DSCANet: Dynamic Snake Convolution with Attention for Crack Segmentation

Wenbo Hu[1], Kuixuan Jiao[1], Kaijian Xia[2,3*], and Rui Yao[1*]

[1] School of Computer Science and Technology, China University of Mining and Technology,
Xuzhou 221116, China {08213113, 08213125, ruiyao}@cumt.edu.cn
[2] Department of Scientific Research, The Changshu Affiliated Hospital of Soochow University,
Jiangsu Suzhou, 215500 China
[3] Changshu Key Laboratory of Medical Artificial Intelligence and Big Data, Jiangsu Suzhou,
215500 China kjxia@suda.edu.cn
*Corresponding author

Abstract. Pixel-wise crack segmentation task plays a crucial role in infrastructure maintenance. However, it poses a significant challenge due to the irregular and slender nature of cracks. The standard convolution kernel struggles to capture crack features accurately due to its fixed square shape. Moreover, shallow information significantly impacts the segmentation results. Inadequacy in local detailed information can lead to segmentation errors. In this paper, we propose a novel crack segmentation method based on encoder-decoder framework. We propose a Dynamic Snake Convolution with Attention (DSCA) module to enhance feature extraction accuracy for cracks. Additionally, we propose a Multi-level Fusion with DSCA and Channel Prior Convolutional Attention (CPCA) (MF-DC) module for local features extraction and a Multi-scale Fusion with CPCA and Atrous Spatial Pyramid Pooling (ASPP) adding Strip Pooling Module (SPM) (MF-CAS) module for global features extraction. Experimental results on two different crack datasets validate the superior performance of our method, surpassing several mainstream methods.

Keywords: Crack segmentation, Feature fusion, Attention mechanism, Convolutional neural network.

1 Introduction

Crack detection plays a key role in infrastructure maintenance. Automatic detection is more efficient and less labor-intensive than manual methods, with pixel-wise crack segmentation standing out as a prominent approach.

Traditional image processing methods for crack segmentation, e.g., edge detection [17], thresholding [14], morphological filters [20], are very sensitive to noise and lack generalization, leading to suboptimal performance. In recent years, Deep Convolutional Neural Networks (DCNNs) and Transformer-based networks are widely used in various computer vision tasks. While these methods have shown superior performance in crack segmentation, certain challenges persist: (1) Difficult-to-capture local features.

As shown in Fig. 1, the structure of cracks is slender and irregular, and their pixels account for a small portion of an image, making it challenging for standard convolutions to accurately capture crack features. This often leads to the extraction of more background information than detailed information, introducing noise that hampers segmentation accuracy. (2) Loss of local features. During forward propagation, there is a reduction in local detailed information while global semantic information becomes more pronounced. In segmentation tasks, retaining detailed local information is crucial, particularly for objects with subtle local features like cracks.

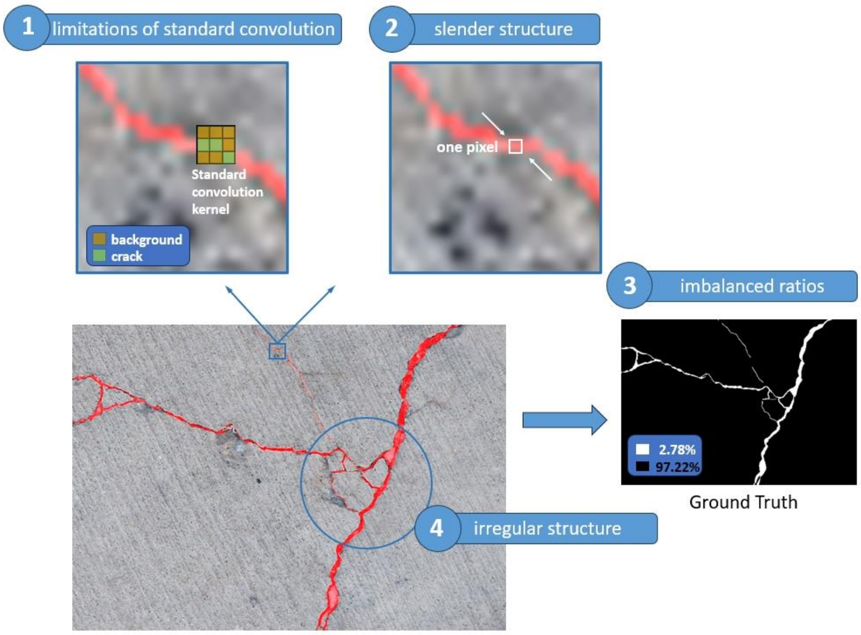

Fig. 1. The special structure of cracks and the limitation of traditional convolution make it difficult to extract local features for cracks.

Some mainstream segmentation methods utilized for crack segmentation struggle to effectively address these challenges. Certain DCNN-based methods, e.g., UNet [23], Deeplabv3+ [4], Deepcrack [15], CrackSegNet [22], and DenseCracks [18], mitigate the inadequacy of local detailed features by incorporating multi-level feature fusion techniques. However, these methods fail to resolve the issue of inaccurate local feature extraction. Some Transformer-based methods, e.g., SERT [28], Swin Transformer [16], Segmenter [24], and SCDeeplab [29], have also been explored. These methods are good at capturing global contextual information through self-attention mechanism. However, they struggle with extracting subtle local features and exhibit high data dependency and computational complexity. Overall, none of the aforementioned methods have introduced a novel feature extraction method based on the characteristics of the crack structure.

To tackle this issue, Qi et al. [21] proposed DSCNet, which utilizes a deformable convolution named Dynamic Snake Convolution (DSConv). DSConv straightens the standard convolution kernel's shape along the x-axis and y-axis to adapt to the slender and irregular crack structures. Typically, DSConv extracts information in both directions from an input feature map, followed by a two-view feature fusion. However, as shown in Fig. 2, capturing features in two directions blindly will introduce additional background noise in both channel and spatial dimension.

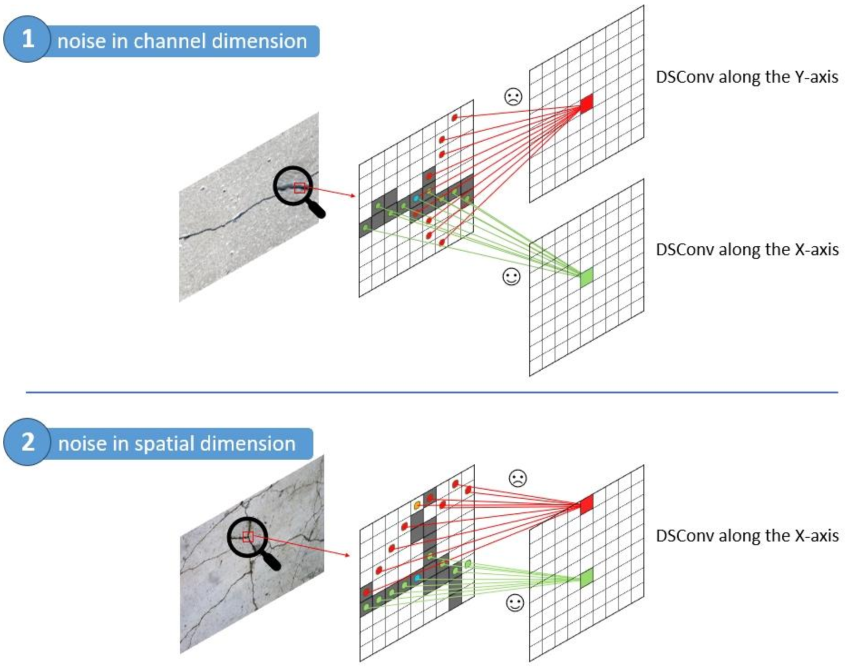

Fig. 2. (1) In the same blue central grid of convolution, DSConv along the x-axis focus on local features while DSConv along the y-axis is outside the target. (2) In different central girds (orange grid and blue grid), DSConv in the same direction yields varying outcomes.

In response to these challenges, we introduce a novel crack segmentation network, named DSCANet, based on encoder-decoder architecture. The main contributions of our work are summarized as follows:

- We propose a DSCA module that enhances the extraction of local features and reduce the noise disturbance. The residual connection is also applied in the module to mitigate gradient vanishing.
- We propose a MF-DC module to address the loss of local detailed information. In MF-DC, the feature maps at different levels processed by our proposed DSCA module will be fused together for comprehensive feature representation.
- We propose a MF-CSA module that integrates CPCA mechanism and SPM. The former directs the network's focus towards crucial features in both channel and

spatial dimensions, enhancing effectiveness in utilizing multi scale features, the latter is applied to ASPP module to extract long-range contextual information.

2 Related work

2.1 Crack Segmentation Algorithms

In the early years, traditional image processing methods, e.g., edge detection [1], morphological operators [11], and thresholding [20] were prevalent for crack detection. However, these methods suffer from various limitations. They are sensitive to background noise and require manual parameter tuning, leading to a lack of generalizability.

Recently, deep learning-based algorithms have gained popularity for pixelwise crack segmentation. CrackSegNet [22] leverages dilated convolutions [27] to expand the receptive field while preserving resolution. However, excessive use of dilated convolutions can result in information continuity loss. Various feature fusion methods [15] [19] [31] [2] [30] [4] are proposed to integrate local detailed and global semantic information. However, standard convolution kernels struggle to adapt to crack structures, making it challenging to capture local details effectively.

To go one step further, some methods attempt to change the convolution kernel's shape to better suit complex object structures. Dai et al. [6] introduce a deformable convolution that can self-adapt to object structures. However, the unlimited deformable offsets may extend the perception field beyond slender structures. In response, DSConv [21] restricts the perception field to line-like regions and divides the offset directions into the x-axis and y-axis, focusing on slender and subtle features. However, there are instances where DSConv may deviate from the target and introduce additional noise. In this paper, we propose a DSCA module that combines DSConv with the CPCA mechanism to highlight crucial features and reduce noise interference. We also incorporate residual connections to mitigate gradient vanishing.

Besides, some methods are proposed to modify the shape of the receptive field in pooling layer. SPM [8] can capture long-range contextual information, particularly suitable for crack structures. However, relying solely on strip pooling will limits the view of feature extraction. To address the limitation, we integrate it into the ASPP module, incorporating multi-scale features.

2.2 Attention Mechanism

Attention mechanisms play a pivotal role in directing networks to focus on essential features, which can be categorized into channel attention and spatial attention. Some methods [9] [13] [25] solely employ channel attention mechanisms to weight feature maps in the channel dimension, but they overlook the differences in the importance of information in the spatial dimension. Convolutional Block Attention Module (CBAM) [26] integrates channel and spatial attention to emphasize critical features in two dimensions. However, the spatial attention weights for each channel exhibit a uniform distribution, hindering the distinction between interest regions in different channels. To overcome this challenge, CPCA [10] computes distinct spatial attention distributions for each channel, enabling flexible focus on essential information. In this paper, we leverage CPCA to guide our network in reducing noise and highlighting crucial information during feature fusion.

3 Method

Our model employs an encoder-decoder architecture. The method we proposed focuses on enhancing the encoder for the extraction of comprehensive and diverse features. Within the encoder, we introduce two novel modules, MF-DC and MFCAS. The MF-DC module incorporates our proposed DSCA module for efficient feature extraction. In this section, we will introduce the details of the DSCA, MF-DC, and MF-CAS modules.

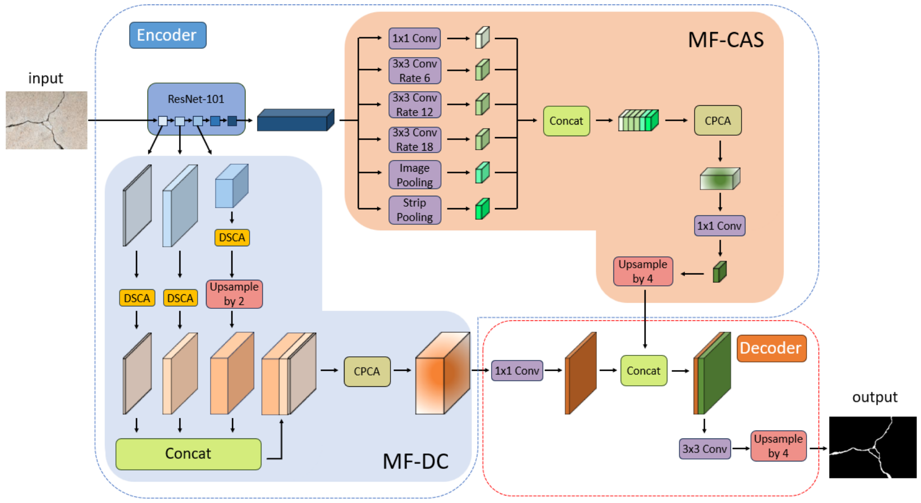

Fig. 3. The framework of the proposed network. In encoder we proposed two modules named MF-DC and MF-CAS. Besides, in MF-DC, we proposed a novel module named DSCA.

3.1 Overview

Fig. 3 illustrates the proposed DSCANet. For a given input image, the network extracts low-level features and high-level features separately before final feature fusion. To capture local detailed information effectively, we propose a MF-DC module leveraging three feature maps from distinct layers in the backbone for encoding low-level features. Our proposed DSCA module is facilitated for accurate detailed feature extraction. For high-level feature encoding, we propose a MF-CAS module employing SPM in ASPP and CPCA in fusion procedure. Semantic features through the backbone are sent to the MF-CAS module for multi-scale feature extraction. During the feature fusion stage, we employ the CPCA mechanism to guide the network's focus towards essential information across various scales of feature maps. Finally, low-level features and high-level features are concatenated together and upsampled to the original image size.

3.2 Dynamic Snake Convolution with Attention module

In previous work, DSConv can be adapted to slender structures for better feature extraction, but it inevitably extends beyond the target and introduce background noise. It is difficult to further introduce prior knowledge to guide DSConv on where to offset. Therefore, we propose DSCA module, leveraging a CPCA module to guide model to focus on important local features. Fig. 4 illustrates our DSCA module.

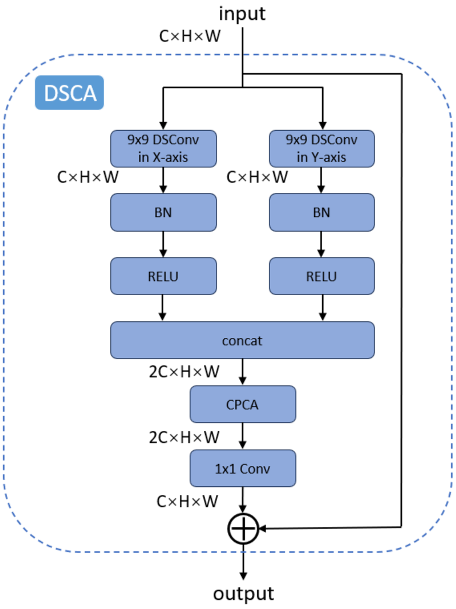

Fig. 4. Our proposed DSCA module employs the CPCA module for multi-view feature fusion and uses residual connection to mitigate gradient vanishing.

Feature Extraction. In feature extraction, we utilize DSConv for the input in x-axis and y-axis directions respectively to obtain comprehensive local information.

Define $\mathrm{F^I} \in \mathrm{R^{C\times H\times W}}$ as the input feature map, as shown in Eq. (1)

$$F^I = [f_1^I, f_2^I, \cdots f_n^I], f_c^I \in R^{H\times W}, 1 \le c \le n, n = C. \tag{1}$$

The DSConv performs feature extraction separately along the x-axis and y-axis on the input, yielding two output feature maps $F^X, F^Y \in R^{C\times H\times W}$, where

$$F^X = [f_1^X, f_2^X, \cdots f_n^X], f_c^X \in R^{H\times W}, 1 \le c \le n, n = C, \tag{2}$$

$$F^Y = [f_1^Y, f_2^Y, \cdots f_n^Y], f_c^Y \in R^{H\times W}, 1 \le c \le n, n = C. \tag{3}$$

Specifically, for each pixel position k in output, the convolution can be expressed as:

$$f_c^X(k) = \sum_i w(k_i) \cdot f_c^I(k_i), \tag{4}$$

$$f_c^Y(k) = \sum_j w\big(k_j\big) \cdot f_c^I\big(k_j\big), \qquad (5)\backslash$$

where k_i and k_j denotes the gird position in DSConv kernel, $f(k)$ denotes the pixel value at position k, and $w(k)$ denotes the weight at position k.

Feature Fusion with Attention. After obtaining the two feature maps extracted along the x-axis and along the y-axis, we concatenate them together to obtain a multi-view detailed feature map, expressed as:

$$\tilde{F} = [F^X, F^Y]. \tag{6}$$

The fused feature maps $\tilde{F} \in R^{2C\times H\times W}$ are rich in local details information, but also contain a lot of noise. We use the CPCA module to weight the feature maps in both channel and spatial dimensions, highlighting important information.

The CPCA module first performs channel attention, expressed as:

$$F^{CA} = CA\big(\tilde{F}\big) \otimes \tilde{F}, \tag{7}$$

where $F^{CA} \in R^{2C\times H\times W}$ is the output, $CA\big(\tilde{F}\big)$ is the channel attention weight, and $\otimes$ represents element-wise product.

Then, the CPCA module performs spatial attention, expressed as:

$$F^{SA} = SA(F^{CA}) \otimes F^{CA}, \tag{8}$$

where $F^{SA} \in R^{2C\times H\times W}$ is the output, $SA(F^{CA})$ is the spatial attention weight.

Following the CPCA module, all the feature values on the feature map $\tilde{F}$ are scaled by an attention weight $\sigma_c(k)$ that contains both channel and spatial attention information, where c denotes the cth channel of $\tilde{F}$ and k denotes the pixel position of a 2D map.

To half the number of channels of F^{SA}, we utilize the 1×1 convolution. The downsized feature map performs residual connection with the input:

$$F^O = F^I \oplus Conv_{1\times 1}(F^{SA}), \tag{9}$$

where $F^O \in R^{C\times H\times W}$ is the output of DSCA, $\oplus$ represents element-wise addition.

3.3 Multi-level Fusion with DSCA and CPCA

Former crack segmentation algorithms suffer from inaccurate and inadequate local information. To address the issue, we propose a new module named MF-DC to enhance the accuracy of local feature extraction at various levels.

The DSCA module is employed in MF-DC to capture local detailed information. Considering that the deformable receptive field of DSConv potentially has a bad effect on global features extraction, causing an offset of the global receptive field, we do not implement the DSCA module in the backbone. Instead, local feature extraction is performed in a separate MF-DC module, leveraging three feature maps from distinct layers in the backbone to extract detailed information by our proposed DSCA module.

Specifically, given an input $F^I \in R^{3\times H\times W}$, the three feature maps selected from our backbone ResNet-101 [7] are:

$$F_1 = MaxPooling2d_{2\times2}\left(Conv_{7\times7}\left(F^I\right)\right), \tag{10}$$

$$F\_2 = Layer\_1(F\_1), \tag{11}$$

$$F\_3 = Layer\_2(F\_2), \tag{12}$$

where F_1, F_2, F_3 denote the selected feature maps, and $Layer1, Layer2$ represent the layer defined in ResNet-101.We do not select higher-level feature maps because they lack accurate local detailed information for further extraction.

Then, the three output maps are concatenated and fed into CPCA module to highlight important information, getting the multi-level local features. We reduce the channel of the output feature maps by 1×1 convolution to prevent the segmentation results from being overly influenced by local features. The fusion can be expressed as:

$$MLF = Conv_{1\times1}\left(CPCA\left(\left[F_1, F_2, \acute{F_3}\right]\right)\right), \tag{13}$$

where MLF denotes the output multi-level local features of MF-DC, $\acute{F_3}$ denotes the upsampled F_3, $CPCA$ represents the previous mentioned CPCA operation.

3.4 Multi-scale Fusion with CPCA and ASPP adding SPM

To improve the model's representation of global semantic information, we propose MF-CAS module. In DeepLabV3+, the ASPP module is a parallel structure containing 1×1 convolution layer, three atrous convolution layers with various expansion rate and a global average pooling layer which is useful for extract object features at different scales. We introduce one strip pooling layer in the original ASPP module to enhance the extraction of long-range contextual information and employ CPCA module during feature fusion.

Specifically, given a input $F^I \in R^{C\times H\times W}$, first perform 1D average pooling in the x-axis direction and y-axis direction on it separately, yielding two outputs $F^X \in R^{C\times H\times \mathbb{1}}, F^Y \in R^{C\times \mathbb{1}\times W}$. Then, F^X and F^Y are expanded into $C \times H \times W$ dimension by bilinear interpolation. Define $\widetilde{F^X}, \widetilde{F^Y}$ as the expanded feature maps, the output F^O of SPM is:

$$\tilde{F} = Conv_{3\times3}\left(RELU\left(\widetilde{F^X} \oplus \widetilde{F^Y}\right)\right), \tag{14}$$

$$F^O = RELU\left(Conv_{1\times1}\left(\tilde{F}\right) \oplus F^I\right), F^O \in R^{C\times H\times W}, \tag{15}$$

where $\oplus$ represents element-wise addition.

To ensure consistency with the channel count from various parallel operations, we downscale the output of the SPM accordingly.

The six feature maps from the ASPP are concatenated and directed to the CPCA module. Subsequently, one 1×1 convolution reduces the channel dimension, followed by bilinear interpolation to match the output size of the MF-DC module.

4 Experiments and Results

Dataset. We evaluate our network on two public crack datasets, including DeepCrack dataset [15] and CFD dataset [5]. The DeepCrack dataset consists of 537 pavement crack images sized 554 × 384 and the CFD dataset contains 118 images sized 480 × 320. Both datasets are divided into training set, validation set, and test set in the ratio of 6:2:2.

Evaluation Metrics. To evaluate the performance of various networks, we chose four evaluation metrics commonly used for segmentation to testing, including Precision, Recall, Dice and mIoU. The formulae for these metrics are as follows:

$$Precision = \frac{TP}{TP+FP}, \tag{16}$$

$$Recall = \frac{TP}{TP+FN}, \tag{17}$$

$$Dice = \frac{2\times TP}{2\times TP+FP+FN}, \tag{18}$$

$$mIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{TP}{TP+FP+FN}, \tag{19}$$

where TP represents true positive, TN represents true negative, FP represents false positive, FN represents false negative, N represents the number of classes.

Implementation Details. We implement our proposed network using PyTorch 1.12.0 + CUDA 11.3 framework in python 3.8. The experiments in this paper were conducted on an NVIDIA V100 GPU. Our model was trained for 200 epochs by SGD optimizer with momentum of 0.9 and weight decay of 5e-4, where the mini batch size is 4. The learning rate was initialized to 5e-3 and the Poly [3] learning rate decay strategy was applied. The cross-entropy loss function was utilized in the training process.

4.1 Comparative Experiment

To demonstrate the superior performance of our model in crack segmentation, we compared it with some advanced segmentation models, including Deeplabv3+ [4], U-Net [23], Deepcrack [31] et al. The result of comparative experiment is shown in Table 1. Our proposed DSCANet model achieves the best segmentation results compared with other methods with Precision of 89.31%, Recall of 91.35%, Dice of 90.32%, mIoU of

90.80% on DeepCrack dataset, and Precision of 56.85%, Recall of 77.07%, Dice of 65.43%, mIoU of 75.39% on CFD dataset.

Table 1. Model comparative experiment.

Dataset	Method	Precision (%)	Recall (%)	Dice (%)	mIoU (%)
DeepCrack	DeepCrack [31]	86.10	86.92	86.51	87.04
	CrackSegNet [22]	85.24	86.66	85.94	86.40
	U-Net [23]	81.95	87.13	84.46	86.11
	U-Net++ [30]	82.86	85.14	83.98	85.74
	Res-UNet++ [12]	83.32	79.21	81.21	83.72
	DeepLabv3+ [4]	87.00	89.81	88.38	89.12
	DSCNet [21]	84.08	88.30	86.14	87.26
	DSCANet	**89.31**	**91.35**	**90.32**	**90.80**
CFD	CrackSegNet [22]	41.54	58.23	48.49	62.31
	U-Net [23]	38.74	61.52	47.54	65.21
	U-Net++ [30]	32.12	63.80	42.73	63.27
	DeepLabv3+ [4]	40.51	69.09	51.07	65.93
	DSCNet [21]	50.31	65.71	56.99	68.75
	DSCANet	**56.85**	**77.07**	**65.43**	**75.39**

Table 2. Ablation study of key components of DSCANet.

Dataset	Method	Precision (%)	Recall (%)	Dice (%)	mIoU (%)
DeepCrack	① baseline(DeepLabv3+)	87.00	89.81	88.38	89.12
	② w/ MF-DC	88.32	90.97	89.63	90.18
	③ w/ MF-CAS	87.68	90.60	89.12	89.74
	④ DSCANet(full model)	**89.31**	**91.35**	**90.32**	**90.80**
CFD	① baseline(DeepLabv3+)	40.51	69.09	51.07	65.93
	② w/ MF-DC	44.02	76.17	55.79	68.23
	③ w/ MF-CAS	**57.43**	70.20	63.18	72.33
	④ DSCANet(full model)	56.85	**77.07**	**65.43**	**75.39**

4.2 Ablation Experiment

Table 2 shows the ablation results to access the individual contributions of each component in our network. As shown in Fig. 3, our model is divided into two main parts for ablation experiments, which are MF-DC and MF-CAS. Method ② and Method ③ respectively introduce only the MF-DC module and the MFCAS module to the corresponding positions in the original DeepLabv3+ model. Compared with the baseline, the dice of Method ② and Method ③ respectively increases by 1.25% and 0.74% on

DeepCrack dataset, and by 4.72% and 12.11% on CFD dataset. Also, the mIoU of Method ② and Method ③ respectively increases by 1.06% and 0.62% on DeepCrack dataset, and by 2.3% and 6.4% on CFD dataset. Moreover, when employing both two components, the effectiveness is better.The result proves both components are effective.

Table 3. Ablation study of different attention mechanisms in DSCA.

Dataset	Method	Precision (%)	Recall (%)	Dice (%)	mIoU (%)
DeepCrack	① w/o attention mechanism	87.61	90.89	89.21	89.83
	② w/ SE	88.52	90.61	89.55	90.11
	③ w/ CBAM	**89.89**	89.39	89.64	90.20
	④ w/ CPCA (DSCANet)	89.31	**91.35**	**90.32**	**90.80**
CFD	① w/o attention mechanism	56.43	70.20	62.57	71.63
	② w/ SE	55.48	75.48	63.95	72.72
	③ w/ CBAM	**57.24**	76.11	65.34	74.24
	④ w/ CPCA (DSCANet)	56.85	**77.07**	**65.43**	**75.39**

Table 4. Ablation study of key components in MF-CAS.

Dataset	Method	Precision (%)	Recall (%)	Dice (%)	mIoU (%)
DeepCrack	① w/o SPM+CPCA	88.32	90.97	89.63	90.18
	② w/ SPM	88.24	91.22	89.70	90.24
	③ w/ CPCA	89.12	91.14	90.11	90.74
	④ w/ SPM+CPCA(DSCANet)	**89.31**	**91.35**	**90.32**	**90.80**
CFD	① w/o SPM+CPCA	44.02	76.17	55.79	68.23
	② w/ SPM	56.00	70.08	62.26	71.82
	③ w/ CPCA	53.85	77.02	63.38	72.38
	④ w/ SPM+CPCA(DSCANet)	**56.85**	**77.07**	**65.43**	**75.39**

In Table 3, we study different attention mechanisms in DSCA module. We replace the CPCA module in DSCA with other attention mechanisms in Method ② and Method ③, including the SE [9] and the CBAM [26]. In Method ①, we test the performance without any attention mechanism in DSCA. Based on the experimental results, we draw the following conclusions: Firstly, compared to Method ② in Table 2, it is evident that MF-DC module enhances the model performance even without attention mechanism. Secondly, DSCA with attention mechanisms make the model exhibit superior performance. Finally, the performance of Method ④ surpasses that of Method ③ and Method ② due to the dynamic channel and spatial weights in CPCA.

In Table 4, we examined the contribution of the SPM and the CPCA in MFCAS module. Method ① denotes the absence of these two components, Method ② only involves the inclusion of the SPM in the ASPP, and Method ③ incorporates the CPCA mechanism solely in the feature fusion after the ASPP. The results indicate that both components contribute to enhancing the model's performance.

4.3 Visualization

Fig. 5 visualizes the segmentation results of different models on the DeepCrack dataset. The visualization demonstrates that the segmentation results of our model are closer to the ground truth in terms of structural morphological, coarseness and connectivity than those of other models. Our model also exhibits stronger resistance to background noise and higher sensitivity in subtle structures.

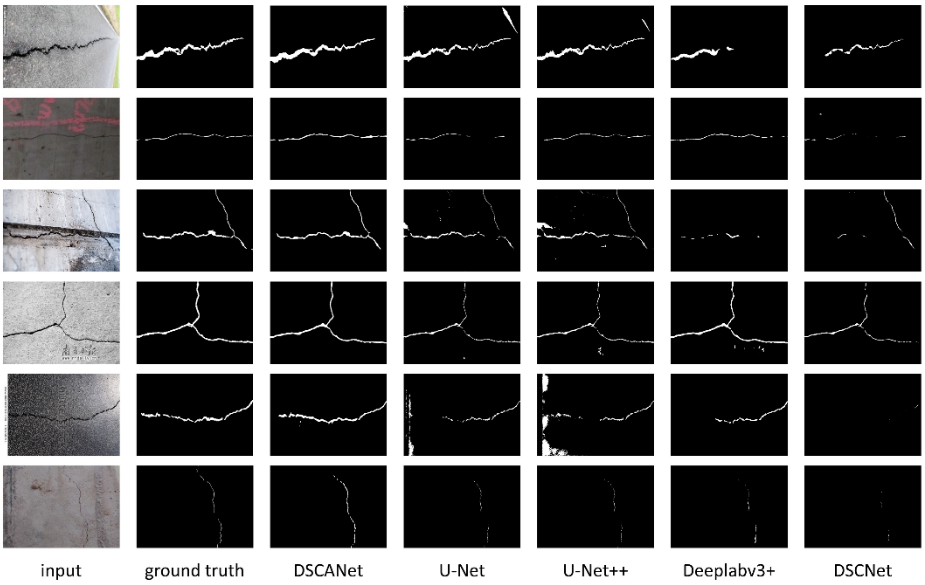

Fig. 5. The segmentation results on DeepCrack dataset.

5 Conclusion

In this paper, we propose a novel network named DSCANet for crack segmentation. We propose a Dynamic Snake Convolution with Attention module to extract slender and irregular features of cracks. Moreover, we propose MF-DC module for local information extraction and MF-CAS module for global information extraction, which can fuse features at various levels and scales. Our model performs exceptionally well and outperforms many advanced segmentation method on two datasets.

References

1. Abdel-Qader, I., Abudayyeh, O., Kelly, M.E.: Analysis of edge-detection techniques for crack identification in bridges. Journal of computing in civil engineering 17(4), 255–263 (2003)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence 39(12), 2481–2495 (2017)
3. 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 transactions on pattern analysis and machine intelligence 40(4), 834–848 (2017)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
5. Cui, L., Qi, Z., Chen, Z., Meng, F., Shi, Y.: Pavement distress detection using random decision forests. In: Data Science: Second International Conference, ICDS 2015, Sydney, Australia, August 8-9, 2015, Proceedings 2. pp. 95–102. Springer (2015)
6. 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. pp. 764–773 (2017)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Hou, Q., Zhang, L., Cheng, M.M., Feng, J.: Strip pooling: Rethinking spatial pooling for scene parsing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4003–4012 (2020)
9. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
10. Huang, H., Chen, Z., Zou, Y., Lu, M., Chen, C.: Channel prior convolutional attention for medical image segmentation. arXiv preprint arXiv:2306.05196 (2023)
11. Iyer, S., Sinha, S.K.: A robust approach for automatic detection and segmentation of cracks in underground pipeline images. Image and Vision Computing 23(10), 921–933 (2005)
12. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D.: Resunet++: An advanced architecture for medical image segmentation. In: 2019 IEEE international symposium on multimedia (ISM). pp. 225– 2255. IEEE (2019)
13. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 510–519 (2019)
14. Liu, F., Xu, G., Yang, Y., Niu, X., Pan, Y.: Novel approach to pavement cracking automatic detection based on segment extending. In: 2008 International Symposium on Knowledge Acquisition and Modeling. pp. 610–614. IEEE (2008)
15. Liu, Y., Yao, J., Lu, X., Xie, R., Li, L.: Deepcrack: A deep hierarchical feature learning architecture for crack segmentation. Neurocomputing 338, 139–153 (2019)
16. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
17. Maode, Y., Shaobo, B., Kun, X., Yuyao, H.: Pavement crack detection and analysis for high-grade highway. In: 2007 8th International Conference on Electronic Measurement and Instruments. pp. 4–548. IEEE (2007)

18. Mei, Q., Gül, M.: Multi-level feature fusion in densely connected deep-learning architecture and depth-first search for crack segmentation on images collected with smartphones. Structural Health Monitoring 19(6), 1726–1744 (2020)
19. Mei, Q., Gül, M.: Multi-level feature fusion in densely connected deep-learning architecture and depth-first search for crack segmentation on images collected with smartphones. Structural Health Monitoring 19(6), 1726–1744 (2020)
20. Oliveira, H., Correia, P.L.: Automatic road crack segmentation using entropy and image dynamic thresholding. In: 2009 17th European Signal Processing Conference. pp. 622–626. IEEE (2009)
21. Qi, Y., He, Y., Qi, X., Zhang, Y., Yang, G.: Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6070– 6079 (2023)
22. 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. Construction and Building Materials 234, 117367 (2020)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
24. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7262–7272 (2021)
25. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11534–11542 (2020)
26. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
27. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
28. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6881–6890 (2021)
29. Zhou, Z., Zhang, J., Gong, C.: Hybrid semantic segmentation for tunnel lining cracks based on swin transformer and convolutional neural network. Computer Aided Civil and Infrastructure Engineering 38(17), 2491–2510 (2023)
30. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer (2018)
31. Zou, Q., Zhang, Z., Li, Q., Qi, X., Wang, Q., Wang, S.: Deepcrack: Learning hierarchical convolutional features for crack detection. IEEE transactions on image processing 28(3), 1498–1512 (2018)