2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

BFMPF-Net: Bidirectional Frequency-domain Modulation Progressive Fusion Network for Road Crack Segmentation

Wen Yang, Yingying Zheng, Hang Sun*, Chao Liang and Lei Fang

Abstract— **Recently, deep learning–based methods for road crack segmentation have achieved promising performance, particularly in robotic vision applications such as automated inspection and maintenance. However, most frequency-domain methods employ a decoupled processing strategy, overlooking the dynamic modulation mechanism between high- and low-frequency components, which constrains the model's effectiveness in detecting cracks within complex environments. Moreover, existing methods suffer from low information fidelity during feature transmission, where critical encoder details are progressively lost in the decoder, making it difficult to reconstruct complete crack structures. To address these issues, we propose a Bidirectional Frequency-domain Modulation Progressive Fusion Network (BFMPF-Net). Specifically, we propose a Bidirectional Frequency-domain Modulation Enhancement (BFME) module that effectively exploits bidirectional modulation between high- and low-frequency components and learns the spatial weights of high-frequency features to attenuate noise and preserve crack edge details, thereby improving the performance of crack segmentation. Furthermore, the Progressive Guidance Fusion module serves as another core component of our framework. It leverages the spatial prior provided by the original low-resolution image to guide feature refinement via stepwise optimization from coarse contours to fine edges, thereby ensuring the integrity of crack segmentation. Evaluation on three publicly available datasets—CrackTree260, CrackLS315, and Crack760—affirms the superior segmentation accuracy of the proposed BFMPF-Net compared to current mainstream methods.**

I. INTRODUCTION

Structural safety monitoring relies on accurate crack identification—as a direct indicator of damage, it plays a central role in robotic inspection tasks for tunnels, bridges, and roads[1], [2], [3]. In robotic vision applications, accurate crack segmentation is critical for autonomous maintenance and safety assessment, as precise detection of crack propagation directly informs maintenance decisions. Recent advances in deep learning–based semantic segmentation have greatly enhanced robotic systems' ability for precise crack detection and delineation [4], [5]. Despite these improvements, existing methods still face two major challenges:

1) Exploiting the modulation relationship between frequency domains is key to improving crack detection accuracy. Some studies [6], [7] indicate that wavelet transforms are highly effective for segmentation tasks. Additionally, to improve segmentation performance, some studies [8], [9] innovatively introduce attention mechanisms to process high- and low-frequency components in a differentiated manner. However, existing methods still inadequately fuse and utilize high- and low-frequency components. Specifically, local edge and detail information is derived primarily from high-frequency components, which facilitates accurate crack localization yet remains susceptible to interference from complex backgrounds. Low-frequency components, on the other hand, capture contextual information, but may be degraded by extraneous background noise. Therefore, effectively integrating frequency-domain information to achieve optimal crack segmentation remains an unresolved challenge.

2) Existing crack segmentation networks fail to fully retain critical feature details during transmission. Most existing crack segmentation methods are based on the U-Net architecture, where image features are extracted using Convolutional Neural Networks (CNNs) and Transformers, with feature fusion strategies employed [10], [11], [12], [13], [14] to enhance segmentation performance and feature representation. As network depth increases, the extracted features often undergo information attenuation, reducing their effectiveness in subsequent processing. In addition, existing approaches remain insufficient in mitigating feature dilution and fail to fully preserve critical information. Consequently, effectively retaining key features during transmission remains a central challenge in crack segmentation research.

To address these challenges, we propose a Bidirectional Frequency-domain Modulation Progressive Fusion Network (BFMPF-Net) for road crack segmentation. Specifically, we propose a Bidirectional Frequency-domain Modulation Enhancement (BFME) module, which utilizes the bidirectional modulation between high- and low-frequency components to alleviate noise interference, while simultaneously enhancing the representation capability of crack edge details, thereby improving segmentation performance. Furthermore, a Progressive Guidance Fusion (PGF) module is developed to improve the accuracy and reliability of information in the encoder-decoder layers. It employs a coarse-to-fine progressive strategy to integrate layer-wise features with low-resolution features.

The major contributions of this study can be encapsulated

This work is partially supported by the National Natural Science Foundation of China (No. 62576192), the Natural Science Foundation of Hubei Province of China (JCZRMS202601159), and the Major Special Project of China Innovation Challenge (Ningbo) (2024T008).

Wen Yang, Yingying Zheng and Hang Sun are with the College of Computer and Information Technology, China Three Gorges University, Yichang, China, and also with the Hubei Key Laboratory of Intelligent Vision Based Monitoring for Hydroelectric Engineering, China Three Gorges University, Yichang, China. (e-mail: yangwen0720@163.com, zhengyingying00@163.com, sunhang@ctgu.edu.cn)

Chao Liang is with the Hubei Provincial Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan, China. (e-mail: cliang@whu.edu.cn)

Lei Fang is with the CAAZ (Zhejiang) Information Technology Co., Ltd., Ningbo, China. (e-mail: icedark@zju.edu.cn)

*: Corresponding author.

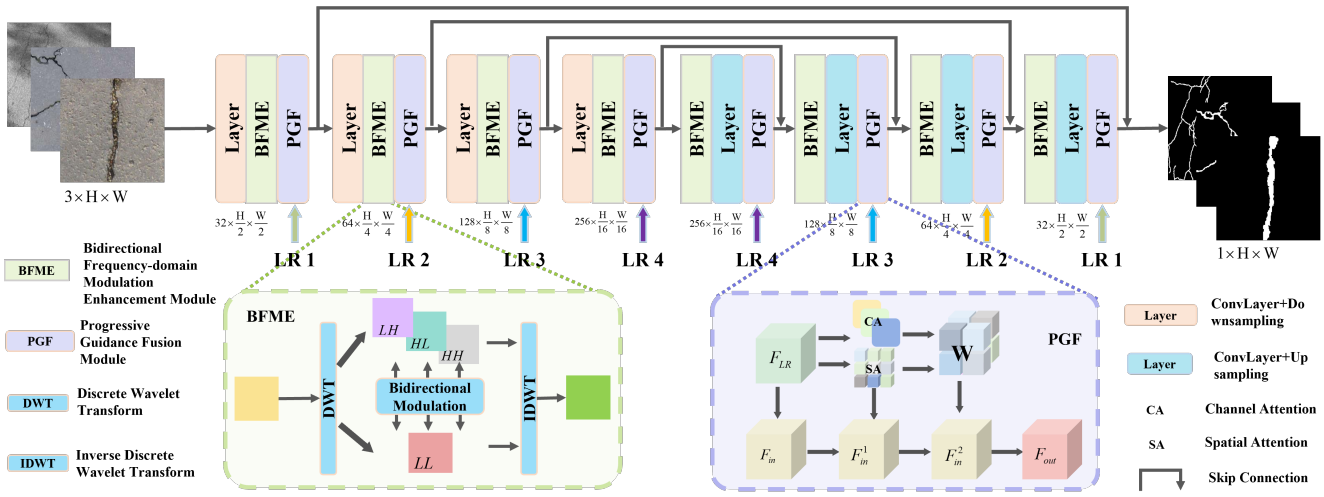

Fig. 1. Architecture of BFMPF-Net with Bidirectional Frequency-domain Modulation Enhancement (BFME) module and the Progressive Guidance Fusion (PGF) module.

as follows:

1) In the BFME module, the modulation mechanism between high- and low-frequency features is fully activated, thereby enriching feature representations and ultimately improving crack segmentation performance.

2) To improve crack feature extraction and enhance segmentation quality, this paper designs the PGF module, which leverages low-resolution features for progressive optimization.

3) The proposed BFMPF-Net innovatively integrates the BFME and PGF modules. Experiments on the CrackTree260 [15], CrackLS315 [16], and Crack760 [17] datasets validate its effectiveness: leveraging superior frequency-domain modulation and a low-resolution progressive fusion mechanism, BFMPF-Net outperforms state-of-the-art segmentation methods across multiple metrics.

II. RELATED WORKS

A. Traditional Methods

Before the widespread adoption of deep learning, the mainstream approaches for crack detection were classical image processing techniques [18]. For instance, Del Río-Barral et al. [19] used a region-growing algorithm for 3D point clouds that suppresses road slope interference, showing greater robustness in complex scenarios. Yuan et al. [20] propose a method for pavement transverse crack identification that reconstructs vehicle vibration signals into two-dimensional images, extracts edge detection features, and integrates them with the RUSBoost classifier. Lei et al. [21] combined segmentation loss with a dynamic threshold binarization objective by introducing a dynamic thresholding branch (DTB) to regress each crack image's optimal threshold. Experiments demonstrate improved segmentation accuracy.

B. Learning-based Segmentation Methods

The rapid progress of deep learning in image processing has led to its widespread adoption in crack segmentation [22], [23], [24], addressing the inherent limitations of conventional methods. For example, Xu et al. [25] use Transformers for long-range dependencies and local modules for fine-grained features, boosting crack detection. Guo et al. [26] incorporated an attention mechanism (STA) for pixel-level pavement crack detection, which outperformed conventional CNNs on multiple public datasets. Liu et al. [27] introduce an upgraded CrackFormer-II network for pavement crack segmentation that integrates Transformer encoders with relative positional embeddings in a SegNet-style encoder-decoder. Subsequently, Zhou et al. [28] proposed SCDeepLab, a hybrid segmentation algorithm combining the Swin Transformer and CNNs in the DeepLabv3+ framework to detect tunnel lining cracks. Yang et al. [8] proposed the two-stage spatial-frequency fusion network SFFNet, which fuses multi-features via WTFD-based frequency decomposition and MDAF-based feature alignment to segment remote sensing images, effectively handling grayscale-varying regions with spatial information preserved. Goo et al. [29] introduce Hybrid-Segmentor, a CNN-Transformer model that combines ResNet-50 and SegFormer to capture local details and global contexts for precise crack segmentation across surfaces. Additionally, Sun et al. [6] utilize the wavelet transform to extract multiple frequency domain components of an image, thereby achieving decent segmentation performance for complex crack images. Qi et al. [30] introduced UltraFastCrackSeg, a lightweight real-time crack segmentation model using task-oriented self-supervised pretraining to achieve high performance with minimal computational cost.

III. THE PROPOSED METHOD

The proposed Bidirectional Frequency-domain Modulation Progressive Fusion Network (BFMPF-Net) is illustrated in Fig. 1. Its overall architecture follows a U-Net framework.

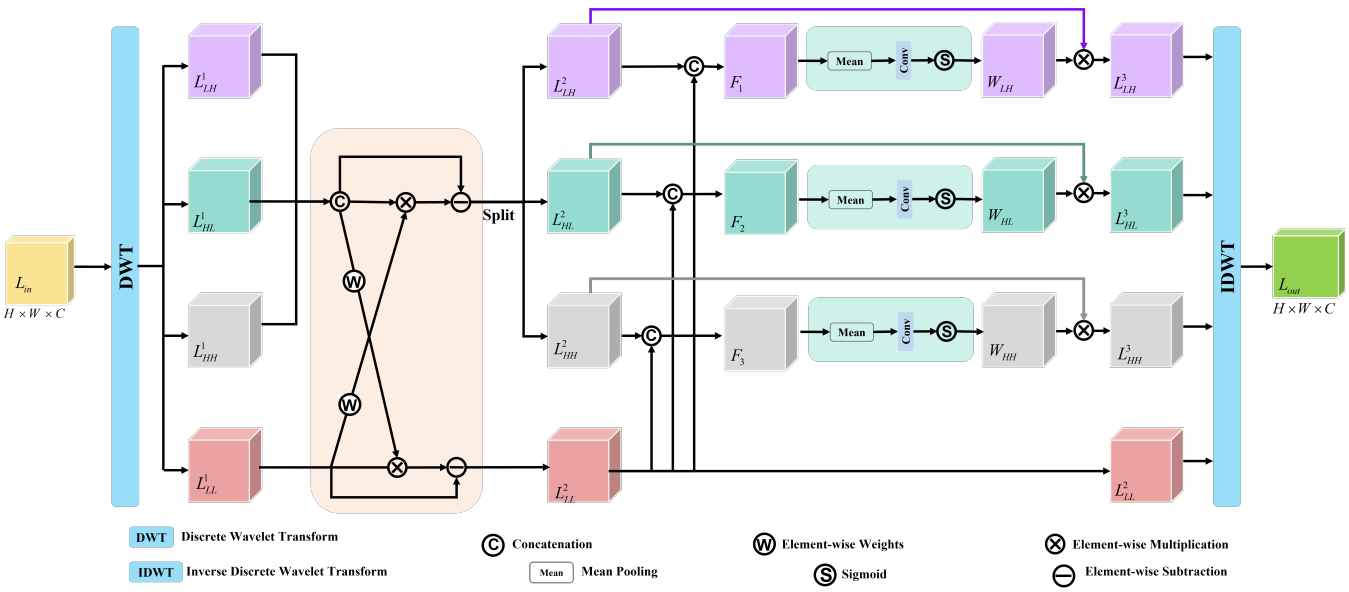

Fig. 2. Bidirectional Frequency-domain Modulation Enhancement (BFME) module.

The input image is first processed by convolutional layers and the BFME module to extract features. The resulting features, together with the downsampled original image, are subsequently passed to the PGF module. Skip connections are employed to link corresponding encoder and decoder layers throughout each encoding–decoding stage. In the following subsections, we detail the two key components: the BFME module and the PGF module.

A. Bidirectional Frequency-domain Modulation Enhancement Module

Wavelet transforms are commonly leveraged in frequency-based crack segmentation methods [8], [9] to extract high-frequency texture information and low-frequency global background. However, they often overlook the correlation between these components, leading to inaccurate detection. To overcome this limitation, the BFME module is developed to enhance feature representation by capturing cross-frequency interactions. Fig. 2 shows the overall architecture of the BFME module.

Applying the Discrete Wavelet Transform (DWT) [31] to the input feature map $L_{in} \in \mathbb{R}^{H\times W\times C}$ (with height H, width W, and channels C) yields four sub-bands, formulated as:

$$L^1_{LL}, \{L^1_{LH}, L^1_{HL}, L^1_{HH}\} = DWT(L_{in}) \tag{1}$$

After wavelet decomposition, the structural information of cracks is mainly concentrated in the low-frequency component L^1_{LL}, while edge and texture details are distributed among the three high-frequency components L^1_{LH}, L^1_{HL}, and L^1_{HH}.

Firstly, the three directional high-frequency sub-bands L^1_{LH}, L^1_{HL}, and L^1_{HH} are concatenated to form the composite feature L_H, which facilitates unified modeling of high-frequency information. This operation is defined as:

$$L_H = Concat(L^1_{LH}, L^1_{HL}, L^1_{HH}) \tag{2}$$

Then, the low-frequency feature L^1_{LL} is processed through a convolutional layer followed by a sigmoid activation to generate weights that modulate L_H, thereby producing a suppression term z_H which attenuates noise-like high-frequency components that are inconsistent with the global semantics. Similarly, the composite high-frequency feature L_H modulates the low-frequency feature L^1_{LL}, yielding a correction term z_{LL} that effectively eliminates redundant background information and thus enhances the crack-region recognition capability of the low-frequency representation. To achieve adaptive control, learnable gating coefficients λ^{LtoH} and λ^{HtoL} are introduced to balance the information exchange between the two domains. This operation is defined as:

$$\begin{cases} z_H = Sigmoid(Conv_{1\times1}(L^1_{LL})) \times \lambda^{LtoH} \times L_H \\ z_{LL} = Sigmoid(Conv_{1\times1}(L_H)) \times \lambda^{HtoL} \times L^1_{LL} \end{cases} \tag{3}$$

here, $\times$ represents element-wise multiplication, with $Sigmoid(\cdot)$ denoting the sigmoid function.

Subsequently, subtracting these modulation terms yields the refined high-frequency feature h_H and low-frequency feature L^2_{LL}, providing more refined representations and a more reliable foundation for the subsequent processing.

$$h_H = L_H - z_H \tag{4}$$

$$L^2_{LL} = L^1_{LL} - z_{LL} \tag{5}$$

To further leverage frequency domain information, we decompose the high-frequency feature h_H into three directional

sub-bands, which are then combined with the low-frequency component to generate the features F_1, F_2, and F_3, which can be expressed as:

$$L^2_{LH}, L^2_{HL}, L^2_{HH} = Split(h_H) \tag{6}$$

$$\begin{cases} F_1 = Concat(L^2_{LH}, L^2_{LL}) \\ F_2 = Concat(L^2_{HL}, L^2_{LL}) \\ F_3 = Concat(L^2_{HH}, L^2_{LL}) \end{cases} \tag{7}$$

Pooling operations are then independently applied to F_1, F_2, and F_3, producing one pixel-wise descriptors for each. The formalization is as follows:

$$\begin{cases} F_{LH} = Conv_{7\times7}(Mean(F_1)) \\ F_{HL} = Conv_{7\times7}(Mean(F_2)) \\ F_{HH} = Conv_{7\times7}(Mean(F_3)) \end{cases} \tag{8}$$

where *Mean*$(\cdot)$ represents mean pooling. *Conv*$_{7\times7}(\cdot)$ denote 7×7 convolutions. Based on this, F_{LH}, F_{HL}, and F_{HH} are converted into spatial weight maps that quantify the importance of each region in the current feature representation, thereby boosting salient information. The detailed procedure is formulated as:

$$\begin{cases} L^3_{LH} = Sigmoid(F_{LH}) \times L^2_{LH} \\ L^3_{HL} = Sigmoid(F_{HL}) \times L^2_{HL} \\ L^3_{HH} = Sigmoid(F_{HH}) \times L^2_{HH} \end{cases} \tag{9}$$

here, by applying $Sigmoid(\cdot)$, the representation of local detail features is enriched, leading to enhanced crack details within the frequency domain.

Finally, the Inverse Discrete Wavelet Transform is employed as a reconstruction method to convert the enhanced sub-bands L^3_{LH}, L^3_{HL}, L^3_{HH}, and L^2_{LL} from the frequency domain back to the spatial domain, generating the output feature map $L_{out} \in \mathbb{R}^{H\times W\times C}$.

$$L_{out} = IDWT\left(L^3_{LH}, L^3_{HL}, L^3_{HH}, L^2_{LL}\right) \tag{10}$$

By effectively leveraging bidirectional frequency-domain modulation and enhancing high-frequency texture details in crack images, the BFME module improves crack recognition capability and thereby enhances segmentation performance.

B. Progressive Guidance Fusion Module

In encoder-decoder architectures, the information preserved tends to diminish as network depth increases. This degradation results from repeated feature transformations and extractions within the network, which can progressively erode fine-grained details and high-level semantic context. To mitigate this drawback, we introduce a Progressive Guidance Fusion (PGF) module. This module utilizes the original low-resolution image and progressively refines crack information through a coarse-to-fine fusion strategy. Specifically, low-resolution features are initially integrated with encoder-decoder layer features, after which the fused representation is gradually refined through a weighting mechanism. This design not only enriches the overall feature representation but also enhances the accuracy of crack segmentation. The detailed architecture is illustrated in Fig. 3.

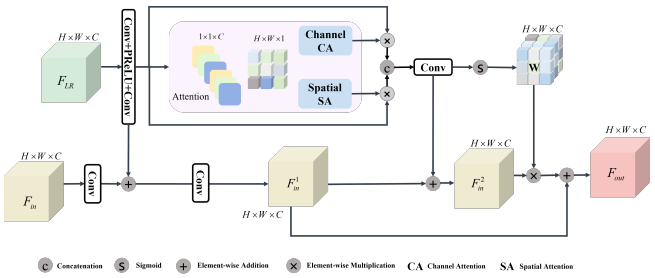

Fig. 3. Progressive Guidance Fusion (PGF) module.

The input to the module comprises two feature sets: $F_{in} \in \mathbb{R}^{H\times W\times C}$ from the encoder or decoder layers, and $F_{LR} \in \mathbb{R}^{H\times W\times C}$, a low-resolution representation produced by repeatedly downsampling the original image. Effective fusion necessitates that F_{LR} and F_{in} share identical spatial dimensions and channel counts.

First, F_{LR} is processed by a 3×3 convolution, PReLU activation, and another 3×3 convolution, resulting in F^1_{LR}. Subsequently, F^1_{LR} is combined with the input feature F_{in} through element-wise addition to obtain F^1_{in}. This operation aligns the downsampled features with the current features to achieve an initial coarse fusion, improving the reliability of the fused representation.

$$F^1_{LR} = Conv_{3\times3}(PReLU(Conv_{3\times3}(F_{LR}))) \tag{11}$$

$$F^1_{in} = Conv_{1\times1}(F^1_{LR} + Conv_{1\times1}(F_{in})) \tag{12}$$

here, $PReLU(\cdot)$ denotes the activation function, whereas $Conv_{3\times3}(\cdot)$ and $Conv_{1\times1}(\cdot)$ represent the 3×3 and 1×1 convolution.

Subsequently, the intermediate feature F^1_{LR} is refined using spatial and channel attention. This mechanism is designed to selectively highlight crack-related salient features and enhance the representation of key regions. The outputs of the spatial and channel attention branches are concatenated to retain complementary multi-dimensional information to support fine-grained feature fusion.

$$F^2_{LR} = Conv_{1\times1}\Big(Concat\big(CA(F^1_{LR}), SA(F^1_{LR})\big)\Big) \tag{13}$$

here, $CA(\cdot)$ and $SA(\cdot)$ denote the channel attention and spatial attention mechanisms, respectively.

Next, the input feature F^1_{in} is fused with F^2_{LR}, generating $F^2_{in} \in \mathbb{R}^{H\times W\times C}$. Through this operation, the representation learned from encoder–decoder layers is complemented. Such integration enables the low-resolution branch to provide effective guidance, encouraging the network to focus on crack-relevant regions.

$$F^2_{in} = F^2_{LR} + F^1_{in} \tag{14}$$

Finally, F_{in}^2 is refined by adaptively weighting each feature according to its relative importance. This step realizes a coarse-to-fine integration process, where informative features are gradually emphasized and merged with the encoder–decoder representations. By integrating F_{in}^1 through residual combination, the final output F_{out} preserves progressively refined representations, thereby enhancing the completeness and accuracy of crack segmentation.

$$F_{out} = \left(Sigmoid(F_{LR}^2) \times F_{in}^2\right) + F_{in}^1 \tag{15}$$

The PGF module incorporates low-resolution features into the encoder–decoder layers, effectively mitigating the information loss resulting from repeated network transformations. This progressive fusion process preserves essential structural cues from the original image, thereby enhancing the accuracy and completeness of crack segmentation.

C. Loss Function

The training objective is formulated as a combination of Binary Cross-Entropy loss and Dice loss, which serves to optimize the segmentation network and address class imbalance. The precise definitions are as follows:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}(1-t_i)\log(1-p_i) - t_i\log(p_i) \tag{16}$$

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i t_i + \varepsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} t_i + \varepsilon} \tag{17}$$

In the above loss functions, L_{BCE} and L_{Dice} denote the binary cross-entropy loss and Dice loss, respectively. N denotes the total number of pixels in the image, t_i denotes the ground truth of the i-th pixel, and p_i denotes the model prediction for that pixel. To avoid a zero denominator, we introduce a smoothing factor ε into the Dice loss and set its value to 1.

The total training objective is formulated by combining the BCE loss and Dice loss, yielding the overall loss function:

$$L = L_{BCE} + L_{Dice} \tag{18}$$

IV. EXPERIMENTS AND ANALYSIS

This section presents quantitative and qualitative comparisons between our method and nine state-of-the-art segmentation approaches, namely FCN [22], Deeplabv3+ [23], DeepCrack [16], UT-Net [32], FAT-Net [33], DeepCrackAT [24], DECS-Net [9], APF-Net [34], and Hybrid-Segmentor [29], on the road crack datasets CrackTree260 [15], CrackLS315 [16], and Crack760 [17].

A. Experimental Configuration

To evaluate the proposed crack segmentation approach, three public pavement crack datasets were used. The CrackTree260 dataset [15] contains 260 images at 640×480 resolution, captured under visible light with a region-array camera, ensuring authentic crack features. CrackLS315 [16] consists of 315 high-resolution images ranging from 640×480 to 1920×1080 pixels, acquired with a line-array camera and laser illumination to enhance contrast and feature extraction. Crack760 [17] includes 760 images with diverse crack morphologies, from simple to complex patterns and wide cracks, some partially obscured by debris—making it more challenging and realistic. An 8:2 random split is applied to each dataset, and all images are cropped to 256×256 pixels for network input.

In the training phase, the Adam optimizer is employed to update network parameters, with an initial learning rate set to 0.0001 and progressively adjusted through a cosine annealing strategy. Training is performed for 200 epochs, processing two samples per iteration. All experiments are carried out using the PyTorch framework. To improve model generalization, data augmentation techniques are applied exclusively to the training set, while the test set is left unaltered.

Model performance is evaluated using four metrics: Precision (Pr), defined as the proportion of correctly predicted crack pixels among all pixels predicted as cracks; Recall (Re), defined as the proportion of correctly identified crack pixels among all ground truth crack pixels; F1 Score (F1), the harmonic mean of Precision and Recall, which comprehensively reflects their trade-off; and Mean Intersection over Union (MIoU), which measures the ratio between the intersection and union of predicted segmentation regions and ground truth annotations, and is calculated only on the crack category, providing an overall assessment of segmentation accuracy.

B. Comparison Experiments

Quantitative comparisons among the evaluated algorithms are presented in Table I and Table II, which report Precision, Recall, F1 Score, and MIoU for each method. The optimal value for each metric is highlighted in bold. Furthermore, the last two columns of Table I provide the Multiply-Accumulate Operations (MACs) and parameter counts for each model.

The data in Table I demonstrate that BFMPF-Net achieves the best performance across all four evaluation metrics on the CrackTree260 dataset: Precision of 81.44%, Recall of 84.30%, F1 Score of 82.85%, and MIoU of 70.72%. Compared with APF-Net and Hybrid-Segmentor, our method outperforms them by 3.91% and 2.90% in Precision, 2.11% and 0.57% in Recall, 3.06% and 1.80% in F1 Score, and 4.34% and 2.58% in MIoU, respectively. This performance improvement stems from two key designs: the BFME module, which effectively exploits the modulation between frequency domains to enhance crack detection accuracy, and the PGF module, which refines segmentation continuity through progressive fusion of low-resolution features. In addition, Table I also provides a comparison of parameter counts and computational costs among the models. BFMPF-Net achieves optimal segmentation performance while maintaining the lowest computational cost and one of the smallest parameter sizes, demonstrating a favorable efficiency-performance balance that makes it well-suited for deployment in robotic vision systems.

TABLE I

QUANTITATIVE PERFORMANCE COMPARISON AND ANALYSIS OF VARIOUS SEGMENTATION METHODS ON THE CRACKTREE260 AND CRACKLS315 DATASETS.

Model	Year	CrackTree260				CrackLS315				MACs	Params
		Pr(%)	Re(%)	F1(%)	MIoU(%)	Pr(%)	Re(%)	F1(%)	MIoU(%)	(G)	(M)
FCN [22]	2015	72.23	70.35	71.28	55.37	54.47	59.72	56.97	39.83	37.132	35.307
Deeplabv3+ [23]	2018	72.54	71.59	72.06	56.33	56.61	53.79	55.17	38.09	43.447	41.994
DeepCrack [16]	2018	71.85	74.42	73.11	57.62	57.27	52.34	54.69	37.64	136.804	30.905
UT-Net [32]	2021	75.16	76.35	75.75	60.97	57.70	58.68	58.19	41.03	20.486	14.407
FAT-Net [33]	2022	73.76	82.60	77.93	63.84	56.63	56.77	56.70	39.57	42.800	29.615
DeepCrackAT [24]	2023	76.85	81.73	79.21	65.58	60.41	57.81	59.08	41.92	93.719	13.328
DECS-Net [9]	2024	76.47	81.25	78.79	65.00	60.27	57.31	58.75	41.60	26.085	68.865
APF-Net [34]	2024	77.53	82.19	79.79	66.38	61.77	58.38	60.02	42.88	**10.106**	7.542
Hybrid-Segmentor [29]	2025	78.54	83.73	81.05	68.14	63.00	58.69	60.77	43.65	268.162	226.805
Proposed	—	**81.44**	**84.30**	**82.85**	**70.72**	**64.48**	**60.14**	**62.24**	**45.18**	11.378	**5.801**

TABLE II

PERFORMANCE COMPARISON OF VARIOUS METHODS ON THE CRACK760 DATASET.

Model	Pr(%)	Re(%)	F1(%)	MIoU(%)
FCN [22]	76.66	76.40	76.53	61.98
Deeplabv3+ [23]	76.86	82.84	79.74	66.30
DeepCrack [16]	75.03	83.63	79.09	65.42
UT-Net [32]	74.55	82.77	78.44	64.53
FAT-Net [33]	78.42	76.33	77.36	63.08
DeepCrackAT [24]	77.47	82.76	80.03	66.71
DECS-Net [9]	80.12	81.52	80.81	67.80
APF-Net [34]	81.24	81.41	81.32	68.53
Hybrid-Segmentor [29]	79.36	83.72	81.48	68.75
Proposed	**82.21**	**84.58**	**83.38**	**71.49**

Experimental Results on CrackLS315 and Crack760: The data in Table I and Table II show that BFMPF-Net ranks first across all evaluation metrics on both datasets. Specifically, on the CrackLS315 dataset, it achieves a Pr of 64.48%, Re of 60.14%, F1 of 62.24%, and MIoU of 45.18%; on the Crack760 dataset, the corresponding metrics are 82.21%, 84.58%, 83.38%, and 71.49%, respectively. When compared to Transformer-based models (UT-Net, FAT-Net), a wavelet-transform-based model (DECS-Net), and an encoder-decoder based approach (Hybrid-Segmentor), BFMPF-Net exhibits notable improvements across all evaluation metrics. These improvements, enabled by the BFME and PGF modules, not only boost segmentation performance but also facilitate practical deployment in robotic vision systems.

To intuitively demonstrate performance, representative results obtained from the proposed method and other top-ranking approaches are displayed in Fig. 4. Compared to DECS-Net, APF-Net, and Hybrid-Segmentor, our method yields fewer false positives and false negatives on the CrackTree260 dataset. On the CrackLS315 dataset, it delivers more continuous segmentation, especially for fine cracks with low contrast. On the Crack760 dataset, it better captures both large and fine cracks. This is due to the use of the BFME module in BFMPF-Net, which facilitates the modulation of crack background information and edge textures, thereby optimizing segmentation performance. Moreover, the PGF module integrates low-resolution representations with encoder-decoder layer output features via progressive fusion, leading to more complete segmentation outcomes.

C. Ablation Study

We conduct ablation studies on CrackTree260, CrackLS315, and Crack760 to assess the effectiveness of each key component in BFMPF-Net. Specifically, four network configurations are examined: (a) the baseline model, (b) baseline with BFME, (c) baseline with PGF, and (d) the complete BFMPF-Net. The quantitative results are presented in Table III.

TABLE III

ABLATION STUDY RESULTS OF THE CORE COMPONENTS OF BFMPF-NET ON THREE DATASETS.

Dataset	Model	Pr(%)	Re(%)	F1(%)	MIoU(%)
CrackTree260	(a)	76.32	78.92	77.60	66.39
	(b)	78.93	81.74	80.31	67.10
	(c)	79.53	82.80	81.13	68.25
	(d)	81.44	84.30	82.85	70.72
CrackLS315	(a)	54.27	51.16	52.67	35.75
	(b)	60.52	59.70	60.10	42.96
	(c)	57.86	54.11	55.92	38.81
	(d)	64.48	60.14	62.24	45.18
Crack760	(a)	76.23	72.10	74.11	58.87
	(b)	76.45	83.80	79.95	66.60
	(c)	80.79	83.73	82.24	69.83
	(d)	82.21	84.58	83.38	71.49

1) Contribution of the BFME module. A comparison between configurations (a) and (b) in Table III reveals the performance gains brought by the BFME module. The incorporation of BFME leads to substantial improvements across Precision, Recall, F1 Score, and MIoU, confirming its efficacy. To further validate the contribution of the BFME module, we incrementally integrate various components (Indices 1–6) corresponding to the encoder and decoder at the same resolution. As shown in Table IV, the gradual incorporation of the BFME module (Indices 2–6) leads to consistent gains across all metrics compared to Index 1. These results demonstrate that the BFME module, through

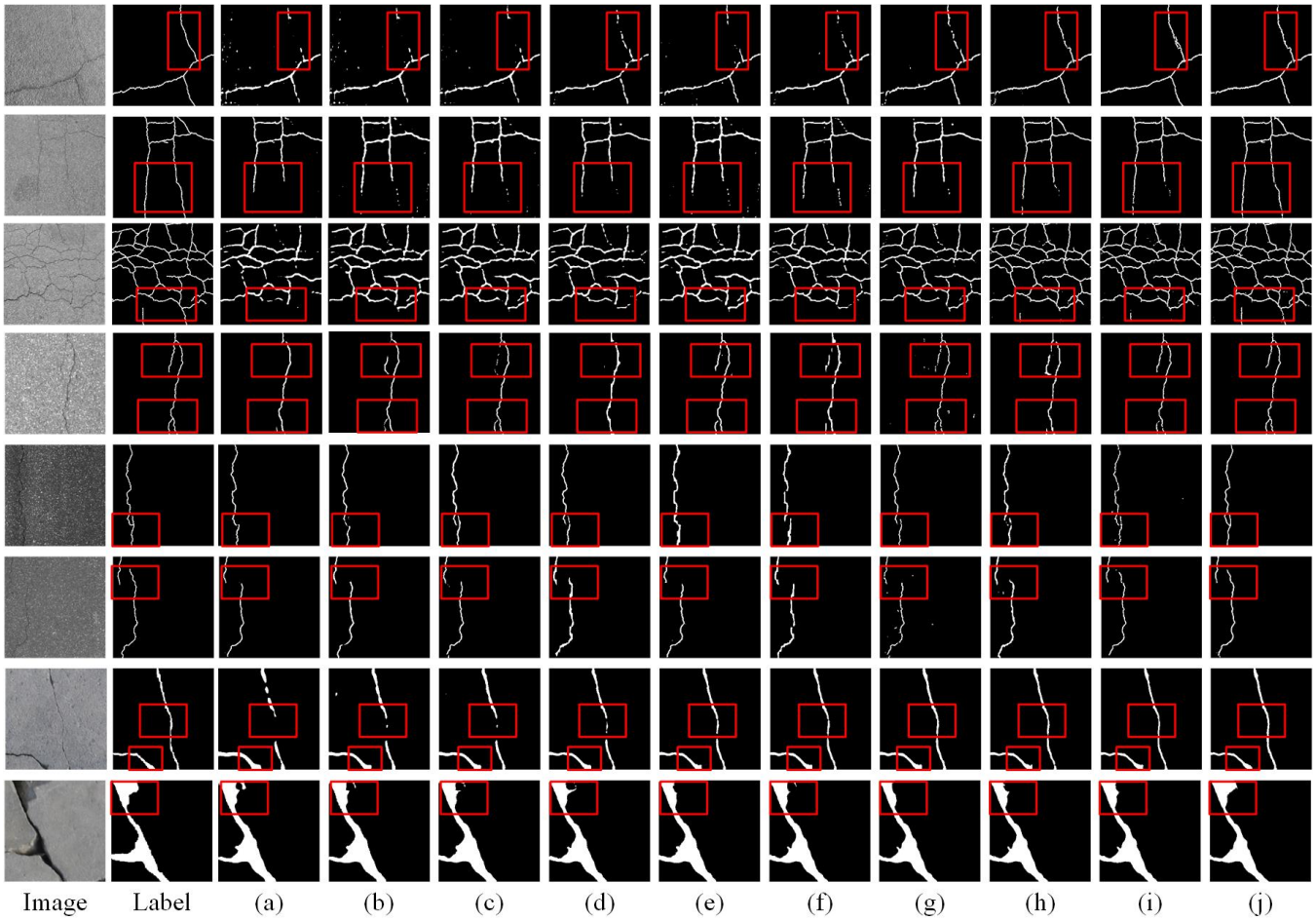

Fig. 4. Segmentation results visualization: Rows 1–3 (CrackTree260), rows 4–6 (CrackLS315), rows 7–8 (Crack760). Columns (a)–(j) correspond to FCN [22], Deeplabv3+ [23], DeepCrack [16], UT-Net [32], FAT-Net [33], DeepCrackAT [24], DECS-Net [9], APF-Net [34], Hybrid-Segmentor [29], and our BFMPF-Net, respectively.

its effective modulation of high- and low-frequency features, substantially improves crack detection accuracy and consequently enhances segmentation outcomes.

TABLE IV
ABLATION EXPERIMENTS OF THE BFME MODULE ON THE CRACKLS315 DATASET.

Index	BFME 1	BFME 2	BFME 3	BFME 4	Pr (%)	Re (%)	F1 (%)	MIoU (%)
1	×	×	×	×	54.27	51.16	52.67	35.75
2	✓	×	×	×	56.84	54.12	55.40	38.31
3	×	✓	×	×	58.75	53.46	55.98	38.87
4	×	×	✓	×	55.98	56.22	56.10	38.99
5	×	×	×	✓	54.91	53.89	54.39	37.35
6	✓	✓	✓	✓	60.52	59.70	60.10	42.96

2) Contribution of the PGF module. A comparison between configuration (a) and configuration (c) reveals substantial performance improvements attributable to the PGF module. With the inclusion of PGF, configuration (c) achieves consistently better results than configuration (a) across all metrics, providing strong evidence for the efficacy of the proposed module. To further validate the contribution of the PGF module, we incrementally integrate various components (Indices 1–6) corresponding to the encoder and decoder at the same resolution. Table V shows that as module components are progressively added (Indices 2 to 6), all evaluation metrics exhibit consistent improvements, steadily surpassing the baseline level. This phenomenon stems from the unique progressive fusion mechanism of the PGF module, which leverages original low-resolution features as spatial priors to

TABLE V
ABLATION EXPERIMENTS OF THE PGF MODULE ON THE CRACKLS315 DATASET.

Index	PGF 1	PGF 2	PGF 3	PGF 4	Pr (%)	Re (%)	F1 (%)	MIoU (%)
1	×	×	×	×	54.27	51.16	52.67	35.75
2	✓	×	×	×	54.41	53.28	53.84	36.84
3	×	✓	×	×	56.55	52.46	54.43	37.39
4	×	×	✓	×	55.61	53.35	54.46	37.42
5	×	×	×	✓	54.75	52.62	53.66	36.67
6	✓	✓	✓	✓	57.86	54.11	55.92	38.81

calibrate encoder-decoder features layer by layer. This allows detailed crack information to be preserved during feature propagation.

V. CONCLUSIONS

The proposed BFMPF-Net presents a novel solution for road crack segmentation. Its core innovation lies in the synergistic design of the BFME and PGF modules: the former enhances the model's sensitivity to crack edges by exploiting the bidirectional modulation between high- and low-frequency features, while the latter optimizes feature transmission efficiency during the encoding-decoding process through progressive fusion of low-resolution representations. Evaluation results on three public datasets demonstrate that the proposed method outperforms existing mainstream segmentation algorithms across multiple metrics. The current work primarily focuses on single-modal image data; future research will extend to crack segmentation in multi-modal scenarios to further improve the model's adaptability in complex environments, laying a foundation for its practical deployment in robotic vision systems.

REFERENCES

[1] Z. Yu, Q. Chen, Y. Shen, and Y. Zhang, "Robust pavement crack segmentation network based on transformer and dual-branch decoder," *Construction and Building Materials*, vol. 453, p. 139026, 2024.

[2] T. S. Tran, S. D. Nguyen, H. J. Lee, and V. P. Tran, "Advanced crack detection and segmentation on bridge decks using deep learning," *Construction and Building Materials*, vol. 400, p. 132839, 2023.

[3] K. Yang, Y. Bao, J. Li, T. Fan, and C. Tang, "Deep learning-based yolo for crack segmentation and measurement in metro tunnels," *Automation in Construction*, vol. 168, p. 105818, 2024.

[4] P. Geng, J. Lu, H. Ma, and G. Yang, "Crack segmentation based on fusing multi-scale wavelet and spatial-channel attention," *Structural Durability & Health Monitoring (SDHM)*, vol. 17, no. 1, 2023.

[5] L. Yang, H. Huang, S. Kong, Y. Liu, and H. Yu, "Paf-net: A progressive and adaptive fusion network for pavement crack segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 12 686–12 700, 2023.

[6] R. Sun, X. Li, L. Zhang, Y. Su, J. Di, and G. Liu, "Wavelet-integrated deep neural network for deblurring and segmentation of crack images," *Mechanical Systems and Signal Processing*, vol. 224, p. 112240, 2025.

[7] G. Wei, J. Xu, W. Yan, Q. Chong, H. Xing, and M. Ni, "Dual-domain fusion network based on wavelet frequency decomposition and fuzzy spatial constraint for remote sensing image segmentation," *Remote Sensing*, vol. 16, no. 19, p. 3594, 2024.

[8] Y. Yang, G. Yuan, and J. Li, "Sffnet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[9] J. Zhang, Z. Zeng, P. K. Sharma, O. Alfarraj, A. Tolba, and J. Wang, "A dual encoder crack segmentation network with haar wavelet-based high–low frequency attention," *Expert Systems with Applications*, vol. 256, p. 124950, 2024.

[10] S. Bai, L. Yang, Y. Liu, and H. Yu, "Dmf-net: A dual-encoding multi-scale fusion network for pavement crack detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 5981–5996, 2024.

[11] J. Wang, Z. Zeng, P. K. Sharma, O. Alfarraj, A. Tolba, J. Zhang, and L. Wang, "Dual-path network combining cnn and transformer for pavement crack segmentation," *Automation in Construction*, vol. 158, p. 105217, 2024.

[12] C. Han, T. Ma, J. Huyan, X. Huang, and Y. Zhang, "Crackw-net: A novel pavement crack image segmentation convolutional neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 22 135–22 144, 2021.

[13] L. Shi, R. Zhang, Y. Wu, D. Cui, N. Yuan, J. Liu, and Z. Ji, "Ahc-net: a road crack segmentation network based on dual attention mechanism and multi-feature fusion," *Signal, Image and Video Processing*, vol. 18, no. 6, pp. 5311–5322, 2024.

[14] J. Wang, H. Yao, J. Hu, Y. Ma, and J. Wang, "Dual-encoder network for pavement concrete crack segmentation with multi-stage supervision," *Automation in Construction*, vol. 169, p. 105884, 2025.

[15] Q. Zou, Y. Cao, Q. Li, Q. Mao, and S. Wang, "Cracktree: Automatic crack detection from pavement images," *Pattern Recognition Letters*, vol. 33, no. 3, pp. 227–238, 2012.

[16] Q. Zou, Z. Zhang, Q. Li, X. Qi, Q. Wang, and S. Wang, "Deepcrack: Learning hierarchical convolutional features for crack detection," *IEEE transactions on image processing*, vol. 28, no. 3, pp. 1498–1512, 2018.

[17] 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.

[18] N. Senthilkumaran and S. Vaithegi, "Image segmentation by using thresholding techniques for medical images," *Computer Science & Engineering: An International Journal*, vol. 6, no. 1, pp. 1–13, 2016.

[19] P. del Río-Barral, M. Soilán, S. M. González-Collazo, and P. Arias, "Pavement crack detection and clustering via region-growing algorithm from 3d mls point clouds," *Remote Sensing*, vol. 14, no. 22, 2022.

[20] W. Yuan and Q. Yang, "Identification of asphalt pavement transverse cracking based on 2d reconstruction of vehicle vibration signal and edge detection algorithm," *Construction and Building Materials*, vol. 408, p. 133788, 2023.

[21] Q. Lei, J. Zhong, C. Wang, Y. Xia, and Y. Zhou, "Dynamic thresholding for accurate crack segmentation using multi-objective optimization," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2023, pp. 389–404.

[22] 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*. IEEE, 2015, pp. 3431–3440.

[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[24] Q. Lin, W. Li, X. Zheng, H. Fan, and Z. Li, "Deepcrackat: An effective crack segmentation framework based on learning multi-scale crack features," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106876, 2023.

[25] "Pavement crack detection from ccd images with a locally enhanced transformer network," *International Journal of Applied Earth Observation and Geoinformation*, vol. 110, p. 102825, 2022.

[26] "A novel transformer-based network with attention mechanism for automatic pavement crack detection," *Construction and Building Materials*, vol. 391, p. 131852, 2023.

[27] H. Liu, J. Yang, X. Miao, C. Mertz, and H. Kong, "Crackformer network for pavement crack segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 9, pp. 9240–9252, 2023.

[28] Z. Zhou, J. Zhang, and C. Gong, "Hybrid semantic segmentation for tunnel lining cracks based on swin transformer and convolutional neural network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 38, no. 17, pp. 2491–2510, 2023.

[29] "Hybrid-segmentor: Hybrid approach for automated fine-grained crack segmentation in civil infrastructure," *Automation in Construction*, vol. 170, p. 105960, 2025.

[30] W. Qi, G. Zhao, F. Ma, M. Liu, and Y. Yang, "Ultrafastcrackseg: A lightweight real-time crack segmentation model with task-oriented pretraining," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 13 486–13 492.

[31] T. Guo, H. Seyed Mousavi, T. Huu Vu, and V. Monga, "Deep wavelet prediction for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. IEEE, 2017, pp. 104–113.

[32] Y. Gao, M. Zhou, and D. N. Metaxas, "Utnet: A hybrid transformer architecture for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 61–71.

[33] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "Fat-net: Feature adaptive transformers for automated skin lesion segmentation," *Medical image analysis*, vol. 76, p. 102327, 2022.

[34] "An attention-based progressive fusion network for pixelwise pavement crack detection," *Measurement*, vol. 226, p. 114159, 2024.