the best performance and the most detailed information, besides the segmentation effect being closest to GT.

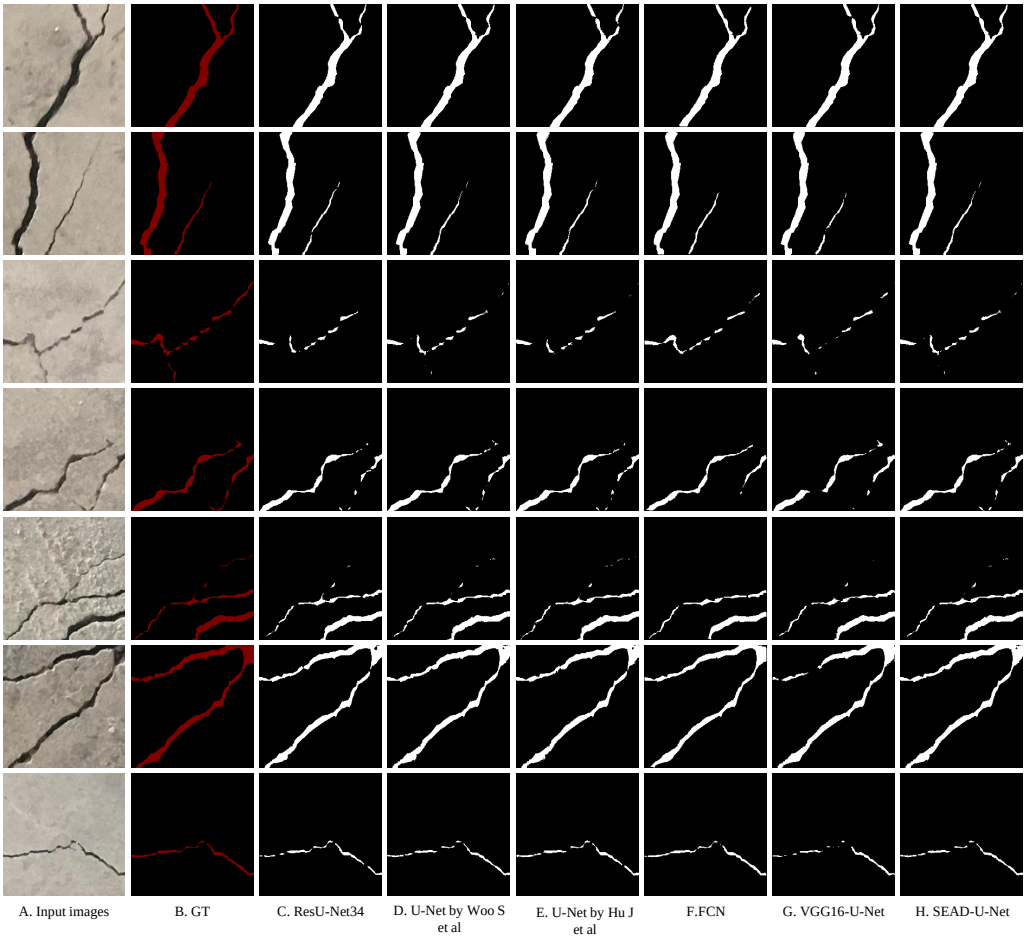


Figure 8. CCIC dataset segmentation result graph.

4.5.2. DeepCrack Dataset Experiment

The training is carried out on the public dataset DeepCrack to further verify the effectiveness of the method proposed in this paper, and the Dice loss function and dynamic learning rate methods have been employed, and the number of iterations is 200 rounds. The experimental quantitative comparison results are listed in Table 3. According to the table, the experimental result from our method is the best. The detection time of the crack image is only 0.060 s, which is shorter than other comparison methods. MIOU has reached 0.872, which proves that the idea proposed in our work has good robustness with respect to different models, and it has an appreciable effect on the crack detection.

Table 3. Experimental results of segmentation on DeepCrack dataset.

Method	P	R	F_1	MIOU	#Params	Time/Image (s)
FCN [6]	0.845	0.835	0.840	0.854	23.9M	0.097
U-Net [8]	0.834	0.862	0.848	0.861	17.2M	0.065
VGG16-U-Net	0.823	0.846	0.835	0.851	28.1M	0.086
U-Net by Woo S et al. [15]	0.849	0.819	0.834	0.849	17.2M	0.070
U-Net by Hu J et al. [11]	0.805	0.865	0.834	0.850	17.2M	0.065
SEAD-U-Net	**0.851**	**0.871**	**0.861**	**0.872**	**15.1M**	**0.060**

Figure 9 is a partial segmentation effect diagram for each model on the DeepCrack dataset. According to the figure, the U-Net by Woo et al. segmentation effect is the worst among all the segmentation results. The segmentation effect of FCN for the method proposed in this paper is the best, and it is closest to GT, under the condition that the crack image is not complex and clear without light and other interference conditions, as shown in the fourth and seventh rows. However, all other models perform better than FCN when the crack structure is not clear, as shown in the first row. Although the models have missed detection, when the crack condition is complex besides a heavy interference, as shown in the fifth and sixth lines, the U-Net and our model have the best segmentation effect, and it can clearly highlight the crack details. Following the overall analysis, the model proposed in this paper can basically maintain the integrity of cracks and capture the details of cracks better than the others in Figure 9.

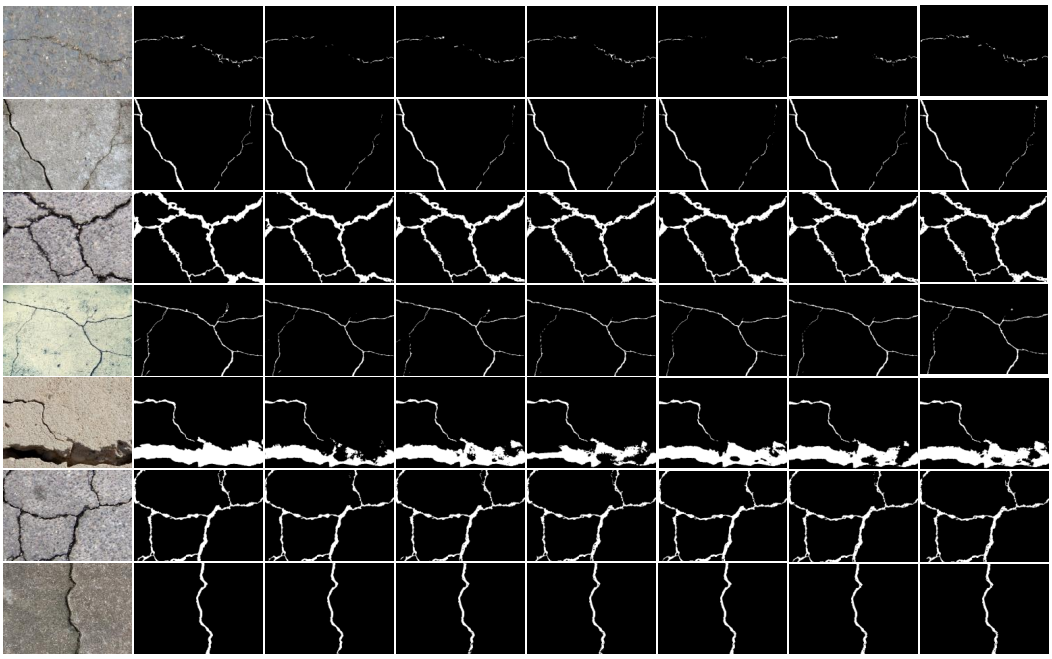


Figure 9. DeepCrack dataset segmentation result graph.

4.5.3. Ablation Experiment

The influence of the skip structure: The skip connection is an important part for recovering the low-level feature information in the U-shaped structure. This experiment intends to verify the impact of different numbers of skip connections on the accuracy of the model. This is completed by changing the number of skip connections to 0-skip, 1-skip (connect at 1/16), 2-skip (1/16 and 1/ 4 connections) and the All-skip adopted by this model for the ablation experiments. Furthermore, the experimental results on the CCIC dataset are shown in Table 4 and Figure 10. Owing to the increase of skip connections, the up-sampled features, and the skip connection structure, the restored low-level features

enable the model to achieve good results with respect to the various evaluation indicators, thus further verifying that the skip connection in the U-shaped structure can help the model can learn the low-level feature information more accurately.

Table 4. Experimental results of different number of skips.

Skip-Connection	P	R	F_1	MIOU
0-skip	0.867	0.842	0.854	0.745
1-skip	0.869	0.856	0.862	0.758
2-skip	0.926	0.870	0.897	0.813
All-skip	0.906	**0.919**	**0.912**	**0.839**

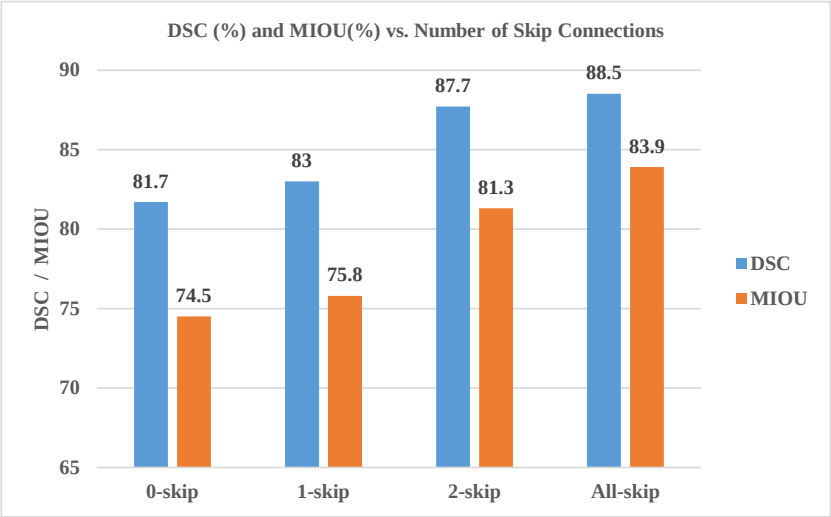


Figure 10. The influence of different number of skips on DSC and MIOU.

The influence of different modules: The ablation experiments have been carried out on the DeepCrack dataset to verify the influence of the fusion attention mechanism and dimensional decoupling algorithm on the model segmentation accuracy. The experiments have included the following, viz., the first DE (Dimensional decoupling) + SE (Squeeze-and-Excitation) + SA (Spatial Attention), no solution Coupling algorithm, decoupling + SA, and decoupling + SE. We believe that following the decoupling, there is a correlation between space and channel in advance, and the model will focus on learning the characteristics of each dimension that lacks the correlation between dimensions. The central idea of this work is to consider the relevant information of space and channel separately instead of ignoring them. Therefore, we focus on the learning of the channel dimension information in the encoding stage without neglecting the relationship between space and channel information. Furthermore, we perform the decoupling operations in the final stage of encoding so that the feature information to be processed in the decoding stage, i.e., the high-dimensional feature information, will be lost, and the spatial information and channel information can be considered separately. Finally, the encoded information is cleverly fused through skip connections so that the spatial attention mechanism can be employed to capture the spatial information in the decoding stage without losing the correlation between space and channels. This enables the network to perform the pixel-level segmentation more accurately. The comparative experimental results are shown in Table 5 and Figure 11.

Table 5. Experimental results of different module combinations on DeepCrack.

Method	P	R	F_1	MIOU
DE + SE + SA	0.859	0.839	0.849	0.862
SE + SA	0.865	0.833	0.849	0.862
DE + SE	0.854	0.822	0.837	0.852
DE + SA	0.827	0.843	0.835	0.851
SEAD-U-Net	0.851	**0.871**	**0.861**	**0.872**

+ It represents a combination.

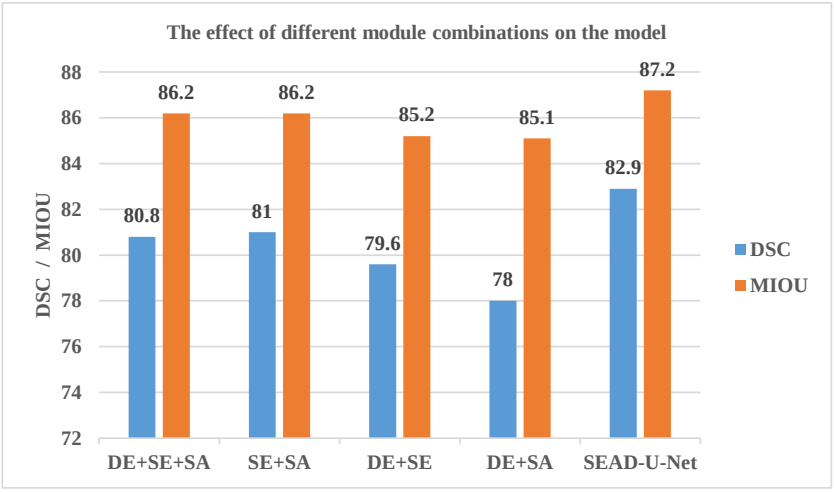


Figure 11. The influence of different modules on model DSC and MIOU.

According to Table 5 and Figure 11, if the channel and spatial correlations are decoupled initially, the network will focus exceedingly on the information of the dimension of the feature itself. Therefore it will ignore the correlation information at the junction of the crack and the background, which is significant for locating the crack position. For a network that does not have the decoupling but only performs the SE+SA operations, even if a plentiful attention mechanism is added to enable the network for an enhanced capture of semantic information, when excess semantic information is combined for cascading the operations, it is difficult to segment the network. Precise processing of the complex information makes MIOU fail to achieve the desired effect. For the network using SE exclusively, the capture effect at the edge of the crack is not ideal, although the information capture inside the crack is relatively complete, since it only focuses on the differences between channels. For the model that only joins the SA network, the context information is focused upon by the global average pooling and the global maximum pooling, and the network has an obvious effect on the detailed information at the edge of the crack. However, the segmentation effect is often not ideal while processing pictures under complex background interference owing to the lack of channels with the benefits of the attention mechanism. Therefore, through the comparison of ablation experiments, it is difficult for alternative models to achieve the accuracy of the method proposed in this paper. The experimental results reveal that MIOU and DSC are 87.2 and 82.9%, respectively. This further verifies the effectiveness of this model with respect to the quantitative comparison of segmentation results.

5. Conclusions

To achieve an accurate segmentation of the cracks under complex background, a crack image detection and segmentation method that integrates the attention mechanism and dimension decoupling has been proposed in this paper. (1) The channel semantic information perception module solves the problem of weak feature extraction ability of cross-channel semantic information. (2) The dimension decoupling module has been employed in the encoding stage to solve the problem of bias in locating the location of cracks. (3) A multimodal feature fusion module has been used to restore the spatial and channel semantic

information of cracks. (4) The Dice loss function, which tends to focus on the difference between pixels, makes the model better locate the segmentation position.

The experimental results illustrate that the proposed method has higher crack segmentation accuracy than the alternative ones in the literature. The mean intersection over union ratio has reached 83.9% and 87.2% on the self-made and public datasets, which is 1.4% and 1.1% higher than the existing methods, respectively.

However, this model inaccurately detects the tiny cracks in the shadow, and while this paper improves the method used to classify different cracks and make corresponding engineering treatment, this subject still needs further research. Owing to the continuous development of deep learning technology in the field of fracture mechanics, the method proposed in this paper has been improved to explore the law of crack growth, which will be the focus of our forthcoming study.

Author Contributions: Data curation, L.H., W.L., Y.L., H.W., S.C. and C.Z.; software, W.L.; methodology, W.L. and L.H.; writing—original draft preparation, W.L. and L.H.; writing—review and editing, L.H. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Talent Foundation of Hefei University (No. 18-19RC28) and the University Outstanding Young Talent Visiting and Research Foundation of Anhui Province (No. gxgnfx2019060).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Kapela, R.; Śniatała, P.; Turkot, A.; Rybarczyk, A.; Pożarycki, A.; Rydzewski, P.; Wyczałek, M.; Błoch, A. Asphalt surfaced pavement cracks detection based on histograms of oriented gradients. In Proceedings of the 2015 22nd International Conference Mixed Design of Integrated Circuits & Systems (MIXDES), Torun, Poland, 25–27 June 2015; pp. 579–584. [CrossRef]
2. Ji, W.; Yu, S.; Wu, J.; Ma, K.; Bian, C.; Bi, Q.; Li, J.; Liu, H.; Cheng, L.; Zheng, Y. Learning Calibrated Medical Image Segmentation via Multi-rater Agreement Modeling. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 12336–12346. [CrossRef]
3. Lu, G.; Duan, S. Deep Learning: New Engine for the Study of Material Microstructures and Physical Properties. *IEEE Trans. Intell. Transp. Syst.* **2019**, *9*, 263–276. [CrossRef]
4. Silva, W.R.L.d.; Lucena, D.S.d. Concrete Cracks Detection Based on Deep Learning Image Classification. *Proceedings* **2018**, *2*, 489. [CrossRef]
5. 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 (ICIVC), Xiamen, China, 5–7 July 2019; pp. 336–340. [CrossRef]
6. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [CrossRef]
7. David Jenkins, M.; Carr, T.A.; Iglesias, M.I.; Buggy, T.; Morison, G. A Deep Convolutional Neural Network for Semantic Pixel-Wise Segmentation of Road and Pavement Surface Cracks. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Roma, Italy, 3–7 September 2018; pp. 2120–2124. [CrossRef]
8. 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; pp. 234–241. ._28. [CrossRef]
9. 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. [CrossRef]
10. Choi, W.; Cha, Y.J. SDDNet: Real-Time Crack Segmentation. *IEEE Trans. Ind. Electron.* **2020**, *67*, 8016–8025. . /TIE.2019.2945265. [CrossRef]
11. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
12. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 510–519. [CrossRef]

13. Qin, Z.; Zhang, Z.; Chen, X.; Wang, C.; Peng, Y. Fd-Mobilenet: Improved Mobilenet with a Fast Downsampling Strategy. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 1363–1367. [CrossRef]
14. Rabano, S.L.; Cabatuan, M.K.; Sybingco, E.; Dadios, E.P.; Calilung, E.J. Common Garbage Classification Using MobileNet. In Proceedings of the 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM), Baguio City, Philippines, 29 November–2 December 2018; pp. 1–4. [CrossRef]
15. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19. ._1. [CrossRef]
16. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
17. Prasanna, P.; Dana, K.J.; Gucunski, N.; Basily, B.B.; La, H.M.; Lim, R.S.; Parvardeh, H. Automated Crack Detection on Concrete Bridges. *IEEE Trans. Autom. Sci. Eng.* **2016**, *13*, 591–599. [CrossRef]
18. 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. [CrossRef]
19. 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.* **2016**, *17*, 2718–2729. [CrossRef]
20. Cord, A.; Chambon, S. Automatic road defect detection by textural pattern recognition based on AdaBoost. *Comput.-Aided Civ. Infrastruct. Eng.* **2012**, *27*, 244–259. [CrossRef]
21. Zhang, L.; Yang, F.; Daniel Zhang, Y.; Zhu, Y.J. Road crack detection using deep convolutional neural network. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3708–3712. [CrossRef]
22. Fang, F.; Li, L.; Gu, Y.; Zhu, H.; Lim, J.H. A novel hybrid approach for crack detection. *Pattern Recognit.* **2020**, *107*, 107474. [CrossRef]
23. Fan, R.; Bocus, M.J.; Zhu, Y.; Jiao, J.; Wang, L.; Ma, F.; Cheng, S.; Liu, M. Road Crack Detection Using Deep Convolutional Neural Network and Adaptive Thresholding. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; pp. 474–479. [CrossRef]
24. Protopapadakis, E.; Voulodimos, A.; Doulamis, A.; Doulamis, N.; Stathaki, T. Automatic crack detection for tunnel inspection using deep learning and heuristic image post-processing. *Appl. Intell.* **2019**, *49*, 2793–2806. [CrossRef]
25. Zou, L.; Liu, Y.; Huang, J. Crack Detection and Classification of a Simulated Beam Using Deep Neural Network. In Proceedings of the 2021 33rd Chinese Control and Decision Conference (CCDC), Kunming, China, 22–24 May 2021; pp. 7112–7117. [CrossRef]
26. Kim, H.; Ahn, E.; Shin, M.; Sim, S.H. Crack and noncrack classification from concrete surface images using machine learning. *Struct. Health Monit.* **2019**, *18*, 725–738. [CrossRef]
27. 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. [CrossRef] [PubMed]
28. Yang, F.; Zhang, L.; Yu, S.; Prokhorov, D.; Mei, X.; Ling, H. Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1525–1535. [CrossRef]
29. Cui, X.; Wang, Q.; Dai, J.; Xue, Y.; Duan, Y. Intelligent crack detection based on attention mechanism in convolution neural network. *Adv. Struct. Eng.* **2021**, *24*, 1859–1868. [CrossRef]
30. 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. [CrossRef]
31. 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.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
32. Özgenel, Ç.F.; Sorguç, A.G. Performance comparison of pretrained convolutional neural networks on crack detection in buildings. In Proceedings of the International Symposium on Automation and Robotics in Construction, Berlin, Germany, 20–25 July 2018; pp. 1–8. [CrossRef]
33. Fu, H.; Meng, D.; Li, W.; Wang, Y. Bridge Crack Semantic Segmentation Based on Improved Deeplabv3+. *J. Mar. Sci. Eng.* **2021**, *9*, 671. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.