scientific reports

Check for updates

OPEN

FCN attention enhancing asphalt pavement crack detection through attention mechanisms and fully convolutional networks

Huiyuan Zhang, Jiawei Liu✉ & Guoping Hu

This paper presents an innovative approach to detecting cracks in asphalt pavement using an FCN-attention model, which integrates attention mechanisms into a fully convolutional network (FCN) for enhanced pixel-level segmentation. The model employs a ResNet-50-based encoder and incorporates channel-wise and spatial attention modules to refine feature extraction and focus on the most relevant image regions. The results demonstrate that the FCN-attention model outperforms traditional models such as VGG-16, AlexNet, MobileNet, and GoogleNet across multiple evaluation metrics. Specifically, the FCN-attention model achieves a global accuracy rate of 90.79%, with a precision of 92.3%, recall of 89.5%, and an F1-score of 90.9%. Additionally, the model achieves an average intersection-over-union (IoU) ratio of 69.7% and a test duration of 109.1 ms per image. The proposed method also excels in crack length and width calculation, providing real-world dimensions for the detected cracks. The model's effectiveness is further validated through an ablation study, which highlights the significant impact of the attention mechanism on model performance.

Keywords Asphalt pavement, Crack detection, Fully convolutional network (FCN), Attention mechanism, Real-time monitoring

Automatic detection of cracks in asphalt pavement is critical for road maintenance and safety, as cracks serve as key indicators of pavement health. Early detection and timely repair can prevent severe damage, reduce maintenance costs, and ensure traffic safety[1–3]. The rapid development of deep learning technologies has led to the prominence of data-driven approaches in this field, significantly enhancing detection efficiency and accuracy[4–7].

Traditional crack monitoring methods rely on manual visual inspections, which are labor-intensive, costly, and subject to subjective interpretation. In contrast, deep learning techniques, particularly convolutional neural networks (CNNs), have demonstrated significant advantages in image processing and crack detection[8–10]. Automated image analysis reduces the manual inspection burden while improving accuracy. For example, Ji et al.[11] proposed an integrated approach using the DeepLabv3 + CNN for crack detection and pixel-level quantification, while Ground Penetrating Radar (GPR) has shown potential for detecting cracks through changes in electromagnetic reflections.

The task of automatically detecting cracks from images is crucial for maintaining the safety and durability of pavements, especially those made from Portland cement concrete (PCC) and asphalt concrete (AC). Advances in deep learning, such as the U-Hierarchical Dilated Network (U-HDN), have enabled end-to-end crack detection by incorporating hierarchical feature learning and dilated convolution[13]. The high labor and cost demands of traditional inspection methods have driven the development of automated systems that utilize CNNs to detect and assess pavement cracks more efficiently[14]. However, challenges persist, particularly in addressing complex noise interference in images. Li et al.[15] developed a novel system for recognizing and analyzing cracks, while Safaei et al.[16] focused on creating an automated method for detecting and categorizing cracks in both 2-D and 3-D pavement images. These systems also measure crack dimensions based on orientation and curve lengths, addressing the various types of cracks caused by harsh weather conditions and prolonged vehicle usage.

Despite these advancements, a gap remains between cutting-edge deep learning technologies and traditional pixel-level detection algorithms. Huyan et al.[19] aimed to bridge this gap by detailing a deep neural network model designed for pixel-wise detection of pavement cracks, using images collected from various sources.

School of Civil and Transportation Engineering, Henan University of Urban Construction, Longxiang Road, Xincheng Area, Pingdingshan 467036, Henan, China. ✉email: liujiawei20250217@163.com

To address ongoing challenges, a machine vision-based method utilizing deep convolutional neural network (DCNN) technology has been proposed, showing effectiveness on publicly accessible benchmark datasets of concrete cracks[20]. Chen et al.[21] examined ARF-Crack, a rotation-invariant deep FCN specifically designed for pixel-level crack detection, tested across multiple benchmark datasets. This system outperforms existing models like FCN and R-CNN in image processing capabilities while using less memory[22]. Additionally, Zhang et al.[23] proposed a method for improved semantic segmentation of high-resolution images, transitioning from CNN to FCN to enhance segmentation accuracy. The introduction of attention mechanisms has further improved feature extraction, particularly in complex environments, enabling the FCN to focus on the most relevant image areas and better generalize to unseen data. The FCN-attention model was chosen for its ability to achieve precise pixel-level segmentation while maintaining computational efficiency. Fully Convolutional Networks (FCNs) are well-suited for crack detection as they retain spatial information and generate dense predictions, which are crucial for accurately segmenting fine crack structures. The integration of attention mechanisms enhances feature extraction, allowing the model to focus on critical regions while reducing interference from background noise. While architectures such as Transformers and YOLO-based segmentation models offer strong feature representation and object detection capabilities, they are less optimized for pixel-wise segmentation tasks, particularly in cases where fine details and structural continuity are essential. Transformer-based models require significantly higher computational resources, making them less suitable for real-time applications. YOLO-based models, designed primarily for object detection, lack the spatial granularity needed for precise crack segmentation. Given these considerations, the FCN-attention model provides a balanced approach, offering high segmentation accuracy, computational efficiency, and adaptability to varying pavement conditions[29]. By distinguishing more accurately between crack and non-crack areas, this approach reduces false detection rates and enhances detection precision and robustness. Thus, the proposed fully convolutional network (FCN) model with an attention mechanism achieves more accurate crack detection across diverse environmental conditions.

Fundamental methods

This study proposes an FCN-attention model that enhances the foundational architecture of Fully Convolutional Networks (FCNs) by integrating an attention mechanism to improve pixel-level segmentation tasks. The model employs a ResNet-50-based encoder, structured into four stages, each with residual blocks containing 3×3 convolutional layers and skip connections. The encoder begins with a 7×7 convolutional layer and a max-pooling layer, with subsequent stages increasing the filter sizes to 512, ultimately producing 256-channel activation maps at 1/16th of the input resolution.

The attention mechanism is incorporated between the encoder and decoder, combining channel-wise and spatial attention to refine the extracted features. The channel-wise attention module reduces spatial dimensions via global average pooling and generates a channel attention map, while the spatial attention module produces a spatial attention map that highlights key regions within the image.

The decoder restores the spatial resolution through upsampling layers and uses skip connections to merge low-level spatial details with high-level semantic information. A final 1×1 convolutional layer with sigmoid activation generates the pixel-wise binary segmentation map to identify cracks.

By integrating the attention mechanism into the encoder-decoder framework, the model emphasizes the most informative features, significantly enhancing its ability to accurately detect cracks in asphalt pavement, thereby improving segmentation accuracy and robustness.

Attention mechanism

The attention mechanism in deep learning enables models to focus on the most relevant parts of the input during data processing. Originally introduced in 2014 for machine translation, this technique allows a model to selectively concentrate on important aspects of the input rather than treating the entire sequence uniformly[30]. Similar to human visual attention, where we focus on specific words while skimming others, the attention mechanism directs the model's resources toward the most useful information for the task at hand. The mechanism works by generating a context vector at each output timestep, which is a weighted average of the input sequence states. These weights represent the importance of each input segment in producing the output, and the alignment between input and output is typically learned jointly during training.

Mathematically, given an input sequence $X = (x_1, x_2, \ldots, x_n)$ and output at time t as y_t, the context vector c_t is computed as:

$$c_t = \sum \alpha_{t,i} * h_i, \tag{1}$$

Where h_i are hidden states from the encoder and $\alpha_{t,i}$ are attention weights denoting the importance of input xi in generating y_t. The weights α allow gradient-based learning and can amplify or attenuate different parts adaptively.

The context vector c_t condenses relevant information from the entire input sequence into a dynamic representation. The decoder can then use c_t to generate y_t. The model learns what to attend to base on the decoding objective during end-to-end training. Attention layers can be inserted into models in various ways. Encoder-decoder models often employ an attention layer between encoding and decoding steps. CNNs and RNNs can also have internal self-attention to amplify important features. Overall, attention improves model accuracy by focusing computation on the most informative parts of the data.

Fully convolutional networks

Fully Convolutional Networks (FCNs)[31] are a specialized type of Convolutional Neural Network (CNN) designed for semantic segmentation, which involves assigning a class label to each pixel in an image. Unlike traditional CNNs that combine convolutional layers with fully-connected layers, FCNs consist entirely of convolutional layers. This architecture allows them to accept inputs of varying sizes and produce correspondingly-sized output segmentation masks. FCNs achieve this by using pooling layers for downsampling to progressively reduce spatial dimensions, followed by upsampling layers, such as transpose convolution or dilated convolution, to restore the original resolution in the output layers. This approach enables precise pixel-level segmentation. Additionally, FCNs support end-to-end learning, where each pixel is mapped to a feature vector indicating class probabilities. The learning process is guided by a loss function, such as cross-entropy, which is supervised by pixel-level ground truth labels. Overall, FCNs excel at tasks requiring detailed and accurate semantic segmentation, as their architecture effectively captures spatial relationships within the input data. This structure is illustrated in Fig. 1.

FCN with attention mechanisms

The FCN-attention model architecture proposed in this study builds upon the foundation of fully convolutional networks (FCNs) by incorporating attention mechanisms to enhance performance in pixel-level segmentation tasks.

The model architecture utilizes a ResNet-50-based encoder, structured into four stages, each comprising residual blocks with 3×3 convolutional layers and skip connections. The encoder begins with a 7×7 convolutional layer containing 64 filters, followed by a max-pooling layer. The subsequent stages progressively increase the filter sizes to 64, 128, 256, and 512, with the final stage outputting 256-channel activation maps at 1/16th of the input size. To expand the receptive field without increasing the number of parameters, atrous (dilated) convolution with a dilation rate of $r = 2$ is applied in the last two stages.

Attention modules are integrated between the encoder and decoder stages, combining channel-wise and spatial attention mechanisms to refine the features extracted by the encoder. The channel-wise attention module compresses the spatial dimensions of the feature maps through global average pooling and generates a channel attention map using a small fully connected network with sigmoid activation. Meanwhile, the spatial attention module applies convolutional operations to the channel-compressed feature maps to produce a spatial attention map, which emphasizes significant regions in the spatial domain.

The decoder is designed to gradually increase the spatial resolution of the feature maps through upsampling layers, restoring them to the original input size. Skip connections are employed to merge low-level spatial information from earlier encoder layers with high-level semantic information from the decoder. Finally, a 1×1 convolutional layer with sigmoid activation generates the pixel-wise binary segmentation map, identifying the presence of cracks.

The attention mechanism is seamlessly integrated within the encoder-decoder framework. The attention-modulated feature maps are passed to the decoder, ensuring that the most informative features are highlighted. This integration enhances the model's capability to focus on relevant image areas, which is critical for accurately detecting cracks in asphalt pavement. The network architecture is shown in Fig. 2.

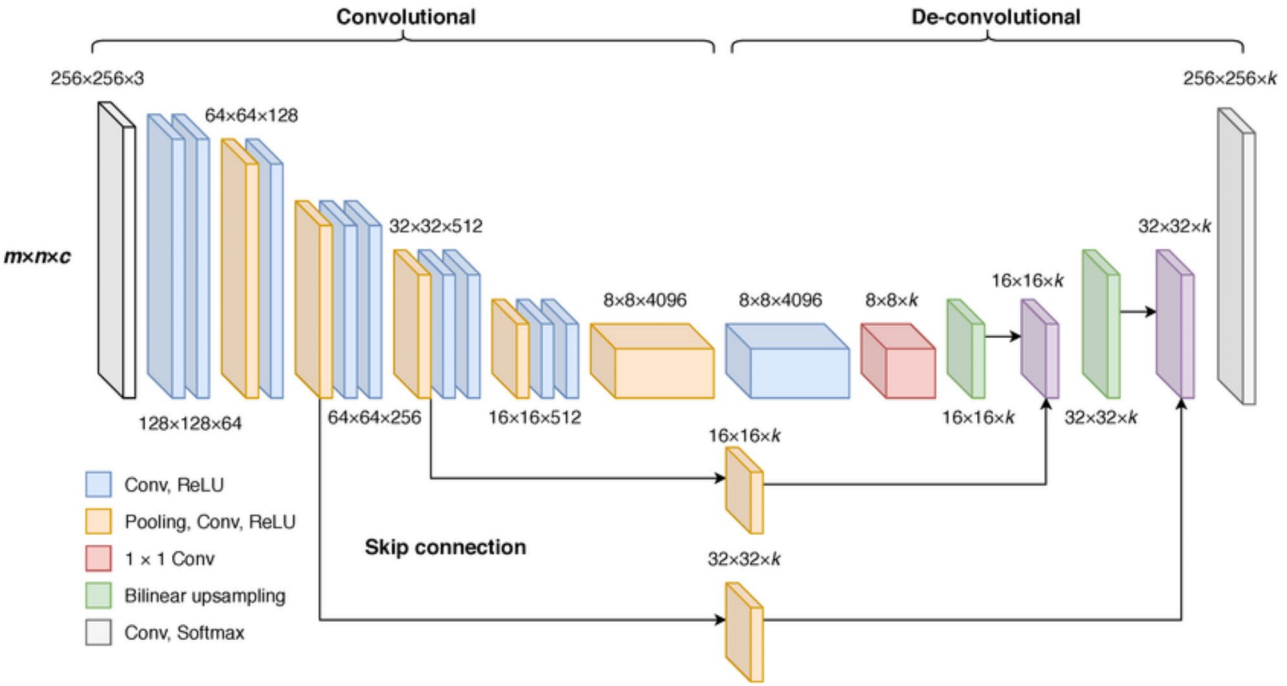


Fig. 1. Network architecture of FCNs.

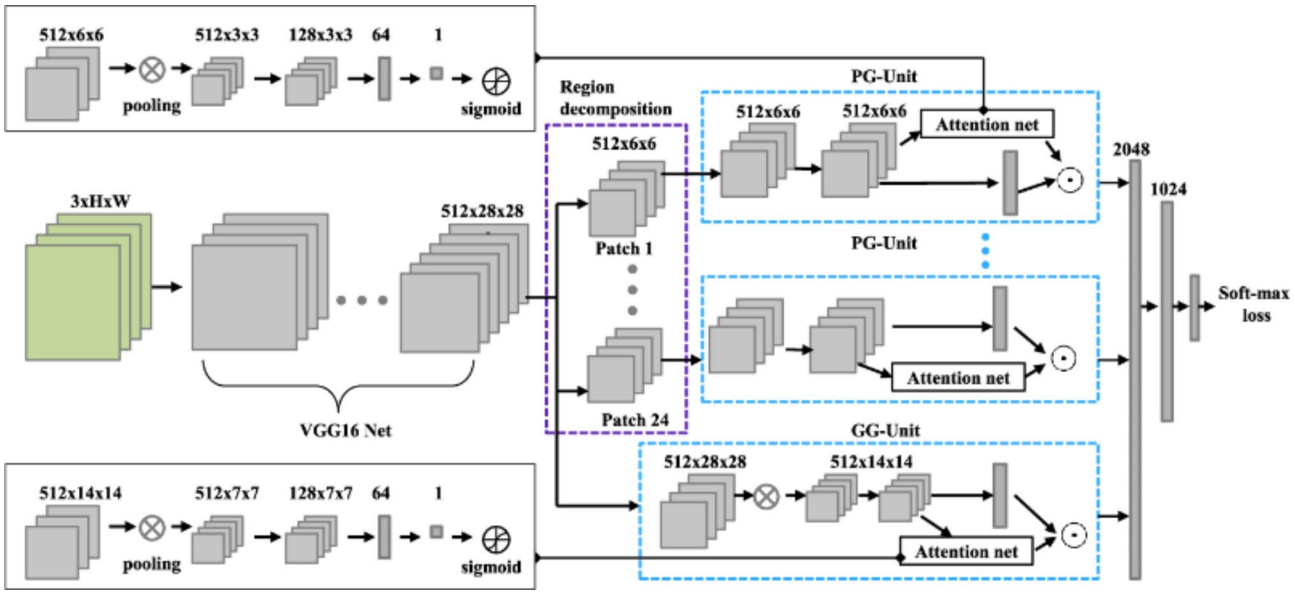


Fig. 2. Network architecture of FCNs with attention mechanisms.

Crack calculation method

1. Threshold segmentation of crack pixel-level identification results.

Defect images are processed through a semantic segmentation network, mapped to a range of (0,1) indicating the probability of a pixel being a crack. To simplify subsequent processing, a threshold of 0.5 is used to binarize the predicted images.

$$\Pr_t(i,j) = \begin{cases} 1 & P(i,j) \geqslant 0.5 \\ 0 & P(i,j) < 0.5 \end{cases}, i \in [1,H], j \in [1,W], t \in [1,T], \tag{2}$$

$\Pr_t$ represents the model's prediction for the tth image, and T is the total number of predicted images. $\Pr_t(i,j) = 1$ indicates that the pixel at position (i,j) in the image is predicted to be a crack, whereas $\Pr_t(i,j) = 0$ indicates that the pixel at that position is predicted to be non-crack (background pixel).

2. Crack skeleton extraction and total length calculation.

Skeletonization converts the multi-pixel representation of a crack into a single-pixel-width skeleton to characterize its topological structure and approximate length. This is a necessary step for calculating crack length. Common algorithms include the medial axis algorithm, morphological thinning algorithm, 3D medial surface axis thinning algorithm, and numerical pattern thinning method. After extracting the crack skeleton, the total crack length can be calculated.

$$\mathcal{L}_c = \int_c G(x,y)d\ell \cong \sum G(x,y)d\ell \tag{3}$$

where $d\ell$ represents the length element of the crack skeleton. The skeleton length can be approximated by the length of the skeleton pixels, so the total crack length is roughly equivalent to the sum of the skeleton pixels' lengths. Assuming the skeleton image of the tth image is Skt. Since the skeleton image is also a binary image of {0,1}, the crack length of the tth image can be calculated using the following formula:

$$\mathcal{L}_{c,t} = \sum_{i=1}^{H}\sum_{j=1}^{W} Sk_t(i,j) \times \ell, \tag{4}$$

where ℓ represents the length or width of a pixel in the image.

3. Calculation of average crack width.

The average width of the crack can be calculated as follows:

$$\mathcal{L}_{c,t} = \sum_{i=1}^{H} \sum_{j=1}^{W} Sk_t(i,j) \times \ell, \tag{5}$$

where $\overline{\mathcal{W}}_{c,t}$ represents the average crack width in the *t*th image.

4. Mapping to real-world space.

The crack parameters obtained in steps (1) to (3) are in pixel units, but engineers are more concerned with the actual physical dimensions of the cracks to assess the structural condition. Based on the total length $\mathcal{L}_{pc,t}$ and average width $\bar{\mathcal{W}}pc,t$, the real-world crack dimensions can be calculated as follows:

$$\mathcal{L}_{pc,t} = \psi \times \mathcal{L}_{c,t}, \tag{6}$$

$$\bar{\mathcal{W}}pc,t = \frac{\mathcal{A}_{pc,t}}{\mathcal{L}_{pc,t}} = \psi \times \frac{\mathcal{A}_{c,t}}{\mathcal{L}_{c,t}}, \tag{7}$$

Model establishment and training

Data sources and data processing

The dataset utilized in this study comprises high-resolution images of asphalt pavement cracks. Specifically, it includes 913 images captured using a Sony ILCE-7R camera equipped with a Sony FE 24–70 mm F4 full-frame lens. The images were taken at a resolution of 7360 × 4144 pixels, ensuring that fine details of the cracks were preserved. Given the large dimensions of the images, each image was cropped to a standardized size of 416 × 416 pixels to create uniform feature maps suitable for training.

The dataset was divided into training and test sets, with 70% of the images (639 images) allocated to the training set and the remaining 30% (274 images) reserved for testing. This split ensures a robust evaluation of the model's performance on unseen data.

To enhance the model's robustness and prevent overfitting, several data augmentation techniques were applied to the training set. These included horizontal and vertical flipping, translation by shifting the images vertically by 1/4 of their length, and adjustments in color and contrast. The color and contrast transformations were particularly important given the variability in lighting conditions and road textures in the images. The augmentation process expanded the diversity of the training data, enabling the model to generalize better across different scenarios.

Blurry, incomplete, or poorly lit images were excluded from the dataset to maintain high-quality input data. After applying these preprocessing steps, the dataset provided a comprehensive foundation for training the deep learning models, as illustrated in Fig. 3.

Given the small sample size of the dataset, using these images alone for training would not be sufficient to evaluate the quality of a neural network model. To address this, several data augmentation techniques were applied to increase the dataset size and enhance the robustness of the model.

1. Image flipping transformations: Image flipping was performed in two directions: horizontal and vertical. Horizontal flips were applied around the y-axis, and vertical flips around the x-axis. Diagonal flips were not used because they can result in incomplete image displays and loss of information. To ensure a one-to-one correspondence between the road crack images and their semantic label images, both horizontal and vertical flips were applied simultaneously to the original and label images. Figure 4 shows the results of these transformations.

2. Image translation transformation: Image translation involves shifting all pixels in an image horizontally or vertically by a specified amount, followed by cropping and stitching. This process was performed vertically, moving the image position by 1/4 of its length. Like image flipping, translation was applied equally to the original and label images to maintain consistency during model training. Figure 5 illustrates the results of the translation transformation.

3. Image color and contrast transformations: Due to variations in the time and environment of image collection, the background colors of the collected images differ, affecting the edge information of the road cracks. To address this and prevent overfitting, the color and contrast of the images were adjusted. While label images, which contain only black and white, were not altered, the original images were adjusted for color and contrast. Two primary methods were employed: contrast adjustment and color adjustment, which involved modifying the hue, brightness, and saturation. The color interval was adjusted from [0,1] to [0.3, 0.5], as shown in Fig. 6.

Model establishment and training

Based on the image analysis in section "Data sources and data processing", we developed an attentional fully convolutional network (FCN) specifically for pixel-level crack detection in asphalt images. The model employs an encoder-attention-decoder architecture optimized for end-to-end learning. The encoder, built on a ResNet-50 backbone, captures hierarchical visual features through four stages of residual blocks, each with 3 × 3 convolutions and skip connections. The initial stage includes a 7 × 7 convolution with 64 channels followed by max pooling, while subsequent stages progressively apply filters of 64, 128, 256, and 512 channels. The final stage generates 256-channel activation maps at 1/16th of the original input size. The model's process flow is illustrated in Fig. 7.

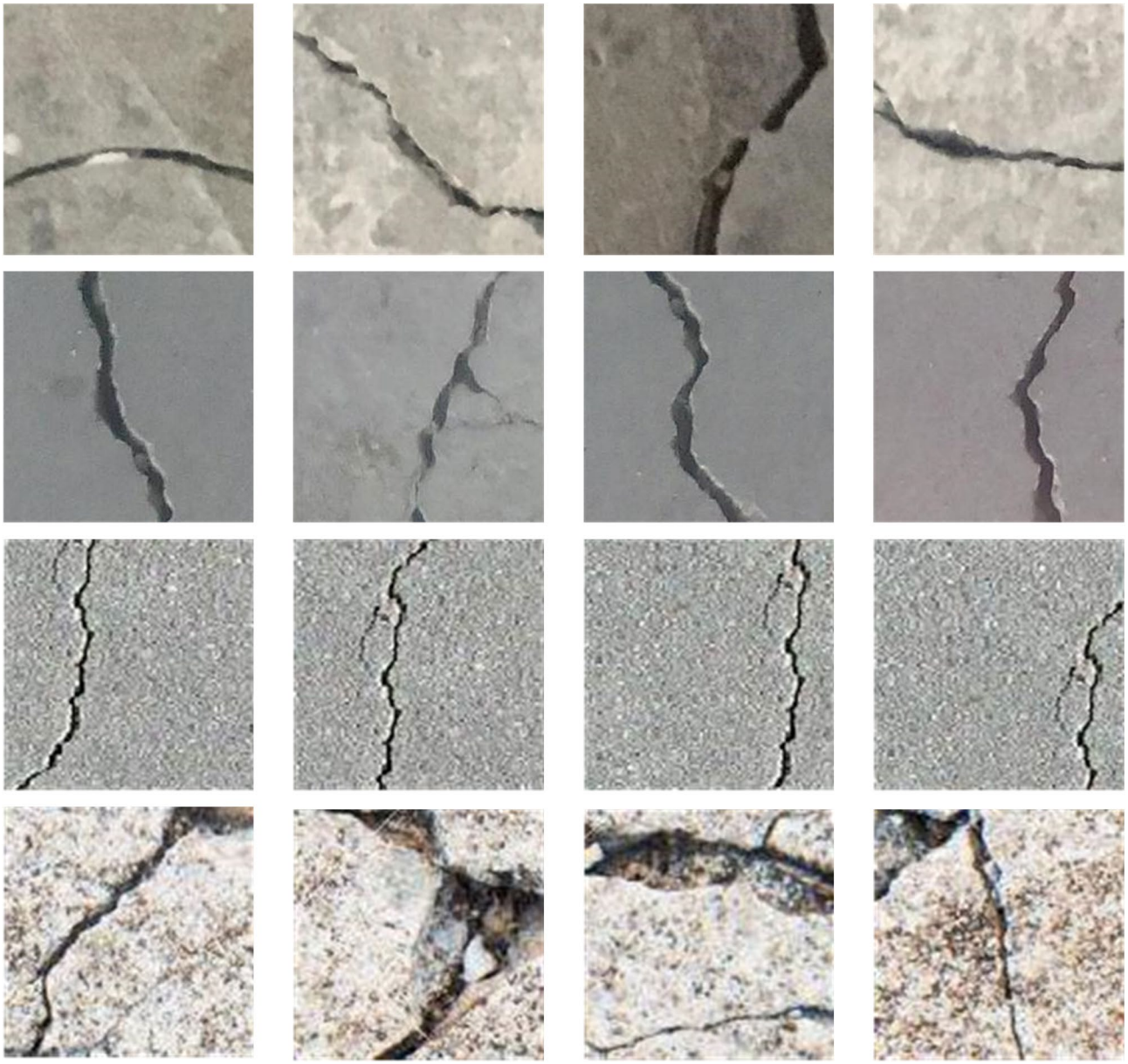


Fig. 3. Sample data of cracks in asphalt pavement.

To densify features, atrous convolution with rate $r=2$ is applied to stages 4 and 5, setting their stride to 1. By inserting holes in kernels, atrous convolution allows enlarging receptive fields without increasing parameters. Thus a larger context is integrated while preserving resolution, beneficial for pixel-accurate crack detection. The encoder passes activation tensors $F \in R^{H \times W \times C}$ to the attention module, which performs squeeze operations across spatial and channel dimensions. Formally, the channel squeeze $F_{sq\_c} \in R^{1 \times 1 \times C}$ has activations:

$$F_{\mathrm{sac}}(z) = \sum i \sum_{di} F(i, i, z). \tag{8}$$

Analogously, the spatial squeeze $F_{sq\_c} \in R^{H \times W \times 1}$ sums channel wise. The squeezes capture global distribution statistics, passed to respective excitation blocks. The excitation units learn additive activations βc, βs through bottleneck convolution and sigmoid layers. Elementwise multiplication of βc and βs with input Fi then incentivizes informative features while suppressing less useful ones across channels and spatial regions. The final output Fi is thus an attention-recalibrated encoder representation for decoding.

The decoder upsamples the encoder's features to the original input resolution using transposed convolutions and convolutional layers. Skip connections from the encoder are employed to merge low-level details with high-level semantic information. Batch normalization is used to stabilize the training process. Finally, a 1×1 convolution layer with sigmoid activation is applied to achieve per-pixel binary crack classification. Our model

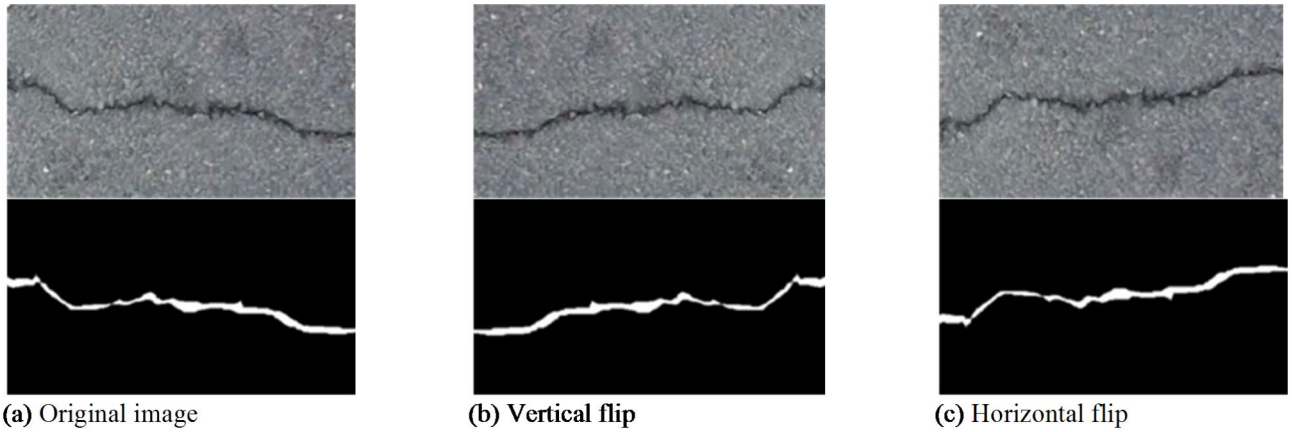


Fig. 4. Image of road cracks and corresponding label after flipping transformation.

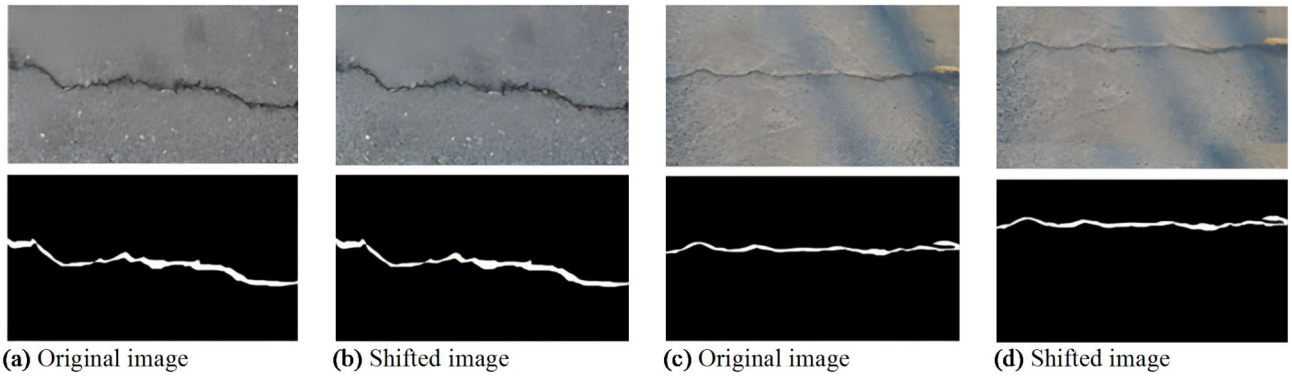


Fig. 5. Image of road cracks and corresponding label after translation transformation.

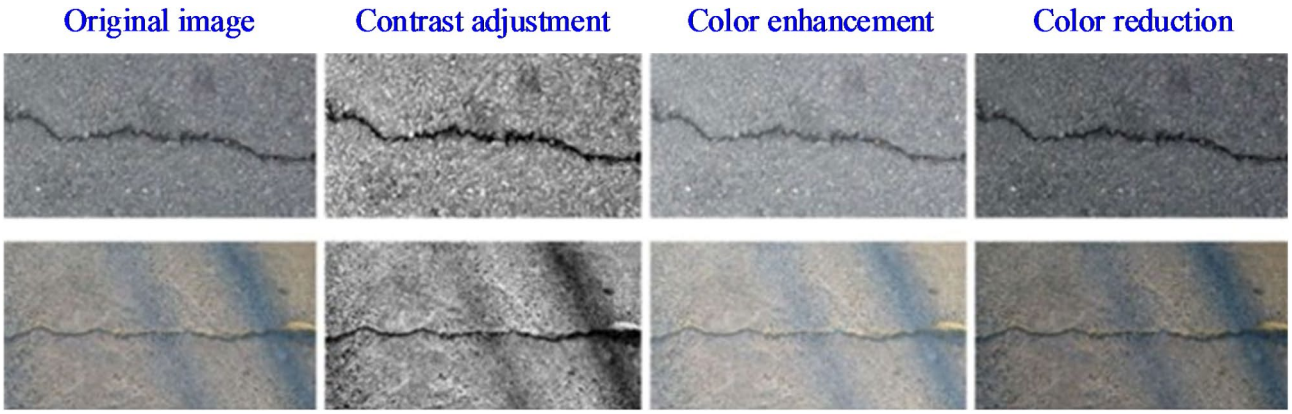


Fig. 6. Image of the original road crack with color and contrast transformations.

was implemented using PyTorch 1.7.1 with CUDA 11.3 for acceleration on an NVIDIA RTX 2060 GPU. The hardware setup includes an Intel i5-12500 CPU and 16GB of RAM running on a 64-bit Windows 11 operating system. The Adam optimizer, with an initial learning rate of 1e-4, was selected based on preliminary experiments to balance convergence speed and model stability. Additionally, the Amsgrad variant of the Adam optimizer was employed to improve convergence properties and mitigate excessive oscillations during training.

The Dice coefficient loss function was chosen for the model, which is particularly effective for segmentation tasks involving class imbalance. This loss function emphasizes the overlap between the predicted and actual segmentation masks, ensuring precise delineation of cracks, even when they constitute a minor portion of the image. The model was trained for a total of 50 epochs, a duration established through cross-validation and an

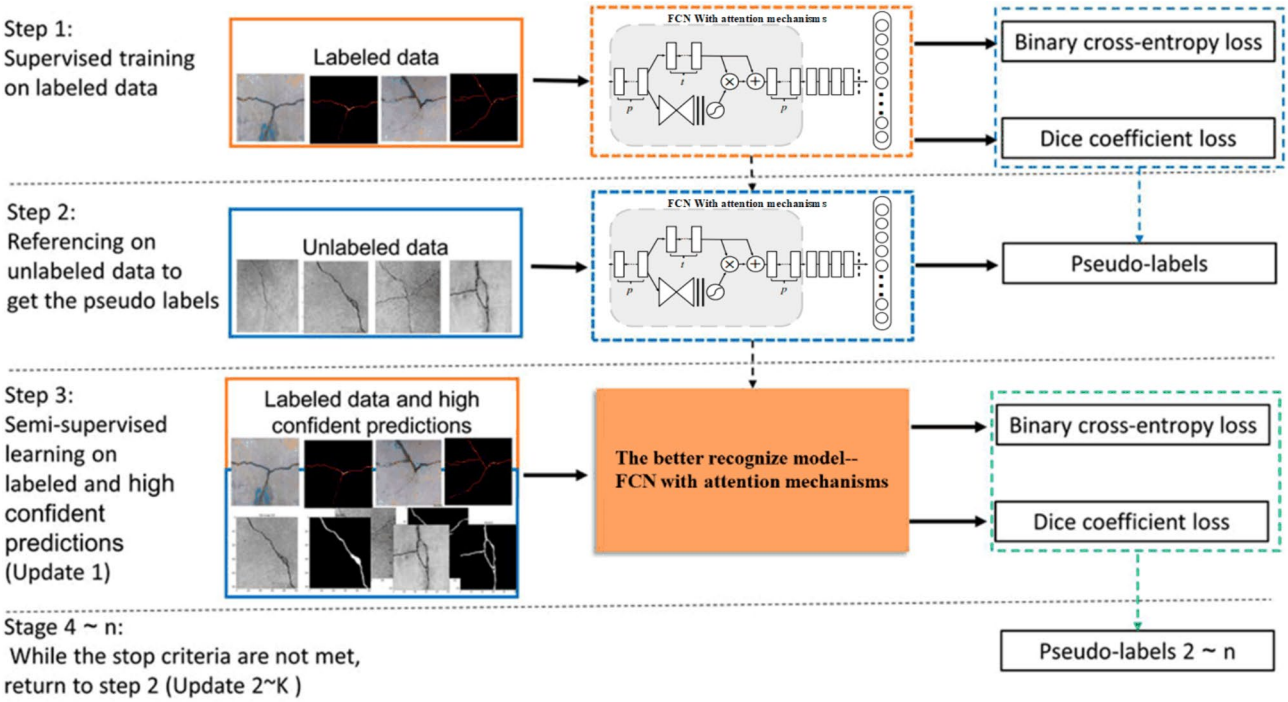


Fig. 7. The running process of the model.

early stopping strategy. Early stopping was implemented to monitor the model's performance on the validation set and prevent overfitting, with training halted if the validation loss failed to improve for 10 consecutive epochs. Additionally, a batch size of 16 was selected, providing an optimal balance between computational efficiency and model performance.

Analysis of test results

Result evaluation indicators

On the test data set, select the pixel accuracy Acc (Pixel Accuracy) and the average intersection-over-union ratio (mIoU) as the performance indicators of the model[32,33]. The pixel accuracy Acc formula is shown in Eq. (9):

$$Acc = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}}, \tag{9}$$

Where, Acc is equal to the number of correctly predicted pixels divided by the total number of predicted pixels.

The average intersection ratio is shown in Eq. (10):

$$M\_IOU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} P_{ji} - p_{ii}}, \tag{10}$$

This formula is equivalent to Eq. (11):

$$M\_IOU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN + FP + TP}, \tag{11}$$

Where, TP means that the real situation and the predicted result are both 1; FN means that the real situation is 1, and the predicted result is 0; FP means that the real situation is 0, and the predicted result is 1; TN means that the real situation and the predicted result are both 0.

Pre: This metric indicates the percentage of correctly predicted crack pixels out of all pixels predicted as cracks, calculated as

$$\mathrm{Pr}\, e = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{12}$$

Rec: This metric measures the percentage of actual crack pixels correctly identified by the model, calculated as

$$\mathrm{Re}\,c = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{13}$$

F1-score: The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both. It is calculated as

$$\text{F1-score} = \frac{2 \times \mathrm{Pre} \times \mathrm{Rec}}{\mathrm{Pre} + \mathrm{Rec}}, \tag{14}$$

Use the trained model to predict the crack image, and a mask image will be generated after prediction. Traverse the mask image pixel by pixel, identify each pixel corresponding to the crack, and use the formula to obtain the identification of the crack.

$$Moraine\_pre = \frac{M\_px}{M\_px + O\_px}, \tag{15}$$

Where, *Moraine_pre* is the identification of cracks obtained, *M_px* is the number of pixels corresponding to cracks in the mask image, and *O_px* is the number of pixels corresponding to the asphalt pavement in the mask image.

Result analysis

In this study, five different deep learning models were evaluated: VGG-16, AlexNet, MobileNet, FCN-attention, and GoogleNet. The same datasets and evaluation metrics were applied consistently across all models during both the training and testing phases. As shown in Table 1, the performance comparison focused on four key indicators: model training loss rate, global accuracy rate, average intersection-over-union (IoU) ratio, and test duration. Each model was tested on the same image 10 times, with the average time taken recorded as the test duration. The results indicate that AlexNet required the most time to predict a single image, while the FCN-attention and GoogleNet models had similar, shorter prediction times of 109.10 ms and 108.10 ms, respectively. Overall, the variation in test duration among the five models was minimal.

Based on the experimental results presented in Table 1, the FCN-attention model clearly outperforms the other models across multiple evaluation metrics. It achieves the lowest loss rate of 9.99% and the highest global accuracy rate of 90.792%, indicating its superior ability to accurately detect and segment cracks in asphalt pavement images. Additionally, the FCN-attention model excels in precision, recall, and F1-score, with values of 92.3%, 89.5%, and 90.9%, respectively, highlighting its robustness in minimizing false positives and false negatives. The model also maintains a competitive test duration of 109.10 milliseconds, demonstrating its efficiency in real-time applications. In contrast, models like VGG-16 and GoogleNet, while showing moderate performance, lag behind in both accuracy and speed. AlexNet and MobileNet also perform reasonably well, but their higher loss rates and lower precision and recall values compared to FCN-attention underscore the latter's superiority. Overall, the FCN-attention model is the most reliable and efficient among the tested models, making it highly suitable for asphalt pavement crack detection tasks.

As shown in Fig. 8a, the loss graph illustrates how the error of the models decreases over time as they learn from the training data. The FCN-attention model (depicted in red) begins with a high loss, similar to the other models, but its loss decreases sharply and remains consistently lower as training progresses. This suggests that the FCN-attention model is learning effectively from the data. By the end of training, it exhibits the lowest loss, indicating a better fit to the training data compared to models like AlexNet, VGG16, MobileNet, and GoogleNet.

Similarly, as shown in Fig. 8b, the accuracy graph demonstrates the percentage of correct predictions made by the models throughout the training epochs. Like the loss graph, the FCN-attention model quickly achieves a high level of accuracy and maintains it as training continues, outperforming the other models. The FCN-attention model appears to plateau around 90% accuracy, which is particularly strong for a complex task. While other models also improve in accuracy over time, none match the performance of the FCN-attention model, with GoogleNet being the closest competitor.

In conclusion, the FCN-attention model outperforms other CNN architectures in this task, as evidenced by the provided graphs. It achieves lower loss and higher accuracy, making it a more suitable model for asphalt pavement crack detection. The attention mechanism's ability to focus the model's learning on important features likely contributes to this improved performance.

Model	Loss rate (%)	Global accuracy rate (%)	Average intersection and union ratio (%)	Pre (%)	Rec (%)	F1-score (%)	Test duration (ms)
VGG-16	38.49	77.864	52.9	75.4	68.2	71.6	129.00
AlexNet	28.86	86.971	59.6	80.5	74.9	77.6	133.00
FCN-attention	9.99	90.792	69.7	92.3	89.5	90.9	109.10
Mobile	27.65	85.351	55.3	78.9	70.6	74.5	118.20
GoogleNet	32.61	85.652	54.3	81.0	72.3	76.4	108.10

Table 1. Four model segmentation experimental results.

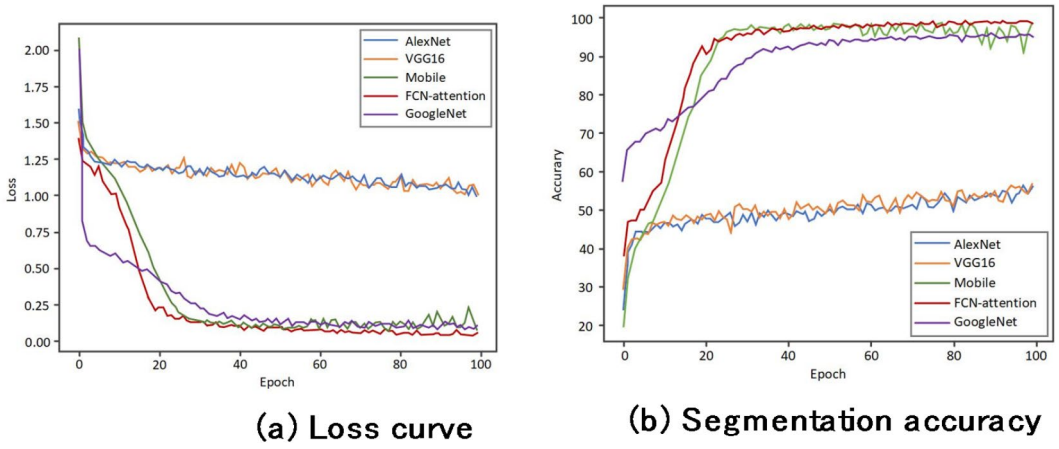


Fig. 8. Learning curve of each model.

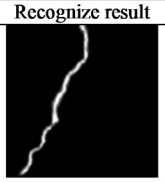
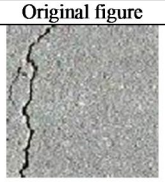
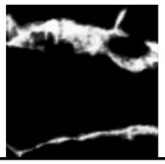
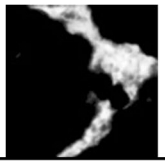
Level	Original figure	Recognize result	Original figure	Recognize result
Easy				
Middle				
Complex				

Table 2. Asphalt pavement crack identification results.

Based on the above analysis, this study uses the FCN-attention model to identify cracks in asphalt pavement. In order to better demonstrate the performance of the model in the recognition process, this study classified the test images according to the complexity of the cracks. The recognition results are shown in Table 2. It can be seen that when using the FCN-attention model to identify simple cracks, very fine cracks can be identified, and the recognition effect is very good. At the same time, when identifying cracks of medium complexity, the details of the cracks can also be easily identified. When identifying relatively complex cracks, although the model cannot completely reflect the entire crack area, it can still identify most areas and the direction of the crack, and the identification results are still satisfactory.

Testing results on the crack dataset

In this section, we evaluate the performance of various models on the Crack dataset, using the same metrics as in previous tests: loss rate, global accuracy rate, average intersection-over-union (IoU) ratio, and test duration. These comparisons aim to assess the effectiveness of each model in handling task-specific datasets.

As summarized in Table 3, the FCN-attention model consistently outperforms the other models. It achieves the lowest loss rate at 10.18% and the highest global accuracy rate at 91.792%, demonstrating its strong capability in accurately recognizing and segmenting cracks. The model also achieves the highest average IoU ratio at 68.9%, reflecting its precise delineation of crack regions compared to the background. Additionally, the FCN-attention model maintains a competitive test duration, making it well-suited for real-time applications. While AlexNet shows minor improvements in loss and accuracy, it still falls short of FCN-attention's performance. VGG-16, Mobile, and GoogleNet exhibit moderate results without significant changes in their metrics.

Model	Loss rate (%)	Global accuracy rate (%)	Average intersection and union ratio (%)	Test duration (ms)
VGG-16	39.49	78.864	53.1	139.12
AlexNet	27.86	84.971	57.4	143.82
FCN-attention	10.18	91.792	68.9	112.47
Mobile	37.65	83.351	51.5	120.98
GoogleNet	36.71	84.652	52.7	111.18

Table 3. Experimental results on the crack dataset.

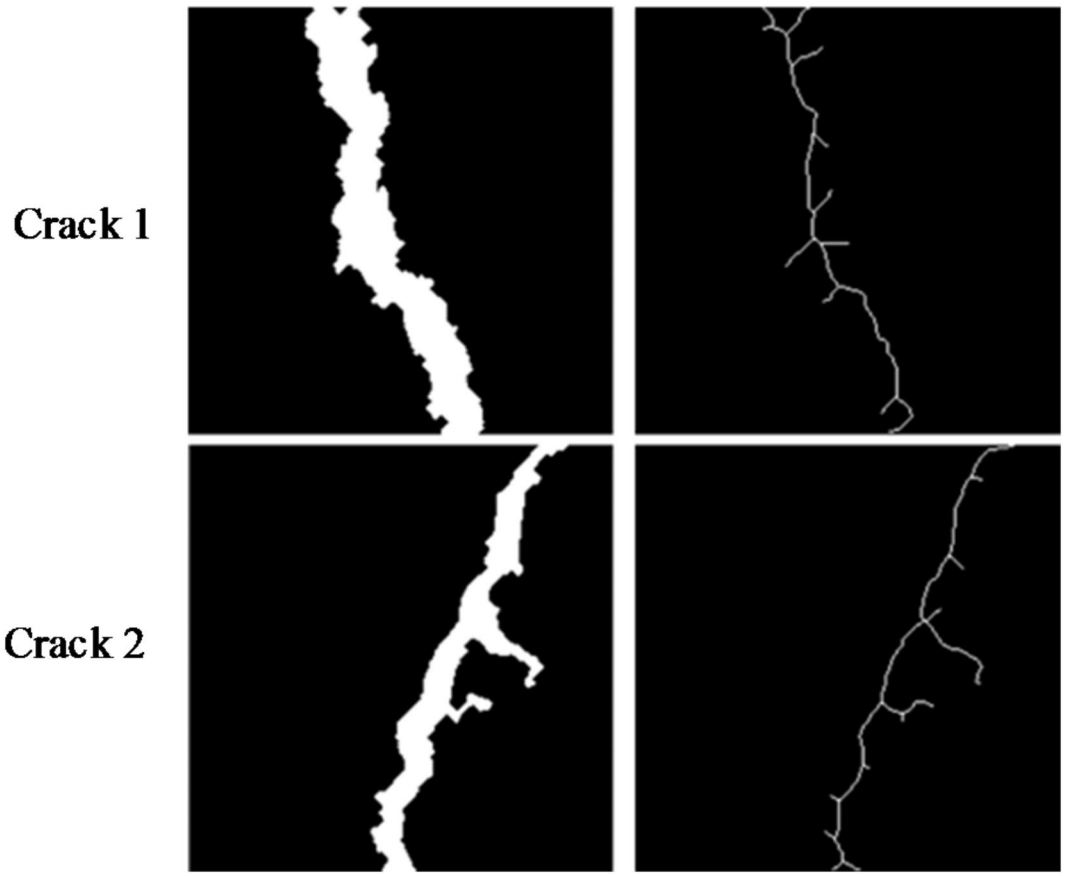


Fig. 9. Part of the crack mask skeleton extraction effect display.

Analysis of crack calculation results

This paper selects length and average width parameters to quantify cracks. The crack mask images detected by the FCN-attention model are binarized, and the skeleton images are generated according to the method in section "Crack calculation method". Some results are shown in Fig. 9.

The detection results indicate that the skeleton extraction algorithm preserves the shape and connectivity features of the cracks, effectively describing the crack characteristics. For example, at a resolution of 100 dpi, there are 100 pixels per inch, and 1 cm = 0.3937008 inches, which converts to 1 cm = 39.37008 pixels. Using this conversion, the actual parameters of the cracks can be calculated as described in section "Crack calculation method". The resolution of the crack images used in this paper is 96 dpi. According to this conversion rule, Crack 1 has a length of 5.23 dm, an average width of 0.11 dm, and a maximum width of 0.21 dm. Crack 2 has a length of 5.17 dm, an average width of 0.09 m, and a maximum width of 0.18 dm.

Ablation study

To further validate the effectiveness of the attention mechanism integrated into our fully convolutional network (FCN) model, an ablation study was conducted. The purpose of this study was to isolate the impact of the attention mechanism by comparing the performance of the FCN model with and without this component. We trained two versions of the model: one with the attention mechanism (FCN-attention) and one without (standard FCN). Both models were trained under identical conditions using the same dataset, optimizer settings, and loss

functions. The evaluation metrics used for comparison included the global accuracy rate, average intersection-over-union (IoU), precision, recall, F1-score, and test duration, as shown in Table 4.

As summarized in Table 4, the inclusion of the attention mechanism significantly improved the model's performance across all evaluation metrics. The FCN-attention model achieved a global accuracy rate of 91.792%, while the standard FCN model achieved 88.345%. Similarly, the average IoU improved from 64.3% in the standard FCN to 68.9% in the FCN-attention model. Precision, recall, and F1-score also showed marked improvements, with the FCN-attention model achieving a precision of 92.3%, recall of 89.5%, and an F1-score of 90.9%, compared to the standard FCN's precision of 88.1%, recall of 85.0%, and F1-score of 86.5%.

The test duration for the FCN-attention model was slightly longer, averaging 112.47 milliseconds per image, compared to 98.54 milliseconds for the standard FCN. However, this increase in processing time is justified by the significant gains in accuracy and precision.

These results clearly indicate that the attention mechanism plays a crucial role in enhancing the model's ability to focus on relevant features, leading to more accurate and robust crack detection. The improvements in IoU, precision, recall, and F1-score demonstrate the effectiveness of the attention mechanism in refining the feature maps generated by the encoder and enabling the decoder to produce more accurate segmentation masks.

Discussion

The findings of this study have significant practical implications for road maintenance and safety. The proposed FCN-attention model demonstrates high accuracy and efficiency in detecting asphalt pavement cracks, which can greatly enhance road maintenance programs. By enabling precise and timely crack detection, this model helps identify potential issues before they escalate into major road damage, thereby reducing maintenance costs and prolonging pavement lifespan. Additionally, automating crack detection alleviates the labor-intensive nature of traditional inspection methods, allowing for more frequent and widespread monitoring of road conditions.

Practical benefits

The FCN-attention model provides several practical advantages that directly impact road maintenance:

(1) Enhanced Accuracy: The model's high precision and recall rates ensure accurate crack identification, minimizing false positives and negatives. This leads to more reliable pavement assessments, enabling maintenance teams to allocate resources efficiently.
(2) Real-Time Monitoring: With its efficient image processing capabilities, the model can be integrated into real-time monitoring systems, continuously updating road surface conditions. This is particularly beneficial in high-traffic areas where road safety is critical.
(3) Cost Efficiency: Automating the crack detection process reduces the reliance on manual inspections, lowering labor costs and accelerating the assessment process. This allows for broader monitoring coverage without increasing operational expenses.

Challenges and future directions

Despite its promising performance, the FCN-attention model has certain limitations and areas for improvement that should be addressed in future research.

While the FCN-attention model achieves superior crack detection, its performance in complex environments still presents challenges. Specifically, the model struggles with images containing heavy texture, overlapping objects, or pavement markings, where the attention mechanism may incorrectly focus on irrelevant features. For example, cracks that intersect with road markings or debris are sometimes misclassified, either as non-cracks or as extraneous objects. Figure 10 illustrates cases where the model encounters such misclassification. To mitigate these issues, potential improvements include:

(1) Enhanced Data Augmentation—Expanding the dataset with synthetic noise, varying lighting conditions, and complex background textures can improve generalization.
(2) Multi-Modal Fusion Approaches—Integrating additional sensor data, such as LiDAR or infrared imaging, may help distinguish cracks from environmental noise.
(3) Refined Attention Mechanisms—Implementing adaptive attention modules that dynamically adjust focus areas based on contextual cues could further improve segmentation accuracy.

Another critical factor for real-world deployment is the model's ability to generalize across diverse environments. The dataset used in this study consists of 913 high-resolution images captured using a single camera setup under

Metric	FCN (without attention)	FCN-attention (with attention)
Global accuracy rate (%)	88.345	91.792
Average IoU (%)	64.3	68.9
Precision (%)	88.1	92.3
Recall (%)	85.0	89.5
F1-Score (%)	86.5	90.9
Test duration (ms per image)	98.54	112.47

Table 4. Ablation study results comparing FCN and FCN-Attention models.

Fig. 10. Examples of model misclassification in complex environments.

specific conditions. While the model performs well on this dataset, its robustness in different geographical regions, lighting conditions, or when using different imaging equipment needs further validation. Several challenges arise when applying the model to data from different sources:

(1) Variations in pavement texture, material composition, and environmental factors (e.g., wet surfaces, shadows, and debris) may introduce inconsistencies affecting segmentation accuracy.
(2) Lighting conditions significantly impact image quality, with extreme brightness or darkness potentially leading to misclassifications.
(3) Differences in camera resolution, focal length, and capture angles can alter the visual representation of cracks, influencing detection performance.

To enhance generalization, future research should focus on:

(1) Expanding the Dataset—Collecting images from multiple regions, road types, and weather conditions to improve model robustness.
(2) Domain Adaptation Techniques—Utilizing transfer learning or domain adaptation methods to fine-tune the model with additional, smaller datasets from different sources.
(3) Advanced Data Augmentation—Implementing techniques such as brightness adjustments, synthetic noise, and style transfer to simulate real-world variations.
(4) Multi-Sensor Integration—Combining visible spectrum images with infrared or LiDAR data for more reliable crack detection, reducing environmental variability effects.

Model complexity and efficiency

While the FCN-attention model demonstrates superior accuracy in crack detection, its computational complexity presents challenges for deployment on resource-constrained devices, such as embedded systems, drones, and mobile applications. The current test duration of 112.47 ms per image, though acceptable for high-performance GPUs, may be a limiting factor in real-time applications where lower latency is required. To improve efficiency, model pruning can remove redundant parameters, and quantization can reduce weight precision to lower memory usage and increase inference speed. Using a lightweight backbone such as MobileNet, EfficientNet, or ShuffleNet can reduce computational cost while maintaining accuracy. Knowledge distillation can train a smaller model to retain the performance of a larger one. Optimizing the model for TensorRT, OpenVINO, or TFLite can enhance inference speed, and using FPGA or EdgeTPU accelerators can further improve real-time performance. Slightly reducing input resolution with adaptive upsampling can balance computational efficiency and detection accuracy. Future research can focus on these aspects to optimize the model.

Real-world application

The FCN-attention model can be deployed in real-world road maintenance by integrating it with unmanned aerial vehicles (UAVs), vehicle-mounted cameras, or mobile applications. UAV-based deployment enables large-scale road inspections with minimal human intervention, improving efficiency and reducing labor costs. Vehicle-mounted systems can perform continuous real-time monitoring on highways, detecting cracks during regular patrols and transmitting data to maintenance teams for timely repairs. Mobile applications allow field inspectors to capture images with smartphones and obtain instant crack detection results, enhancing on-site decision-making. Optimizing the model for embedded AI chips can enable real-time inference on edge devices, reducing dependence on cloud computing and improving response speed. These deployment strategies enhance crack detection automation, making road maintenance more efficient and cost-effective.

Conclusion

This study proposes a fully convolutional network (FCN) enhanced with an attention mechanism for the automated detection and analysis of asphalt pavement cracks. The introduction of the attention mechanism significantly improves the model's ability to focus on relevant features, leading to more precise crack identification and segmentation, especially in complex environments. The results demonstrate that the FCN-attention model achieves superior performance compared to other models such as VGG-16, AlexNet, MobileNet, and GoogleNet. Specifically, the FCN-attention model achieves a global accuracy rate of 91.792% and an average intersection-over-union (IoU) of 68.9%, with a test duration of 112.47 ms. These results highlight its robustness and efficiency, making it highly suitable for real-time applications in road maintenance. The crack parameter

calculations, including length and average width, further validate the effectiveness of the proposed model. The crack skeleton extraction preserves the shape and connectivity features, enabling accurate quantification of cracks. For instance, the calculated lengths and widths of cracks in test images align well with expected values, showcasing the model's practical utility.

Data availability

The dataset can be found in the supporting files. https://github.com/liujiawei214/Crack-Data.git.

Received: 17 February 2025; Accepted: 4 March 2025
Published online: 12 July 2025

References

1. Luo, X., Gu, F., Ling, M. & Lytton, R. L. Review of mechanistic-empirical modeling of top-down cracking in asphalt pavements. *Constr. Build. Mater.* **191**, 1053–1070. https://doi.org/10.1016/j.conbuildmat.2018.10.005 (2018).
2. Karlaftis, A. G. & Badr, A. Predicting asphalt pavement crack initiation following rehabilitation treatments. *Transp. Res. Part. C: Emerg. Technol.* **55**, 510–517. https://doi.org/10.1016/j.trc.2015.03.031 (2015).
3. Yin, H. M. Opening-Mode cracking in asphalt pavements: crack initiation and saturation. *Road. Mater. Pavement Des.* **11**, 435–457. https://doi.org/10.1080/14680629.2010.9690283 (2010).
4. 4. Recognition of asphalt pavement crack length using deep convolutional neural networks: road materials and pavement design: vol 19, No 6. (2017, accessed 4 Mar 2024). https://www.tandfonline.com/doi/abs/10.1080/14680629.2017.1308265.
5. Pixel-level cracking detection on 3D asphalt pavement images through deep-learning—based CrackNet-V|IEEE Journals & Magazine|IEEE Xplore. (2024, accessed on 4 Mar 2024). https://ieeexplore.ieee.org/abstract/document/8620557.
6. Que, Y. et al. Automatic classification of asphalt pavement cracks using a novel integrated generative adversarial networks and improved VGG model. *Eng. Struct.* **277**, 115406. https://doi.org/10.1016/j.engstruct.2022.115406 (2023).
7. Zakeri, H., Nejad, F. M. & Fahimifar, A. Image based techniques for crack detection, classification and quantification in asphalt pavement: A review. *Arch. Computat. Methods Eng.* **24**, 935–977. https://doi.org/10.1007/s11831-016-9194-z (2017).
8. Behnia, B., Buttlar, W. & Reis, H. Evaluation of low-temperature cracking performance of asphalt pavements using acoustic emission: a review. *Appl. Sci.* **8**, 306. https://doi.org/10.3390/app8020306 (2018).
9. Liu, Z. et al. Automatic Pixel-Level detection of vertical cracks in asphalt pavement based on GPR investigation and improved mask R-CNN. *Autom. Constr.* **146**, 104689. https://doi.org/10.1016/j.autcon.2022.104689 (2023).
10. Liu, F., Liu, J. & Wang, L. Deep learning and infrared thermography for asphalt pavement crack severity classification. *Autom. Constr.* **140**, 104383. https://doi.org/10.1016/j.autcon.2022.104383 (2022).
11. Ji, A., Xue, X., Wang, Y., Luo, X. & Xue, W. An integrated approach to automatic Pixel-Level crack detection and quantification of asphalt pavement. *Autom. Constr.* **114**, 103176. https://doi.org/10.1016/j.autcon.2020.103176 (2020).
12. Eskandari Torbaghan, M. et al. Automated detection of cracks in roads using ground penetrating radar. *J. Appl. Geophys.* **179**, 104118, https://doi.org/10.1016/j.jappgeo.2020.104118 (2020).
13. Fan, Z. et al. Automatic crack detection on road pavements using Encoder-Decoder architecture. *Mater. (Basel).* **13**, 2960. https://doi.org/10.3390/ma13132960 (2020).
14. Chun, P., Yamane, T. & Tsuzuki, Y. Automatic detection of cracks in asphalt pavement using deep learning to overcome weaknesses in images and GIS visualization. *Appl. Sci.* **11**, 892. https://doi.org/10.3390/app11030892 (2021).
15. Li, G. et al. Automatic recognition and analysis system of asphalt pavement cracks using interleaved Low-Rank group Convolution hybrid deep network and SegNet fusing dense condition random field. *Measurement* **170**, 108693. https://doi.org/10.1016/j.measurement.2020.108693 (2021).
16. Safaei, N., Smadi, O., Masoud, A. & Safaei, B. An automatic image processing algorithm based on crack pixel density for pavement crack detection and classification. *Int. J. Pavement Res. Technol.* **15**, 159–172. https://doi.org/10.1007/s42947-021-00006-4 (2022).
17. Hu, G. X., Hu, B. L., Yang, Z., Huang, L. & Li, P. Pavement crack detection method based on deep learning models. *Wirel. Commun. Mobile Comput.* **2021**, e5573590. https://doi.org/10.1155/2021/5573590 (2021).
18. Abbas, I. H. & Ismael, M. Q. Automated pavement distress detection using image processing techniques. *Eng. Technol. Appl. Sci. Res.* **11**, 7702–7708. https://doi.org/10.48084/etasr.4450 (2021).
19. Huyan, J., Ma, T., Li, W., Yang, H. & Xu, Z. Pixelwise asphalt concrete pavement crack detection via deep Learning-Based semantic segmentation method. *Struct. Control Health Monit.* **29**, e2974. https://doi.org/10.1002/stc.2974 (2022).
20. Islam, M. M. M. & Kim, J. M. Vision-based autonomous crack detection of concrete structures using a fully convolutional encoder-decoder network. *Sens. (Basel)* **19**, 4251. https://doi.org/10.3390/s19194251 (2019).
21. Chen, F. C., Jahanshahi, M. R. & ARF-Crack: rotation invariant deep fully convolutional network for pixel-level crack detection. *Mach. Vis. Appl.* **31**, 47. https://doi.org/10.1007/s00138-020-01098-x (2020).
22. Zheng, M., Lei, Z. & Zhang, K. Intelligent detection of building cracks based on deep learning. *Image Vis. Comput.* **103**, 103987. https://doi.org/10.1016/j.imavis.2020.103987 (2020).
23. Zhang, J. & Zhang, J. An improved nondestructive semantic segmentation method for concrete dam surface crack images with high resolution. *Math. Probl. Eng.* **2020**, e5054740. https://doi.org/10.1155/2020/5054740 (2020).
24. Li, S. & Zhao, X. Pixel-Level detection and measurement of concrete crack using faster Region-Based convolutional neural network and morphological feature extraction. *Meas. Sci. Technol.* **32**, 065010. https://doi.org/10.1088/1361-6501/abb274 (2021).
25. Deng, J. Image-based crack assessment of concrete structures using artificial intelligence for bridge health monitoring. thesis, Monash University (2021).
26. An, Q. et al. Segmentation of concrete cracks by using fractal dimension and UHK-Net. *Fractal Fract.* **6**, 95. https://doi.org/10.3390/fractalfract6020095 (2022).
27. Li, S. & Zhao, X. A. Performance improvement strategy for concrete damage detection using stacking ensemble learning of multiple semantic segmentation networks. *Sens. (Basel).* **22**, 3341. https://doi.org/10.3390/s22093341 (2022).
28. Chaiyasarn, K. et al. Integrated Pixel-Level CNN-FCN crack detection via photogrammetric 3D texture mapping of concrete structures. *Autom. Constr.* **140**, 104388. https://doi.org/10.1016/j.autcon.2022.104388 (2022).
29. Qiao, W., Liu, Q., Wu, X., Ma, B. & Li, G. Automatic Pixel-Level pavement crack recognition using a deep feature aggregation segmentation network with a ScSE. *Atten. Mechanism Module Sens.* **21**, 2902. https://doi.org/10.3390/s21092902 (2021).
30. Niu, Z., Zhong, G. & Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **452**, 48–62. https://doi.org/10.1016/j.neucom.2021.03.091 (2021).
31. Chen, Q., Xu, J. & Koltun, V. *Fast Image Processing With Fully-Convolutional Networks* 2497–2506 (2017).
32. Qiu, Z. et al. Variety identification of single rice seed using hyperspectral imaging combined with convolutional neural network. *Appl. Sci.* **8**, 212. https://doi.org/10.3390/app8020212 (2018).
33. Nam, S., Park, H., Seo, C. & Choi, D. Forged signature distinction using convolutional neural network for feature extraction. *Appl. Sci.* **8**, 153. https://doi.org/10.3390/app8020153 (2018).

Acknowledgements
This research was funded by the Natural Science Foundation of Henan Province, grant number 232300421408. This research was also funded by the Science and Technology Project of Henan Province, grant number 222102320260.

Author contributions
Conceptualization, Zhang Huiyuan and Liu Jiawei.; methodology, Liu Jiawei ; writing—original draft preparation, Hu Guoping and Liu Jiawei. All authors have read and agreed to the published version of the manuscript.

Funding
This research was funded by the Natural Science Foundation of Henan Province, grant number 232300421408. This research was also funded by the Science and Technology Project of Henan Province, grant number 222102320260.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-92971-0.

Correspondence and requests for materials should be addressed to J.L.

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

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

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

© The Author(s) 2025