

A CNN–Transformer Dual-Encoder Network for Precise Building Crack Detection [†]

Kang Chen and Lingzhi Li *

Department of Structural Disaster Prevention and Mitigation Engineering, College of Civil Engineering, Tongji University, Siping Road Campus, Shanghai 200092, China; c1432580014@163.com

* Correspondence: lilingzhi@tongji.edu.cn

[†] Presented at the 7th International Conference on Civil, Architecture and Disaster Prevention and Control, Dali, China, 30 January–1 February 2026.

Abstract

Detecting wall damage is crucial for ensuring building safety, as undetected cracks may result in serious structural issues. Early inspections are therefore critical to maintaining structural integrity and avoiding expensive repairs down the line. However, building cracks exhibit highly diverse characteristics. Fine, surface-level cracks require precise local information for reliable identification. Large cracks demand global contextual information to be properly recognized. Convolutional neural networks specialize in capturing detailed local features, making them ideal for detecting small cracks. Transformers are adept at modeling long-range interactions and global pixel correlations, which are crucial for segmenting large and intricate crack structures. Despite these complementary strengths, most existing approaches rely exclusively on either CNN-based or Transformer-based encoders. They often fail to balance local detail and global context, leading to incomplete or inaccurate segmentation results. To overcome this challenge, we introduce DEF-Net, an innovative dual-encoder architecture. This model combines the strengths of CNNs and Transformers. A CNN encoder is employed to extract rich local representations. A Transformer encoder is used to provide global structural context. To fully integrate the complementary features from both encoders, we present an attention-based feature fusion module that aligns and merges features at different levels, boosting the network's overall representation ability. This design allows DEF-Net to precisely identify both minor, intricate cracks and large, complicated ones across varied conditions. Comprehensive experiments on the CrackSeg9k and DeepCrack datasets show that DEF-Net consistently surpasses existing methods, delivering its exceptional segmentation performance and generalization ability. These findings underscore the value of integrating local and global modeling for assessing structural damage and lay a solid groundwork for future advancements in intelligent crack detection.

Keywords: crack detection; dual-encoder architecture; feature fusion; image segmentation; DEF-Net



Academic Editor: Mohamed Abdelkader Ismail

Published: 26 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

As urbanization progresses and cities expand, the number of buildings has increased significantly. With the aging of these structures, cracks in concrete walls become more prevalent, posing serious safety risks [1]. Failure to detect these cracks may cause them to deteriorate over time, ultimately resulting in structural damage and posing risks to both lives and property [2]. Regular inspections are essential for the early detection of cracks.

This proactive approach not only ensures the safety and stability of buildings but also extends their lifespan and enhances overall safety.

Traditional detection methods [3–5] typically rely on threshold segmentation and edge detection algorithms. These techniques primarily involve threshold-based segmentation and edge-detection algorithms. Threshold-based segmentation works by comparing pixel values to a set threshold and dividing the image into distinct regions. In contrast, edge-detection algorithms rely on differential operators to calculate grayscale intensity changes between crack boundaries and the background. Oliveira et al. [4] introduced a dynamic decision boundary method for the automatic detection of concrete cracks. Peng et al. [5] developed a secondary threshold segmentation technique specifically for crack detection on airport runways.

Methods based on deep learning have attracted considerable focus in recent times and shown impressive performance in the domain of computer vision. Unlike traditional methods, deep learning techniques can learn the semantic features of cracks and have exhibited enhanced robustness and scalability for real-world applications. Long et al. [6] presented a fully convolutional architecture for end-to-end pixel-wise segmentation. Ronneberger et al. [7] developed an encoder–decoder framework, which has since become a standard benchmark for many segmentation methods. Zou et al. [8] introduced the Deepcrack model for crack detection by incorporating a multi-level feature integration approach. Lin et al. [9] utilized a ResNet-34 backbone, a pyramid edge component, and a multi-scale fusion mechanism to achieve efficient multi-level feature integration. The CNN-based methods mentioned above perform well in detecting small cracks. However, they struggle to accurately segment large cracks because of their restricted capacity to capture broad semantic context. In contrast, Transformer excels at modeling long-range contextual relationships and offers superior feature extraction capabilities for complex, elongated cracks. Alexey et al. [10] extracted feature tokens through a feature embedding module and a series of self-attention modules. Liu et al. [11] introduced a layered Transformer framework, which generates multi-level, multi-resolution feature maps with rich semantic details. Xie et al. [12] introduced a compact multi-layer perceptron design that removes position encoding. Liu et al. [13] integrated the Transformer encoder module into an encoder–decoder architecture to extract global semantic features for complex crack segmentation.

CNNs are highly capable of capturing fine details and edge features, making them well-suited for identifying small cracks. In contrast, Transformers can capture long-range dependencies and model complex, elongated cracks. However, existing approaches fail to fully leverage the unique advantages of both feature extraction techniques for building crack segmentation. To address this challenge, we present a dual-encoder segmentation network that leverages the advantages of both CNNs and Transformers. This model is specifically tailored for accurate building crack segmentation under complex conditions. The key contributions of this study are outlined below:

- (1) We present DEF-Net, an innovative dual-encoder feature fusion network that combines CNNs with Transformers. This design effectively captures global contextual information while preserving key local features, allowing for the detection of both fine, surface-level cracks and large, intricate fractures.
- (2) We introduce a novel attention-based feature fusion module that enables the smooth combination of features from both the CNN and Transformer encoders. This component allows the model to effectively balance and integrate both local and global information.
- (3) We conduct comprehensive experiments on two benchmark datasets, CrackSeg9k and DeepCrack. Extensive evidence shows that DEF-Net surpasses leading crack

detection models and delivers exceptional segmentation precision and adaptability across various crack patterns and conditions.

2. Related Work

2.1. Crack Segmentation

The aim of crack segmentation is to precisely detect possible fractures in images of infrastructure such as buildings, roads, or pavements to prevent concrete degradation or collapse. Traditional methods [3–5] often rely on manual feature extraction to detect cracks, followed by threshold segmentation or edge detection algorithms to differentiate between healthy concrete and cracked regions. However, these methods often face difficulties when dealing with intricate or noisy images and may overlook the fine details necessary for accurate crack identification. They also tend to be sensitive to environmental conditions, lack durability, and find it difficult to expand in a scalable way. One of the pioneering deep learning methods for crack identification utilized a pixel-wise CNN model based on a fully convolutional network (FCN) [14]. This approach considerably shortened the training cost for crack detection, in contrast to CrackNet [15]. CrackNet reached top-tier performance during its period by removing the pooling layer, enabling accurate detection of fine cracks. However, its real-time prediction efficiency was limited. DeepCrack [16] extended the FCN framework by adding batch normalization and side networks, which enhanced the speed of convergence and boosted generalization performance. Furthermore, the DeepCrack dataset [16] served as an important foundation for further progress in crack detection research. Cheng et al. [17] presented a crack segmentation model built on the UNet architecture. This design was further validated by subsequent studies [17–21] for its effectiveness in crack segmentation tasks. It was found that using pre-trained image classification backbones can enhance feature extraction [19]. Several other encoder–decoder models have also been explored. Among them, DeepCrack2 [8] (distinguished from Liu et al.'s previous DeepCrack model [16]) enables automatic crack identification via end-to-end learning. The model emphasizes the extraction of advanced crack features and integrates multi-scale information from hierarchical convolution layers. This multi-scale fusion strategy captures finer crack details in large-scale feature maps while preserving structural information in smaller-scale maps. Like SegNet [22], DeepCrack2 adopts an encoder–decoder structure and carries out pairwise feature integration between matching encoder and decoder layers, positioning it as one of the most extensively evaluated models in crack segmentation. Although CNN-based methods for crack detection have progressed significantly, many continue to encounter issues related to efficiency and robustness in practical applications. HrSegNet [23] was introduced to preserve high-resolution features across the entire network. In contrast to traditional methods, HrSegNet utilizes low-resolution semantic information to direct the reconstruction of high-resolution details. The HrSegNet-B64 variant, in particular, has demonstrated new cutting-edge performance in both crack segmentation precision and real-time speed.

2.2. Convolutional Neural Network

Deep learning has recently demonstrated significant advancements and established itself as a key paradigm in computer vision. Owing to their convolutional kernel design, CNNs effectively capture local details and have demonstrated strong performance across diverse application domains. Furthermore, CNNs are capable of capturing features across multiple semantic levels and are applied extensively in domains including image segmentation and crack detection. Long et al. [6] pioneered the use of deep learning methods for semantic segmentation by proposing a fully convolutional network (FCN) with an end-to-end architecture. Building upon earlier advances, Ronneberger et al. [7] introduced

a U-Net with a symmetric encoder–decoder design, incorporating skip links that help preserve fine-grained spatial details at different scales while alleviating the degradation of information. SegNet [22], an additional FCN-based architecture, incorporates a mechanism for transferring pooling indices. Chen et al. [24] incorporated ASPP into the network and proposed the Deeplab series.

To address crack detection challenges, Zou et al. [8] extended SegNet by introducing a feature fusion module and introduced DeepCrack. Xu et al. [25] further enhanced segmentation precision through a fusion CNN architecture. Ren et al. [26] applied dilated convolutions with fixed expansion rates, enabling improved representation of global context within feature maps. EMRA-Net integrates a self-adaptive weighting mechanism, which enhances its capability to detect fine-grained defects. More recently, Ma et al. [27] applied depthwise separable convolution operations to reduce the number of trainable parameters while enhancing multi-scale feature representation via cross-layer integration. In addition, they incorporated a hybrid attention framework to model long-range correlations in crack structures. Dong et al. [28] introduced a novel multi-layer feature association fusion network that integrates local and global features through a dual-branch feature association module, improving segmentation precision significantly.

In general, CNN-based approaches have demonstrated high effectiveness in crack detection. However, the limited capability of these methods in modeling the global context can lead to segmentation outputs that are partial or structurally broken. This limitation underscores the importance of developing more sophisticated architectural designs.

2.3. Transformer

Transformers [10,11,29] are recognized for their strong capability in modeling long-range dependencies, with notable applications in natural language processing. They accomplish this by capturing global interactions among tokens. However, due to their causal design that operates on data in a sequential manner, these models are not optimal for visual tasks. In such tasks, spatial dependencies and local contextual cues are essential for accurate representation. ViT [10] adapts the Transformer architecture and utilizes self-attention mechanisms to model long-range dependencies, providing a perfect demo for vision applications. ViT [10] modifies the Transformer framework and employs self-attention to capture long-range dependencies, serving as a compelling demonstration of its applicability to visual tasks.

Xie et al. [12] eliminated the use of positional encoding, thereby mitigating potential declines in model performance. A compact multilayer perceptron (MLP) is then integrated to facilitate effective inter-layer feature aggregation, ultimately achieving high-quality segmentation outcomes. Chen et al. [30] integrated the complementary capabilities of Transformers and CNNs by upgrading the U-Net into a hybrid design, which boosts segmentation performance through the effective coordination of global contextual information and fine-grained local details. Liu et al. [13] further advanced this direction by designing a novel Transformer encoder module within an encoder–decoder framework. Bai et al. [31] integrated convolutional neural networks with Transformer modules, proposing a framework capable of capturing fine-grained spatial features alongside broader contextual relationships while consolidating outputs derived from diverse processing pathways.

Transformer-based models have demonstrated a strong capability in crack detection through their ability to model long-range dependencies. However, the inherent quadratic time complexity imposes substantial computational overhead and may impair the accurate capture of fine-grained local features.

3. Method

This section describes the proposed DEF-Net. First, we employed a dual-encoder design. Next, the attention-based feature fusion module was used to effectively fuse the CNN and Transformer features. Finally, a lightweight decoder was employed to generate the final crack segmentation map.

3.1. Overview of DEF-Net

As shown in Figure 1, DEF-Net is composed of a CNN encoder, a Transformer encoder, an attention-based fusion module, and a hybrid decoder. Initially, two parallel encoder networks receive the raw image and generate feature representations at multiple spatial scales. These features are then integrated via a feature interaction module. Finally, the hybrid decoder enhances the fused features and produces the final crack segmentation map.

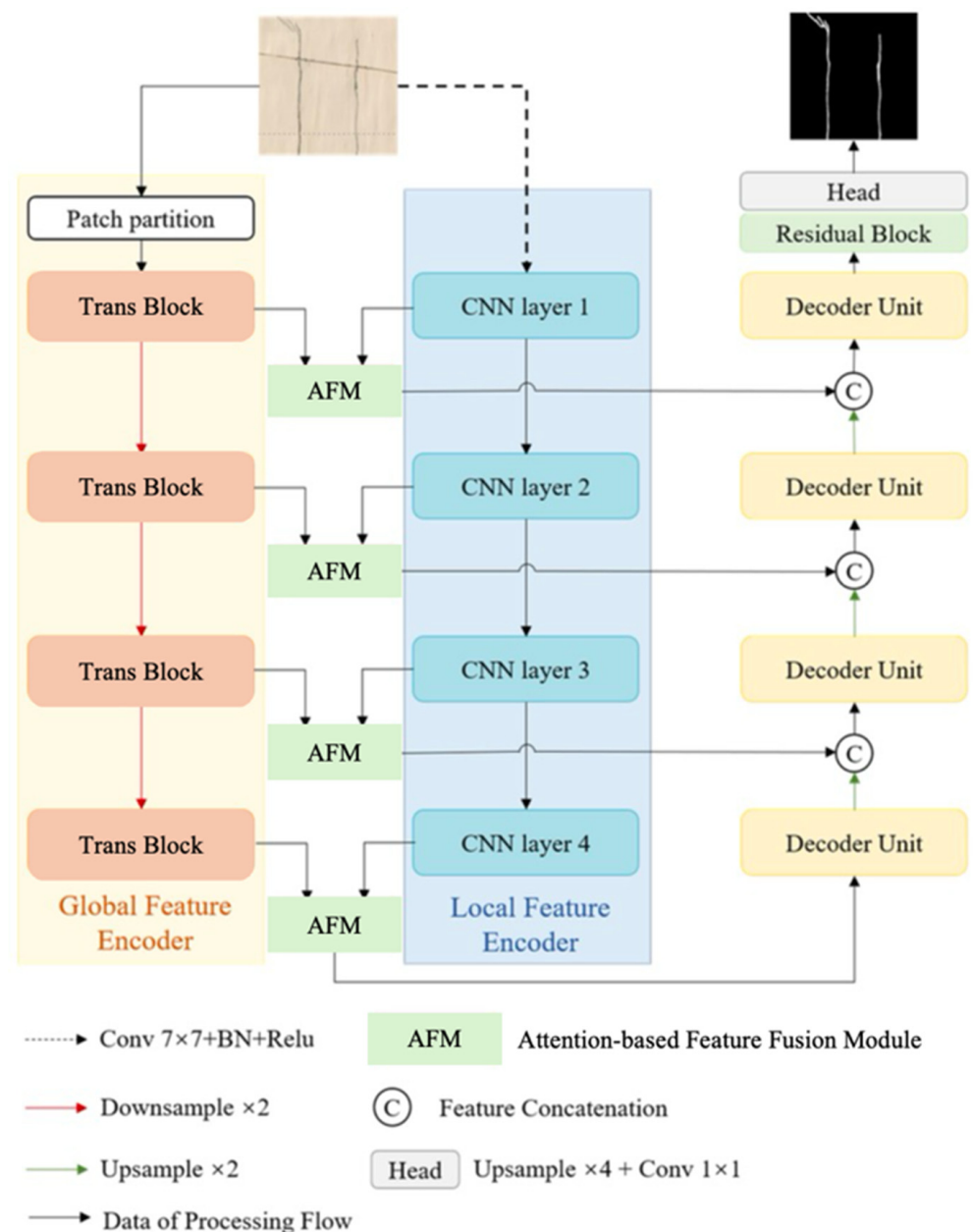


Figure 1. Overview of DEF-Net.

3.2. Dual-Path Encoding Stages

Detecting small cracks requires attention to fine-grained local features, whereas segmenting large and complex cracks necessitates leveraging broader contextual cues. Based on this finding, we adopted a dual-encoder strategy that integrates convolutional neural networks with Transformer modules. This design aims to improve building crack segmentation accuracy and to maintain robustness in challenging environmental conditions.

3.2.1. CNN Encoder

The spatial locality inherent to convolutional kernels enables CNNs to capture detailed patterns with high precision. This property renders them advantageous for the detection of small-scale targets, including micro-cracks. ResNet mitigates problems such as the vanishing and exploding of gradients, thereby enhancing the robustness and learning efficiency of the model. In this paper, ResNet-50 [31] is used to construct the CNN encoder. Initially, a convolutional layer with a 7×7 kernel is employed to downsample the input image. Subsequently, the extracted features pass through multiple residual blocks, allowing hierarchical representations to be formed and generating multi-scale feature maps for later fusion.

3.2.2. Transformer Encoder

Transformers capture global context by attending to all spatial positions within an image, thereby delivering outstanding results in dense vision applications. The researchers partitioned images to generate multiple semantic tokens through feature embedding and utilized a global self-attention mechanism to model inter-token relationships, enabling effective global feature modeling and yielding strong results. Here, the Transformer [11] encoder was utilized to facilitate global representation modeling by performing window-level partitioning and capturing long-range dependencies between local windows.

3.3. Attention-Based Feature Fusion Module

To combine the hierarchical features generated by the dual encoders, we designed an attention-based fusion module (AFM) incorporating a self-attention mechanism together with depthwise-separable convolution layers. A comprehensive overview of AFM is shown in Figure 2. Initially, the CNN-derived features F_c and Transformer-derived features F_t are fed into the AFM. They are subsequently element-wise summed and refined through a residual self-attention module, which optimizes the feature interactions. A channel-wise recalibration mechanism adjusts the relative significance of diverse feature activations, thereby suppressing redundancy in the representation. A feedforward neural network is further utilized to extract fine-grained features of cracks, thereby improving the model's sensitivity to nuanced structural variations. In the final stage, the refined features and the original feature maps are merged via element-wise summation, resulting in an enriched feature representation for accurate crack segmentation. Specifically, the Swin Transformer begins by dividing the input image into fixed-size patches according to the predefined window dimensions. To maintain spatial coherence, overlapping windows are employed to capture relationships across patches. The encoder is composed of four hierarchical stages. Each stage integrates several attention blocks operating on sliding windows alongside feedforward neural layers. The resulting multi-scale features are later aggregated to strengthen the overall representation.

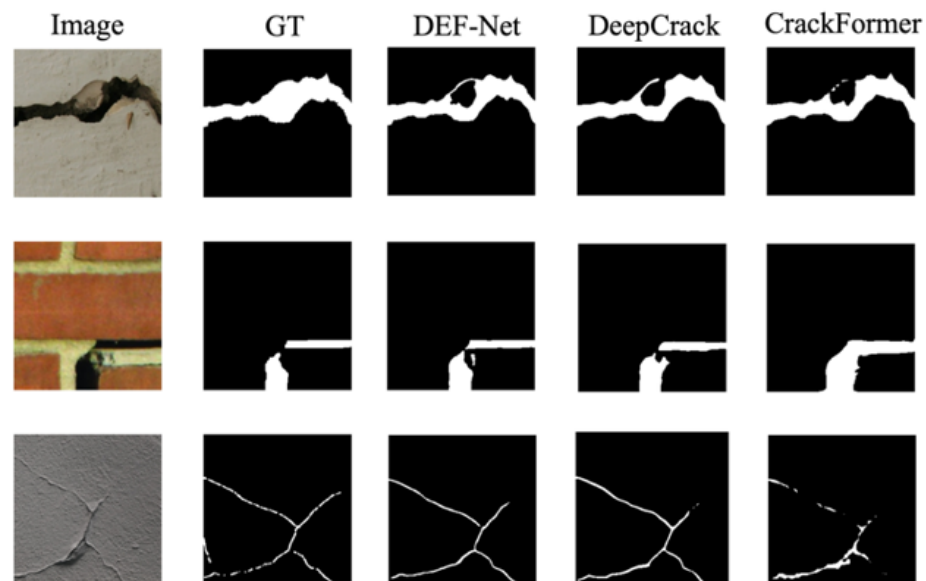


Figure 2. Qualitative results of different methods on the CrackSeg9k dataset.

3.4. Hybrid Decoder

To leverage the Transformer's capacity for capturing long-range dependencies and the CNN's ability to encode fine-grained local patterns, we introduce a hybrid decoder that integrates Transformer and CNN blocks. Initially, the enhanced feature map is processed by the Transformer block and subsequently refined using a 1×1 convolution, batch normalization, and ReLU nonlinearity. A residual pathway is applied in the subsequent stage to enhance the feature representation, reducing irrelevant data and retaining key crack-related characteristics. Finally, a segmentation head is applied to generate the final segmentation output, ensuring accurate segmentation by employing enhanced features.

4. Experiments

4.1. Datasets

CrackSeg9k [32]: The dataset is collected from public sources such as Crack500 and GAPS384, featuring various crack types and sizes. For training, the input image was resized to 256×256 , and the dataset's default data partitioning scheme was adopted. The dataset consists of 7332 training images and 1827 test images, each with a resolution of 400×400 pixels.

DeepCrack [16]: The dataset was collected from a variety of scenes, including building walls and pavements. It comprises 537 images, each with a resolution of 544×384 pixels.

4.2. Implementation Details

DEF-Net was developed and trained on an experimental platform equipped with Ubuntu 20.04, utilizing a single NVIDIA GTX 3090 GPU. We employed the AdamW optimizer and set the learning rate to 1×10^{-4} . We trained DEF-Net for 500 epochs, incorporating an early stopping strategy with a patience of 50. Furthermore, to achieve stable and robust model training, we standardized the input data to normalize the pixel values and ensured consistent input distributions. Additionally, we employed data augmentation to increase diversity, including random rotations and scaling. These augmentations can simulate various real scenes, where cracks may appear at different orientations and scales.

4.3. Evaluation Metrics

Standard evaluation metrics for building crack segmentation typically involve the mean Dice coefficient (mDice), mean Intersection over Union (mIoU), precision, and

recall. Both mDice and mIoU assess the extent of overlap between predictions and expert-annotated labels. Precision refers to the fraction of true positive pixels among all predicted positive pixels, whereas recall represents the fraction of true positive pixels among all actual positive pixels.

4.4. Comparative Experiments

CrackSeg9k: To evaluate the effectiveness of DEF-Net, we conducted comprehensive comparative experiments on the CrackSeg9k and DeepCrack datasets, benchmarking its performance against four SOTA methods, including UNet [7], DeepLabv3 [33], DeepCrack [9], and Crackformer [13]. Table 1 shows the quantitative performance of these methods on the CrackSeg9k dataset. DEF-Net achieves SOTA results across all four metrics—mDice, mIoU, precision, and recall—outperforming existing approaches. Specifically, DEF-Net attains an mDice of 71.2% and an mIoU of 75.5%. Compared with CrackFormer, these scores represent improvements of 2.4% and 1.8%, respectively. Additionally, in Figure 2, we further provide the visualized results of different methods. In the first row, our model delivers a finer boundary of the damaged wall, demonstrating superior crack detection performance. In the third row, DEF-Net stands out by identifying even the subtlest cracks that other models fail to detect, highlighting its comprehensive and fine-grained detection capabilities. The results confirm that our method is capable of accurately segmenting complex cracks across diverse scenarios, underscoring its robustness and effectiveness.

Table 1. Performance of different models on the CrackSeg9k dataset.

Model	meanDice	IoU	Precision	Recall
U-Net [7]	0.535	0.564	0.559	0.571
DeepLabv3 [33]	0.581	0.621	0.601	0.612
DeepCrack [9]	0.652	0.692	0.672	0.670
CrackFormer [13]	0.688	0.737	0.727	0.711
Ours	0.712	0.755	0.753	0.732

DeepCrack: We performed extensive comparative experiments on the DeepCrack dataset. In Table 2, the quantitative results show that DEF-Net outperforms other SOTA methods. DEF-Net achieves meanDice and IoU scores of 86.7% and 88.2%, respectively, surpassing CrackFormer by 1.9% and 0.5%. These results further substantiate the superior efficacy of DEF-Net in crack segmentation tasks. Obviously, compared to other SOTA methods, DEF-Net delivers highly accurate segmentation in complex scenes.

Table 2. Performance of different models on the DeepCrack dataset.

Model	meanDice	IoU	Precision	Recall
U-Net [7]	0.685	0.709	0.694	0.673
DeepLabv3 [33]	0.731	0.754	0.742	0.701
DeepCrack [9]	0.793	0.811	0.804	0.775
CrackFormer [13]	0.848	0.877	0.859	0.823
Ours	0.867	0.882	0.872	0.855

4.5. Ablation Studies

To assess the performance of the proposed component in crack segmentation, we performed extensive ablation experiments using the CrackSeg9k dataset. The quantitative results of these experiments are presented in Table 3. Expr. #1 serves as the baseline model, consisting solely of a CNN encoder and a simple decoder. Expr. #4 corresponds to the proposed DEF-Net model, which incorporates a CNN encoder, a Transformer encoder,

an AFM module, and a hybrid decoder. Expr. #2 represents the model without the AFM module, while Expr. #3 excludes the hybrid decoder. Removing the attention-based feature fusion module leads to a 3.9% decrease in meanDice, highlighting the critical role of this module in the model's architecture. The results clearly demonstrate that the combination of the dual encoder, attention-based feature fusion, and hybrid decoder leads to the best performance, significantly enhancing crack detection accuracy.

Table 3. Ablation experiments on the CrackSeg9k dataset.

Expr.	meanDice	IoU	Precision	Recall
#1	0.626	0.663	0.651	0.633
#2	0.673	0.713	0.698	0.677
#3	0.688	0.721	0.711	0.702
#4	0.712	0.755	0.753	0.732

4.6. Discussion

Convolutional neural networks (CNNs), with their fixed convolutional kernels, are highly effective at extracting local features. However, this local focus limits their ability to perform global modeling and leads to irrational results when segmenting long or deep cracks. To address this limitation, we integrated the hierarchical features of CNNs with those of Transformers. Although this strategy alleviates the issue to some extent, the significant representational gap between dual encoders necessitates carefully designed feature fusion modules. The AFM module proposed in this study shows promising benefits in mitigating information loss, but it still has a shortcoming—the trade-off of reduced sensitivity to small cracks in favor of overall robustness. In future work, we plan to develop robust feature aggregation modules to achieve effective integration between CNN and Transformer features for crack segmentation.

5. Conclusions

We proposed DEF-Net, a hybrid architecture for accurate building crack detection. The network adopts a dual-encoder design, combining the Transformer's long-range modeling ability with the CNN's proficiency in capturing local information. To fuse multi-scale features from dual encoders, we introduced the AFF module that enhances these features via a self-attention mechanism and reduces redundancy via a compressed excitation module. Additionally, we proposed a hybrid decoder to optimize the feature map and improve the segmentation precision of building crack detection. Extensive comparative and ablation experiments validate the performance of the DEF-Net architecture in building crack detection. Although DEF-Net has demonstrated state-of-the-art performance in building crack detection, it requires substantial computational resources, which limits its deployment on mobile devices. Future work will focus on developing lightweight segmentation approaches for crack detection in buildings. Moreover, hybrid CNN–Mamba architectures present another promising avenue for investigation.

Author Contributions: Conceptualization, L.L. and K.C.; investigation, K.C.; data curation, K.C.; formal analysis, K.C.; writing—original draft preparation, K.C.; writing—review and editing, L.L.; supervision, L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: This work is supported by Tongji University.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Chen, K.; Reichard, G.; Xu, X.; Akanmu, A. Automated Crack Segmentation in Close-Range Building Facade Inspection Images Using Deep Learning Techniques. *J. Build. Eng.* **2021**, *43*, 102913. [\[CrossRef\]](#)
- Dais, D.; Bal, I.E.; Smyrou, E.; Sarhosis, V. Automatic Crack Classification and Segmentation on Masonry Surfaces Using Convolutional Neural Networks and Transfer Learning. *Autom. Constr.* **2021**, *125*, 103606. [\[CrossRef\]](#)
- Xie, S.; Tu, Z. Holistically-Nested Edge Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1395–1403. [\[CrossRef\]](#)
- Ai, D.; Jiang, G.; Lam, S.-K.; He, P.; Li, C. Computer vision framework for crack detection of civil infrastructure—A review. *Eng. Appl. Artif. Intell.* **2023**, *117*, 105478. [\[CrossRef\]](#)
- Peng, L.; Chao, W.; Shuangmiao, L.; Baocai, F. Research on Crack Detection Method of Airport Runway Based on Twice-Threshold Segmentation. In Proceedings of the International Conference on Instrumentation and Measurement, Computer, Communication and Control (IMCCC), Qinhuaogdao, China, 18–20 September 2015; pp. 1716–1720. [\[CrossRef\]](#)
- Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [\[CrossRef\]](#)
- Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, Germany, 5–9 October 2015; pp. 234–241.
- Zou, Q.; Zhang, Z.; Li, Q.; Qi, X.; Wang, Q.; Wang, S. DeepCrack: Learning Hierarchical Convolutional Features for Crack Detection. *IEEE Trans. Image Process.* **2018**, *28*, 1498–1512. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lin, Q.; Zhou, J.; Ma, Q.; Ma, Y.; Kang, L.; Wang, J. EMRA-Net: A Pixel-Wise Network Fusing Local and Global Features for Tiny and Low-Contrast Surface Defect Detection. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 2504314. [\[CrossRef\]](#)
- Dosovitskiy, A. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022. [\[CrossRef\]](#)
- Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090. [\[CrossRef\]](#)
- Liu, H.; Yang, J.; Miao, X.; Mertz, C.; Kong, H. Crackformer Network for Pavement Crack Segmentation. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 9240–9252. [\[CrossRef\]](#)
- Yang, X.; Li, H.; Yu, Y.; Luo, X.; Huang, T.; Yang, X. Automatic pixel-level crack detection and measurement using fully convolutional network. *Comput.-Aided Civ. Infrastruct. Eng.* **2018**, *33*, 1090–1109. [\[CrossRef\]](#)
- Zhang, A.; Wang, K.; Fei, Y.; Liu, Y.; Chen, C.; Yang, G.; Li, J.; Yang, E.; Qiu, S. Automated pixel-level pavement crack detection on 3D asphalt surfaces with a recurrent neural network: Automated pixel-level pavement crack detection on 3D asphalt surfaces using CrackNet-R. *Comput.-Aided Civ. Infrastruct. Eng.* **2018**, *34*, 213–229. [\[CrossRef\]](#)
- Liu, Y.; Yao, J.; Lu, X.; Xie, R.; Li, L. DeepCrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing* **2019**, *338*, 139–153. [\[CrossRef\]](#)
- Cheng, J.; Xiong, W.; Chen, W.; Gu, Y.; Li, Y. Pixel-level crack detection using U-net. In Proceedings of the TENCON 2018—2018 IEEE Region 10 Conference, Jeju, Republic of Korea, 28–31 October 2018; pp. 0462–0466. [\[CrossRef\]](#)
- Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
- König, J.; Jenkins, M.D.; Mannion, M.; Barrie, P.; Morison, G. Optimized deep encoder-decoder methods for crack segmentation. *Digit. Signal Process.* **2021**, *108*, 102907. [\[CrossRef\]](#)
- Pal, D.; Reddy, P.B.; Roy, S. Attention UW-Net: A fully connected model for automatic segmentation and annotation of chest X-ray. *Comput. Biol. Med.* **2022**, *150*, 106083. [\[CrossRef\]](#) [\[PubMed\]](#)
- Panella, F.; Lipani, A.; Boehm, J. Semantic segmentation of cracks: Data challenges and architecture. *Autom. Constr.* **2022**, *135*, 104110. [\[CrossRef\]](#)
- Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [\[CrossRef\]](#) [\[PubMed\]](#)
- Li, Y.; Ma, R.; Liu, H.; Cheng, G. Real-time high-resolution neural network with semantic guidance for crack segmentation. *Autom. Constr.* **2023**, *156*, 105112. [\[CrossRef\]](#)

24. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818. [\[CrossRef\]](#)
25. Baduge, S.K.; Thilakarathna, S.; Perera, J.S.; Arashpour, M.; Sharafi, P.; Teodosio, B.; Shringi, A.; Mendis, P. Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications. *Autom. Constr.* **2022**, *141*, 104440. [\[CrossRef\]](#)
26. Ren, Y.; Huang, J.; Hong, Z.; Lu, W.; Yin, J.; Zou, L.; Shen, X. Image-based concrete crack detection in tunnels using deep fully convolutional networks. *Constr. Build. Mater.* **2020**, *234*, 117367. [\[CrossRef\]](#)
27. Ma, M.; Yang, L.; Liu, Y.; Yu, H. An attention-based progressive fusion network for pixelwise pavement crack detection. *Measurement* **2024**, *226*, 114159. [\[CrossRef\]](#)
28. Dong, J.; Wang, N.; Fang, H.; Guo, W.; Li, B.; Zhai, K. MFAFNet: An innovative crack intelligent segmentation method based on multi-layer feature association fusion network. *Adv. Eng. Inform.* **2024**, *62*, 102584. [\[CrossRef\]](#)
29. Mehta, S.; Rastegari, M. 2: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
30. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
31. Bai, S.; Yang, L.; Liu, Y.; Yu, H. DMF-Net: A Dual-Encoding Multi-Scale Fusion Network for Pavement Crack Detection. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 5981–5996. [\[CrossRef\]](#)
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [\[CrossRef\]](#)
33. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.