scientific reports

Check for updates

OPEN

A crack detection model fusing local details and global context for nuclear cladding coating surfaces

Wei Quan✉, Xiwen Li, Jie Yang & Zhaoyu Chen

Crack detection on the surface of nuclear cladding coatings is critical for ensuring the safe operation of nuclear power plants. However, due to the imbalance between crack and background pixels, complex crack morphology, numerous interfering factors, and the subtle features of fine cracks in nuclear cladding coating surface images, the detection performance of existing methods remains unsatisfactory. To address these issues, this paper proposes a novel crack detection model for nuclear cladding coatings surfaces, named CrackCTFuse. This model effectively captures both local detailed features and global context in crack images. Additionally, a crack local feature enhancement module is designed to supplement and enhance the edge details information of cracks, and a crack feature fusion module is proposed to facilitate the effective integration of local and global features. Moreover, a multi-scale convolutional attention module based on channel segmentation is developed to aggregate multi-scale contextual information, enhance skip connections, and improve the model's ability to perceive and represent crack features at various scales. Experiments conducted on the constructed nuclear cladding coating surfaces crack dataset demonstrate the effectiveness and accuracy of the CrackCTFuse model, achieving a MIoU of 92.70% and an F1-score of 92.54%.

In the operation of nuclear power plants, the nuclear cladding coating is an important component of the nuclear fuel assembly. The safety performance of the surface coating directly impacts the antioxidant performance, corrosion resistance, and mechanical strength of the nuclear cladding coating material, thereby influencing the operational safety of the entire reactor[1,2]. Under actual working conditions, the nuclear cladding coating may face mechanical deformation and other conditions, leading to cracks of various shapes and sizes on the coating surfaces. These cracks can significantly compromise the performance of coating and pose substantial safety risks[3]. Therefore, studying crack detection technology on nuclear cladding coating surfaces and achieving accurate detection and identification of cracks on nuclear cladding coatings is of significant importance and practical value for the safe operation and routine maintenance of reactors in the nuclear industry.

In recent years, numerous researchers have focused on crack detection and identification techniques, and a large number of crack detection methods have emerged. In terms of the development trend of the detection methods, crack detection methods for nuclear cladding coating surfaces can be categorized into classical crack detection methods and deep learning-based crack detection methods. Classical crack detection methods are mainly divided into physical detection techniques based crack detection methods[4,5], image processing techniques based crack detection methods[6–8] and traditional machine learning based crack detection methods[9–11]. Physical detection technology-based crack detection methods employ ultrasonic, optical, and other physical detection technologies to identify cracks on the coating surface. However, these methods usually require specialized and expensive detection equipment, involve complex operation processes, and demand high levels of operator skill, resulting in low detection efficiency. Image processing-based crack detection methods are image preprocessing, threshold segmentation, feature extraction, and other operations to achieve crack recognition, segmentation, and detection. Despite their utility, these methods typically extract only shallow visual features such as grayscale, edge, and texture information. They lack the capability to extract high-level semantic features, which hinders their ability to achieve accurate and precise crack detection and identification results. Traditional machine learning-based methods learn patterns from existing crack image data and apply these patterns to new crack image data for crack detection and identification. However, these algorithms mainly rely on low-level handmade features and struggle to fully explore the high-level semantic information and contextual relationships of crack images. Consequently, their performance in detecting small cracks in complex scenes is often inadequate.

Deep learning-based crack detection methods have garnered significant attention and adoption due to their exceptional accuracy and robustness. By automatically extracting features from large-scale datasets,

School of Electrical Engineering, Southwest Jiaotong University, No. 999, Xi'an Road, Pidu District, Chengdu, Sichuan 611756, China. ✉email: wquan@swjtu.edu.cn

these models overcome the limitations of traditional approaches, leading to improved detection performance. Several methods have demonstrated promising results on road and concrete crack datasets[12–14]. However, unlike cracks in roads or concrete structures, crack detection on nuclear fuel cladding coatings presents the following challenges due to the harsh operational environment:

1. Subtle and weak crack features: The fine cracks on nuclear cladding surfaces exhibit low contrast and blurry edges, making them difficult to distinguish from the background.
2. Complex and noisy background: The presence of bubbles, crystalline structures, and surface scratches leads to frequent false detections due to visual similarity with actual cracks.
3. Diverse crack morphology and scale variance: Cracks vary in shape and size—from minute, thin fissures to large discontinuous patterns—posing significant challenges for single-scale feature extraction.

Existing crack detection techniques struggle to maintain effectiveness under these conditions. To address these challenges, we propose CrackCTFuse, a two-branch, parallel, pixel-level crack detection model that integrates local detail refinement with global contextual awareness. The key contributions of this work are as follows:

1. We propose a two-branch parallel pixel-level nuclear cladding surfaces crack detection model, named CrackCTFuse. This model demonstrated excellent performance on the nuclear cladding coating surfaces crack dataset, achieving an MIoU of 92.70% and an F1-score of 92.54%. It also performed well on the public crack dataset.
2. A crack local feature enhancement module (CLFE) is designed, incorporating dilated convolution, edge detection and squeeze excitation block to extract the local features of cracks. This module enhances edge detail information and increases the local receptive field.
3. A crack feature fusion module (CTSF) is proposed, utilizing the attention mechanism to enhance feature expression and perform fine-grained feature interaction, achieving effective fusion of local and global features.
4. A multi-scale convolutional attention module (MSCA) based on channel segmentation is designed to effectively aggregate multi-scale contextual information, enhance skip connections, and improve the model's ability to perceive and represent crack features at different scales. Its multiscale feature fusion helps the model distinguish true-scale patterns of the target while suppressing crack-independent interference.

Related work

Classical crack detection methods

In the past decades, researchers have relied on classical crack detection methods for crack detection. For example, Liu et al.[15] used scattering interferometry to detect and identify surface cracks in metal parts. However, rough surface textures may cause scattering noise similar to cracks, resulting in lower accuracy for surface and crack detection on rough metal parts. Vincitorio et al.[16] proposed a lensless Fourier-based optical crack detection method, which effectively improves the accuracy on rough surfaces. Tao et al.[17] proposed a method using linear structured light to detect cracks on the surface of nuclear fuel rods, which employs grey-scale two-dimensional search, horizontal projection, and first-order differentiation techniques to locate crack location regions, and uses the seed-point judgment method to determine the upper and lower boundaries of the cracks. Zhang et al.[18] addressed the complexity of detecting defects on the end faces of the nuclear fuel core by designing a left-right symmetric raster illumination image acquisition system, which detects defects by fusing the left and right structured light images and extracts the cross points, classifying them with a Gaussian mixture model, and performing morphological operations such as dilation. Shi et al.[19] proposed a crack detection method based on the CrackForest model, which uses a random forest to process crack multilevel gradient information and color information, redefining crack features and combining them with classification algorithms to significantly improve crack detection accuracy. Although the above methods have achieved certain detection effects, they suffer from low efficiency and accuracy when faced with complex noise interference. Additionally, the extracted crack feature information is limited, and the robustness of these methods is insufficient.

Deep learning based crack detection methods

In recent years, deep learning-based crack detection methods have been extensively researched and applied. Zhang et al.[20] designed a lightweight deep-learning model for crack image classification. Firstly, the original image is divided into multiple image blocks, which are then classified using the designed neural network, and finally, crack detection is achieved by merging the classified crack image blocks. Nie et al.[21] improved the YOLOv3 network by fusing the multi-scale crack information and enhancing the classifier to achieve higher crack detection accuracy and faster detection speed. Li et al.[22] proposed an improved YOLOv4 target detection model called YOLO-attention for real-time online detection of cracks on metal surfaces. The model maintains the high detection efficiency of YOLOv4 while introducing techniques such as the attention mechanism, multi-scale spatial pyramid pooling, and exponential moving average to improve the detection accuracy of small target defects on metal surfaces. Xu et al.[23]proposed a bridge crack detection algorithm based on an improved YOLOv8n model. To enhance the model's feature extraction capabilities, a global attention mechanism was integrated into the backbone and neck of the network, enabling the capture of more discriminative crack features. Additionally, the original feature fusion module was optimized using Gam-Concat, further improving the effectiveness of multi-scale feature integration and enhancing overall detection performance. Ali et al.[24] proposed a triple-stage method for crack detection in stone masonry, involving YOLO-based ensemble detection, MobileNetV2-enhanced U-Net segmentation, and spectral clustering for decision support. The approach significantly improves accuracy and robustness in identifying complex crack patterns compared to single-stage methods. However, detection-based methods can only provide coarse positional information

about cracks, as they output only bounding boxes, neglecting important details such as crack length and area. Consequently, segmentation-based approaches are now more commonly employed for crack detection. Pan et al.[25] proposed a semantic segmentation model that employs multi-feature fusion and the focal loss function to address challenges such as small surface cracks, high noise levels, and low contrast in nuclear cladding. This approach improves crack segmentation without increasing network depth, enabling precise detection of minute cracks. Zuo et al.[26] introduced MMPA-Net, a surface defect detection model that extracts multi-scale features and fuses global and local information to enhance the accuracy of defect identification. Although these methods perform well on public datasets, a major drawback is their inability to capture remote features within a global context. This constraint significantly reduces detection accuracy when dealing with extensive and complex crack patterns. Zhang et al.[27] proposed CPCDNet, a model designed to improve the continuity and accuracy of crack extraction in complex environments. The network incorporates a Crack Alignment Module (CAM) to better align discontinuous crack features, and introduces a Weighted Edge Cross-Entropy Loss (WECEL) to emphasize edge information during training, thereby enhancing the model's ability to detect fine and continuous crack structures.

To address this issue, researchers introduced the Transformer to leverage its strong performance in modeling long-range dependencies. While the Transformer excels at capturing global context, it has limitations in detecting fine-grained detail. Moreover, purely Transformer-based models are rarely used for crack image segmentation due to the relatively smaller size of crack image datasets compared to natural image datasets. Models applied to crack image segmentation are typically hybrid approaches that combine CNNs with Transformers. Liu et al.[12] proposed Crackformer, a network model for fine-grained crack detection, which utilizes the advantages of the Transformer model to capture long-range contextual information and employs small convolutional kernels for fine-grained attentional awareness, which achieves excellent performance in fine-grained crack detection. Shamsabadi et al.[13] proposed a Vision Transformer (ViT)-based framework for crack detection on asphalt and concrete surfaces, utilizing TransUnet[28] as the backbone network. By leveraging transfer learning and a micro-intersecting concatenated set loss function, their approach significantly enhanced real-world crack segmentation performance. Ju et al.[14] introduced a Transformer-based multiscale fusion model, TransMF, which captures crack images from both local and global perspectives by combining convolutional and Swin Transformer blocks. TransMF enhances the correlation between contexts and mitigates the impact of noise by fusing outputs from each layer of both the encoder and decoder modules at unique scales using a mixing module. Guo et al.[29] employed Swin Transformer as an encoder to learn global and remote semantic features of pavement cracks, leveraging its layered architecture. They used UperNet with an Attention Module as the decoder to recover more crack detail, improving segmentation accuracy. Yadav et al.[30] incorporated a dual-stream Transformer module into a single 2D CNN layer for damage recognition, enabling pixel-level accurate views of global and local features through bi-directional routing. Li et al.[31] proposed a semi-supervised instance segmentation method using the Segment Anything Model (SAM)[32] and deep transfer learning for pavement crack detection. The approach efficiently reduces annotation efforts, maintains high accuracy with limited labeled data, and demonstrates robust performance across diverse detection scenarios.

Methodology

In this section, the crack detection network model CrackCTFuse proposed in this paper will be introduced first. Then, the designed Crack Localized Feature Enhancement (CLFE) module, Crack Feature Fusion (CTSF) module, and Multi-scale Convolutional Attention (MSCA) module will be described in detail.

Model architecture

For the task of detecting cracks on the surface of the nuclear cladding coating, this paper proposes CrackCTFuse, a two-branch parallel pixel-level crack detection model that fuses local details and global context. As shown in Fig. 1, the model consists of five key components. The first component is the global context branch, which first slices the original cracked image into 12×16 image blocks as the input sequence of the Transformer. The high-dimensional image input data is transformed into a low-dimensional feature vector matrix through the Embedding layer, and then it is fed into the Transformer encoder, which outputs the feature vector sequence with the global information after eight coding modules. This sequence is combined with the up-sampling for the decoding to get four feature maps with different scales and containing global information. The second component is the local details branch, whose input is the original crack image. In this paper, we use the classical Resnet-34 pre-training model to extract features from the original crack data and obtain four feature maps of the same size as the global context branch. We employ a ResNet-34 backbone to extract hierarchical features, generating four feature maps of the same dimensions as those in the global context branch. To further enhance fine-grained crack details, we introduce a Crack Local Feature Enhancement (CLFE) module, which refines edge representations by enhancing the shallowest CNN-derived feature maps. The component part is the CTSF module, which combines multimodal fusion strategies and attention mechanisms to facilitate the effective integration of global and local features. The fourth component is the skip connection part, which introduces a multi-scale convolutional attention module MSCA into the traditional skip connection to efficiently aggregate multi-scale contextual information and enhance the model's ability to perceive and express crack features at different scales. The fifth component is segmentation prediction, performs convolution and upsampling on the fused feature maps to generate an output of the same dimensions as the ground-truth crack labels. Each pixel in this feature map is classified to obtain the final segmentation result map. Additionally, the gradient flow is enhanced through supplementary supervision of the global context and fusion branches, improving the training process and boosting model performance.

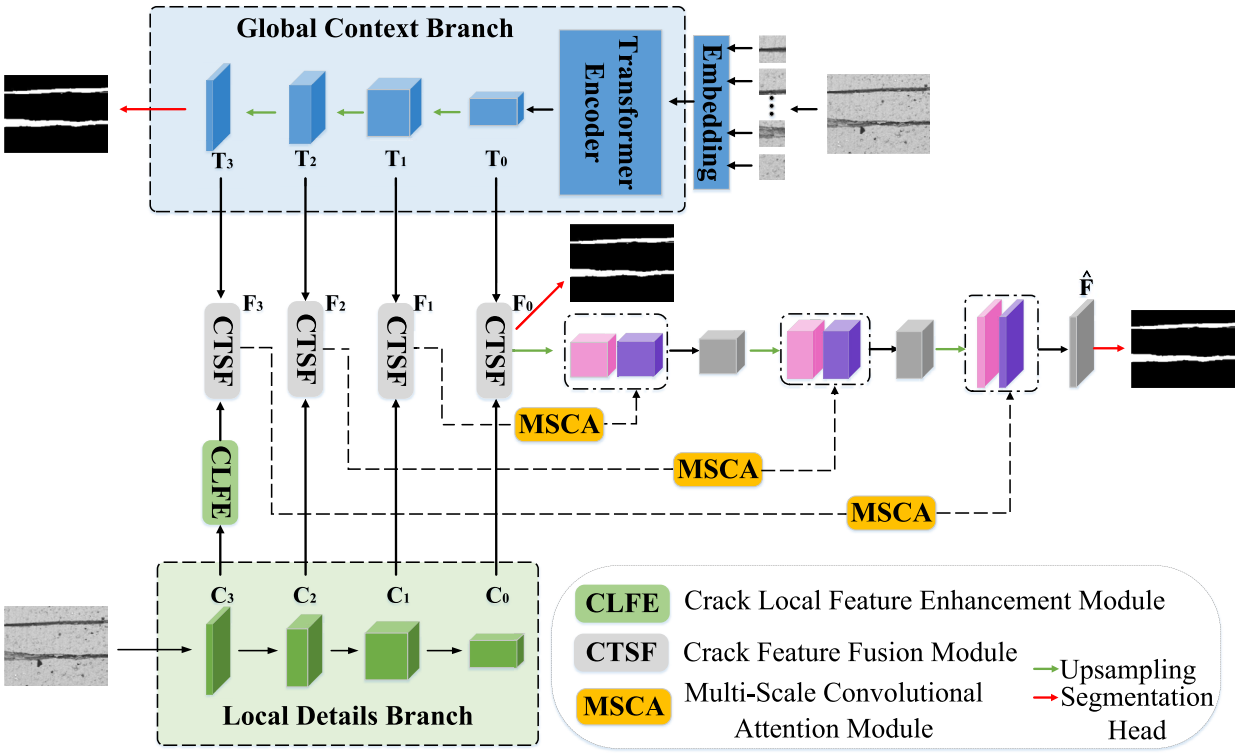

Fig. 1. Model architecture of the proposed CrackCTFuse.

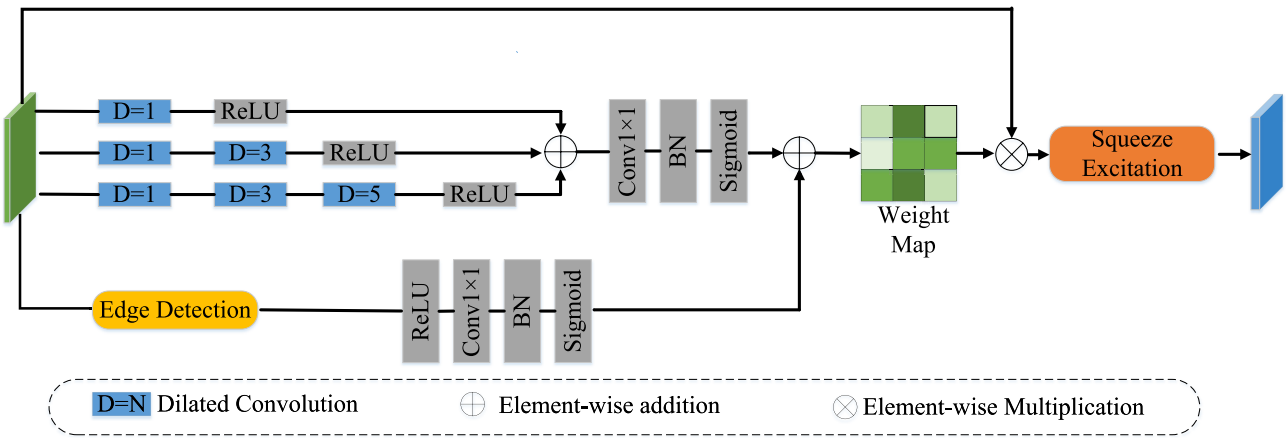

Fig. 2. Architecture of CLFE module.

Crack local feature enhancement module

Most of the commonly used CNN crack detection methods utilize multilayer convolution and pooling operations to extract the feature information of the crack image. However, high-frequency information, such as the edges and subtle branches of the cracks, which are crucial for accurate crack segmentation, are inevitably lost in the downsampling process. Inspired by Zhang et al.[33], who used the Sobel edge detector to enhance boundaries in medical image segmentation, in this paper, we design a crack localized feature enhancement module (CLFE), which combines edge detection with CNN to complement the edge detail information of enhanced cracks. The overall structure of the CLFE is shown in Fig. 2, which is divided into three main parts, which are multiscale feature extraction, edge detection module, and squeezing excitation module. The multi-scale feature extraction part sums up the input crack image features extracted by three convolutional branches to obtain multi-scale local information. Since the receptive field of ordinary convolution is small and cannot fully capture the local details in the crack image, dilated convolution is introduced to expand the receptive field. Specifically, the features F_{LD} extracted from the local details branch are used as input to this module, and the outputs from the three convolutional branches are fused via element-wise addition to D. This process can be defined as:

$$D = \sum_{p=0}^{2} \sum_{r\in\{1,3,\ldots,2p+1\}} \mathrm{ReLU}(\text{dilated conv}_r(F_{LD})) \tag{1}$$

where $\text{dilated conv}_r(\cdot)$ represents the dilated convolution with an dilation rate of r, and ReLU represents the ReLU activation function.

The edge detection module uses the multidirectional Sobel operator to compute the gradient of the input feature map, capturing the edge information in the crack image from different directions and generating more complete and accurate edge detection results. The squeeze excitation module learns the correlation between channels to adaptively adjust the weights of each channel, prompting the model to better focus on important features and suppress irrelevant ones. The output of the multi-scale feature extraction and edge detection module is processed by ReLU, 1×1 convolution, BN, and sigmoid functions to produce a spatial weight map that emphasizes key features and local detail information. Multiplying the input feature map by the generated spatial weight map weights the input to highlight the key areas and local detail information. Finally, after processing by the squeeze excitation module, the feature map with enhanced local information is output, and the edge detail of the crack is consolidated.

The edge detection module utilizes the traditional Sobel edge detection operator to process the input feature map and extract the crack edge information. In order to obtain richer edge information, the module employs four Sobel operators with different directions. These operators are applied to the input feature maps to compute gradient values in the horizontal, vertical, and diagonal directions. The absolute values of these gradients are then combined to produce the final edge detection results. The output of the edge detection module is as follows: The output of the edge detection module is as follows:

$$\begin{gathered} G = \sqrt{{G_0}^2 + {G_1}^2 + {G_2}^2 + {G_3}^2} \\ G_0 = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * I, G_1 = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * I \\ G_2 = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{bmatrix} * I, G_3 = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -2 \end{bmatrix} * I \end{gathered} \tag{2}$$

where G is the final edge detection feature map, G_0 is the result of gradient computation in the horizontal direction, G_1 is the result of gradient computation in the vertical direction, G_2 and G_3 are the results of gradient computation in the two diagonal directions, I is the input feature map, and $*$ denotes the convolution operation. Therefore, the weight map W_{map} can be computed as follows:

$$Wmap = \sigma(BN(conv_{1\times1}(\mathrm{ReLU}(D)))) + \sigma(BN(conv_{1\times1}(\mathrm{ReLU}(G)))) \tag{3}$$

where $\sigma(\cdot)$ denotes sigmoid function, $BN(\cdot)$ denotes batch normalization, $conv_{1\times1}$ denotes 1×1 convolution and ReLU represents the ReLU activation function.

The squeeze excitation module[34] mainly consists of a global average pooling and a fully connected layer that selectively enhances useful features and suppresses unimportant ones, and its structure is shown in Fig. 3. In the squeezing phase, the squeezing excitation block utilizes a global average pooling operation to reduce the input feature map to a channel vector, which encapsulates the aggregated features of each channel. In the excitation phase, this channel vector is processed through two fully connected layers followed by a sigmoid function to produce a weight vector. Each element of this weight vector represents the significance of the corresponding channel and acts as an excitation factor, adaptively adjusting the feature response of each channel in the feature map. Finally, the input feature map is multiplied by this weight vector, thereby weighting the feature map according to channel importance and enhancing relevant features while suppressing less significant ones.

Therefore, the output of CLFE can be calculated as:

$$CLFE(F_{LD}) = SE(F_{LD} \otimes Wmap) \tag{4}$$

where $SE(\cdot)$ denotes the squeezed excitation module and $\otimes$ denotes element-wise multiplication.

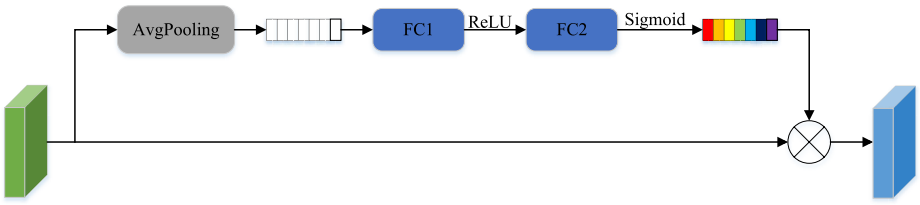

Fig. 3. Architecture of Squeeze Excitation Module.

Crack feature fusion module

To enhance the model's ability to represent crack features, it is necessary to effectively integrate the global and local features extracted from the global context branch and the local details branch. Many existing feature fusion methods typically use a sequential approach, achieving feature fusion by element-wise addition or channel concatenation of the two feature. However, these methods often overlook the correlation and importance differences between the features, which may lead to some useful feature information being masked or lost, thereby reducing the expressive power of the fused features. To address this issue, a new crack feature fusion module CTSF is designed in this paper. The overall structure of CTSF is shown in Fig. 4. This module leverages an attention mechanism to enhance feature representation while ensuring an effective integration of local and global information.

As shown in Fig. 4, where the input features are represented as the feature T_i extracted from the global context branch and the feature C_i extracted from the local details branch. Firstly, the Hadamard product is used to interact T_i and C_i with fine-grained information, and then the feature information is further extracted and integrated by convolutional operation to obtain the preliminary fused feature map B_i. Meanwhile, the scSE module[35] is introduced to enhance the feature representation. After T_i and C_i are processed by the scSE module, two enhanced feature maps are obtained. Finally, the enhanced feature maps from the global context branch, the enhanced feature maps from the local details branch, and the fused feature map Bi are concatenated along the channel direction. The feature map F_i with rich information is output by the residual block. The specific fusion strategy is shown as follows:

$$\begin{aligned} B_i &= Conv(T_i \odot C_i) \\ \hat{T}_i &= SCSE(T_i) \\ \hat{C}_i &= SCSE(C_i) \\ F_i &= Residual([B_i, \hat{T}_i, \hat{C}_i]) \end{aligned} \tag{5}$$

where $i \in [0, 3]$, conv is a 3×3 convolution, $\odot$ represents the Hadamard product, residual denotes the residual block, and SCSE denotes the operation processed utilizing the scSE module. The scSE module improves on the squeeze excitation module by introducing a spatial attention mechanism that fuses both the channel and spatial dimensions of the attention, which enhances meaningful features and suppresses useless ones. Its structure is shown in Fig. 5, the upper part uses 1×1 convolution to compress the channel dimension of the input features, activated by the sigmoid function to obtain the spatial weight map, and then multiply the spatial weight map with the input feature map to complete the spatial information calibration. The lower part is the squeeze excitation module introduced earlier, and superimpose the outputs of the two parts that is the final enhanced feature map. Specifically, assuming that the feature map U is fed into the scSE module, the computation process is as in 6,

$$\begin{aligned} U_1 &= \sigma(conv_{1\times1}(U)) \otimes U \\ U_2 &= \sigma(\mathrm{ReLU}(Avgpooling(U))) \otimes U \\ \hat{U} &= U_1 + U_2 \end{aligned} \tag{6}$$

where $\sigma(\cdot)$ denotes sigmoid function, ReLU denotes the ReLU activation function, $Avgpooling(\cdot)$ denotes the average pooling operation and $\otimes$ respresents element-wise multiplication.

Multi-scale convolutional attention module

The traditional skip connection does not consider the scale and semantic differences between different feature layers, and the high-level features cannot effectively utilize the detailed information of the low-level features, which limits the expressive power of the network. In order to aggregate multi-scale contextual information more effectively, a multi-scale convolutional attention module (MSCA) based on channel segmentation is designed in this paper to enhance the jump connection and enrich the extracted feature representations of the cracked

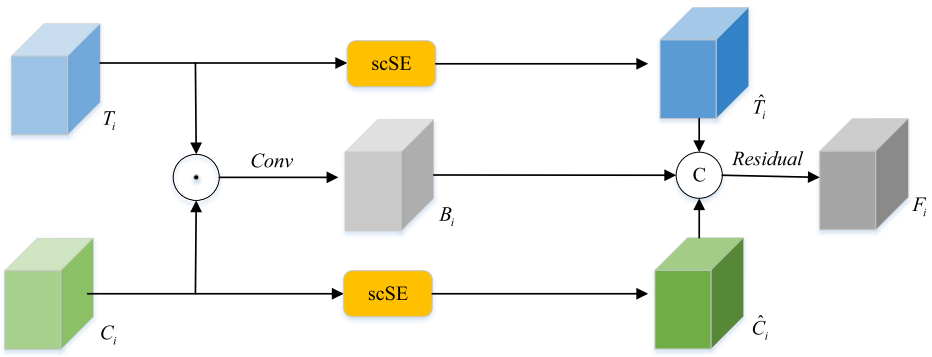

Fig. 4. Architecture of CTSF Module.

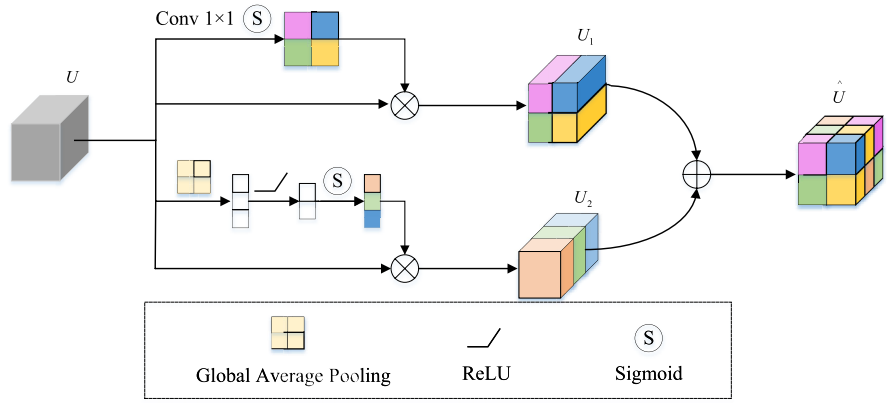

Fig. 5. Architecture of scSE Module

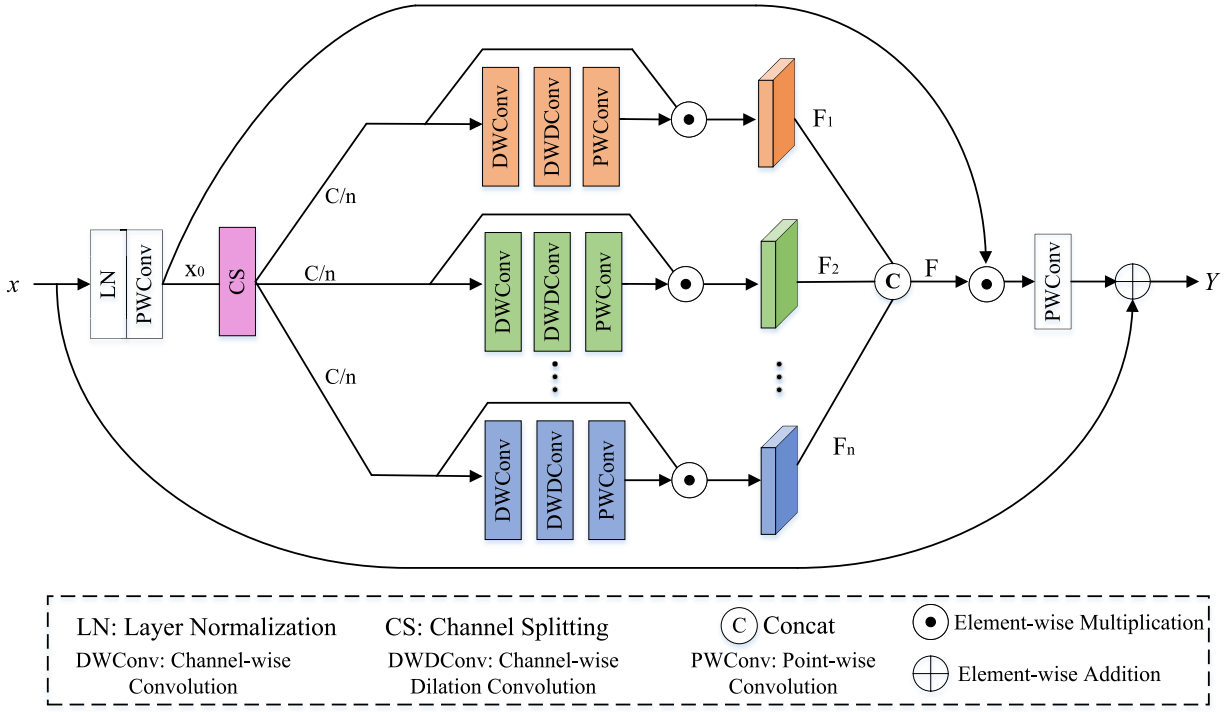

Fig. 6. Architecture of MSCA Module.

image. The overall structure of MSCA is shown in Fig. 6, which is mainly composed of combined convolutional blocks and residual connections.

As shown in Fig. 6, where the input feature map is denoted as *x*, the input features are first preprocessed using layer normalization and point-wise convolution, and then the preprocessed feature map x_0 is segmented along the channel dimension and divided into n sub-features, each denoted as x_i. Each x_i is processed by the combinatorial convolution block to obtain F_i, and F_i from different paths are concatenated along the channel to generate the fusion feature *F*. *F* is then multiplied element-wise with x_0 to achieve the enhanced representation of important features. Finally, through residual concatenation, the enhanced features are summed with the original input features to obtain a richer and more accurate crack information feature map *Y*. The detailed procedure is shown below:

$$\begin{aligned} x_0 = PWConv(LN(x))x_1, x_2, ..., x_n = Channelsplit(x_0) \\ F_i = x_i * PWConv(DWDConv(DWConv(x_i)))i = 1, ..., n \\ F = concat([F_1, F_2, ..., F_n], dim = 1) \\ Y = PWConv(x_0 * F) + x \end{aligned} \tag{7}$$

where *PWConv* ,*DWConv*, and *DWDConv* denote point-wise convolution, channel-wise convolution and channel dilation convolution, respectively. *LN* denotes layer normalization, and *Channelsplit* denotes the channel splitting. Channel-wise convolution learns spatial features within individual channels by performing convolution operations independently on each channel, thereby capturing local spatial information. In contrast, channel-wise dilation convolution broadens the receptive field, facilitating the capture of a larger range of contextual information. Point-wise convolution, on the other hand, enables information interaction and feature fusion across different channels. When combined, these convolution methods enhance the model's feature representation while keeping it lightweight, allowing for improved capture of both spatial and semantic information from the input data.

Loss function

In this paper, additional loss functions are introduced in the last layer of the global context branch and the first layer of the fusion block together for supervised training. This deeply supervised training method can accelerates the gradient propagation, enhances feature learning and improves the model's training effect. The overall expression of the training loss function of the model is shown in the following equation:

$$L = \lambda \times L(G, head(\hat{F})) + \alpha \times L(G, head(T_3)) + \beta \times L(G, head(F_0)) \tag{8}$$

where $L(G, head(\hat{F}))$ denotes the loss function between the final prediction result and the true label, $L(G, head(T_3))$ denotes the loss function between the global context branch prediction result and the label, $L(G, head(F_0))$ denotes the loss function between the output prediction of the first layer of the fusion block and the label, $\lambda = 0.5$, $\alpha = 0.3$, $\beta = 0.2$, all three are set empirically.

In the crack detection task, the effective expression of the crack edge information is crucial for accurate detection and segmentation. Inspired by the loss functions proposed in the study of Fan et al.[36], this paper introduces the structured loss function to represent three loss functions. This approach enhances the model's focus on edge regions, thereby improving the accuracy of the edge segmentation. Taking $L(G, head(\hat{F}))$ as a example to illustrate the introduced structured loss function, its calculation formula is shown below:

$$L(G, head(\hat{F})) = L_{IoU}^{\omega} + L_{BCE}^{\omega} \tag{9}$$

where L_{IoU}^{ω} is a weighted cross-merge ratio loss that motivates the model to better localize the crack boundary, and L_{BCE}^{ω} is a binary cross-entropy loss that strengthens the model's ability to identify the crack target and edge details. The calculations for L_{IoU}^{ω} and L_{BCE}^{ω} are shown below:

$$L_{IoU}^{\omega} = 1 - \frac{\omega \times \sum_{r=1}^{H}\sum_{c=1}^{W}(F(r,c)G(r,c)+1)}{\omega \times \sum_{r=1}^{H}\sum_{c=1}^{W}(F(r,c)+G(r,c)-F(r,c)G(r,c)+1)} \tag{10}$$

$$L_{BCE}^{\omega} = -\sum_{(r,c)}[G(r,c)log(F(r,c))+(1-G(r,c))log(1-F(r,c))] \tag{11}$$

where ω is the weight added by the loss function, adjusting its value can change how much attention the model pays to the edges of the image. The value of ω is the absolute value of the difference between the average pooled true labels and the original true labels. $G(r, c) \in \{0, 1\}$ represents the true value of each pixel point of the image, and $F(r, c)$ is the predicted pixel value.

Experiments

Datasets and evaluation metrics

In order to simulate the sprouting and expansion under of coating crack under actual working conditions, the nuclear cladding coating specimen after grinding, polishing and cleaning was placed the electronic mechanical testing machine to perform the annular compression experiment, as shown in Fig. 7.

After completing the annular compression test, the nuclear cladding coating material undergoes deformation, resulting in cracks of various shapes and sizes on the coating surface. Using an electron microscope, a small

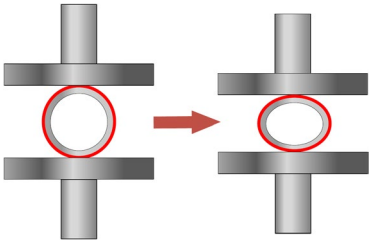

Fig. 7. Experimental simulation of annular compression.

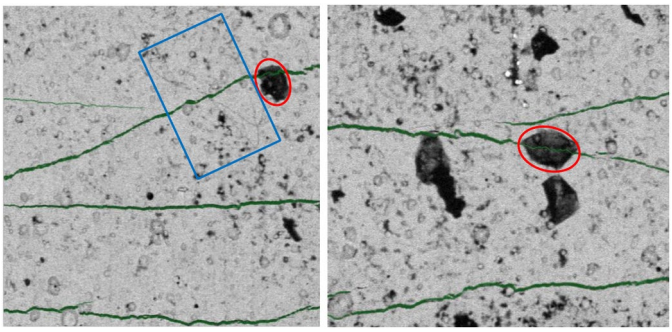

Fig. 8. Crack labeling diagram.

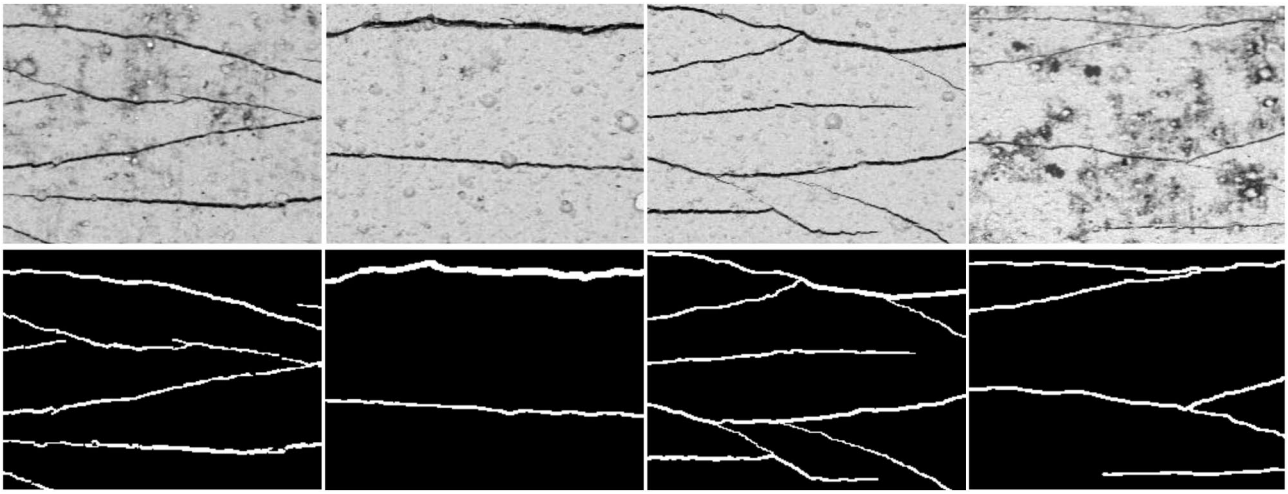

Fig. 9. The nuclear cladding coating surfaces cracks dataset.

amount of original crack image data was collected by photographing the surface of the nuclear cladding coating. The crack targets in the nuclear cladding coating surface images are labeled at the pixel level using Colabeler, as shown in Fig. 8.

To ensure high-quality labeled images, the following labeling standards and procedures were established prior to the labeling process: 1) All cracks, both coarse and fine, should be labeled. 2) For continuous cracks obscured by dust (as indicated by the red circle in Fig. 8), consult relevant professionals to label them as continuous cracks based on their trend. 3) Scratches on the coating surface (shown in the blue box in 8) should not be labeled as cracks. 4) Label all cracks with the category name "crack."

The collected crack images were processed through data enhancement to construct the crack image dataset used in this paper, containing a total of 2,750 images, as illustrated in Fig. 9. Furthermore, the crack dataset was divided into training, validation, and test sets consisting of 2,000, 150, and 600 images, respectively.

To evaluate the generalizability of CrackTFuse, the method described in this paper will be trained and assessed on commonly used crack datasets, including CrackLS315, DeepCrack537, and CFD. We compared the data variability of these datasets as shown in Table 1.

The CrackLS315[37] consists of 315 asphalt pavement images captured using a line-array camera under active illumination. Of these, 250 images are designated for training, 15 for validation, and 50 for testing. This dataset poses significant challenges for model training due to the presence of low-contrast images and elongated cracks, which are recorded under laser illumination.

The DeepCrack537[38] dataset comprises 537 images, with 400 allocated for training, 37 for validation, and 100 for testing. It presents a complex detection scenario, as its background encompasses diverse materials, including concrete and asphalt pavement. Additionally, the cracks exhibit significant variation in shape and size, with some partially obscured, further complicating the segmentation task.

The CFD(CrackForest-Dataset)[19] includes 118 road images, with 75 used for training, 18 for validation, and 25 for testing. These images are captured under uneven illumination and contain various types of noise, such as shadows, grease stains, and lane markings. The presence of extremely thin cracks further increases the difficulty of accurate detection.

Dataset	Background	Occlusion	Crack morphology	Noise
CrackLS315	Asphalt pavement	✗	Slender	✗
DeepCrack537	Concrete Asphalt pavement etc.	✓	Different sizes	✗
CFD	Asphalt pavement	✗	Slender	✓(Lane line)
Nuclear Cladding Coating Surfaces Crack	Nuclear Cladding Coating Surfaces	✓	Complex	✓(Scratch)

Table 1. Comparison of data variability in crack dataset.

Method	MIoU(%)	P(%)	R(%)	F1-Score(%)
U-Net	82.84	81.91	79.21	80.54
DeepLabv3+	80.65	72.75	82.84	77.47
SETR	78.95	74.62	75.63	75.12
SegFormer	80.52	79.25	75.87	77.52
Mask2Former	82.22	83.82	84.60	84.20
CrackFormer-II	87.19	91.09	80.75	90.92
SAM	83.23	87.79	85.32	86.53
CrackCTFuse	**92.70**	**88.47**	**97.01**	**92.54**

Table 2. Results of different segmentation methods on the nuclear cladding coating surface crack dataset.

Due to the complexity and variability of the surface cracks of the nuclear cladding coating, a single evaluation index cannot fully capture the algorithm's performance. Therefore, this paper adopts multiple commonly used semantic segmentation metrics, including MIoU, Precision (P), Recall (R), and F1-Score, for validation.

Experimental details

The environment for this experiment is pytorch1.8.1, the CNN encoder uses ResNet34 pre-trained on the ImageNet dataset, and the Transformer encoder uses the DeiT[39] pre-trained model. In this experiment, Adam is chosen as the optimizer, the initial learning rate is set to 0.0001, the batch size is set to 16, the number of training rounds is set to 1500, the number of training iterations per round is 125, and the resolution of the input image is 256×192.

Experimental results

In order to verify the effectiveness of the CrackCTFuse model designed in this paper, we compared it with several semantic segmentation models such as U-Net[40], Deeplabv3+[41] based on convolutional neural network and SETR[42], SegFormer[43] and Mask2Former[44] based on Transformer. In addition, we also compared with the specifically designed models for crack detection tasks, CrackFormer-II[45] and the latest visual large-scale model, SAM. Under identical experimental conditions, each model was trained and tested using the nuclear cladding coating surface crack dataset constructed in this paper. The evaluation metrics for each model are presented in Table 2. Compared to the above models, the CrackCTFuse showed substantial improvements across all evaluation metrics. Specifically, the MIoU is improved by 9.86%, 12.05%, 13.75%, 12.18%, 10.48%, 5.51% and 9.47% respectively, while the F1-Score is improved by 12%, 15.07%, 17.42%, 15.02%, 8.34%, 1.62% and 6.01%. This comparative analysis demonstrates that the CrackCTFuse model excels in detecting cracks on nuclear cladding coating surfaces. It significantly outperforms segmentation models that rely solely on either CNN or Transformer architectures. This confirms that the fusion architecture effectively leverages the complementary strengths of both networks.

Figure 10 compares crack segmentation results of various models on complex test images. The U-Net model produces smooth but imprecise results with severe leakage and false detections. SegFormer struggles with discontinuous segmentation and misses small cracks. Mask2Former retains main crack features but loses fine detail, especially in branch cracks. CrackFormer improves overall detection but fails on long, thin cracks and edge accuracy. While SAM has shown improvements in detection accuracy, it still produces false detections and requires extended training times and substantial GPU memory due to its large model size. In contrast, as shown in the last column of Fig. 10, the CrackCTFuse model produces results closest to the ground truth. It accurately detects and segments most crack regions, effectively captures edge details, and significantly reduces both false detection and leakage rates, thereby enhancing the accuracy and reliability of the segmentation.

Table 3 presents a comparative efficiency analysis of various methods evaluated on the nuclear cladding coating surface crack dataset, focusing on computational complexity (FLOPs), model size (parameters), and inference speed (FPS). U-Net exhibits the highest computational cost (84.8G FLOPs) with moderate inference speed (45 FPS). Transformer-based methods such as SETR and Mask2Former show relatively high computational overhead and larger model sizes, resulting in lower inference speeds (16 FPS and 27 FPS, respectively). SegFormer achieves optimal computational efficiency, with low FLOPs (2.57G), minimal parameters (3.72M), and high

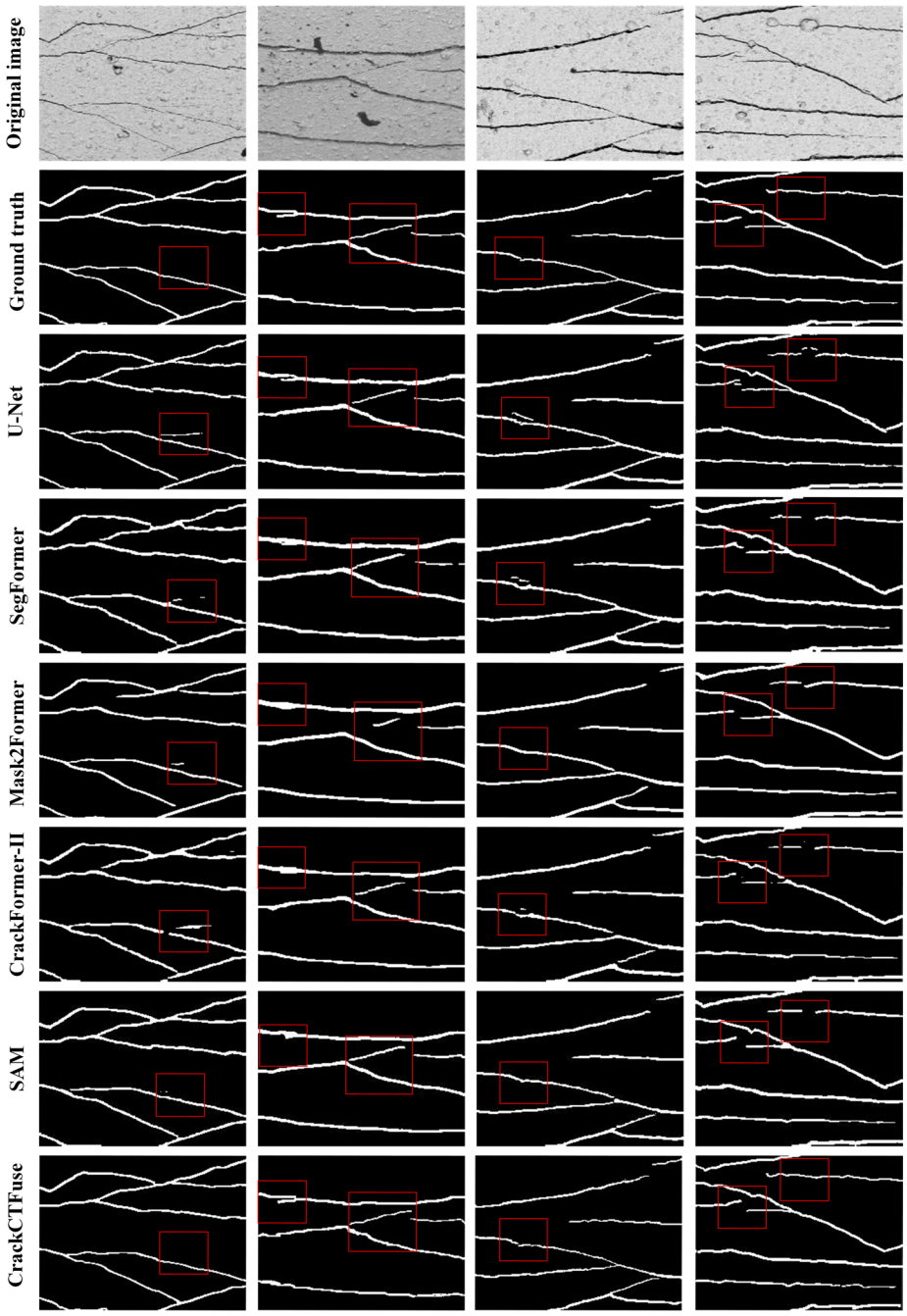

Fig. 10. Comparison of results of different segmentation methods on the nuclear cladding coating surfaces crack dataset.

inference speed (47 FPS). The proposed CrackCTFuse strikes a balanced trade-off, demonstrating moderate computational complexity (10.21G FLOPs), reasonable parameter count (26.11M), and acceptable inference speed (33 FPS), suggesting its suitability for effective and efficient crack segmentation in practical scenarios. Since SAM is a large-scale foundation model, its computational cost and parameter size were not included in the analysis.

Given that nuclear cladding coating typically operates in harsh environments, we processed the nuclear cladding coating surface crack dataset constructed in this paper to evaluate the performance of CrackCTFuse in real-world applications. The processing included simulated dust, reduced luminance, and added noise. The test was then repeated under the same experimental conditions.

As shown in the Table 4, when dealing with a more complex background that includes occlusion, each evaluation metric for each method is slightly reduced compared to the original image. However, compared to other models, the MIoU has improved by 2.04% to 17.09%, while the F1-Score has increased by 3.22% to 21.01%, respectively. From the Fig. 11, it is evident that the limitations of each method become more pronounced when

Method	FLOPs(G)	Paras(M)	FPS
U-Net	84.8	24.89	45
DeepLabv3+	9.94	5.81	46
SETR	6.51	63.99	16
SegFormer	2.57	3.72	47
Mask2Former	22.73	47.32	27
CrackFormer-II	15.34	4.55	46
SAM	-	-	8
CrackCTFuse	10.21	26.11	33

Table 3. The performance efficiency analysis on the nuclear cladding coating surface crack dataset.

Method	MIoU(%)	P(%)	R(%)	F1-Score(%)
U-Net	73.31	79.74	80.68	80.20
DeepLabv3+	72.36	77.51	75.21	76.34
SETR	67.42	71.11	67.14	69.07
SegFormer	72.18	77.24	72.46	74.77
Mask2Former	75.95	80.52	79.06	79.78
CrackFormer-II	82.47	84.97	88.84	86.86
SAM	79.72	83.00	84.28	83.63
CrackCTFuse	**84.51**	**87.20**	**93.17**	**90.08**

Table 4. Results of different segmentation methods on the **PROCESSED** nuclear cladding coating surfaces crack dataset.

processing occluded images. The results predicted by U-Net, SegFormer, Mask2Former, CrackFormer, and SAM show breaks in the cracks when occlusion occurs, with SegFormer exhibiting this phenomenon the most severely. Additionally, the crack boundaries predicted by SegFormer and Mask2Former are relatively rough, with some expansion. CrackFormer-II and SAM perform better, but there are still false detection issues. In contrast, the CrackCTFuse model proposed in this paper balances crack width and detail retention more effectively, capturing finer crack boundaries.

In terms of model generalizability, the results of CrackCTFuse on the CrackLS315, DeepCrack537, and CFD datasets are presented in the Tab. 5, where CrackCTFuse consistently achieves either optimal or suboptimal performance in each metric across all three datasets.On the CrackLS315 dataset, the MIoU improved by 0.15%–27.36%, and the F1-Score increased by 7.35%–22.86%. On the DeepCrack dataset, the MIoU improved by 0.57%–6.4%, and the F1-Score increased by 0.09%–9.61%. On the CFD dataset, the MIoU improved by 0.22%–21.56%, and the F1-Score increased by 3.17%–20.85%. This provides ample evidence that CrackCTFuse possesses strong generalization capabilities.

Ablation experiment

In order to validate the effectiveness of each module in the designed CrackCTFuse model, we constructed a baseline model (Baseline) for comparison. This Baseline model removes three key modules in the CrackCTFuse model: CLFE, CTSF (replacing it with a simple stitching operation) and MSCA. Based on this setup, a series of ablation experiments were conducted to evaluate the role of each module in the crack detection task and to assess the impact of their combination on the overall performance of the model. The results of the ablation experiments are presented in Table 6.

As shown in Table 6, adding each module individually improves the performance of the baseline model to varying degrees. Specifically, adding the CLFE module alone results in a 1.63% improvement in IoU and a 2.14% increase in Recall compared to the baseline model. CTSF module alone enhances IoU by 2.13% and Precision by 2.28%. When the MSCA module is added alone, Recall improves by 2.40%. Moreover, combining different modules further enhances the model's overall performance compared to adding individual modules. The most significant improvement is observed when the MSCA and CTSF modules are used together. When all modules are incorporated into the baseline model, forming the complete CrackCTFuse model, all performance metrics reach their optimal values. This demonstrates the strong complementarity and synergy among the modules, verifying the crucial role of their combination in enhancing the performance of crack detection segmentation. Although the FPS decreases slightly from 42 to 33 with the addition of modules, the model maintains efficient inference speed, indicating an effective balance between computational complexity and segmentation precision.

Table 7 presents ablation studies examining the impact of different configurations of global context and local detail branches on segmentation performance, measured by MIoU, Recall, and FPS. Removing either branch significantly reduces segmentation accuracy, highlighting the necessity of both global context and local detail information. Incrementally adding layers to each branch steadily improves performance, with the highest MIoU (92.70%) and Recall (97.01%) achieved when fully utilizing all layers. Additionally, the model maintains stable

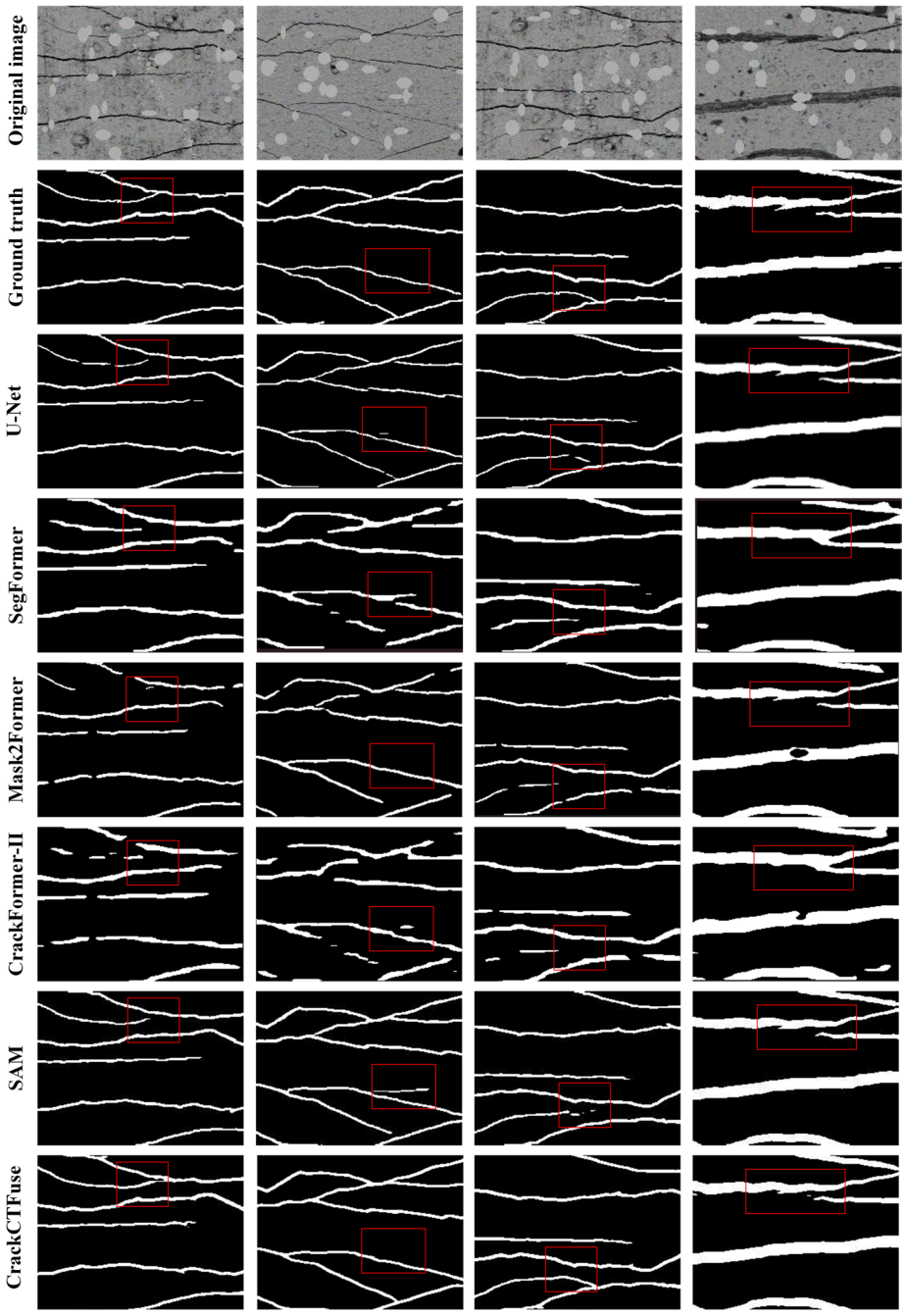

Fig. 11. Comparison of results of different segmentation methods on the **PROCESSED** nuclear cladding coating surfaces crack dataset.

inference efficiency, with minimal FPS reduction (from 34 to 33), demonstrating an effective trade-off between accuracy and computational cost.

Conclusion

In this paper, we propose two-branch parallel pixel-level crack detection model, CrackCTFuse, to achieve accurate detection of cracks on the surface of nuclear cladding coatings. CrackCTFuse effectively captures both local details features and global background information within crack images. In addition, in order to further improve the performance of the model, we designed three key modules: a crack local feature enhancement module to supplement and enhance the edge detail information of the cracks, a crack feature fusion module to effectively fuse local and global features, and a multi-scale convolutional attention module to efficiently aggregate multi-scale contextual information, enhance skip connections, and improve the model's ability of perceiving and expressing the crack features at different scales. The effectiveness and accuracy of the proposed CrackCTFuse model are validated through comparative experiments. On the nuclear cladding coating surfaces

Method	CrackLS315		DeepCrack537		CFD	
	MIoU	F1-Score	MIoU	F1-Score	MIoU	F1-Score
U-Net	65.30	62.54	83.70	81.52	66.79	72.44
DeepLabv3+	63.52	67.59	82.84	79.82	74.75	83.13
SETR	70.51	70.08	81.20	81.34	63.26	69.74
SegFormer	62.29	68.99	85.52	89.31	75.98	83.87
Mask2Former	74.65	79.47	85.92	87.67	77.86	82.25
CrackFormer-II	89.50	**88.68**	**87.60**	86.97	83.98	87.42
SAM	69.20	78.05	84.59	83.86	72.44	76.40
CrackCTFuse	**89.65**	85.40	87.03	**89.43**	**84.82**	**90.59**

Table 5. Results of different segmentation methods on commonly used crack datasets.

Case	IoU(%)	MIoU(%)	R(%)	P(%)	F1-Score(%)	FPS
Baseline	80.84	89.21	94.53	83.14	88.47	42
Baseline+CLFE	82.47	90.76	96.67	84.88	90.39	38
Baseline+CTSF	82.97	91.03	96.65	85.42	90.69	39
Baseline+MSCA	82.25	90.65	96.93	84.45	90.26	41
Baseline+MSCA+CTSF	85.13	91.98	96.76	87.13	91.69	38
Baseline+MSCA+CLFE	84.47	91.83	96.69	86.98	91.58	38
Baseline+CTSF+CLFE	84.01	91.59	96.73	86.47	91.31	34
CrackCTFuse	**86.12**	**92.70**	**97.01**	**88.47**	**92.54**	33

Table 6. Results of ablation experiment on the nuclear cladding coating surface crack dataset.

Global Context Branch	Local Details Branch	MIoU(%)	R(%)	FPS
$\{T_1, T_2, T_3, T_4\}$	✘	89.25	94.39	34
$\{T_1, T_2, T_3, T_4\}$	$\{C_1\}$	90.51	95.24	34
$\{T_1, T_2, T_3, T_4\}$	$\{C_1, C_2\}$	90.97	95.37	33
$\{T_1, T_2, T_3, T_4\}$	$\{C_1, C_2, C_3\}$	91.38	95.54	33
✘	$\{C_1, C_2, C_3, C_4\}$	89.42	94.84	35
$\{T_1\}$	$\{C_1, C_2, C_3, C_4\}$	90.65	95.23	34
$\{T_1, T_2\}$	$\{C_1, C_2, C_3, C_4\}$	91.78	96.16	34
$\{T_1, T_2, T_3\}$	$\{C_1, C_2, C_3, C_4\}$	92.18	96.33	34
$\{T_1, T_2, T_3, T_4\}$	$\{C_1, C_2, C_3, C_4\}$	**92.70**	**97.01**	33

Table 7. Ablation experiment of global context and local details branches

cracking dataset, CrackCTFuse achieved a MIoU of 92.70% and an F1-Score of 92.54%. The MIoU is improved by 9.86%, 12.05%, 13.75%, 12.18%, 10.48%, 5.51% and 9.47% respectively compared to U-Net, DeepLabv3+, SETR, SegFormer, Mask2Former, CrackFormer-II and SAM. The F1-Score is improved by 12%, 15.07%, 17.42%, 15.02%, 8.34%, 1.62% and 6.01%, respectively. Additionally, the robustness of CrackCTFuse in real-world applications was demonstrated by adding noise masking and other perturbations to the original nuclear cladding coating surfaces crack dataset, and repeating the experiments, which highlighted the model's resilience under harsh conditions. Besides, the model was trained and evaluated on three widely-used crack datasets—CrackLS315, DeepCrack537, and CFD—where it consistently achieved optimal or suboptimal performance, validating its generalization capability to some extent. Finally, ablation experiments were conducted to verify the contribution of each module to the model's performance improvement and to demonstrate the effective synergy between the models.

Data availability

The datasets generated and analysed during the current study are not publicly available due to the requirement of the dataset provider institution but are available from the corresponding author on reasonable request. The data presented in this study are available upon request from the corresponding author.

Received: 18 April 2025; Accepted: 19 August 2025
Published online: 22 August 2025

References

1. Haixia, J., Zewen, D., Pengxiang, M. & Peng, W. Research progress on fretting wear behavior of fuel cladding materials in nuclear reactor. *Tribology* **41**, 423–436 (2021).
2. Tang, C. et al. High-temperature oxidation and hydrothermal corrosion of textured cr2alc-based coatings on zirconium alloy fuel cladding. *Surface and Coatings Technology* **419**, 127263 (2021).
3. Yang, Z. et al. High temperature and high pressure flowing water corrosion resistance of multi-arc ion plating cr/tialn and cr/tialsin coatings. *Materials Research Express* **6**, 086449 (2019).
4. Abou-Khousa, M. A., Rahman, M. S. U., Donnell, K. M. & Al Qaseer, M. T. Detection of surface cracks in metals using microwave and millimeter-wave nondestructive testing techniques–a review. *IEEE Transactions on Instrumentation and Measurement* **72**, 1–18 (2023).
5. Petersen, C. R. et al. Non-destructive subsurface inspection of marine and protective coatings using near-and mid-infrared optical coherence tomography. *Coatings* **11**, 877 (2021).
6. Gupta, P. & Dixit, M. Image-based crack detection approaches: a comprehensive survey. *Multimedia Tools and Applications* **81**, 40181–40229 (2022).
7. Oliveira, H. & Correia, P. L. Automatic road crack detection and characterization. *IEEE Transactions on Intelligent Transportation Systems* **14**, 155–168 (2012).
8. Hanzaei, S. H., Afshar, A. & Barazandeh, F. Automatic detection and classification of the ceramic tiles' surface defects. *Pattern recognition* **66**, 174–189 (2017).
9. Hittawe, M. M., Muddamsetty, S. M., Sidibé, D. & Mériaudeau, F. Multiple features extraction for timber defects detection and classification using svm. In *2015 IEEE International Conference on Image Processing (ICIP)*, 427–431 (IEEE, 2015).
10. Martins, L. A., Pádua, F. L. & Almeida, P. E. Automatic detection of surface defects on rolled steel using computer vision and artificial neural networks. In *IECON 2010-36th Annual Conference on IEEE Industrial Electronics Society*, 1081–1086 (IEEE, 2010).
11. Cubero-Fernandez, A., Rodriguez-Lozano, F. J., Villatoro, R., Olivares, J. & Palomares, J. M. Efficient pavement crack detection and classification. *EURASIP Journal on Image and Video Processing* **2017**, 1–11 (2017).
12. Liu, H., Miao, X., Mertz, C., Xu, C. & Kong, H. Crackformer: Transformer network for fine-grained crack detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3783–3792 (2021).
13. Shamsabadi, E. A., Xu, C. & Dias-da Costa, D. Robust crack detection in masonry structures with transformers. *Measurement* **200**, 111590 (2022).
14. Ju, X., Zhao, X. & Qian, S. Transmf: Transformer-based multi-scale fusion model for crack detection. *Mathematics* **10**, 2354 (2022).
15. Liu, K., Wu, S. J., Gao, X. Y. & Yang, L. X. Simultaneous measurement of in-plane and out-of-plane deformations using dual-beam spatial-carrier digital speckle pattern interferometry. *Applied mechanics and materials* **782**, 316–325 (2015).
16. Vincitorio, F., Bahuer, L., Fiorucci, M., López, A. & Ramil, A. Improvement of crack detection on rough materials by digital holographic interferometry in combination with non-uniform thermal loads. *Optik* **163**, 43–48 (2018).
17. Tao, X. et al. A survey of surface defect detection methods based on deep learning. *Acta Automatica Sinica* **47**, 1017–1034 (2021).
18. Zhang, B. et al. Defect inspection system of nuclear fuel pellet end faces based on machine vision. *Journal of Nuclear Science and Technology* **57**, 617–623 (2020).
19. Shi, Y., Cui, L., Qi, Z., Meng, F. & Chen, Z. Automatic road crack detection using random structured forests. *IEEE Transactions on Intelligent Transportation Systems* **17**, 3434–3445 (2016).
20. Zhang, L., Yang, F., Zhang, Y. D. & Zhu, Y. J. Road crack detection using deep convolutional neural network. In *2016 IEEE international conference on image processing (ICIP)*, 3708–3712 (IEEE, 2016).
21. Nie, M. & Wang, C. Pavement crack detection based on yolo v3. In *2019 2nd international conference on safety produce informatization (IICSPI)*, 327–330 (IEEE, 2019).
22. Li, W. et al. Deep learning based online metallic surface defect detection method for wire and arc additive manufacturing. *Robotics and Computer-Integrated Manufacturing* **80**, 102470 (2023).
23. Xu, W. et al. Improved YOLOv8n-based bridge crack detection algorithm under complex background conditions. *Scientific Reports* **15**, 13074 (2025).
24. Mayya, A. M. & Alkayem, N. F. Triple-stage crack detection in stone masonry using yolo-ensemble, mobilenetv2u-net, and spectral clustering. *Automation in Construction* **172**, 106045 (2025).
25. Pan, P., Xu, Y., Xing, C. & Chen, Y. Crack detection for nuclear containments based on multi-feature fused semantic segmentation. *Construction and Building Materials* **329**, 127137 (2022).
26. Zuo, L., Xiao, H., Wen, L. & Gao, L. A pixel-level segmentation convolutional neural network based on global and local feature fusion for surface defect detection. *IEEE Transactions on Instrumentation and Measurement* (2023).
27. Zhang, J., Sun, S., Song, W., Li, Y. & Teng, Q. A novel convolutional neural network for enhancing the continuity of pavement crack detection. *Scientific Reports* **14**, 30376 (2024).
28. Chen, J. et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021).
29. Guo, F., Liu, J., Lv, C. & Yu, H. A novel transformer-based network with attention mechanism for automatic pavement crack detection. *Construction and Building Materials* **391**, 131852 (2023).
30. Yadav, D. P., Chauhan, S., Kada, B. & Kumar, A. Spatial attention-based dual stream transformer for concrete defect identification. *Measurement* **218**, 113137 (2023).
31. Li, J., Yuan, C., Wang, X., Chen, G. & Ma, G. Semi-supervised crack detection using segment anything model and deep transfer learning. *Automation in Construction* **170**, 105899 (2025).
32. Kirillov, A. et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026 (2023).
33. Zhang, M., Yu, F., Zhao, J., Zhang, L. & Li, Q. Befd: Boundary enhancement and feature denoising for vessel segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, 775–785 (Springer, 2020).
34. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141 (2018).
35. Roy, A. G., Navab, N. & Wachinger, C. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, 421–429 (Springer, 2018).
36. Fan, D.-P. et al. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, 263–273 (Springer, 2020).
37. Zou, Q. et al. Deepcrack: Learning hierarchical convolutional features for crack detection. *IEEE transactions on image processing* **28**, 1498–1512 (2018).
38. 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).

39. Touvron, H. et al. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357 (PMLR, 2021).
40. 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*, 234–241 (Springer, 2015).
41. 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)*, 801–818 (2018).
42. Zheng, S. 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*, 6881–6890 (2021).
43. Xie, E. et al. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems* **34**, 12077–12090 (2021).
44. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A. & Girdhar, R. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299 (2022).
45. Liu, H., Yang, J., Miao, X., Mertz, C. & Kong, H. Crackformer network for pavement crack segmentation. *IEEE Transactions on Intelligent Transportation Systems* **24**, 9240–9252 (2023).

Author contributions

All authors contributed significantly to this paper. W.Q. and J.Y. developed the research idea of this paper and collected the experimental data. X.L and J.Y. carried out the experiments and analyzed the results. X.L. wrote the paper and revised the manuscript. W.Q. and Z.C. supervised the revision of the manuscript. All authors have read and agreed to the publication of this paper.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 52277127), Science and Technology Innovation Talent Project of Sichuan Province (Grant No. 2021JDRC0012), Key Interdisci pli-nary Basic Research Project of Southwest Jiaotong University (Grant No. 2682021ZTPY089).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.Q.

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

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025