

## RESEARCH ARTICLE

# Research on Crack Disease Identification Based on Visible Spectrum in Harsh Tunnel Environment

RUIJUN BAI<sup>1</sup>, JING GAO<sup>1</sup>, ZHONG LI<sup>2</sup>, DONGHANG LIU<sup>1</sup>, AND XUEKUI SHANGGUAN<sup>1</sup><sup>1</sup>Shanxi Information Industry Technology Research Institute Company Ltd., Taiyuan 030012, China<sup>2</sup>School of Software, North University of China, Taiyuan 030051, China

Corresponding author: Jing Gao (whiterj@foxmail.com)

This work was supported in part by the Shanxi Province Key Research and Development Program (International Science and Technology Cooperation) of China under Project 201703D421010.

**ABSTRACT** In recent years, deep learning-based crack detection techniques have been widely used in ground crack detection, urban street crack detection, ordinary wall crack detection, and road tunnel crack detection. However, due to the scarcity of data, crack detection in railway tunnels is temporarily rare, and at the same time, some existing railway tunnels of relatively old age have extremely limited lighting conditions, which are subject to the dark conditions in railway tunnels, as well as the structural surface noise and crack-like interferences that can cause great challenges to the identification of cracks in railway tunnels. Based on this, this paper collects images inside real-world railway tunnels, produces a dataset, and proposes a novel and effective hybrid neural network tunnel crack disease recognition iFormer Unet model, which is based on the iFormer block module that can extract high-frequency features and low-frequency features at the same time, and constructs a U-shape network consisting of an encoder, a Bottleneck, a decoder, and a jump connection U-shaped network composed of encoder, Bottleneck, decoder and jump connection. The results of 10-fold cross-validation in the experiments show that the proposed method has a relatively low misdetection rate of about 7.56%, with about 30.31M Params and 34.84G FLOPs. iFormer Unet model has the lowest misdetection rate compared to the Swin Unet and Unet models, which are 5.28% and 8.58% lower, respectively, when tested on six image categories. 5.28% and 8.58% respectively. The proposed iFormer Unet algorithm realises the automatic identification of cracks in railway tunnels under harsh environments, which provides a certain reference and basis for the maintenance of railway tunnels.

**INDEX TERMS** Harsh environment, railway tunnel, hybrid neural networks, high-frequency characteristics, low-frequency characteristics, crack identification.

## I. INTRODUCTION

When modern road transportation falls short of meeting transportation needs, railway transportation becomes particularly important. However, when constructing long-distance railway lines in natural and man-made terrain, it is often necessary to build railway tunnels in areas such as mountains and hills. The construction of a large number of railway tunnels improves transportation conditions, but it also inevitably leads to a significant increase in the amount of tunnel construction and maintenance work. As railway tunnels are used

year after year, cracks, deformations, and other ailments are prone to occur, which pose a great threat to traffic safety. It is particularly important to carry out ailment detection for newly built and existing tunnels to ensure the safe operation of tunnels.

Cracking is one of the major defects in railway tunnels. Traditional methods for crack detection in railway tunnels rely mainly on the visual inspection of maintenance personnel or simple devices, which are based on personal experience and subjective judgment. However, these methods lack both reliability and efficiency, failing to meet the demands for safe operation and development of railway tunnels. As a result, the development of automated crack detection using intelligent

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin<sup>1</sup>.

systems has become an evolving technological direction in this field. Currently, many scholars have proposed some detection methods for crack lesions for different application scenarios. In terms of target detection, Zhou et al. [1] proposed a data augmentation-based deep learning detection method for YOLOv4 cracks, which firstly enriches the dataset by enabling automatic generation of crack images through data augmentation methods, and secondly selects YOLOv4 as the basic model for training, and introduces a pruning algorithm to reduce the size of the model so as to efficiently perform crack detection. Liu et al. [2] also used an image enhancement algorithm to increase the crack image data, and then achieved the detection of road tunnel cracks through a transfer learning method. Zhou et al. [3] propose an improved YOLOX algorithm with an improved backbone network, an increased attention module, and also a replacement of a more suitable loss function for tunnel crack detection.

In the area of image segmentation, Kang and Cha [4] proposed a new deep encoder and decoder based network to detect pixel level cracks in complex scenes by improving/enhancing the dataset and performance. Ali and Cha. [5] GAN was used to generate synthetic image data in order to multiply the dataset and to segment the internal damage of concrete components at pixel level using active thermography, but this method cannot be used when the concrete is wet or other disturbing factors. Choi and Cha [6] proposed an original convolutional neural network, the model consists of standard convolution, densely connected separable convolutional modules, a modified spatial pyramid module, and a decoder module, and verified to have a good performance in recognising cracks in urban streets by collecting the produced dataset. Protopapadakis et al. [7] proposed a crack detection mechanism for concrete tunnel surfaces that utilises deep convolutional neural networks and domain-specific heuristic post-processing techniques for data processing and was validated at the Egnatia motorway tunnel in Metsovo, Greece. Makantasis et al. [8] proposed a deep learning based approach for the detection of concrete defects in tunnels using a convolutional neural network to hierarchically construct high-level features from low-level features for describing the defects, as well as a multilayer perceptron to perform the detection task. Shuangxi et al. [9] proposed a deep learning target detection framework combining texture features and concrete crack data by merging texture features and pre-processed concrete data to increase the number of feature channels. Zhou et al. [10] Based on a deep learning approach, proposed a crack detection network consisting of a hybrid attention module based on effective embedded channel and positional information, as well as an integrated RFE and a multiscale feature fusion module for the detection of cracks on the surface of tunnel linings.

Some scholars have integrated intelligent algorithms into robotic platforms, Liao et al. [11] proposed a new fast tunnel crack detection device, which consists of a novel mobile imaging module and an automatic crack detection

module. The imaging module consists of a high-resolution charge-coupled device (CCD) camera array, a mobile laser scanner and an illumination array. The core of the crack detection module utilises a novel lightweight convolutional neural network for tunnel crack detection. In-Ho et al. [12] acquired bridge images from an unmanned aerial vehicle (UAV) equipped with a high-performance vision sensor and trained and recognised bridge cracks based on a convolutional neural network. Protopapadakis et al. [13] proposed a working prototype for visual inspection of tunnels. Firstly, it was crack detection by deep learning method. Then, a detailed 3D model of the cracked area was created using photogrammetry. Finally, laser profiling of tunnels close to a narrow region of detected cracks was performed and validated on the Egnatia motorway and on underground infrastructure in London. Loupos et al. [14] proposed a robotic platform that automates tunnel inspection. This robotic platform consists of a crane arm, a high-precision robotic arm, a computer vision system, a 3D laser scanner and ultrasonic sensors, and utilises a multidisciplinary and multimodal approach to automate the inspection of transport tunnels and analyse potential defects. However, the robotic platform is installed and deployed on a 5 tonne crane vehicle and the cost of tunnel inspection is significantly higher.

Some other authors have also applied deep learning methods to other areas of disease detection. Katsamenis et al. [15] used the unet method to identify rust lesions on metallic structures. Lewis et al. [16] proposed a dual codec network for segmentation of colorectal polyp lesions. Wang et al. [17] proposed a two-stage approach using edge detection and convolutional neural networks for crack identification on railway sleepers. Yang and Mei [18] proposed a deep transfer learning method for crack identification on mountain slopes to guard against geological hazards such as landslides.

The application of crack detection by the above scholars can be mainly divided into ground crack detection, city street crack detection, ordinary wall crack detection, road tunnel crack detection and other aspects. Ground cracks, city street cracks, ordinary wall cracks of these three types of data lighting conditions are better, cracks and non-crack texture and other characteristics are obvious, easy to detect. Road tunnels are generally fitted with lighting equipment due to the high number of vehicles travelling through the tunnel, so the lighting conditions for data image acquisition in road tunnels are also relatively good. The data used in this study are railway tunnel image data, and some of the existing railway tunnels of relatively old age have extremely limited lighting conditions, and the tunnels are generally only fitted with emergency lights that cannot support the lighting of the railway tunnels. Therefore, the essential difference between this study and the above scholars is that the dark conditions inside railway tunnels, as well as structural surface noise and crack-like interferences can pose a significant challenge to the identification of cracks in railway tunnels.

Our contribution can be summarised as follows:

1. Due to the current lack of readily available railway tunnel crack datasets, as well as the fact that many scholars' crack datasets are augmented by image synthesis techniques such as GAN, which do not accurately reflect the real-world data situation, for this reason, a visible-spectrum image acquisition device for railway tunnels was designed to collect images of real-world tunnels inside the Taoping Tunnel of the Houyue Line of the Zhengzhou Railway Bureau, China, and produce a dataset which of harsh lighting conditions, and in addition to crack damage, the tunnel walls are accompanied by structural surface noise such as concave holes, wall bulges, shadows, and seepage flow marks. The focus of this research is to process the acquired images to identify cracks in railway tunnels.

2. After analysing the divided six tunnel crack images, a novel and effective hybrid neural network tunnel crack disease recognition iFormer Unet model is proposed, which is based on the iFormer block module that can extract both high-frequency features and low-frequency features, and constructs a U-shape network consisting of Encoder, Bottleneck, Decoder, and Jump Connection, in which The main structure of the iFormer block module consists of a hybrid module with parallel maximum pooling, parallel deep convolution, and parallel self-attention mechanism. The experimental results show that the model can effectively identify the railway tunnel crack disease.

3. Through the verification of the experiment, this study can provide certain reference for the scientific maintenance of railway tunnels.

## II. MATERIALS AND METHOD

### A. DATA COLLECTION AND CLASSIFICATION OF RAILWAY TUNNEL CRACKS

The experimental data were collected from Taoping Tunnel on the Houyue line of Zhengzhou Railway in China, at a tunnel depth of approximately 1km. Data was collected using a railway tunnel visible spectral image acquisition device equipped with 8 sets of industrial high-definition CCD line array cameras (each camera has a resolution of 4096 pixels x 1 pixel), with shooting angles ranging from  $-45^\circ$  to  $225^\circ$ . The railway tunnel visible spectral image acquisition device is shown in Figure 1.

Railway tunnels are one-dimensional arch structures in low-light environments. The surface of the lining has a special curvature and is accompanied by complex and interfering backgrounds and obstacles such as wall joints, cables, cavities, and protrusions formed under pressure, which affect the recognition of crack images. Based on these interfering factors and complex and adverse backgrounds, crack images in railway tunnels are classified into six categories as shown in Table 1. Category 1 crack images contain only cavities; Category 2 crack images have both joint gaps and cavities; Category 3 crack images contain shadows and cavities; Category 4 crack images contain joint gaps, cavities, and shadows;

Category 5 crack images have wall protrusions, cavities, and shadows; and category 6 crack images contain water seepage marks, shadows, and cavities.

### B. TUNNEL IMAGE CHARACTERISTICS IN COMPLEX AND HARSH ENVIRONMENTS

Determining the crack area in complex and harsh environments is critical for identifying crack diseases in tunnels. Figure 2 shows the grayscale value curve of low-frequency features in the tunnel crack images of six categories, which simultaneously include crack and interference section images. Among them, I-I represents the section that contains both cracks and interference, A-A area represents the crack area at the section, B-B area represents the concave area at the section, C-C area represents the splicing area at the section, D-D area represents the area where shadow and concavity coexist at the section, E-E area represents the wall protrusion area at the section, and F-F area represents the water seepage area at the section. From the grayscale value curves of the six categories' low-frequency features, it can be observed that both the crack area and the interference area have valleys in the curve, and the lowest peak of the grayscale value is also similar. These similar low-frequency features pose a challenge to crack identification based on semantic segmentation.

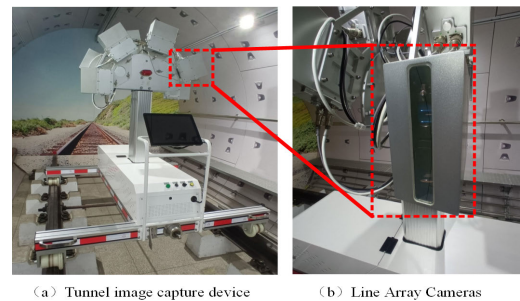


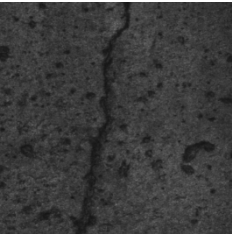
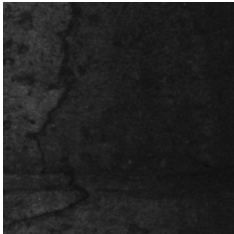
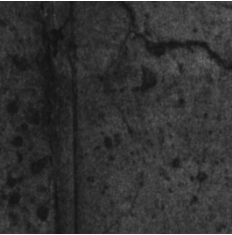
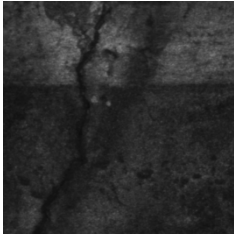
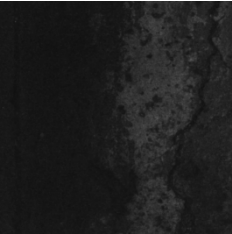
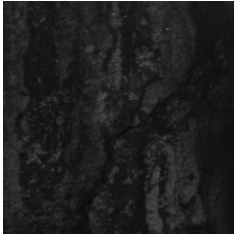
FIGURE 1. Visible spectrum image acquisition device for railway tunnels.

### C. MODEL CONSTRUCTION FOR TUNNEL CRACK DISEASE IDENTIFICATION

#### 1) U-NET

U-net [19] is a type of deep convolutional neural network primarily used for medical image segmentation. The network is characterized by a U-shaped architecture consisting of two main parts: an encoder and a decoder. The encoder, made up of convolutions and pooling, is used to extract local features from the image while decreasing resolution. On the other hand, the decoder utilizes upsampling and skip connections to combine extracted features with the corresponding pixel features in the original image. This process gradually generates high-resolution segmentation results from low-resolution features while utilizing more contextual information, improving the accuracy and robustness of the segmentation. U-net has thus become a classic algorithm in the field of medical image segmentation.

TABLE 1. Classification of tunnel crack image.

Category	Image	Description	Category	Image	Description
1		cavities	4		joint gaps+cavities+shadows
2		joint gaps+cavities	5		wall protrusions+cavities+shadows
3		shadows+cavities	6		water seepage marks+shadows+cavities

### 2) INCEPTION TRANSFORMER

The Inception Transformer [20] model is a combination of Transformers [21] and Inception structures [22]. First, the input is transformed like Inception, then it is fed into the transformer encoder for processing. This combination allows the model to better handle multi-scale image information while simultaneously building the ability to construct remote dependency relationships in parallel.

### 3) IFORMER UNET

Inspired by literatures [19] and [20], an iFormer Unet model is proposed as shown in Figure 3. The iFormer Unet consists of Encoder, Bottleneck, Decoder, and skip connections, with iFormer blocks as their basic unit. The workflow of the iFormer Unet model is as follows: (1) Firstly, in the encoder, the tunnel crack disease image is segmented into non-overlapping  $4 \times 4$  small blocks through the Patch Partition layer, converting the minimum unit pixel of the image into a small block. If the input is a three-channel image, the feature dimension becomes  $4 \times 4 \times 3 = 48$  dimensions. Then, the 48 dimensions are mapped to any dimension  $C$  through the Linear Embedding layer. Next, the representation learning and down-sampling and dimension increasing of features are accomplished respectively by 3 sets of iFormer Block and Patch Merging. The Patch Merging layer will increase the dimension to twice the previous layer while

down-sampling the feature. (2) In the bottleneck part, the features and dimensions of the encoder are transitioned to the decoder. (3) In the decoder part, corresponding to the encoder, three sets of iFormer Blocks and Patch Expanding layers are used to perform feature upsampling and down-sampling. At the same time, the three sets of different scale features in the encoder are fused with the corresponding three sets of upsampling features in the decoder through the Skip Connection layer. Similar to the Patch Expanding layer in step (2), when upsampling the features, the dimension will also be reduced to half of the previous layer. (4) The last layer of Patch Expanding performs a  $4x$  upsampling to restore the output to the same resolution as the input. Then, the output goes through a Linear Projection layer to achieve output pixel-level segmentation prediction

The detailed structure of the iFormer Block in the iFormer Unet network is shown in Figure 4(a). The most important unit in the iFormer Block is the Inception Mixer module, which is composed of high-frequency and low-frequency mixing modules, as shown in Figure 4(b). The Inception Mixer first divides the input feature map along the channel direction and then inputs the divided feature maps into the high-frequency and low-frequency feature extraction units, respectively. The high-frequency feature extraction unit consists of a Max Pooling layer and parallel depthwise convolutional layers (DWConv), while the low-frequency feature

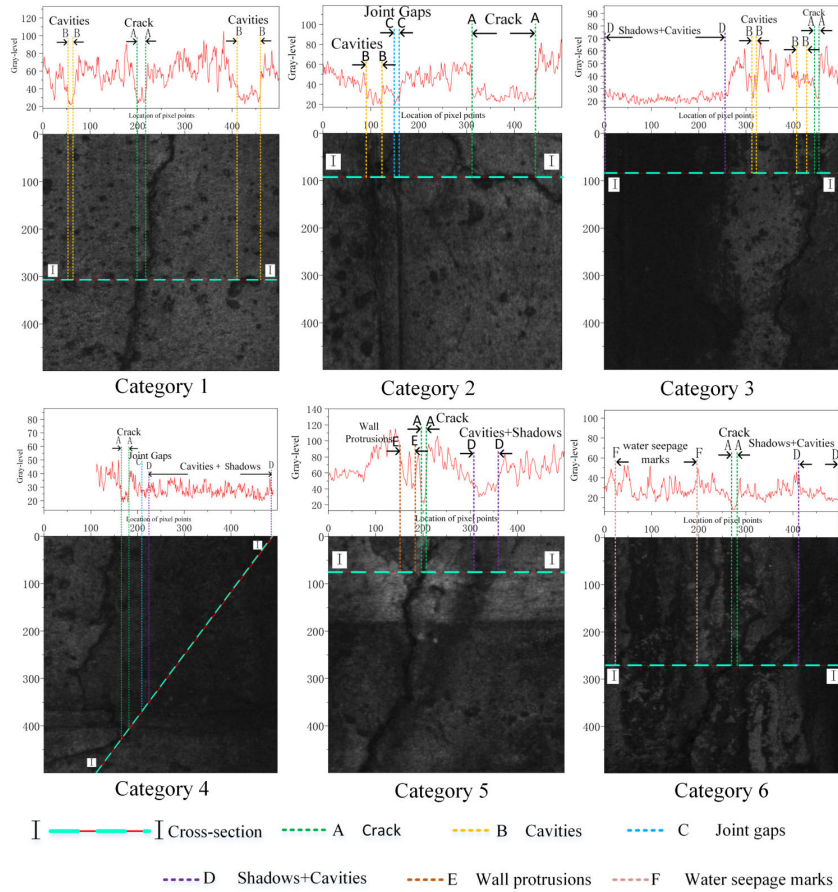


FIGURE 2. Gray value maps of different types of images and corresponding I-I sections.

extraction unit is implemented by self-attention. If the feature map of the Inception Mixer input is set to  $X \in R^{N \times C}$ , The feature map will be split in the direction of the channel dimension into  $X_{high} \in R^{N \times C_{high}}$  and  $X_{low} \in R^{N \times C_{low}}$ , where  $C_{high} + C_{low} = C$ , Feature maps  $X_{high}$  and  $X_{low}$  will be fed into the high frequency feature extraction unit and the low frequency feature extraction unit respectively.

*a: HIGH-FREQUENCY FEATURE EXTRACTION UNIT*

In the high-frequency feature extraction unit, the feature map  $X_{high}$  will be continued to be split into  $X_{high1} \in R^{N \times C_{high1}}$  and  $X_{high2} \in R^{N \times C_{high2}}$  along the channel dimension direction, where  $X_{high1}$  will be fed into a structure consisting of MaxPool with linearised Linear and  $X_{high2}$  will be fed into a structure consisting of linearised Linear and depth convolution DWConv, calculated as shown in equations (1) and (2) respectively.

$$Y_{high1} = FC(MaxPool(X_{high1})) \quad (1)$$

$$Y_{high2} = DWConv(FC(X_{high2})) \quad (2)$$

where  $Y_{high1}$  and  $Y_{high2}$  are high-frequency features output by the high-frequency feature extraction unit and  $FC()$  denotes

the fully connected function that completes the linearisation operation.

*b: LOW-FREQUENCY FEATURE EXTRACTION U*

The low-frequency feature extraction unit consists of Ave-Pool, Multihead Self-Attention (MSA) and Upsample, which can be represented by Equation (3).

$$Y_{low} = Upsample(MSA(AvePool(X_{low}))) \quad (3)$$

where  $Y_{low}$  is the low-frequency feature output from the low-frequency feature extraction unit. The calculation process of the multi-headed self-attentive MSA is shown in equation (4) equation (5) equation (6).

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^o \quad (4)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (5)$$

$$Attention(Q, K, V) = soft \max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (6)$$

where Q,K are the Query and Key matrices of dimension  $d_k$  and V is the Value matrix of dimension  $d_v$ . head is the head of self-attention, h is the number of heads of multi-headed self-attention,  $W^o$  is the weight matrix when multi-headed

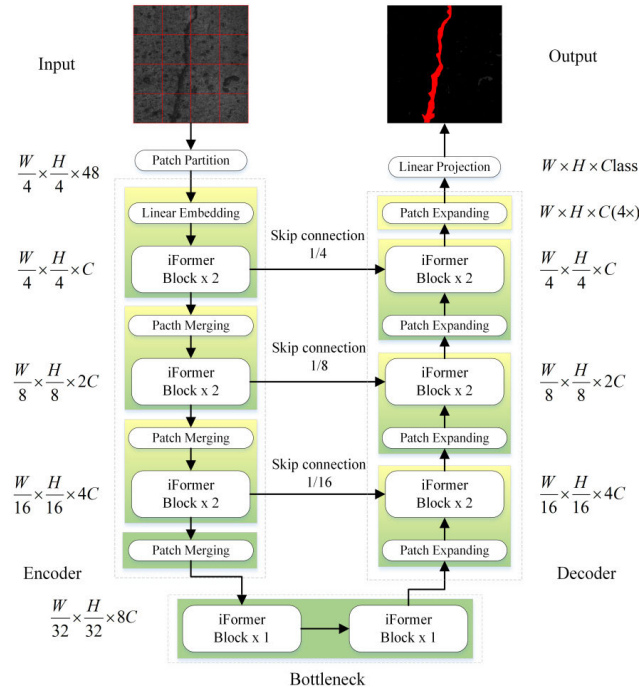
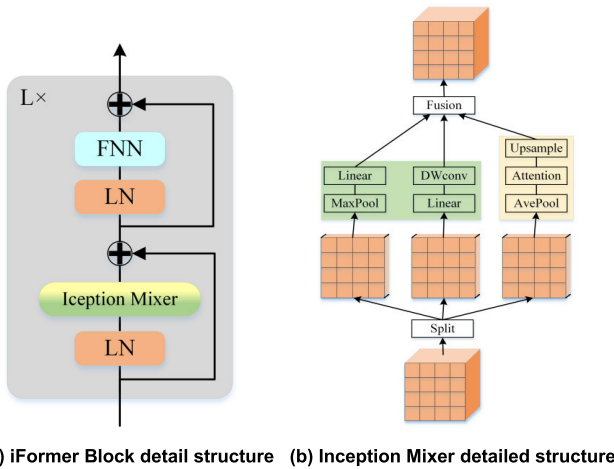


FIGURE 3. iFormer Unet network structure.



(a) iFormer Block detail structure (b) Inception Mixer detailed structure

FIGURE 4. iFormer Block structure.

self-attention is spliced,  $W_i^Q, W_i^K, W_i^V$  are the weight matrices of Query, Key, Value of the  $i$ -th self-attention head, and  $\text{Concat}()$  is the splicing function.

Therefore, the final output of the Inception Mixer high and low frequency mixing module shown in Figure 4(b) is  $Y_c = \text{Concat}(Y_{high1}, Y_{high2}, Y_{low})$ , where  $Y_c$  is the high and low frequency mixing characteristics of the Inception Mixer output.

Then, the calculation process of the iFormer Block shown in Figure 4(a) can be represented by equation (7) equation (8).

$$Y = X + \text{Inception Mixer}(\text{LN}(X)) \quad (7)$$

$$H = Y + \text{FFN}(\text{LN}(Y)) \quad (8)$$

where  $X$  is the input features,  $\text{LN}()$  denotes the normalisation function,  $Y$  is the output transition features,  $\text{FFN}$  is the feed-forward neural network and  $H$  denotes the final output features.

### III. EXPERIMENTAL RESULTS

An NVIDIA GTX1060 graphics card with 6GB of memory was used to train the tunnel crack disease recognition model, based on Windows 10-64 bit operating system, Python 3.6 as the training environment, and Pytorch version 1.10.0 deep learning framework. The dataset is the data collected in Taoping Tunnel of Houyue Line of Zhengzhou Railway Bureau, China, with an image size of 224 pixels  $\times$  224 pixels, and a total of 10,000 pieces of valid data containing cracks and diseases. The model is trained and tested using a 10-fold cross-validation method, in which the dataset is divided into 10 copies, 9 of which are used for training and the remaining 1 is used for validation in turn, and the corresponding misdiagnosis rate is obtained in each experiment.

#### A. TRAINING PARAMETER SETTINGS

The experiments were conducted to compare the training of this paper's methods iFormer Unet, Swin Unet [23] and Unet model respectively, the training parameters were set as shown in the table below, a gradient stochastic gradient descent optimiser was used with an initial learning rate of 0.01, and after 1/3 and 2/3 of the total number of iterations were reached, the learning rate was adjusted to 0.001 and 0.0001 respectively, the weights were decayed to  $1 \times 10^{-4}$ , Batch size was set to 8 and iteration Epochs were 300.

TABLE 2. Training parameter settings.

Training parameters		Parameter settings
Input size		224x224x3
Optimizer		SDG
learning rate	Initial learning rate	0.01
	1/3 of total iterations	0.001
	2/3 of total iterations	0.0001
Batch size		8
Epochs		300

#### B. COMPARATIVE ANALYSIS OF FAULT DETECTION RATES OF CRACKING DISEASES

Crack disease identification is a single target image segmentation task, and using the disease error detection rate as an evaluation criterion is a better test of the model's ability to identify crack disease. The disease misdetection rate is calculated as shown below:

$$P_i = \frac{N_i}{N_{total}} \times 100\% \quad (9)$$

where  $P_i$  is the disease misdetection rate of algorithm model  $i$ ,  $i$  is the three cases of method iFormer Unet, Swin Unet, and Unet respectively in this paper,  $N_i$  is the number of pixels error detection by algorithm model  $i$ , and  $N_{total}$  is the total

TABLE 3. Comparison of params and FLOPs.

Model	iFormer unet	Swin unet	Unet
Params	30.31M	40.35M	34.56M
FLOPs	34.84G	33.94G	57.23G

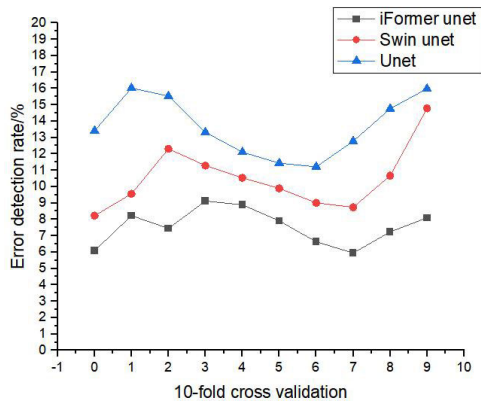


FIGURE 5. 10-fold cross-validation.

number of image pixels. The input images in this paper are all 224 pixels by 224 pixels, so  $N_{total} = 50176$  pixels.

C. 10-FOLD CROSS-VALIDATION EXPERIMENTS

In order to avoid training overfitting, the experiments were carried out 10-fold cross-validation experiments on this paper’s method iFormer Unet, Swin Unet and Unet model, and the experimental results are shown in Figure 5, from the experimental results, it can be seen that this paper’s method has the lowest misdetection rate, which is at the bottom of the folding diagram of Swin Unet and Unet algorithms, and at the same time, it is illustrated by the 10-fold cross-validation experiments that The method of this paper has good robustness. Table 3 shows the quantitative comparison between the different methods. Floating point operations (FLOPs) and network parameters (Params) are used to compare the computational cost of the network. In the comparison of the results, this paper’s method has the least Params, which is about 30.31M, while FLOPs are slightly higher than Swin unet and much lower than the Unet algorithm model. The main reason why the FLOPs of this paper’s method are higher than the Swin unet algorithm is that the attention mechanism used in this paper’s method is the MSA, whereas the attention mechanism used in Swin unet is the W-MSA ( Window Multi-head Self-Attention), and the FLOPs of W-MSA itself are lower than those of MSA, from this point of view, this is also a place that needs to be improved in the future work of this paper. In conclusion, the algorithmic model in this paper performs better in terms of Params and FLOPs.

D. IDENTIFICATION OF THE SIX CATEGORIES OF CRACKS

The number of pixels error detection and the error detection rate of the three crack disease recognition algorithms for the six categories of cracks are shown in Table 4 and

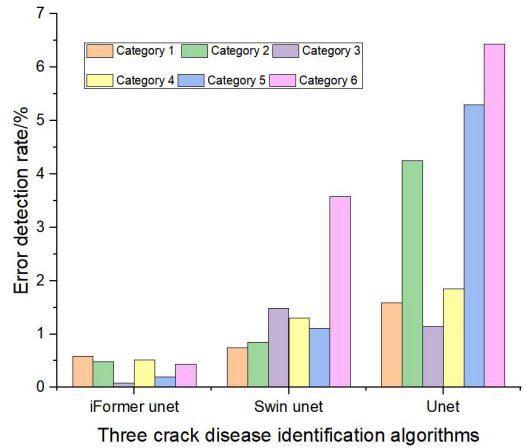


FIGURE 6. Comparison of error detection rates of different recognition algorithms.

the corresponding bar charts are shown in Figure 6. The experimental results show that for the six categories of crack disease images in a harsh environment tunnel, the iFormer Unet algorithm proposed in this paper has the lowest error detection rate, followed by the Swin Unet algorithm which performs better and the Unet algorithm which has the highest error detection rate.

The iFormer Unet algorithm proposed in this paper is a hybrid network consisting of low-frequency and high-frequency feature extraction modules that can automatically learn the high and low frequency features of crack lesions. The Swin Unet algorithm is only capable of low frequency feature extraction and is susceptible to interference from low frequency feature grey scale values such as cavities, wall protrusions, shadows, water seepage marks, etc. that are similar to cracks, but Swin Unet has the ability to capture features over long distances due to the presence of a self-attentive mechanism, and can perform better in the identification of crack lesions in harsh tunnel environments. unet algorithm mainly consists of convolutional neural network, and the convolutional neural network has the ability to generalise local features, also due to this ability, it is highly susceptible to the similarity of local features such as cavities, wall protrusions, shadows, water seepage marks and other local features with cracks in a harsh tunnel environment, resulting in the highest final error detection rate.

Figure 7 shows the loss plots of the three different algorithms during the training phase. It can be seen that the algorithm proposed in this paper, iFormer Unet, also converges the fastest during training, has the lowest loss and performs the best.

E. HEAT MAP ANALYSIS OF DIFFERENT ALGORITHMS

In order to better validate and understand the algorithm model proposed in this paper, the Grad-CAM [24] technique was introduced to the last layer of output of iFormer Unet, Swin Unet and Unet networks to generate heat maps

TABLE 4. Error detection rate results of different crack disease identification algorithms.

Category number	Number of pixels detected by error / pixels			Total number of pixels/pixels	Error detection rate/%		
	$N_{iFormer-Unet}$	$N_{Swin-unet}$	$N_{Unet}$		$P_{iFormer-unet}$	$P_{Swin-unet}$	$P_{Unet}$
1	3116	3919	6101	50176	6.21	7.81	12.16
2	2604	4461	7290		5.19	8.89	14.53
3	1786	5760	5590		3.56	11.48	11.14
4	3085	5168	6297		6.15	10.3	12.55
5	2052	4962	7577		4.09	9.89	15.10
6	2539	6808	8148		5.06	13.57	16.24
Average error detection rate/%					5.04	10.32	13.62

TABLE 5. Error detection rate results of different hybrid network structures for crack disease identification.

Category number	Number of pixels detected by error / pixels			Total number of pixels/pixels	Error detection rate/%		
	Attention	Attention+MaxPool	Attention+MaxPool+DwConv		Attention	Attention+MaxPool	Attention+MaxPool+DwConv
1	3909	3347	3116	50176	7.79	6.67	6.21
2	4481	3808	2604		8.93	7.59	5.19
3	5936	4666	1786		11.83	9.30	3.56
4	5073	4907	3085		10.11	9.78	6.15
5	5037	4762	2052		10.04	9.49	4.09
6	6508	5414	2539		12.97	10.79	5.06
Average error detection rate/%					10.27	8.94	5.04

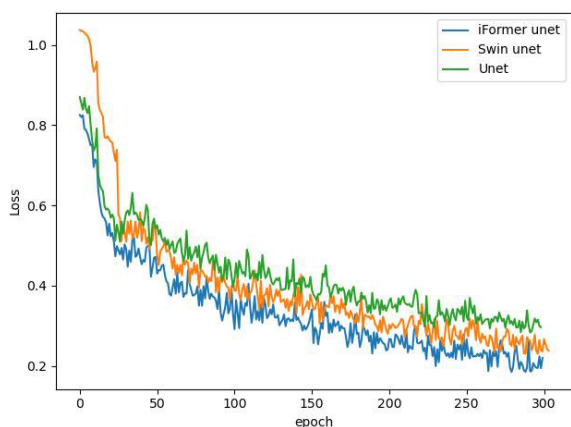


FIGURE 7. Loss diagrams for training stages of different algorithms.

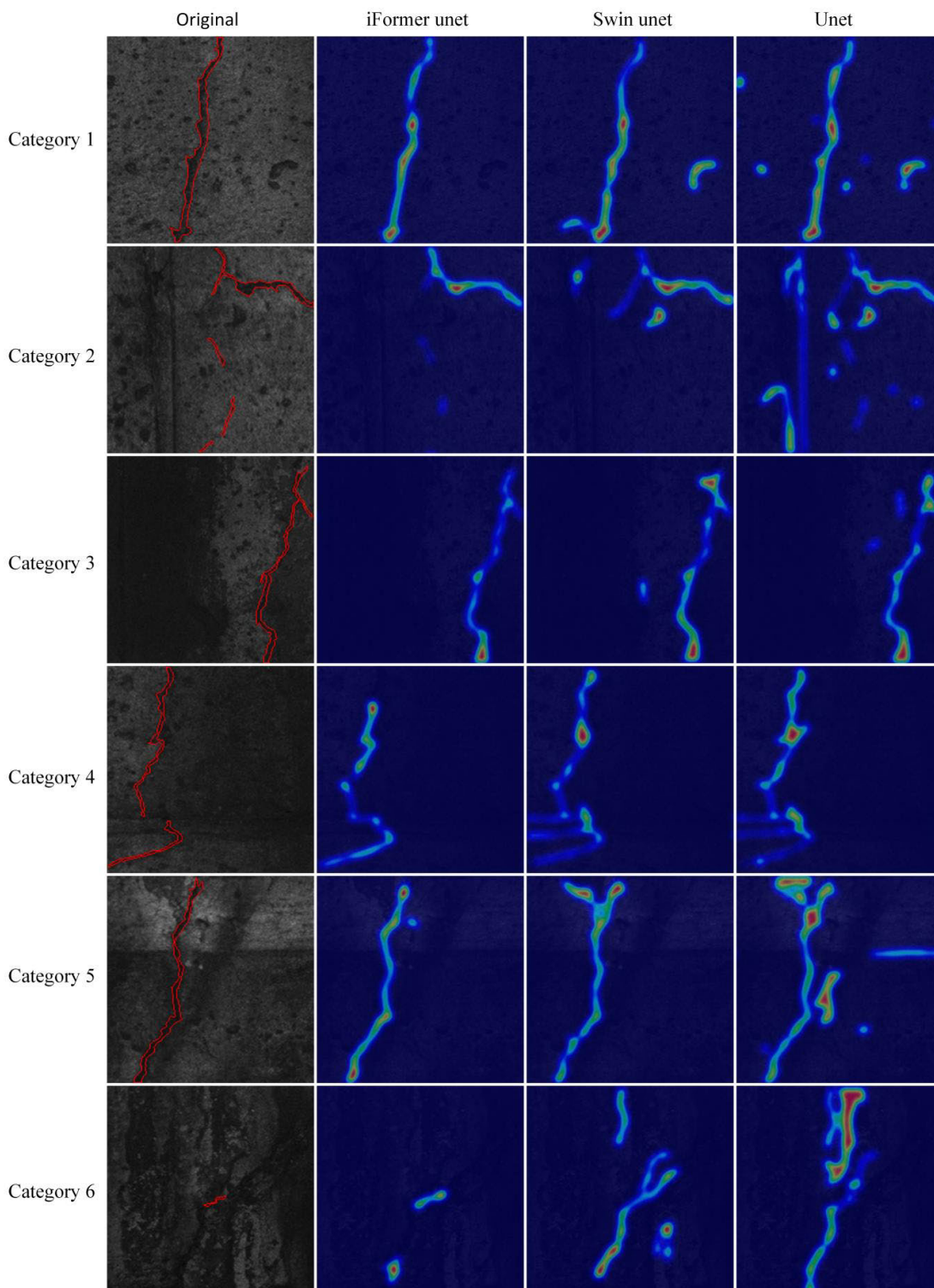
for six representative categories of fracture disease samples, as shown in Figure 9. In the original figure, the area outlined by the red line indicates the fracture disease. In the heat map of the three algorithms, the darker the red colour indicates that the algorithm model is more interested in this area, followed by the yellow colour, and the darker the blue colour indicates



FIGURE 8. Application site.

that the algorithm model is less interested in this area. From Figure 8, it can be seen that the algorithm proposed in this paper, which can focus well on the feature areas of crack lesions, can identify crack lesions more accurately, while the Swin Unet and Unet algorithms, when focusing on crack lesion features, are also disturbed by cavities, wall protrusions, shadows and seepage flow marks, and focus on these





**FIGURE 9.** Thermal diagram for identifying six kinds of crack diseases under different algorithms.

disturbed areas, which are incorrectly identified as crack areas. Taking category V as an example, due to the high frequency and low frequency feature extraction capability of the algorithm in this paper, it is able to avoid the influence of tunnel wall protrusions, shadows, etc. in terms of features

such as greyscale values and textures very well; Swin Unet, due to its low frequency feature extraction capability only, identifies the edges formed by tunnel wall protrusions as crack regions; Unet, due to its ability to generalise local features, identifies the crack regions with similar the cavities,

**TABLE 6. Segmentation performance of different methods on the COVID-QU-Ex dataset.**

Method	Precision(%)	Recall(%)	F1-score(%)
Unet	87.45	85.79	86.61
Swin unet	90.25	89.67	89.96
iFormer unet	90.38	89.98	90.18

shadows, and edges formed by projections on the tunnel wall, which have similar features such as greyscale, are incorrectly generalised as crack features.

It can be seen that the iFormer Unet algorithm model proposed in this paper is effective in identifying crack disease in harsh tunnel environments.

#### F. ABLATION EXPERIMENT

In order to verify the effectiveness of the hybrid network Inception Mixer module structure of the iFormer block in the model, three sets of control experiments were set up, namely, a hybrid network structure with only the self-attentive mechanism “Attention”, a mix of the self-attentive mechanism and maximum pooling “Attention+MaxPool”, and a mix of the attention mechanism, maximum pooling and deep convolution “Attention+MaxPool+DwConv”. “The training setup and data for the three sets of experiments and the test data were the same as those of the TRAINING PARAMETER SETTINGS. The training settings and data for the three sets of experiments as well as the test data are the same as in the TRAINING PARAMETER SETTINGS section, and the results are shown in Table 5.

From the results of the ablation experiments obtained in Table 5, it can be seen that when the Inception Mixer module consists of a mixture of self-attentive mechanisms, maximum pooling and deep convolutional structures, the minimum error detection rate for crack recognition can be obtained, which is the best performance among the three sets of control experiments, indicating that this hybrid network structure approach is effective.

#### G. OTHER EXPERIMENTS

To further validate the robustness of the proposed model, we selected other types of public dataset COVID-QU-Ex dataset [25] for our experiments. COVID-QU-Ex dataset is a dataset compiled by the researchers at Qatar University, which has a total of 33920 chest X-ray images, and we used the COVID-19 Infection Segmentation Data sub-dataset (1456 Normal and 1457 Non-COVID-19 CXRs with corresponding lung mask, plus 2913 COVID-19 CXRs with corresponding lung mask) for validation. The data set consists of 3728 training images, 932 validation images, and 1166 test images.

To evaluate the metrics we use Precision, Recall, F1-score, which are associated with three values, i.e. true-positive (TP), false-positive (FP), and false-negative (FN). These metrics are calculated as follows.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

$$F1 = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Precision} + \text{Recall}} \quad (12)$$

The experimental results are shown in Table 6, on the public dataset COVID-QU-Ex dataset, this paper’s method performs the best in Precision, Recall, and F1-score compared to Unet and Swin unet methods, which shows that this paper’s method also performs well on new datasets, and also shows that this paper’s method has good robustness.

#### IV. CONCLUSION

Due to the extremely limited lighting conditions in some existing railway tunnels of relatively old age, the dark conditions inside the railway tunnels, as well as the structural surface noise and crack-like interfering objects can pose a great challenge to the identification of railway tunnel cracks. At the same time, due to the current lack of readily available railway tunnel crack datasets, as well as the fact that many scholars’ crack datasets are datasets augmented by image synthesis techniques such as GAN, they cannot accurately reflect the real-world data situation. To address this problem, we collected images inside real-world railway tunnels, produced a dataset, and proposed a novel and effective hybrid neural network tunnel crack disease recognition iFormer Unet model, which is based on the iFormer block module that can extract high-frequency features and low-frequency features at the same time, and constructed a U-shape network consisting of encoder, Bottleneck, decoder and jumping Connections consisting of a U-shaped network. The results of 10-fold cross-validation in the experiment show that the method proposed in this paper has a relatively low misdetection rate of about 7.56%, with about 30.31M params and 34.84G FLOPs, as shown in Figure 8. The method proposed in this paper has been experimented in Taoping Tunnel of Houyue Line of Zhengzhou Railway Bureau of China, and it can provide a certain reference for the scientific maintenance of railway tunnels.

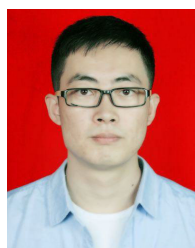
There is still room for improvement of the algorithm in this paper, because of the limitation of resources and time, this paper only uses the collected dataset as the data source for this experiment, and in the subsequent research, it will be compared with more excellent algorithmic models, and some public datasets will be used to enhance the generalisation ability of the algorithmic models and obtain more accurate and effective models.

#### ACKNOWLEDGMENT

The authors would like to thank the editor and the anonymous reviewers for their valuable comments and suggestions. These comments and suggestions improved the quality of this paper. (Rui-Jun Bai and Zhong Li are co-first authors.)

## REFERENCES

- [1] Z. Zhou, J. Zhang, C. Gong, and W. Wu, "Automatic tunnel lining crack detection via deep learning with generative adversarial network-based data augmentation," *Underground Space*, vol. 9, pp. 140–154, Apr. 2023.
- [2] J. Liu, Z. Zhao, C. Lv, Y. Ding, H. Chang, and Q. Xie, "An image enhancement algorithm to improve road tunnel crack transfer detection," *Construct. Building Mater.*, vol. 348, Sep. 2022, Art. no. 128583.
- [3] Z. Zhou, L. Yan, J. Zhang, and H. Yang, "Real-time tunnel lining crack detection based on an improved you only look once version X algorithm," *Georisk, Assessment Manage. Risk Engineered Syst. Geohazards*, vol. 17, no. 1, pp. 181–195, Jan. 2023.
- [4] D. H. Kang and Y.-J. Cha, "Efficient attention-based deep encoder and decoder for automatic crack segmentation," *Struct. Health Monitor.*, vol. 21, no. 5, pp. 2190–2205, Sep. 2022.
- [5] R. Ali and Y.-J. Cha, "Attention-based generative adversarial network with internal damage segmentation using thermography," *Autom. Construct.*, vol. 141, Sep. 2022, Art. no. 104412.
- [6] W. Choi and Y.-J. Cha, "SDDNet: Real-time crack segmentation," *IEEE Trans. Ind. Electron.*, vol. 67, no. 9, pp. 8016–8025, Sep. 2020.
- [7] E. Protopapadakis, A. Voulodimos, A. Doulamis, N. Doulamis, and T. Stathaki, "Automatic crack detection for tunnel inspection using deep learning and heuristic image post-processing," *Int. J. Speech Technol.*, vol. 49, no. 7, pp. 2793–2806, Jul. 2019.
- [8] K. Makantasis, E. Protopapadakis, A. Doulamis, N. Doulamis, and C. Loupos, "Deep convolutional neural networks for efficient vision based tunnel inspection," in *Proc. IEEE Int. Conf. Intell. Comput. Commun. Process. (ICCP)*, Sep. 2015, pp. 335–342.
- [9] S. Zhou, Y. Pan, X. Huang, D. Yang, Y. Ding, and R. Duan, "Crack texture feature identification of fiber reinforced concrete based on deep learning," *Materials*, vol. 15, no. 11, p. 3940, Jun. 2022.
- [10] Q. Zhou, Z. Qu, Y.-X. Li, and F.-R. Ju, "Tunnel crack detection with linear seam based on mixed attention and multiscale feature fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.
- [11] J. Liao, Y. Yue, D. Zhang, W. Tu, R. Cao, Q. Zou, and Q. Li, "Automatic tunnel crack inspection using an efficient mobile imaging module and a lightweight CNN," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15190–15203, Sep. 2022.
- [12] I.-H. Kim, H. Jeon, S.-C. Baek, W.-H. Hong, and H.-J. Jung, "Application of crack identification techniques for an aging concrete bridge inspection using an unmanned aerial vehicle," *Sensors*, vol. 18, no. 6, p. 1881, Jun. 2018.
- [13] E. Protopapadakis, C. Stentoumis, N. Doulamis, A. Doulamis, K. Loupos, K. Makantasis, G. Kopsiaftis, and A. Amditis, "Autonomous robotic inspection in tunnels," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 5, pp. 167–174, Jun. 2016.
- [14] K. Loupos, A. D. Doulamis, C. Stentoumis, E. Protopapadakis, K. Makantasis, N. D. Doulamis, A. Amditis, P. Chrobocinski, J. Victores, R. Montero, E. Menendez, C. Balaguer, R. Lopez, M. Cantero, R. Navarro, A. Roncaglia, L. Belsito, S. Camarinopoulos, N. Komodakis, and P. Singh, "Autonomous robotic system for tunnel structural inspection and assessment," *Int. J. Intell. Robot. Appl.*, vol. 2, no. 1, pp. 43–66, Mar. 2018.
- [15] I. Katsamenis, N. Doulamis, A. Doulamis, E. Protopapadakis, and A. Voulodimos, "Simultaneous precise localization and classification of metal rust defects for robotic-driven maintenance and prefabrication using residual attention U-Net," *Autom. Construct.*, vol. 137, May 2022, Art. no. 104182.
- [16] J. Lewis, Y.-J. Cha, and J. Kim, "Dual encoder–decoder-based deep polyp segmentation network for colonoscopy images," *Sci. Rep.*, vol. 13, no. 1, p. 1183, Jan. 2023.
- [17] G. Wang, Y. Liu, and J. Xiang, "A two-stage algorithm of railway sleeper crack detection based on edge detection and CNN," in *Proc. Asia-Pacific Int. Symp. Adv. Rel. Maintenance Modeling (APARM)*, Aug. 2020, pp. 1–5.
- [18] Y. Yang and G. Mei, "Deep transfer learning approach for identifying slope surface cracks," *Appl. Sci.*, vol. 11, no. 23, p. 11193, Nov. 2021.
- [19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
- [20] C. Si, W. Yu, P. Zhou, Y. Zhou, X. Wang, and S. Yan, "Inception transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 23495–23509.
- [21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.
- [22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [23] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 205–218.
- [24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
- [25] M. T. Anas, M. E. H. Chowdhury, and Y. Qiblawey, "COVID-QU-Ex," Kaggle, San Francisco, CA, USA, Tech. Rep., 2021, doi: [10.34740/KAGGLE/DSV/3122958](https://doi.org/10.34740/KAGGLE/DSV/3122958).



**RUIJUN BAI** received the master's degree from the North University of China, in 2020. He is currently an Engineer with Shanxi Information Industry Technology Research Institute Company Ltd. His current research interests include computer vision and deep learning.



**JING GAO** received the master's degree from the North University of China, in 2013. He is currently a Senior Engineer with Shanxi Information Industry Technology Research Institute Company Ltd. His current research interest includes machine vision.



**ZHONG LI** received the Ph.D. degree from the North University of China, in 2014. He is currently the Associate Dean, a Professor, and a Graduate Advisor with the School of Software, North University of China. His current research interests include image processing and system reliability theory.



**DONGHANG LIU** received the master's degree from Xidian University, in 2018. He is currently an Engineer with Shanxi Information Industry Technology Research Institute Company Ltd. His current research interest includes blockchain.



**XUEKUI SHANGGUAN** received the bachelor's degree from the Taiyuan University of Technology, in 2008. He is currently a Senior Engineer with Shanxi Information Industry Technology Research Institute Company Ltd. His current research interest includes computer science.

...