Automatic Crack Detection Based On Attention U-Net

Zhongyang Rao, *Member, IAENG,* Chunyuan Feng

***Abstract*—Crack identification is critical for structural health monitoring and damage assessment in concrete structures. While numerous automated inspection methods have been developed to replace manual approaches, most exhibit limitations in adapting to diverse environmental conditions and achieving precise crack localization. In this paper, an end-to-end semantic segmentation neural network based on Attention U-Net. To mitigate the challenge of limited annotated data, sophisticated data augmentation techniques were employed to prevent overfitting. The proposed architecture maintains the original input dimensions while performing pixel-level classification (crack vs. non-crack) with high precision. Comparative experimental results demonstrate that the Attention U-Net model significantly outperforms conventional U-Net approaches across various complex scenarios, eliminating the need for manual feature extraction.**

***Index Terms*—Crack Detection, Machine Learning (ML), Attention U-Net, Semantic Segmentation**

I. Introduction

Highway infrastructure is subject to significant environmental stressors and heavy traffic loads, leading to premature deterioration and reduced service life compared to international standards. The assessment and maintenance of pavement surface distress, particularly crack detection, constitute critical components of infrastructure management. Conventional manual inspection methods present substantial limitations, including labor intensiveness and suboptimal detection accuracy. Systematic crack monitoring is indispensable for structural health evaluation, as crack morphology and spatial distribution provide crucial diagnostic information regarding material degradation mechanisms and potential failure modes. Fracture characteristics, including dimensional parameters and propagation patterns, serve as essential indicators for structural condition assessment. However, the inherent limitations of the artificial crack characterization method lie in its time-consuming and subjective interpretation, which heavily depends on the inspector's expertise and may damage the reliability of quantitative analysis. To address these limitations, this study proposes the implementation of computer vision-based automated crack detection systems as a viable alternative, using advanced imaging technologies for enhanced accuracy and efficiency.

The structure of the article is as follows: Section II provides a comprehensive literature review of relevant research in this field. Section III details the proposed attention-based UNET architecture and its implementation. Section IV systematically analyzes the experimental results and their significance. Finally, the paper concludes with Section V, which outlines future research avenues and practical applications derived from the study's findings.

II. Related Work

Conventional manual crack inspection methods are known to suffer from multiple drawbacks, such as being time-consuming, labor-intensive, posing safety risks, and yielding subjective evaluation results.[1]. Consequently, automated crack detection systems are increasingly supplanting traditional manual methods due to their superior efficiency, consistency, and rapid analytical capabilities in smart transportation infrastructure applications[2].

Crack detection methodologies can be categorized into two primary approaches: Destructive Testing (DT) and Non-Destructive Testing (NDT). Although ultrasonic techniques are widely utilized in NDT, conventional contact-based ultrasound methods demonstrate limited effectiveness across diverse structural configurations. Automated crack detection systems have emerged to overcome the inefficiencies of manual inspection methods, providing objective and efficient surface defect evaluation [3]. Over the past decade, researchers have proposed numerous computational algorithms for automated crack identification in various infrastructure elements, including concrete surfaces and pavement systems. However, most existing methodologies are constrained to addressing specific crack detection challenges. For instance, threshold-based techniques utilize local and global intensity thresholds to identify cracks through image illumination normalization, while segmentation-based approaches employ edge detection and region-based methods to partition images into discrete segments for crack localization based on predefined morphological features. The accuracy and reliability of these methods are fundamentally dependent on the precise characterization of crack features. Furthermore, while conventional approaches primarily focus on crack visualization, the accurate quantification of crack morphological properties for subsequent structural analysis remains a significant challenge in the field[4].

Deep learning has revolutionized computer vision by enabling deep neural networks (DNNs) to achieve state-of-the-art results in multiple visual processing domains such as image categorization, object identification, and

Manuscript received January 22, 2025; revised August 9, 2025.

This research was partly supported by the National Natural Science Foundation of China (61101225, 61601264) and the Doctoral Research Startup Foundation of Shandong Jiaotong University.

Zhongyang Rao is a professor of Shandong Key Laboratory of Technologies and Systems for Intelligent Construction Equipment, Shandong Jiaotong University, Jinan, Shandong Province, China (Corresponding author phone: 086-0531-6720; fax: 086-0531-6720; e-mail: raozhongyang@sdjtu.edu.cn).

Chunyuan Feng is an associate professor of Shandong Jiaotong University, Jinan, Shandong Province, China (email: fengchunyuan@sdjtu.edu.cn).

pattern analysis. While DNN architectures exhibit greater complexity through multiple hierarchical layers and extensive parameters compared to traditional methods, they substantially improve detection accuracy by enabling pixel-level analysis rather than conventional image-level interpretation. This sophisticated analytical capability allows for exact identification of target object pixels, establishing a reliable framework for precise crack detection and quantitative assessment. Consequently, the application of deep neural networks supports detailed fracture characterization at the microscopic level, successfully overcoming the challenges associated with diverse crack patterns and morphologies[5, 6].

Contemporary image processing techniques offer diverse approaches for automated fracture detection and characterization, as documented in reference[7]. The fundamental framework for crack identification through digital image analysis is comprehensively described in reference[8]. However, automated crack detection presents significant technical challenges due to several factors: the inherent variability in crack morphology and dimensional characteristics, coupled with various imaging artifacts. These artifacts include illumination inconsistencies, shadow effects, surface imperfections, and concrete spalling phenomena, all of which can substantially affect detection accuracy in images.

Modern visual inspection systems utilize diverse computational approaches, mainly including six key methodologies: gradient-based edge detection, morphological operations, intensity thresholding, porous media modeling, machine learning-based decision systems, and advanced algorithmic solutions[9, 10].

Edge detection algorithms exhibit robust performance in high-contrast image scenarios but demonstrate significant vulnerability to noise interference, frequently producing fragmented crack patterns[11]. These computational approaches encompass various transform-based methods, including the Haar Transform (HT), Fast Fourier Transform (FFT), as well as gradient-based operators such as Sobel and Canny edge detectors. Morphological processing techniques have been effectively implemented for road surface image analysis[12]. Threshold-based segmentation methods provide effective mechanisms for crack isolation from background elements[13] , with advanced implementations incorporating Fuzzy C-Means (FCM) clustering for adaptive threshold determination[14].

The advancement of machine learning technologies has facilitated the development of numerous sophisticated methodologies for crack detection, particularly emphasizing automated feature extraction and pattern recognition capabilities.

Recent advancements in machine learning-based crack detection have yielded various methodological approaches. Oliveira et al. developed CrackIT, an unsupervised learning framework utilizing standard deviation analysis to differentiate between cracked and intact image blocks[15]. Cord et al. implemented an AdaBoost algorithm for optimal selection of textural descriptors in crack image characterization[16]. Convolutional Neural Networks (CNNs) have emerged as a predominant architecture for concrete crack identification[17-21], with Cha et al. introducing a sliding window-based partitioning technique that segments images into discrete regions for CNN-based crack presence classification[21]. Zhang et al. extended CNN applications to pixel-level analysis, initially developing a framework for single-pixel classification based on local contextual information[22], followed by a comprehensive pixel-wise classification system[23]. However, these approaches exhibit limitations in capturing spatial relationships between pixels and tend to overestimate crack dimensions. Additionally, the requirement for manual feature engineering and the network's dependency on input image dimensions constrain the method's generalizability. In a distinct application domain, Elhariri et al. successfully adapted U-Net architecture for crack detection in historical surface preservation[24].

III. Attention U-Net Model

This section provides a concise overview of the

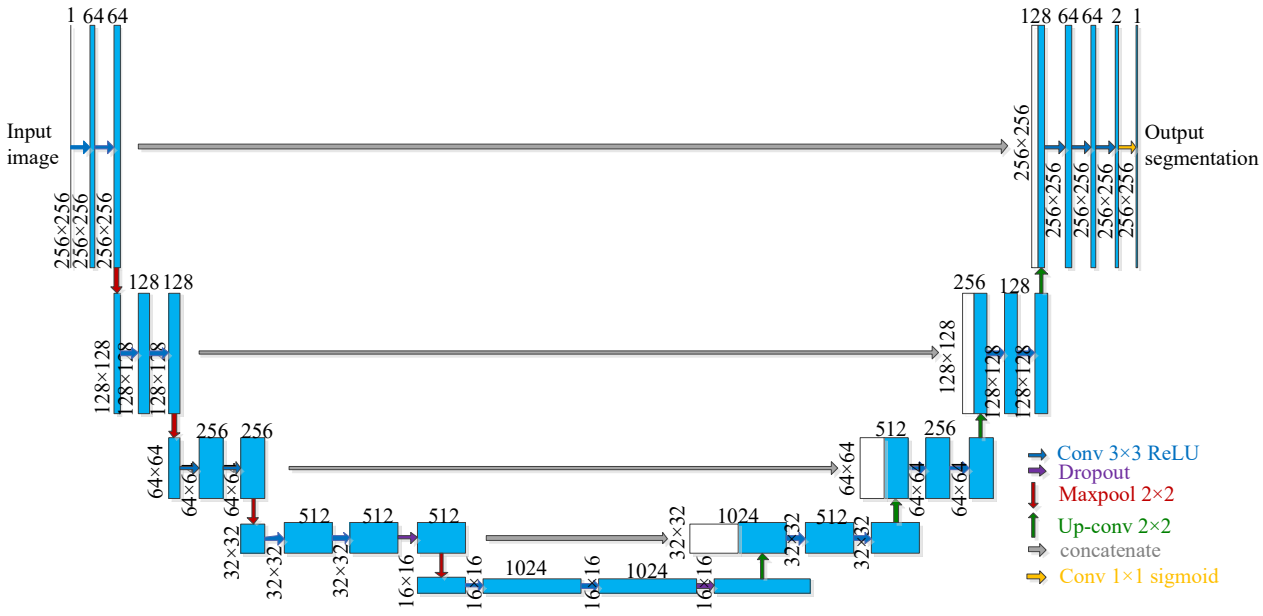

Fig. 1. Block diagram of U-Net segmentation model

attention-based U-Net architecture and its implementation in the proposed methodology.

A. Block Diagram of U-Net Model

In 2015, Ronneberger et al. proposed a U-Net architecture, a specialized convolutional neural network (CNN) framework designed for image segmentation tasks, particularly in medical imaging applications[25]. This network architecture extends the fundamental principles of fully convolutional networks[26], leveraging their inherent capability to generate hierarchical feature representations. Experimental results demonstrate that when trained end-to-end using pixel-level annotations, this architecture achieves superior performance in semantic segmentation tasks compared to previous state-of-the-art methods. The block diagram of the U-Net model is presented in Fig.1.

The U-Net architecture uses a symmetrical encoder-decoder structure characterized by its U-shaped topology, optimized for biomedical image segmentation tasks. The encoder (contracting path) hierarchically extracts semantic features through successive processing blocks, each comprising dual 3×3 convolutional layers with ReLU activation, followed by 2×2 max-pooling (stride=2) for progressive spatial dimensionality reduction. This down-sampling mechanism systematically halves feature map resolutions while doubling channel depth at each stage, capturing high-level semantic representations.

Conversely, the decoder (expansive path) reconstructs segmentation masks through 2×2 transposed convolutions (stride=2), restoring spatial resolution while halving channel depth. Up-sampled features are concatenated with skip-connected encoder outputs via channel-wise fusion, enabling precise boundary recovery by integrating multi-scale contextual information. Subsequent 3×3 convolutional operations refine the merged feature maps, reducing channel dimensionality to generate pixel-level segmentation outputs.

The contracting path in U-Net follows conventional CNN architecture principles, employing successive convolutional layers, ReLU activations, and max-pooling operations. The architectural pipeline establishes hierarchical feature learning through successive down-sampling operations, where systematic augmentation of channel dimensions counterbalances spatial resolution reduction to maintain representational fidelity. The symmetric expanding pathway enhances feature resolution through successive up-sampling operations. Skip connections between corresponding encoder and decoder layers integrate high-resolution features from the contracting path with contextual information from the expanding path, thereby maintaining both global context and precise localization.

The network architecture employs 3×3 convolutional filters throughout, with the exception of the final layer which utilizes a 1×1 convolution to reduce feature channels to the required number of output classes. ReLU activation functions are implemented after each convolutional layer (except the final layer), with sigmoid activation used for binary classification tasks and soft-max for multi-class segmentation. The architecture consistently applies 2×2 max-pooling for down-sampling and 2×2 up-sampling for feature map expansion. The network output consists of pixel-wise classification masks that precisely delineate object boundaries and categories.

B. Block Diagram of Attention U-Net Model

The Attention U-Net architecture builds upon the conventional U-Net framework by incorporating an attention mechanism into its structural design[27], as illustrated in Fig.2. Specifically, the enhancement integrates the attention mechanism within the skip-connection pathways, enabling more effective feature aggregation and spatial attention weighting during the feature fusion process.

The Attention U-Net architecture introduces a hierarchical attention mechanism within the U-Net framework, specifically integrating attention gate (AG) modules. These modules dynamically weight multi-scale features through soft attention coefficients, enabling selective amplification of crack-related patterns while suppressing irrelevant

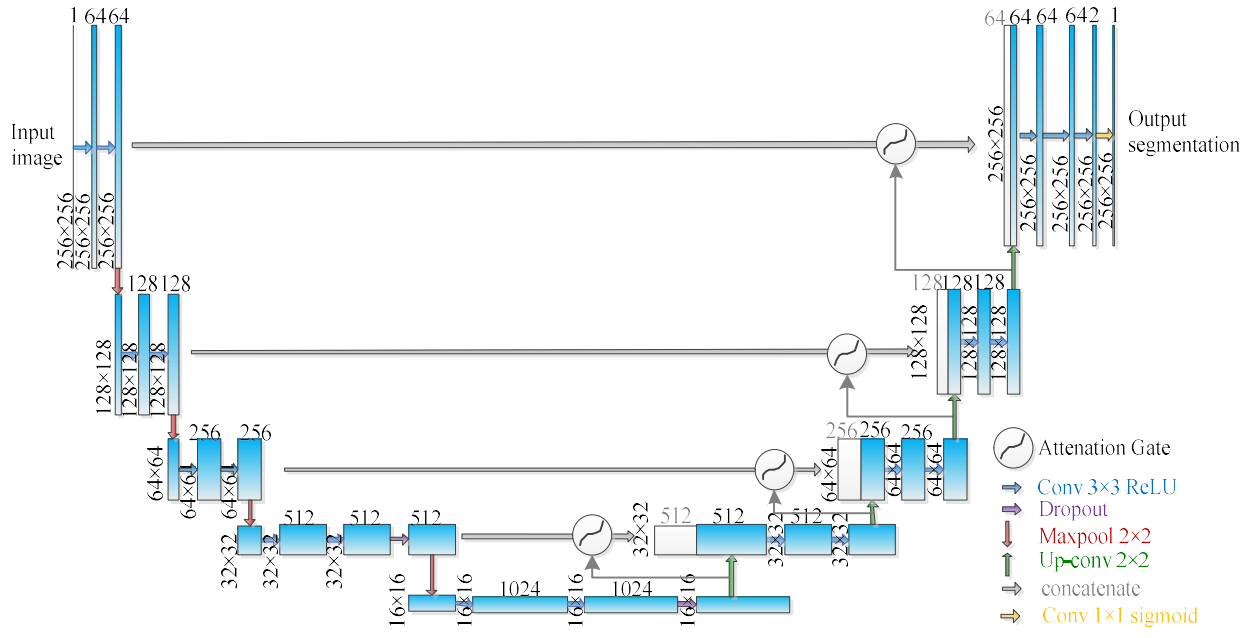

Fig. 2. Block diagram of Attention U-Net segmentation model

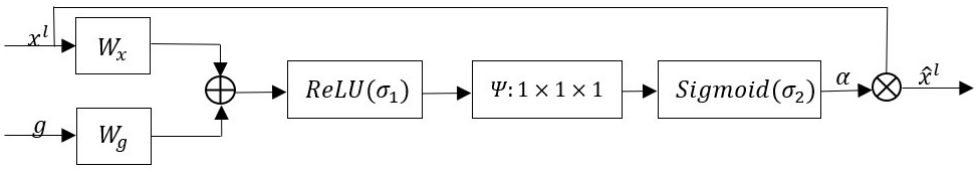

Fig. 3. Attention Gate

background noise[28]. This localized attention coefficient generation demonstrates superior performance compared to conventional global feature vector-based gating approaches, effectively enhancing segmentation precision and computational efficiency.

The block diagram of Attention U-Net is shown in Fig.2. The encoder part is basically the same as the U-Net encoder, with the main change being the decoder part. The structure is briefly described as follows: in the encoder part, the input image undergoes two sets of 3×3 convolution and ReLU activation, then undergoes max pooling down-sampling. Following a sequence of four convolutional and pooling units, the architecture transitions into the decoding phase. The final encoder layer's feature map undergoes direct upscaling and also engages in an attention gating process with the encoder's feature map before its integration with the upscaled map. With the completion of four upscaling units like this, the ultimate segmented image output is produced.

Attention U-Net enhances feature fusion by implementing cross-scale attention mechanisms between encoder and decoder pathways. This architecture first computes attention weights for skip-connected features, then performs weighted fusion before upsampling operations. The resultant attention-augmented feature maps exhibit spatial-adaptive receptive fields, enabling dynamic region-of-interest emphasis.

Fig.3 presents the attention gate mechanism in the Attention U-Net architecture. The output feature map of the layer l is represented as x^l. Meanwhile, the feature map g represents the up-sampling of the decoder. It is used to calculate the attention gating signal parameters and x^l. So, the size of g is half of x^l. It needs to down-sample x^l or up-sample g to ensure consistent size. The feature x^l convolves 1×1 to obtain $W_x^T x^l$. The feature map of the previous layer in up-sampling is g, and after 1×1 convolution, $W_g^T g$ is obtained. After the size is adjusted, $W_x^T x^l$ and $W_g^T g$ perform a point by point add operation, then pass through ReLU to obtain $\sigma_1(W_x^T x^l + W_g^T g + b_g)$. Then 1×1 convolution the result is q_{att}^l. Subsequently, by applying the sigmoid function for activation processing, the ultimate attention score α^l is derived from the convoluted result.

$$q_{att}^l = \psi^T(\sigma_1(W_x^T x^l + W_g^T g + b_g)) + b_\psi \quad (1)$$

$$\alpha^l = \sigma_2(q_{att}^l(x^l, g; \theta_{att})) \quad (2)$$

Multiply the attention coefficients α^l and x^l, it can be to obtain $\hat{x}^l$. $\hat{x}^l$ is the feature map after attention gating calculation. The attention mechanism scales the feature map, diminishing the influence of extraneous areas while amplifying the importance of the focal region, which in turn enhances the network's predictive efficiency and the precision of the image segmentation. Furthermore, the study's findings demonstrate an improvement in the performance of the U-Net with the integration of attention gates over the baseline U-Net model.

IV. Experiments And Results

A. Experiments

A publicly available crack dataset is CFD[29], which contains 118 annotated crack images with a resolution of 480×320. The dataset is an annotated road crack image database that can roughly reflect the condition of urban road surfaces.

It is apparent that these pictures are marred by interference like shadows, oily marks, and damp patches. For the images, we apply a division of 60% for training and 40% for testing.

To enhance model generalization and mitigate overfitting, augmented versions of crack images and their corresponding annotations were generated through geometric and photometric transformations. These included such as rotating, width shifting, height adjusting, shearing, increasing or decreasing luminosity, scaling randomly, and flipping horizontally. This approach was taken to prevent the model from overfitting and to bolster the network's capacity to generalize. Enhanced crack images and their corresponding annotated images were generated based on the original crack images through methods such as rotation angle, width offset, height offset, shear strength, brightness enhancement, brightness reduction, random scaling, and horizontal mirroring. The study outlines the employed methods in Table I. In this research, the table summarizes the range of data augmentation techniques utilized, with 20 degrees for rotation, 0.1 for width shifting, 0.1 for height shifting, 0.05 for shearing, 0.05 for zooming, and a 0.25 probability for horizontal flipping.

TABLE I
DATA AUGMENTATION METHODS APPLIED IN THIS STUDY

Methods	Range
Rotation range	20
Width shift range	0.1
Height shift range	0.1
Shear range	0.05
Zoom range	0.05
Horizontal flip	0.25

B. Results

Within the research, the assessment of crack identification methods employs metrics such as accuracy, F1 Score, and the Intersection over Union (IoU). The calculation of precision and recall is based on true positives (TP), false negatives (FN), and false positives (FP). For gauging the accuracy of

image segmentation, these criteria proficiently quantify the correlation between the algorithm's output and the actual reference data. The formulae may be articulated in the subsequent manner:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{F1\_Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\text{IoU} = \frac{\text{GroundTruth} \cap \text{Prediction}}{\text{GroundTruth} \cup \text{Prediction}} \tag{6}$$

Fig.4(a) presents the raw crack specimen, with its corresponding manually annotated ground truth displayed in Fig.4(b). For enhanced visual interpretation, Fig.4(c) superimposes the crack segmentation results (highlighted in red) onto the original image.

This comparative study quantitatively evaluates the crack detection performance between standard U-Net and its attention-enhanced variant. While Fig.4(d) presents U-Net's segmentation output, Fig.4(e) demonstrates Attention U-Net's superior prediction clarity. The visual comparison in Fig.4(f) further highlights the attention mechanism's effectiveness through red-highlighted crack regions, showing significant reduction in both false positives and negatives.

In order to facilitate intuitive analysis of the segmentation effect of Attention U-Net, take an image with oil stains for crack segmentation. The original image and ground truth of cracks are shown in Fig.5(a) and Fig.5(b) respectively. The crack segmentation result of Attention U-Net is shown in Fig.5(d). The binarized image of Fig.5(d) is shown in Fig.5(e).

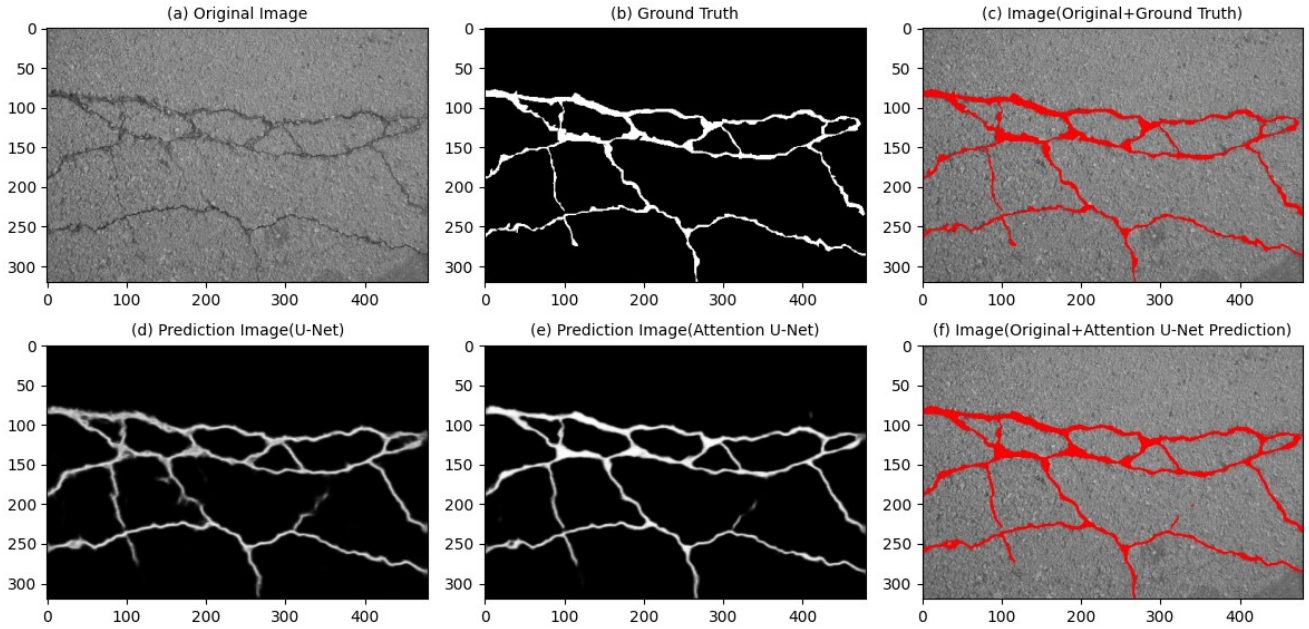

Fig. 4. Crack images and predicted results

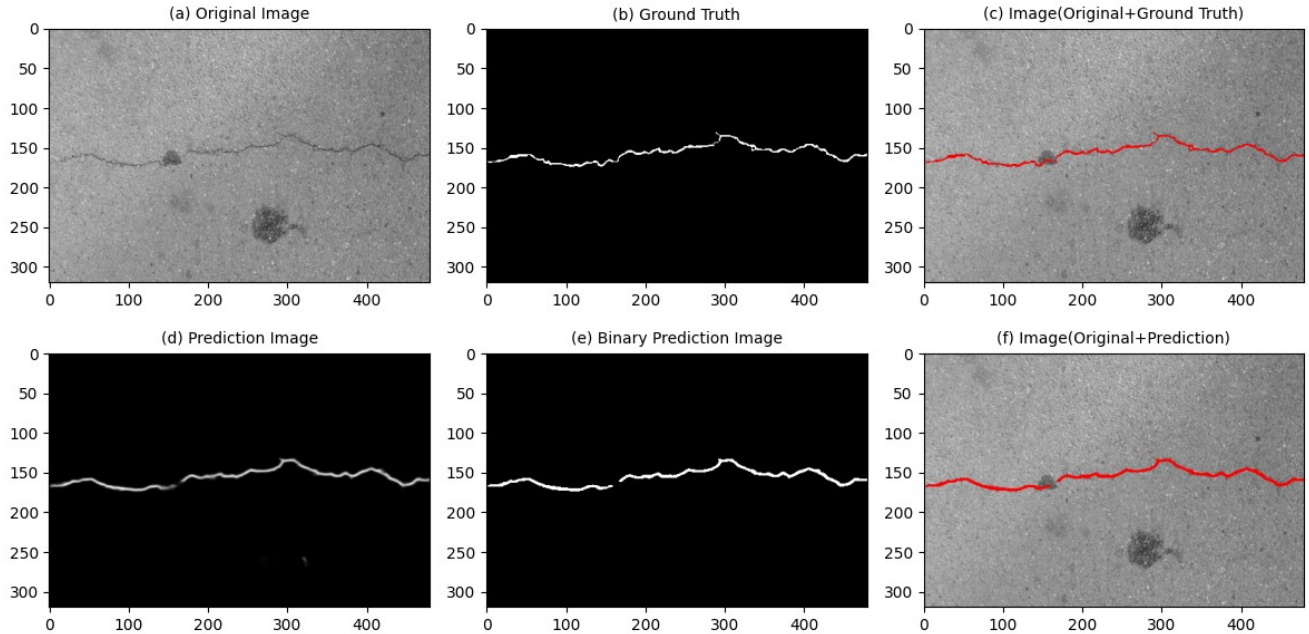

Fig. 5. Crack images and Attention U-Net predicted results

From Fig.5(e), it can be seen that the adverse effects of noise such as oil pollution on segmentation have been effectively suppressed.

The crack segmentation results of U-Net and Attention U-Net are shown in Table II.

TABLE II
CRACK DETECTION RESULTS

Method	Precision	F1 Score	Recall	IOU
U-Net	91.1%	72.5%	79.9%	57.0%
Attention U-Net	95.2%	76.8%	85.1%	62.8%

From Table II, it can be seen that the segmentation performance of Attention U-Net is better than that of U-Net.

V. CONCLUSIONS

The paper conducts a systematic performance evaluation between conventional U-Net and its attention-enhanced variant, with particular emphasis on architectural differences. Utilizing the standardized CFD crack detection dataset, both models undergo rigorous comparative testing. Experimental findings demonstrate the superior performance of Attention U-Net, which exhibits significantly enhanced robustness against common industrial noise interference. Notably, the attention-based architecture achieves satisfactory detection accuracy across diverse complex backgrounds without requiring post-processing procedures.

REFERENCES

[1] T. S. Nguyen, M. Avila, and S. Begot, "Automatic detection and classification of defect on road pavement using anisotropy measure," in *2009 17th European Signal Processing Conference*, 2009, pp. 617-621: IEEE.

[2] H. Oliveira and P. L. Correia, "Automatic road crack detection and characterization," *IEEE Transactions on Intelligent Transportation Systems,* vol. 14, no. 1, pp. 155-168, 2012.

[3] D. Dhital and J.-R. Lee, "A fully non-contact ultrasonic propagation imaging system for closed surface crack evaluation," *Experimental mechanics,* vol. 52, no. 8, pp. 1111-1122, 2012.

[4] X. Yang, H. Li, Y. Yu, X. Luo, T. Huang, and X. Yang, "Automatic pixel - level crack detection and measurement using fully convolutional network," *Computer - Aided Civil and Infrastructure Engineering,* vol. 33, no. 12, pp. 1090-1109, 2018.

[5] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *International journal of multimedia information retrieval,* vol. 7, no. 2, pp. 87-93, 2018.

[6] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience,* vol. 2018, 2018.

[7] A. Mohan and S. Poobal, "Crack detection using image processing: A critical review and analysis," *Alexandria Engineering Journal,* vol. 57, no. 2, pp. 787-798, 2018.

[8] P. Wang and H. Huang, "Comparison analysis on present image-based crack detection methods in concrete structures," in *2010 3rd International Congress on Image and Signal Processing*, 2010, vol. 5, pp. 2530-2533: IEEE.

[9] C. Koch, K. Georgieva, V. Kasireddy, B. Akinci, and P. Fieguth, "A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure," *Advanced Engineering Informatics,* vol. 29, no. 2, pp. 196-210, 2015.

[10] M. R. Jahanshahi, J. S. Kelly, S. F. Masri, and G. S. Sukhatme, "A survey and evaluation of promising approaches for automatic image-based defect detection of bridge structures," *Structure and Infrastructure Engineering,* vol. 5, no. 6, pp. 455-486, 2009.

[11] I. Abdel-Qader, O. Abudayyeh, and M. E. Kelly, "Analysis of edge-detection techniques for crack identification in bridges," *Journal of Computing in Civil Engineering,* vol. 17, no. 4, pp. 255-263, 2003.

[12] N. Tanaka and K. Uematsu, "A crack detection method in road surface images using morphology," *MVA,* vol. 98, pp. 17-19, 1998.

[13] Y. Fujita, Y. Mitani, and Y. Hamamoto, "A method for crack detection on a concrete structure," in *18th International Conference on Pattern Recognition (ICPR'06)*, 2006, vol. 3, pp. 901-904: IEEE.

[14] N. B. C. Ahmed, S. Lahouar, C. Souani, and K. Besbes, "Automatic crack detection from pavement images using fuzzy thresholding," in *2017 international conference on control, automation and diagnosis (ICCAD)*, 2017, pp. 528-537: IEEE.

[15] C. Koch and I. Brilakis, "Pothole detection in asphalt pavement images," *Advanced Engineering Informatics,* vol. 25, no. 3, pp. 507-515, 2011.

[16] R. Amhaz, S. Chambon, J. Idier, and V. Baltazart, "A new minimal path selection algorithm for automatic crack detection on pavement images," in *2014 IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 788-792: IEEE.

[17] Y.-J. Cha and W. Choi, "Vision-based concrete crack detection using a convolutional neural network," in *Dynamics of Civil Structures, Volume 2*: Springer, 2017, pp. 71-73.

[18] X. Zhao and S. Li, "A method of crack detection based on convolutional neural networks," *Proceedings of the structural health monitoring,* 2017.

[19] K. Wang, A. Zhang, J. Q. Li, Y. Fei, C. Chen, and B. Li, "Deep learning for asphalt pavement cracking recognition using convolutional neural network," in *Proc. Int. Conf. Airfield Highway Pavements*, 2017, pp. 166-177.

[20] C. V. Dung, "Autonomous concrete crack detection using deep fully convolutional neural network," *Automation in Construction,* vol. 99, pp. 52-58, 2019.

[21] Y. J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning - based crack damage detection using convolutional neural networks," *Computer - Aided Civil and Infrastructure Engineering,* vol. 32, no. 5, pp. 361-378, 2017.

[22] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, "Road crack detection using deep convolutional neural network," in *2016 IEEE international conference on image processing (ICIP)*, 2016, pp. 3708-3712: IEEE.

[23] A. Zhang *et al.*, "Automated pixel - level pavement crack detection on 3D asphalt surfaces with a recurrent neural network," *Computer - Aided Civil and Infrastructure Engineering,* vol. 34, no. 3, pp. 213-229, 2019.

[24] E. Elhariri, N. El-Bendary, and S. A. Taie, "Automated pixel-level deep crack segmentation on historical surfaces using U-Net models," *Algorithms,* vol. 15, no. 8, p. 281, 2022.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.

[27] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999,* 2018.

[28] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," *arXiv preprint arXiv:1804.02391,* 2018.

[29] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Transactions on Intelligent Transportation Systems,* vol. 17, no. 12, pp. 3434-3445, 2016.