IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 July 2024, accepted 29 July 2024, date of publication 5 August 2024, date of current version 20 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3438112

RESEARCH ARTICLE

MixSegNet: A Novel Crack Segmentation Network Combining CNN and Transformer

YANG ZHOU[1], RAZA ALI[2], (Senior Member, IEEE), NORRIMA MOKHTAR[1], SULAIMAN WADI HARUN[1], AND MASAHIRO IWAHASHI[3], (Senior Member, IEEE)

[1]Department of Electrical Engineering, Faculty of Engineering, Universiti Malaya, Kuala Lumpur 50603, Malaysia
[2]Department of Electrical Engineering, Faculty of Information and Communication Technology (FICT), Balochistan University of Information Technology, Engineering and Management Sciences (BUITEMS), Quetta 87300, Pakistan
[3]Department of Electrical, Electronics and Information Engineering, Nagaoka University of Technology, Nagaoka 940-2188, Japan

Corresponding author: Norrima Mokhtar (norrimamokhtar@um.edu.my)

This work was supported in part by JSPS KAKENHI under Grant 24K02975.

ABSTRACT In the domain of road inspection and structural health monitoring, precise crack identification and segmentation are essential for structural safety and disaster prediction. Traditional image processing technologies encounter difficulties in detecting cracks due to their morphological diversity and complex background noise. This results in low detection accuracy and poor generalization. To overcome these challenges, this paper introduces MixSegNet, a novel deep learning model that enhances crack recognition and segmentation by integrating multi-scale features and deep feature learning. MixSegNet integrates convolutional neural networks (CNNs) and transformer architectures to enhance the detection of small cracks through the extraction and fusion of fine-grained features. Comparative evaluations against mainstream models, including LRASPP, U-Net, Deeplabv3, Swin-UNet, AttuNet, and FCN, demonstrate that MixSegNet achieves superior performance on open-source datasets. Specifically, the model achieved a precision of 95.2%, a recall of 88.2%, an F1 score of 91.5%, and a mean intersection over union (mIoU) of 84.8%, thereby demonstrating its effectiveness and reliability for crack segmentation tasks.

INDEX TERMS Crack segmentation network, crack images, convolutional neural network, transformer model, image processing, deep learning, self-attention mechanism.

I. INTRODUCTION

Crack identification occupies a vital position in the field of structural health monitoring because it is directly related to the safety and reliability of building structures. With the development of technology, crack detection methods have gradually transformed from traditional manual inspection to automatic identification using modern technologies such as advanced image processing, artificial intelligence, and machine learning [1]. These methods not only improve identification accuracy and efficiency, but also enable potential structural problems to be discovered at an early stage, enabling preventive maintenance and extending the life of the building. The existence of cracks may be caused by a variety of reasons, including structural aging, environmental erosion, excessive loads, and natural disasters. If these cracks are not discovered and treated in time, they may lead to a decrease in structural performance and even threaten personnel safety. Therefore, developing effective crack detection and identification systems is crucial to ensure structural safety.

Traditional crack detection methods, such as threshold techniques [2], demonstrate limited adaptability. To address this, Yang et al. [3] introduced a novel approach utilizing a fully convolutional network (FCN), enhancing the detection process. This technique employs single-pixel width skeletons for crack segmentations, allowing for the detailed analysis of crack features—like topology, length, and widths—thus offering critical indicators for practical assessments. However, the scarcity of training data for crack segmentation presents a challenge. In response, König et al. [4] developed a method to streamline the annotation process for semantic segmentation of surface cracks. They utilized a U-Net architecture based on a fully convolutional

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

network, optimized for small datasets through patch-based training, leading to unprecedented results on various crack datasets. Ren et al. [5] explored the application of deep fully convolutional networks for concrete crack detection in tunnel images, proposing CrackSegNet, an advanced network for comprehensive crack segmentation. This innovation improves feature extraction, aggregation, and resolution reconstruction, significantly boosting segmentation performance. Kang et al. [6] introduced an automated method combining Faster R-CNN and a modified TuFF algorithm for precise crack detection, localization, and quantification, overcoming the limitations posed by varying environmental conditions. Similarly, Lau et al. [7] applied convolutional neural networks for segmenting pavement crack images, marking a significant advancement in the field. Liu et al. [8] proposed a two-step convolutional neural network method for enhanced crack detection and segmentation. Following this, Guan et al. [9] aimed to refine the accuracy and speed of 3D crack segmentation models, pushing the boundaries of current methodologies. Ali et al. [10] proposed an additive attention gate-based network architecture called Crack Segmentation Network-II (CSN-II).

A. RESEARCH GAP

Recent research has led to further improvements in various aspects of crack segmentation. Wang et al. [11] introduced a lightweight crack segmentation network based on knowledge distillation. Liu et al. [12] presented an upgraded CrackFormer network for pavement crack segmentation. This network achieved higher accuracy with fewer floating-point operations (FLOPs) and parameters compared to previous methods. Wu et al. [13] developed a lightweight MobileNetV2-DeepLabV3 network for enhanced precision in dam crack width measurement. Yao et al. [14] developed a CrackResU-Net model with a pyramid region attention module for pixel-level pavement crack recognition. Lin et al. [15] proposed DeepCrackAT, a framework for crack segmentation based on learning multi-scale crack features. Tang et al. [16] introduced a novel lightweight concrete crack segmentation method based on DeeplabV3+. This method reduces the number of model parameters and enhances segmentation accuracy. Chen et al. [17] introduced a dynamic semantic segmentation algorithm with an encoder-crossor-decoder structure for pixel-level building crack segmentation. Li et al. [18] concentrated on crack segmentation in asphalt pavement using an enhanced YOLOv5s model. Moreover, Sohaib et al. [19] proposed an ensemble approach for robust automated crack detection and segmentation in concrete structures, achieving high precision and an intersection over union score. Collectively, these studies contribute to the advancement of crack segmentation algorithms, addressing various challenges and improving the accuracy and efficiency of crack detection and segmentation processes.

However, the aforementioned models fail to fully leverage the respective strengths of CNN and Transformer. Therefore, we propose the MixSegNet model as a means of enhancing the accuracy of crack segmentation.

II. RELATED WORK

A. SEMANTIC SEGMENTATION

Semantic segmentation is derived from the further refinement of classification problems. It requires pixel-level classification tasks and puts forward higher requirements for architecture and algorithms. At present, semantic segmentation technology has been widely used in different fields of computer vision. Among them, semantic segmentation is applied in various fields, including satellite imagery [20], medicine [21], material science [22], and meteorology [23]. It can be seen that semantic segmentation technology is crucial. FCN (Fully Convolutional Networks) [24], which was first proposed by Jonathan Long, Evan Shelhamer, and Trevor Darrell in 2015, aims to classify each pixel in the image into the corresponding category. The core idea of FCN is to use a fully convolutional layer to replace the fully connected layer in the traditional convolutional neural network, so that the network can accept input images of any size and output a spatial map of corresponding size. The spatial map can be directly applied to pixel-level prediction tasks. U-Net [25], a deep learning model specifically designed for medical image segmentation, was initially introduced by Olaf Ronneberger, Philipp Fischer, and Thomas Brox in 2015. The architectural design of U-Net is particularly well-suited for tasks that necessitate high-precision localization, such as the segmentation of organs and tissues within medical imagery. This model is named for its unique "U"-shaped structure, which effectively combines shallow (high-resolution) features and deep (high-level semantics) features to improve segmentation accuracy. LRASPP (Lite R-ASPP, or Lightweight Residual Atrous Space Pyramid Pooling) [26] is a deep learning architecture optimized for mobile devices and edge computing, especially for semantic segmentation tasks. It is improved and simplified based on the original ASPP (Atrous Spatial Pyramid Pooling) and DeepLab series models. ASPP captures multi-scale information by using different dilation rates in parallel convolutional layers, thereby improving the model's ability to understand different areas of the image. LRASPP aims to reduce the computational complexity and number of parameters to adapt to environments with limited computing resources. DeepLabv3 [27] is an advanced deep learning architecture designed specifically for image semantic segmentation tasks. It is the third version of the DeepLab series model, developed by Liang-Chieh Chen and others, aiming to further improve the segmentation accuracy of the model in complex image scenes. The core contributions of DeepLabv3 include the improved atrous spatial pyramid pooling (ASPP) module and the systematic application of atrous convolution. These features enable the model to effectively capture multi-scale information and handle different object sizes in images. AttuNet [28] is a recently proposed semantic segmentation architecture. It is an improved version of U-Net. It better integrates shallow and deep semantic information through a special attention module.

B. ATTENTION MECHANISM

As Transformer [29] based on self-attention mechanisms have gained advantages in NLP in recent years, more and more people are trying to use Transformers in visual models. The ViT (Vision Transformer) [30] model is a deep learning model based on the Transformer architecture, specially designed to handle image recognition tasks. It was originally proposed in 2020 by Alexey Dosovitskiy and others at Google Research. By partitioning images directly into serially arranged patches and subsequently processing these serialized image patches using Transformers, Vision Transformer (ViT) has demonstrated performance that matches or even surpasses the state-of-the-art on multiple image recognition tasks, providing results comparable to those of Convolutional Neural Networks (CNN) models. However, ViT cannot be directly used for semantic segmentation. The original ViT can only be used for classification tasks. Subsequently, many people have proposed variants based on the ViT model for semantic segmentation.

C. SEGMENTATION BASED ON ATTENTION MECHANISM

The SETR (Semantic Segmentation Transformer) [31] model is a deep learning model specially designed for semantic segmentation tasks. It combines the powerful capabilities of Transformer with the advantages of traditional semantic segmentation methods. SETR was originally proposed by Zheng et al. in 2020. Its core idea is to apply Transformer to the global feature extraction of images to directly classify at the pixel level. This is an idea borrowed from the NLP field, and it is the first time that it has been widely used. Scale applied to semantic segmentation tasks in computer vision. Swin Transformer [32] is a deep learning model designed based on the Transformer architecture and optimized for processing computer vision tasks. It was proposed by Ze Liu et al. of Microsoft Research in 2021. The core innovation of Swin Transformer is the introduction of a Transformer structure called ''hierarchy'', which effectively manages global dependencies and computational complexity in images by using variable window sizes, allowing the model to be more efficient Process large-scale image data. SegFormer [33] is an advanced deep learning model designed for semantic segmentation tasks, which combines the power of Transformer with the efficiency of Convolutional Neural Networks (CNN). Introduced by Xie et al. in 2021, SegFormer achieves precise segmentation of objects of varying sizes within images by incorporating a lightweight Transformer encoder and an efficient multi-scale feature fusion strategy, all while preserving the model's efficiency and adaptability. The Swin-UNet [34] model is a deep learning model that combines the characteristics of Swin Transformer and U-Net architecture. It is specially designed for fine segmentation tasks such as medical image segmentation. It was proposed by Hao Chen et al. and aims to utilize the hierarchical and self-attention mechanism of Swin Transformer to capture the details and contextual information of the image, while achieving high-precision pixel-level segmentation through the encoder-decoder structure of U-Net. Through this combination, Swin-UNet aims to improve the model's ability to understand details and structures in complex images such as medical images, thereby improving the accuracy and efficiency of segmentation.

The models described above are either based on convolutional neural networks (CNNs) or transformers, or combine the advantages of both. However, in the context of crack segmentation, where the need for fine segmentation of the scene and the ability to deal with a variety of background noise, environmental impact, and other factors is paramount, the aforementioned models are not optimal. Consequently, this research paper proposes the MixSegNet model as a solution to address the shortcomings of existing models in the context of more complex segmented images. In summary, the main contributions of this article include: (1) We adopt a similar structure to U-Net, while using the innovative UC Block module to obtain more details while increasing the receptive field, and through the proposed multi-scale fusion module for crack segmentation (Fuse Block) to enhance it. Ablation experiments show that all the proposed modules, including parallel CNN and Transformer architecture, help the model combine multi-scale features more effectively and generate more accurate crack segmentation masks. (2) The developed model demonstrates satisfactory segmentation accuracy on a benchmark datasets (cracks-APCGAN [28]).

III. METHODOLOGY

In our research, the MixSegNet crack segmentation model uses two major deep learning technologies, Convolutional Neural Network (CNN) and Transformer, to take full advantage of their respective strengths to achieve highly accurate crack image segmentation. CNN is a powerful deep learning tool specifically designed to process data with a grid structure (such as images). In MixSegNet, we use CNN to extract local and low-level features from images, taking advantage of its excellent spatial feature extraction capabilities. The advantage of CNN is that it can automatically learn basic features such as edges and textures of images through convolutional layers, and capture more complex image features through deep network structures, providing a solid foundation for accurate crack segmentation. Transformer technology is based on the self-attention mechanism and can process sequence data and is particularly good at capturing long-range dependencies. In the MixSegNet model, we introduce Transformer to complement the limitations of CNN, especially in understanding the global context of images and capturing long-range dependencies. The advantage of Transformer is that it can dynamically weigh the importance of each part in the image through the self-attention mechanism, thereby better understanding the global structure of the image. This is particularly important

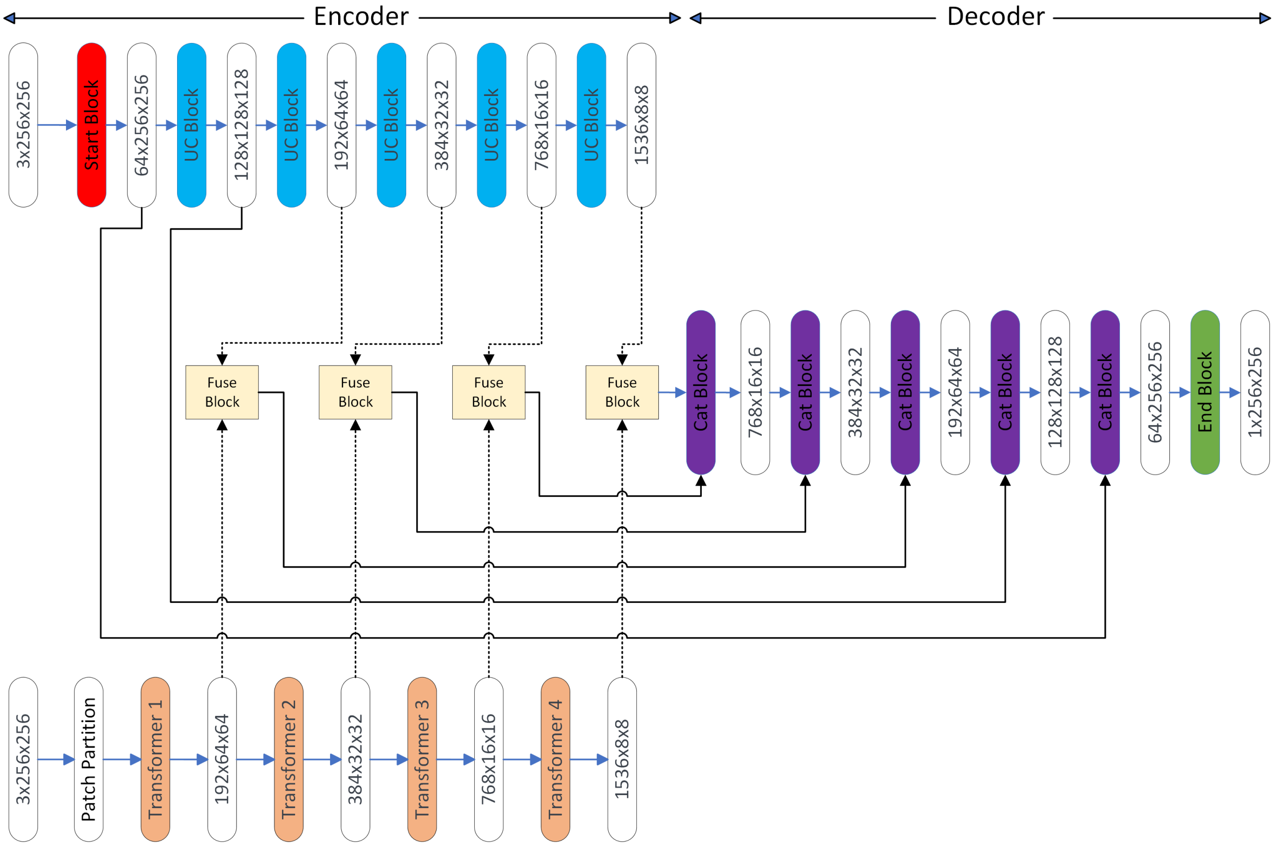

FIGURE 1. The proposed MixSegNet framework.

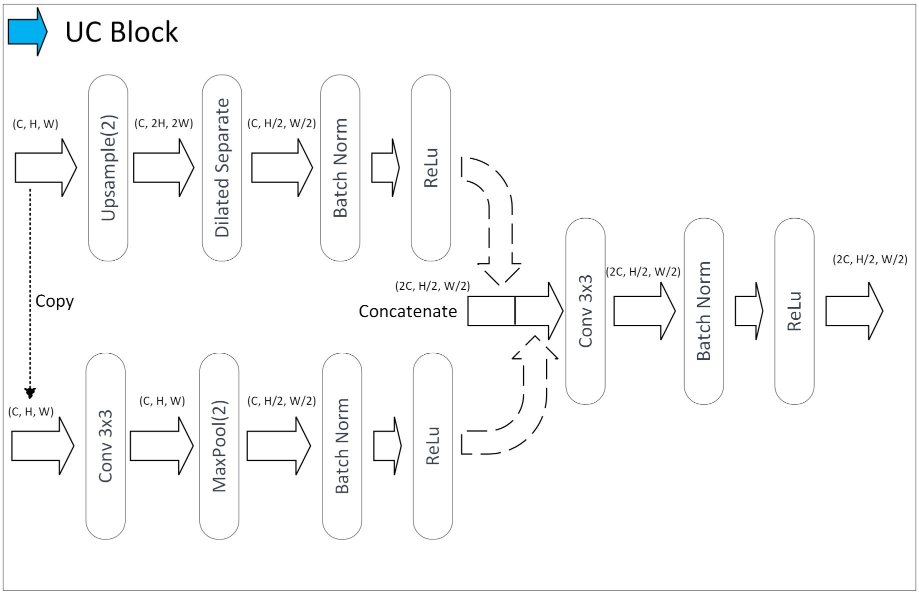

FIGURE 2. UC block.

for the crack segmentation task, where crack identification requires not only the accurate extraction of local features, but also a comprehensive understanding of their location and morphology in the entire image. By combining the benefits of CNN and Transformer, MixSegNet is able to simultaneously leverage the advantages of CNN in extracting powerful local features and Transformer in understanding the global context. This combination not only improves the accuracy of fracture

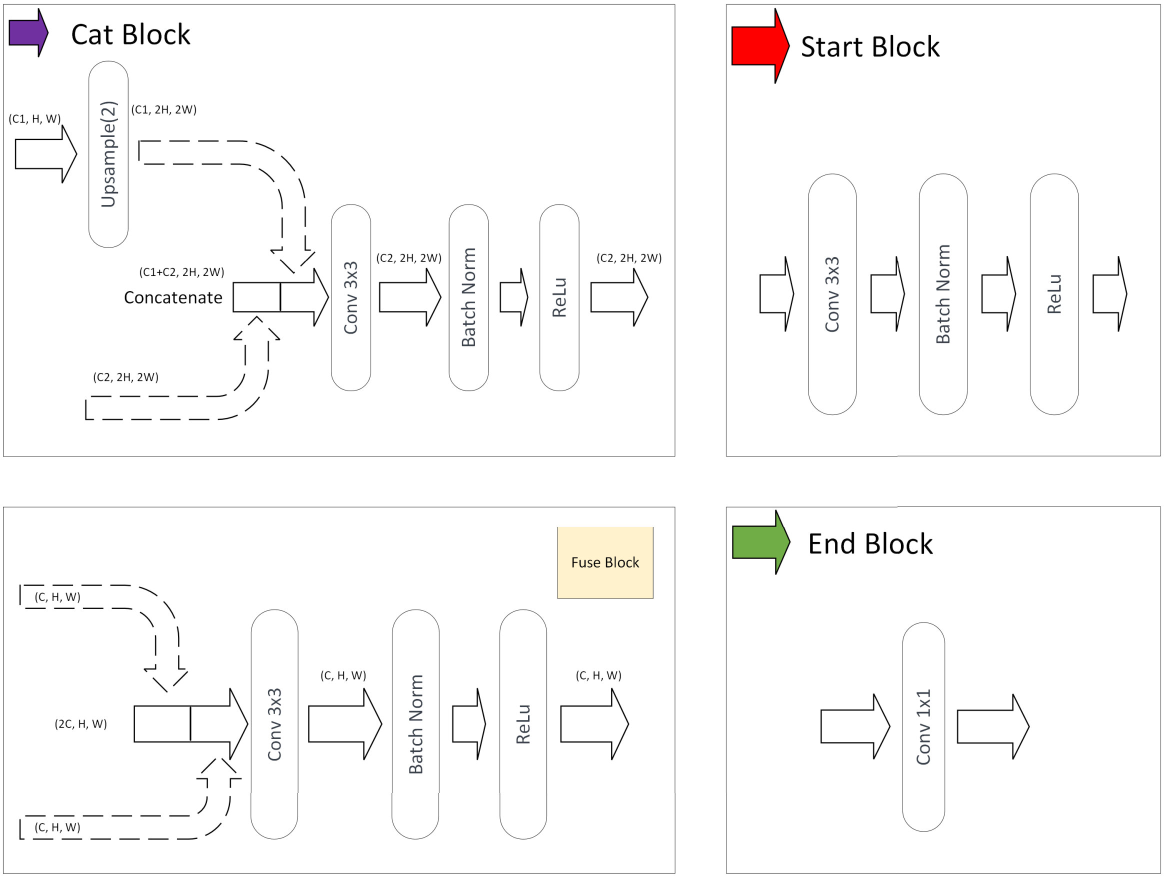

FIGURE 3. Cat Block. Fuse Block. Start Block. End Block.

segmentation, but also enhances the adaptability and generalisation of the model to different fracture types and complex backgrounds. Our method enables an efficient exchange of information between CNN and Transformer through a carefully designed network structure, ensuring the model's excellent performance in the fracture segmentation task.

A. MIXSEGNET

As seen in Figure 1, the overall architecture is divided into two parts, one is Encoder and the other is Decoder, where the Encoder part is divided into three parts, the top part is the tandem CNN architecture, the middle part is the Fuse Block part, and the bottom part is the architecture that uses part of the Swin Transformer [32] model(If you want more details, refer to the original paper). The Decoder part is the tandem CNN module that accepts the output of the Fuse Block module, which in turn recovers the feature map step by step and extracts the segmentation part we need from it. When given an image 3 × 256 × 256, we first pass through the upper part of the Encoder (such as the consecutive blue part of the figure, which we call UC Block, and we will explain it in detail later) and the lower part at the same time by means of the concatenated architecture, so that we can get the hierarchical feature maps. Then the layered features from CNN and Transformer are fused by the proposed Fuse Block module fusion. The most important use of Fuse Block is in fusing the features extracted from Swin Transformer using transformer and the features extracted from CNN network. With the feature fusion technique, the global and local features can be captured better, and the segmentation accuracy can be improved. Finally, these layered features are passed into the continuous Decoder part (as shown in the continuous purple part of the figure, which we call Cat Block, and will be explained in detail later), which in turn recovers the mask information of the image step by step.

1) UC BLOCK

As shown in the Figure 2, this is our proposed UC Block module, which differs from the previous approach of changing the image width and height to obtain features at different levels through Maxpool in that we cleverly first enlarge the input image to more than twice its size through the Upsample operation, and then obtain more sensory fields through oversized convolutions (in this research, the size of convolution is 1.5 times the size of the image in the input UC Block, and the dilation rate is 6) to obtain more sensory fields, although this causes a computational burden, we reduce the computational burden by using a null depth-separable convolution. At the same time, we combine the previous Maxpool method to acquire detailed features. With

UC Block, we acquire feature maps with larger receptive fields and detailed information.

The formula for upsampling using nearest neighbor interpolation is as follows Equation 1.

$$U(x,y) = F\left(\left\lfloor \frac{x}{s} \right\rfloor, \left\lfloor \frac{y}{s} \right\rfloor\right) \tag{1}$$

$U(x, y)$ represents the value at coordinates (x, y) in the upsampled feature map. F is the original feature map. s is the scaling factor for upsampling. $\lfloor . \rfloor$ denotes the floor function. This formula indicates that for each point (x, y) in the output feature map, we find the value in the original feature map F at the point closest to $\frac{x}{s}$, $\frac{y}{s}$, and use it as the new pixel value.

Dilated Depthwise Separable Convolution consists of two main steps, Dilated Convolution, calculated in Equation 2.

$$G[i,j] = \sum_{k,l} F[i + r \cdot k, j + r \cdot l] \cdot K[k,l] \tag{2}$$

$G[i, j]$ is the output feature map, F is the input feature map, K is the convolutional kernel, r is the dilation rate(In this paper, we take the value 6), and i, j denote positions in the output feature map, while k, l denote positions in the convolutional kernel.

Depthwise Separable Convolution, calculated in Equation 3 and Equation 4.

$$H[i,j,m] = \sum_{k,l} G[i + k, j + l] \cdot D_m[k,l] \tag{3}$$

$$O[i,j,n] = \sum_{m} H[i,j,m] \cdot P_{mn} \tag{4}$$

H is the feature map after Depthwise convolution, D_m is the m-th Depthwise convolutional kernel, O is the final output feature map, and P is the pointWise convolutional kernel.

2) CAT BLOCK

The Cat Block is shown in the Figure 3. This module accepts inputs from two sources, on the one hand the input obtained by fusing the feature maps through Fuse Block and on the other hand the input processed through UC Block. Through this fusion module, we first concatenate the two, and then further fuse the feature maps by CNN, so that we can get the feature maps with the advantages of CNN and the feature maps from the Transformer, which is able to directly calculate the dependency between any two positions in the sequence through the mechanism of self-attention, which makes the model more efficient in dealing with the long distance dependency information, and thus able to capture the long distance dependency information. The integration of this mechanism enhances the model's effectiveness in handling long-range dependency information, thereby enabling it to capture more complex data patterns. Since crack segmentation is an intensive task, the Fuse Block module allows us to obtain feature maps with more detail while maintaining accuracy at large scales.

B. MODEL TRAINING DETAILS

1) DATASETS

In this study, we have opted to utilize the secondary open-source dataset, cracks-APCGAN [28], which has recently been supplemented with additional data from the DeepCrack [35] dataset. This choice was made in light of the open-source nature of the DeepCrack dataset, which we have found to be a valuable resource in our research. cracks-APCGAN was developed via the APCGAN enhanced dataset, the principle is to generate more similar images by GAN based on the training set, and then further enhance the training dataset by manually annotating the GAN-generated images, the enhanced training dataset in the original paper has a great enhancement for the training process, so we chose cracks-APCGAN as our benchmark dataset.

2) LOSS FUNCTION

The loss function plays a crucial role in deep learning as a measure of the difference between the predicted and actual values of the model. During training, the main purpose of the loss function is to guide the model learning and adjust the model parameters by minimising the loss function values to make the model predictions more accurate. The loss function not only affects the efficiency and effectiveness of model training, but also relates to whether the model can effectively learn the complex patterns and structures in the data. Therefore, choosing an appropriate loss function is crucial for the performance optimisation of deep learning models. Three common loss functions are listed below and analysed one by one.

$$BCE_{loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] \tag{5}$$

The Binary Cross-Entropy (BCE) loss function defined as Equation 5. N is the total number of samples, representing all the samples considered when computing the loss. y_i is the actual label of the i-th sample, which can be 0 or 1, representing the two categories in binary classification. p_i is the model's predicted probability that the i-th sample belongs to class 1, with a value between 0 and 1. log denotes the natural logarithm, a function to measure the discrepancy between the predicted probabilities and the actual labels. This formula averages the discrepancies between the predicted probabilities and the actual labels across all samples to obtain the overall loss value.

$$WBCE_{loss} = -\frac{1}{N}\sum_{i=1}^{N}[w_{pos} \cdot y_i \log(p_i) + w_{neg} \cdot (1 - y_i)\log(1 - p_i)] \tag{6}$$

The Weighted Binary Cross-Entropy (WBCE) loss function [25] defined as Equation 6. N is the total number of samples, indicating all samples are considered when computing the loss. y_i is the actual label of the i-th sample, which can be 0 or 1, representing the two categories in binary classification. p_i is the model's predicted probability that the i-th sample belongs to class 1, with a value between 0 and 1. w_{pos} and w_{neg} are the weights for positive and negative classes, respectively. These weights are used to handle class imbalance by adjusting the loss contribution of each class. log denotes the natural logarithm, used to measure the discrepancy between the predicted probabilities and the actual labels. This formula computes the weighted average of the discrepancies between the predicted probabilities and the actual labels across all samples, thus obtaining the overall loss value.

$$
\begin{aligned}
Focal_{loss} = -\frac{1}{N}\sum_{i=1}^{N}[\alpha y_i(1-p_i)^{\gamma}\log(p_i) \\
+(1-\alpha)(1-y_i)p_i^{\gamma}\log(1-p_i)]
\end{aligned}
\tag{7}
$$

The Focal loss function [36] defined as Equation 7. N is the total number of samples, indicating that all samples are considered when computing the loss. y_i is the actual label of the i-th sample, which can be 0 or 1, representing the two categories in binary classification. p_i is the model's predicted probability that the i-th sample belongs to class 1, with a value between 0 and 1. α is a weighting factor for the positive class, used to address class imbalance by weighting the importance of positive/negative examples differently. The exact calculation can be obtained by counting the proportion of positive and negative classes in the datasets. γ is the focusing parameter, a hyperparameter that adjusts the rate at which easy examples are down-weighted, thus allowing the model to focus more on hard, misclassified examples. $(1-p_i)^{\gamma}$ and p_i^{γ} are factors that adjust the contribution of each sample to the loss based on the prediction confidence. These factors reduce the loss for well-classified examples, thereby focusing the model's learning on hard examples. log denotes the natural logarithm, used to measure the discrepancy between the predicted probabilities and the actual labels. This formula aims to reduce the loss contribution from easy examples and increase the influence of hard examples, improving model performance on difficult classification tasks by modulating the effect of the sample based on its prediction confidence and actual class.

In the case of crack segmentation, the default and most commonly used choice is the cross-entropy loss, which is applied pixel by pixel. The loss function evaluates the class prediction for each pixel independently and averages over all pixels. However, it can be biased by an unbalanced dataset, causing most classes to dominate. To overcome this problem when the dataset is unbalanced, they introduced weighted cross-entropy loss. As a further improvement to cross-entropy loss, the focal loss technique was introduced. This is achieved by changing the structure of the cross-entropy loss. When the focal loss is applied to samples with accurate classification, the scaling factor values are weighted down. This ensures that more difficult samples are emphasised and therefore advanced imbalances do not bias the overall computation. Therefore, in this paper we have chosen the focal loss as the loss function.

3) OPTIMIZER

In our crack segmentation task, we selected AdamW [37] as the optimisation algorithm to better adjust model weights and mitigate overfitting. The AdamW optimiser is a variant of the Adam optimiser that primarily improves model generalisation by modifying the weight decay strategy. Traditional L2 regularisation methods may not be effective in adaptive learning rate optimisation algorithms, as such algorithms automatically adjust the update step for each parameter, potentially conflicting with the goals of L2 regularisation. In contrast, AdamW decouples weight decay from the optimiser's adaptive learning rate adjustments, allowing weight decay to operate independently of the adaptive learning rate mechanism, thereby implementing regularisation more effectively.

In this study, we used an initial learning rate of 6×10^{-5} chosen on the basis of experience and the results of several experiments. This learning rate is intended to strike a balance between convergence speed and stability during the training process, avoiding excessively large update steps early in training that could cause the model to fail to stabilise on an optimal solution. The AdamW optimiser allows us to finely control the learning rate for each parameter in an adaptive manner, while using weight decay to suppress overfitting, providing strong support for deep learning models in crack segmentation tasks.

Furthermore, by using AdamW's weight decay mechanism, we can more effectively manage model complexity and prevent the occurrence of overfitting, which is particularly important for tasks such as crack segmentation that require a high level of detail and precision. By considering both training efficiency and model generalisation capabilities, we are confident that the choice of the AdamW optimiser and its configuration will provide optimal training results for our crack segmentation model.

IV. RESULTS AND DISCUSSION

A. EVALUATION METRICS

In the table 1, we first define the variables commonly used to evaluate the metrics. In crack segmentation, the commonly used evaluation indicators are as follows:

- **Precision** in crack segmentation assesses the ratio of correctly predicted positive areas to all areas predicted as positive, highlighting the accuracy of positive class predictions. The Precision can be calculated in Equation 8.

$$
Precision = \frac{\sum p}{\sum p + \sum \overline{p}} \tag{8}
$$

- **Recall** in crack segmentation quantifies the fraction of true positive areas correctly identified, reflecting the model's sensitivity to actual positives. The Recall can be calculated in Equation 9.

$$Recall = \frac{\sum p}{\sum p + \sum \overline{g}} \tag{9}$$

- **F1 score** in crack segmentation is the harmonic mean of precision and recall, balancing both metrics. It evaluates the model's accuracy and sensitivity, with higher values indicating better overall performance. The F1 score can be calculated in Equation 10.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{10}$$

- **mIoU** (mean Intersection over Union) in semantic segmentation calculates the average IoU, which is the overlap versus total area of predicted and true regions, across all classes. It assesses the model's segmentation accuracy, with higher mIoU indicating better performance. The mIoU can be calculated in Equation 11.

$$mIoU = \frac{1}{C}\sum_{c=1}^{C}\frac{\sum p}{\sum p + \sum \overline{p} + \sum \overline{g}} \tag{11}$$

TABLE 1. Definitions in crack segmentation.

	Prediction Crack	Prediction Non-Crack
Ground Truth Crack	p	$\overline{g}$
Ground Truth Non-Crack	$\overline{p}$	g

B. RESULTS

In this paper, we describe a novel deep learning model for crack segmentation, MixSegNet, and provide an in-depth analysis of its performance. As shown in Table 2, MixSegNet outperforms current mainstream segmentation models, including LRASPP, FCN, DeepLabV3, U-Net, AttuNet, and Swin-UNet, in a number of key metrics. Specifically, MixSegNet attains a precision of 0.952, a recall of 0.882, an F1 score of 0.915, and a mean intersection over union (mIoU) scores of 0.848. These results show that MixSegNet has a significant leading performance on the crack segmentation task. When analysing these results in more detail, we can see that MixSegNet is only slightly higher in precision by 0.001 compared to U-Net, but the improvement in recall is even more significant, being 0.040 higher than that of U-Net. This suggests that MixSegNet is able to maintain a high level of detection accuracy while avoiding missing actual cracks. In addition, the F1 score, which is a reconciled average of precision and recall, also shows that MixSegNet outperforms all compared models on this metric, further demonstrating its superiority in correctly identifying and segmenting cracks. MixSegNet also performs well on the mIoU metric, outperforming the second highest model, AttuNet, by 0.012, demonstrating better consistency and overall performance in the crack segmentation task. mIoU is an important metric for assessing the quality of a model's segmentation, and its high value underscores MixSegNet's reliability and robustness in the crack detection task. These performance improvements of MixSegNet are due to its unique model design. The model combines the use of a CNN-based UC block module and an attention mechanism that improves segmentation accuracy by focusing on key information in the image. The introduction of the UC block module improves the capture of crack features, while the attention mechanism ensures that the model is able to distinguish between foreground cracks and complex backgrounds. In addition, the parallel and serial structure we adopt allows the model to respond effectively to crack variations in different scenes, improving the model's ability to adapt to crack morphology. We also employ a focal loss function and an adaptive learning strategy to address the problem of category imbalance in the dataset. The focal loss function can reduce the weight of easy to classify samples, allowing the model to focus more on crack regions that are difficult to segment. The adaptive learning strategy further optimises the training process and ensures the model's performance in a variety of complex scenarios and conditions.

TABLE 2. Comparison of different models on the test data.

Models	Precision	Recall	F1 score	mIoU
LRASPP	0.933	0.775	0.847	0.758
FCN	0.942	0.824	0.879	0.783
DeepLabV3	0.938	0.839	0.886	0.795
U-Net	0.951	0.842	0.893	0.818
AttuNet	0.947	0.868	0.906	0.836
Swin-UNet	0.937	0.852	0.892	0.815
MixSegNet	**0.952**	**0.882**	**0.915**	**0.848**

As illustrated in Figure 4, this study randomly selected seven images from diverse scenes and employed distinct segmentation models to illustrate the outcomes. The segmentation outcomes depicted in the figure demonstrate that the results produced by the MixSegNet model align with those presented in Table 2. Additionally, the MixSegNet model exhibits consistent and continuous segmentation outcomes when compared to other models. This consistency can be attributed to the fact that MixSegNet integrates the strengths of CNN and Transformer. A comparison of the details of the various models reveals that the MixSegNet model also maintains a leading level of detail processing, which is crucial for the refined crack segmentation scene.

In summary, MixSegNet's innovative design and strategy set a new performance benchmark for the crack segmentation task. Its outstanding performance bodes well for the model's wide application and far-reaching impact on future crack segmentation. We are excited about MixSegNet's ability to handle complex problems and look forward to seeing its performance in real-world applications.

C. DISCUSSION

Table 3 shows the results of the ablation experiments performed on the MixSegNet model, where the contribution of each part to the overall model performance is verified by incrementally adding UC Block and Transformer modules.

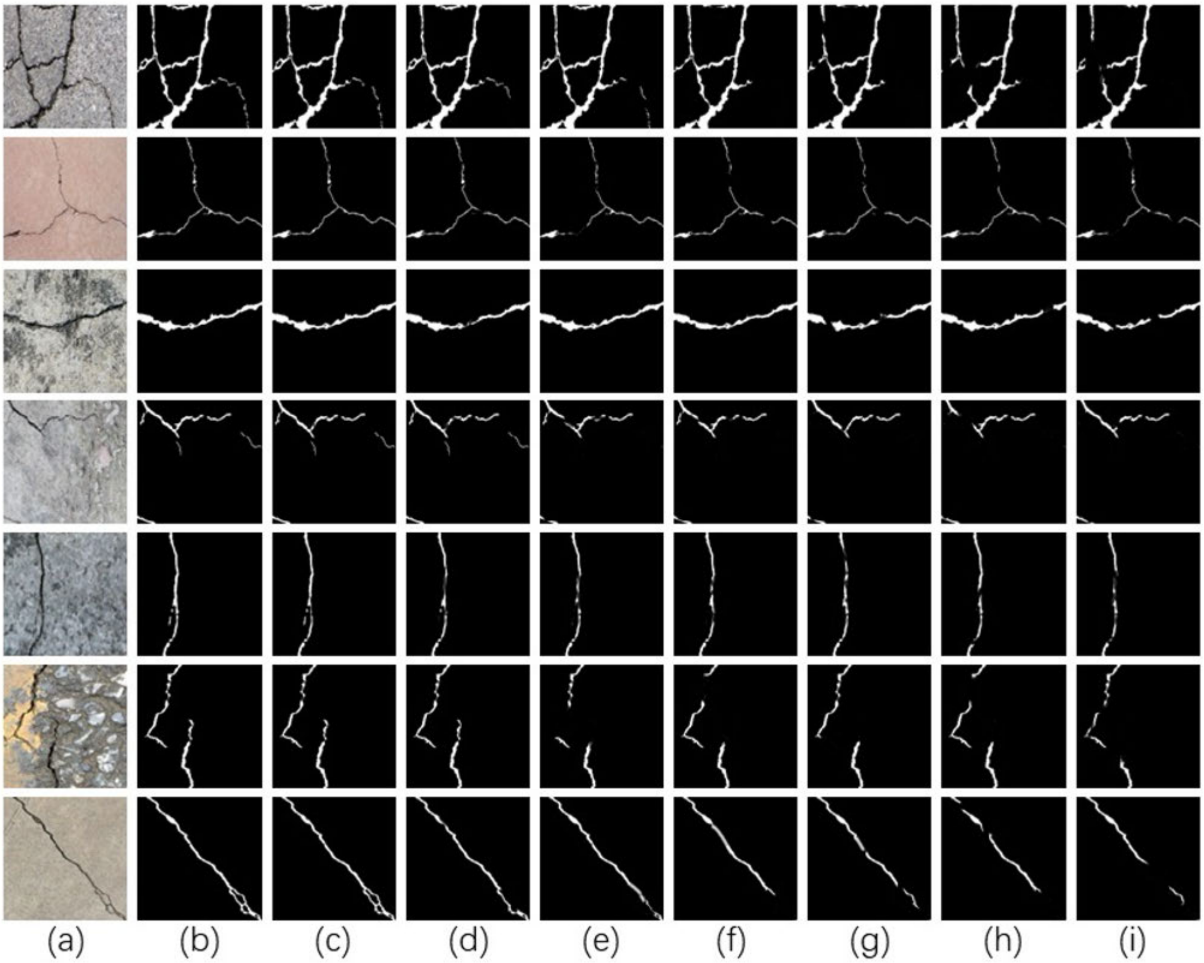

FIGURE 4. A comparison of the results obtained using different segmentation models in different scenarios. (a) Original Image (b) Ground Truth (c) MixSegNet (d) AttuNet (e) Swin-UNet (f) U-Net (g) Deeplabv3 (h) FCN (i) LRASPP.

Ablation experiments are a method of assessing the importance of model components by removing or adding specific sections and observing changes in model performance. The results of these experiments are analysed in detail below. The base model serves as a frame of reference with Precision, Recall, F1 Score and mean Intersection over Union (mIoU) of 0.931, 0.836, 0.881 and 0.813 respectively. This model already has good performance on its own, providing a solid foundation for adding new modules. Adding the UC Block module to the base model improves all performance metrics, including precision to 0.935, recall to 0.858, F1 score to 0.895 and mIoU to 0.827. This shows that the UC Block module plays a key role in improving the performance of crack segmentation, especially in the significant increase in recall, suggesting that the addition of UC Block helps the model to reduce the cases of missed crack detection. When the Transformer module is added to the base model, the recall rate improves from 0.836 to 0.863, showing the effectiveness of the Transformer module in capturing the global information of the crack image and understanding the relationship between the crack and the background. However, the accuracy decreased slightly to 0.928, which may be due to the fact that the Transformer module's emphasis on global features may lead to some local noise being misidentified. Nevertheless, the slight improvement in F1 and mIoU confirms the positive contribution of the Transformer module in the model. By combining the UC Block and Transformer modules into MixSegNet, the model was significantly improved in all metrics. Precision reached a maximum of 0.952, recall also reached a maximum of 0.882, and F1 score and mIoU reached 0.915 and 0.848 respectively. This all-round performance improvement fully demonstrates the positive impact of combining the UC Block and Transformer modules on the model's performance, especially on recall and mIoU, which shows the model's efficiency in crack detection and accuracy in segmenting the crack region. The contribution of each component to the MixSegNet model was experimentally demonstrated: the UC Block module significantly improved the recall rate, showing that it is effective in avoiding missed crack detection. While the Transformer module enhances the global understanding of the model and improves recall. When these two modules are combined, they work in synergy to significantly improve the overall performance of the model, especially in terms

TABLE 3. Ablation experiment.

Models	Precision	Recall	F1 score	mIoU
Base	0.931	0.836	0.881	0.813
Base+UC	0.935	0.858	0.895	0.827
Base+Transformer	0.928	0.863	0.894	0.822
MixSegNet	**0.952**	**0.882**	**0.915**	**0.848**

of precision and recall, enabling MixSegNet to excel in the field of crack segmentation. These results validate the effectiveness of our proposed model design and provide a new powerful tool for crack segmentation tasks.

Although the results indicate that MixSegNet has achieved a leading level of performance, it is important to note that the model combines the CNN and Transformer architectures, which inevitably increases the computational complexity. However, the focus of this study is on high-precision segmentation, and computational complexity is not the primary consideration. In the next research, we will optimize the computational efficiency of the model and improve its real-time performance. The current work focuses on improving the accuracy of crack segmentation, and subsequent work will be used in real drone scenarios to realize real-time crack segmentation warning using a drone and an onboard computer.

V. CONCLUSION

This research proposes an innovative crack segmentation model, MixSegNet, which represents a major breakthrough in the field of crack segmentation. MixSegNet not only improves the perceptual capability of the model, but also strengthens the capture of details and the maintenance of long-range dependencies by combining an innovative UC block and a parallel CNN and Transformer design. This unique two-pronged approach effectively overcomes the limitations of previous single-architecture designs and achieves significant improvements in key performance metrics such as 95.2% precision, 88.2% recall, 91.5% F1 score and 84.8% mIoU. The performance advantages of MixSegNet are fully demonstrated by comparing it to existing state-of-the-art models LRASPP, FCN, DeepLabV3, U-Net, AttuNet and Swin-UNet. The model not only improves in all indices, but also shows better generalisation ability in the experiments, predicting its wide applicability and potential value in practical applications. In future research, we plan to extend the scope of application of MixSegNet, improve its generalisation ability, and verify its robustness by testing it on more diverse and complex datasets. At the same time, we will work on optimising the computational efficiency of the model to meet real-time processing requirements and applications in real industrial scenarios. We will also explore the potential of MixSegNet in cross-domain image segmentation tasks such as medical image analysis and remote sensing image processing. Improving the interpretability of the model and adapting it to small-sample learning environments to maintain excellent performance in data-constrained situations will also be the focus of our future work. With these efforts, we expect to open new avenues for research and practical applications of crack segmentation.

REFERENCES

[1] S. Zhou, C. Canchila, and W. Song, "Deep learning-based crack segmentation for civil infrastructure: Data types, architectures, and benchmarked performance," *Autom. Construct.*, vol. 146, Feb. 2023, Art. no. 104678.

[2] A. Akagic, E. Buza, S. Omanovic, and A. Karabegovic, "Pavement crack detection using OTSU thresholding for image segmentation," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, Aug. 2018, pp. 1092–1097.

[3] X. Yang, H. Li, Y. Yu, X. Luo, T. Huang, and X. Yang, "Automatic pixel-level crack detection and measurement using fully convolutional network," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, no. 12, pp. 1090–1109, Dec. 2018.

[4] J. König, M. David Jenkins, P. Barrie, M. Mannion, and G. Morison, "A convolutional neural network for pavement surface crack segmentation using residual connections and attention gating," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1460–1464.

[5] Y. Ren, J. Huang, Z. Hong, W. Lu, J. Yin, L. Zou, and X. Shen, "Image-based concrete crack detection in tunnels using deep fully convolutional networks," *Construct. Building Mater.*, vol. 234, Feb. 2020, Art. no. 117367.

[6] D. Kang, S. S. Benipal, D. L. Gopal, and Y.-J. Cha, "Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning," *Autom. Construct.*, vol. 118, Oct. 2020, Art. no. 103291.

[7] S. L. H. Lau, E. K. P. Chong, X. Yang, and X. Wang, "Automated pavement crack segmentation using U-Net-based convolutional neural network," *IEEE Access*, vol. 8, pp. 114892–114899, 2020.

[8] J. Liu, X. Yang, S. Lau, X. Wang, S. Luo, V. C. Lee, and L. Ding, "Automated pavement crack detection and segmentation based on two-step convolutional neural network," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 35, no. 11, pp. 1291–1305, Nov. 2020.

[9] J. Guan, X. Yang, L. Ding, X. Cheng, V. C. S. Lee, and C. Jin, "Automated pixel-level pavement distress detection based on stereo vision and deep learning," *Autom. Construct.*, vol. 129, Sep. 2021, Art. no. 103788.

[10] R. Ali, J. H. Chuah, M. S. A. Talip, N. Mokhtar, and M. A. Shoaib, "Crack segmentation network using additive attention gate—CSN-II," *Eng. Appl. Artif. Intell.*, vol. 114, Sep. 2022, Art. no. 105130.

[11] W. Wang, C. Su, G. Han, and H. Zhang, "A lightweight crack segmentation network based on knowledge distillation," *J. Building Eng.*, vol. 76, Oct. 2023, Art. no. 107200.

[12] H. Liu, J. Yang, X. Miao, C. Mertz, and H. Kong, "CrackFormer network for pavement crack segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 1, no. 1, pp. 1–13, Aug. 2023.

[13] Z. Wu, Y. Tang, B. Hong, B. Liang, and Y. Liu, "Enhanced precision in dam crack width measurement: Leveraging advanced lightweight network identification for pixel-level accuracy," *Int. J. Intell. Syst.*, vol. 2023, pp. 1–16, Sep. 2023.

[14] H. Yao, Y. Liu, H. Lv, J. Huyan, Z. You, and Y. Hou, "Encoder–decoder with pyramid region attention for pixel-level pavement crack recognition," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 39, no. 10, pp. 1490–1506, May 2024.

[15] Q. Lin, W. Li, X. Zheng, H. Fan, and Z. Li, "DeepCrackAT: An effective crack segmentation framework based on learning multi-scale crack features," *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 106876.

[16] C. Tang, S. Jiang, H. Li, D. Huang, X. Huang, and Y. Xiong, "Lightweight concrete crack segmentation method based on Deeplabv3+," in *Proc. 3rd Int. Conf. Comput. Vis. Pattern Anal.*, Aug. 2023, pp. 399–404.

[17] Y. Chen, S. Dong, B. Hu, Q. Liu, and Y. Qu, "A dynamic semantic segmentation algorithm with encoder-crossor-decoder structure for pixel-level building cracks," *Meas. Sci. Technol.*, vol. 35, no. 2, Feb. 2024, Art. no. 025139.

[18] Z. Li, C. Yin, and X. Zhang, "Crack segmentation extraction and parameter calculation of asphalt pavement based on image processing," *Sensors*, vol. 23, no. 22, p. 9161, Nov. 2023.

[19] M. Sohaib, S. Jamil, and J.-M. Kim, "An ensemble approach for robust automated crack detection and segmentation in concrete structures," *Sensors*, vol. 24, no. 1, p. 257, Jan. 2024.

[20] Y. A. Lumban-Gaol, A. Rizaldy, and A. Murtiyoso, "Comparison of deep learning architectures for the semantic segmentation of slum areas from satellite images," in *Proc. Int. Archives Photogramm., Remote Sens. Spatial Inf. Sci.*, 2023, pp. 1439–1444.

[21] E. Kot, Z. Krawczyk, K. Siwek, L. Królicki, and P. Czwarnowski, ''Deep learning-based framework for tumour detection and semantic segmentation,'' *Bull. Polish Acad. Sci. Tech. Sci.*, vol. 69, Mar. 2021, Art. no. 136750.

[22] S. Agarwal, A. Sawant, M. Faisal, S. E. Copp, J. Reyes-Zacarias, Y.-R. Lin, and S. J. Zinkle, ''Application of a deep learning semantic segmentation model to helium bubbles and voids in nuclear materials,'' *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 106747.

[23] L. Fan and C. Zhou, ''Cloud-to-Ground and intra-cloud nowcasting lightning using a semantic segmentation deep learning network,'' *Remote Sens.*, vol. 15, no. 20, p. 4981, Oct. 2023.

[24] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[25] O. Ronneberger, P. Fischer, and T. Brox, ''U-Net: Convolutional networks for biomedical image segmentation,'' in *Proc. 18th Int. Conf.*, 2015, pp. 234–241.

[26] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, ''Searching for MobileNetV3,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[27] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, ''Rethinking atrous convolution for semantic image segmentation,'' 2017, *arXiv:1706.05587*.

[28] T. Zhang, D. Wang, A. Mullins, and Y. Lu, ''Integrated APC-GAN and AttuNet framework for automated pavement crack pixel-level segmentation: A new solution to small training datasets,'' *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 4474–4481, Apr. 2023.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–20.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, ''An image is worth 16×16 words: Transformers for image recognition at scale,'' 2021, *arXiv:2010.11929*.

[31] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, ''Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886.

[32] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ''Swin transformer: Hierarchical vision transformer using shifted windows,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[33] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, ''SegFormer: Simple and efficient design for semantic segmentation with transformers,'' in *Proc. Adv. Neural Inf. Process. Sys. (NIPS)*, vol. 34, Dec. 2021, pp. 12077–12090.

[34] H. Cao and Y. Wang, ''Swin-Unet: Unet-like pure transformer for medical image segmentation,'' in *Computer Vision–ECCV*. Cham, Switzerland: Springer, 1007, pp. 205–218.

[35] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, ''DeepCrack: A deep hierarchical feature learning architecture for crack segmentation,'' *Neurocomputing*, vol. 338, pp. 139–153, Apr. 2019.

[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, ''Focal loss for dense object detection,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[37] I. Loshchilov and F. Hutter, ''Decoupled weight decay regularization,'' 2019, *arXiv:1711.05101*.

YANG ZHOU received the B.Eng. degree in automation from Hubei Engineering University. He is currently pursuing the M.Eng.Sc. degree with the Faculty of Engineering, Universiti Malaya, under the supervision of Dr. Mokhtar from the Department of Electrical Engineering.

During his previous projects involving machine learning applications, he gained expertise in various computer vision and natural language processing models and techniques, such as convolutional neural networks (CNNs), transformers, model fine-tuning, and data augmentation. His research interests include computer vision, deep learning, and image segmentation, with a focus on crack segmentation using imbalanced data.

RAZA ALI (Senior Member, IEEE) received the B.S. degree in telecommunication engineering from Balochistan University of Information Technology, Engineering and Management Sciences (BUITEMS), Quetta, Pakistan, the M.S. degree in electrical engineering (communication) from UET, Lahore, and the Ph.D. degree from the University of Malaya, Malaysia, in 2022. During the Ph.D. studies, he was associated with the VIP Laboratory, University of Malaya. He is currently an Assistant Professor with the Faculty of Information and Communication Technology, BUITEMS. His research interests include signal processing, computer vision, machine learning, and deep learning.

NORRIMA MOKHTAR received the B.Eng. degree in electrical engineering from the University of Malaya, in 2000, and the M.Eng. degree from Oita, Japan, in 2006. After working two years with the International Telecommunication Industry with attachment at Echo Broadband GmbH, she managed to secure a Panasonic Scholarship which required intensive screening at the national level, in 2002. To date, she has successfully supervised seven Ph.D. and four M.Eng.Sc. students (by research). She is the author and co-author of more than 50 publications in international journals and proceedings in *Sensors*, *Automation*, IEEE Transactions on Image Processing, *Human-Computer Interface*, *Brain-Computer Interface*, *UAV*, and *Robotics*. She received financial support from a Panasonic Scholarship for her M.Eng. degree. She is active as a reviewer for many reputable journals and several international conferences.

SULAIMAN WADI HARUN received the B.E. degree in electrical and electronics system engineering from Nagaoka University of Technology, Japan, in 1996, and the M.Sc. and Ph.D. degrees in photonics technology from the University of Malaya, in 2001 and 2004, respectively. He was an Adjunct Professor at Airlangga University, Indonesia, and Ton Duc Thang University, Vietnam. He has nearly 20 years of research experience in the development of optical fiber devices, including fiber amplifiers, fiber lasers, and fiber optic sensors. He was also involved in exploiting new nanomaterials, such as graphene, carbon nanotubes, black phosphorous, topological insulators for various fiber lasers, and sensor applications. He has received about ten research grants of value over RM4M from the Ministry of Education and the Ministry of Science, Technology, and Innovation. He has published more than 700 articles in ISI journals and his papers have been cited more than 7000 times with an H-index of 37, showing the impact on the community. He is a fellow of Malaysian Academic of Science and the Founder and Honorary Advisor for the Optical Society of Malaysia. He received the prestigious award of Malaysian Rising Star from the Ministry of Higher Education, in 2016, for his contribution to international collaboration.

MASAHIRO IWAHASHI (Senior Member, IEEE) received the B.Eng., M.Eng., and D.Eng. degrees in electrical engineering from Tokyo Metropolitan University, Tokyo, Japan, in 1988, 1990, and 1996, respectively. In 1990, he joined Nippon Steel Company Ltd. Since 1993, he has been with Nagaoka University of Technology, Nagaoka, Japan, where he is currently a Professor with the Department of Electrical, Electronics, and Information Engineering. His research interests include digital signal processing, multirate systems, and image compression. He is a Senior Member of IEICE and a member of the Asia–Pacific Signal and Information Processing Association and the Institute of Image Information and Television Engineers.