SFPNet-UIE: Spatial-Frequency Perception Network for Underwater Image Enhancement

1st Xiaopeng Liu
College of Computer Science and Engineering
Shandong University of Science and Technology
Qingdao, China
xiaopengliu@sdust.edu.cn

1st Hui Xin
College of Computer Science and Engineering
Shandong University of Science and Technology
Qingdao, China
202482060062@sdust.edu.cn

2nd Cong Liu
School of Science
Nanjing University of Science and Technology
Nanjing, China
LiuC@njust.edu.cn

3rd Long Chen*
Department of Medical Physics and Biomedical Engineering
University College London
London, England, UK
chenlongcv@gmail.com

4th Junyu Dong
Faculty of Information Science and Engineering
Ocean University of China
Qingdao, China
dongjunyu@ouc.edu.cn

***Abstract*—Underwater Image Enhancement (UIE) aims to restore the visual quality of degraded underwater images and is essential for various underwater applications. While most deep learning approaches focus solely on spatial domain restoration, overlooking the valuable information in the frequency domain. To address this limitation, we propose SFPNet-UIE, a Spatial-Frequency Perception Network for Underwater Image Enhancement. SFPNet-UIE incorporates multiple Dynamic Fusion Blocks (DFBs), each containing a Spatial-Frequency Perception Module (SFPM) and Multi-Scale Receptive Field Module (MSRFM). The SFPM employs a dual-path structure, where the spatial branch captures local details and the frequency branch models global features. The MSRFM applies multi-kernel receptive field aggregation to fuse information from both domains while preserving fine details. Experiments on four real-world underwater datasets demonstrate that SFPNet-UIE delivers state-of-the-art (SOTA) or competitive performance across multiple benchmarks. Code is available at https://github.com/xnogjlnla/SFPNet-UIE.**

***Index Terms*—Spatial-Frequency Feature Fusion, Underwater Image Enhancement, Fourier Transform**

I. INTRODUCTION

Underwater Image Enhancement (UIE) [1] plays a vital role in marine engineering and scientific exploration. It provides indispensable data support for seafloor topography mapping [2], marine biology research [3], and remote sensing applications [4]. As underwater images frequently suffer from complex degradation [5] caused by light scattering, absorption, and suspended particles, which pose significant challenges for downstream tasks.

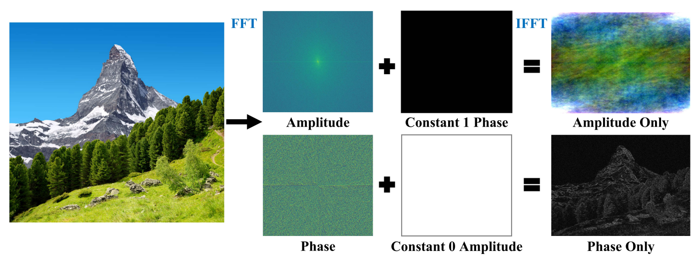

Fig. 1. Reconstructions using real amplitude with a constant phase and real phase with a constant amplitude confirm that, in the frequency domain, amplitude encodes global color and illumination information, whereas the phase component captures global structural details.

Traditional UIE methods based on the physical characteristics of underwater imaging can be broadly divided into two categories: visual prior-based [6], [7] and physical model-based [8]–[10] approaches. However, visual prior-based methods focus on pixel-level adjustments like contrast and saturation, overlook underlying optical degradation mechanisms. Physics model-based methods invert degradation processes by estimating parameters such as transmission maps, but these models often fail under diverse underwater conditions. These methods exhibit limited expressive power and cannot fully capture the complex physical and optical factors of underwater scenes, often leading to suboptimal results.

This work was supported in part by the Ministry of Education of Humanities and Social Science Project under Grant (24YJAZH092), the Key R&D Program of Shandong Province (Grant No. 2024ZLGX06), the Natural Science Foundation of Shandong Province (Grant No. ZR2024ZD04), the Fundamental Research Funds for the Central Universities (30923011022), the Natural Science Foundation of Jiangsu Province (BK20231454), the National Natural Science Foundation of China (62176183) and Jiangsu Planned Projects for Postdoctoral Research Funds. (Corresponding author: Long Chen, chenlongcv@gmail.com.)

In recent years, deep learning-based UIE methods have been introduced [11]–[14]. Some studies integrate deep learning techniques with physical models, leveraging the efficient feature extraction capabilities of convolutional neural networks (CNNs) to estimate parameters within imaging models. Other approaches rely on generative architectures such as generative adversarial networks (GANs) and diffusion models, or the global modeling capabilities of transformers, to directly learn enhancement mappings from data. With the application of large-scale datasets and the introduction of novel network architectures, these methods have demonstrated significantly superior performance and robustness compared to previous techniques. However, most of these methods operate solely in the spatial domain, with limited exploration of frequency domain representations and cross-domain with task-oriented feature fusion interaction, despite the promising progress of frequency-based approaches.

Driven by above challenges, we explore image decoupling and representation learning from the frequency domain perspective. As shown in Fig. 1, an image is transformed using the Fast Fourier Transform (FFT) to obtain amplitude and phase components. Reconstructing with the real amplitude and a constant phase through the Inverse Fast Fourier Transform (IFFT) preserves global color and illumination, while reconstructing with the real phase and a constant amplitude retains global structural details. This confirms that amplitude encodes color and illumination, while phase encodes global spatial structure. These findings highlight that FFT provides inherent global modeling capabilities, suggesting that frequency domain prior offer an efficient way to address underwater degradation.

Inspired by these insights, we propose SFPNet-UIE, a spatial-frequency perception network for underwater image enhancement that integrates both spatial domain cues and frequency domain prior to better handle underwater image degradation. SFPNet-UIE incorporates novel Dynamic Fusion Blocks (DFBs) to jointly capture and fuse spatial and frequency features. Each DFB contains a Spatial–Frequency Perception Module (SFPM), where the spatial branch captures local details and the frequency branch models global structure, illumination, and color, as well as a Multi-Scale Receptive Field Module (MSRFM) that uses multi-kernel receptive fields to preserve fine textures. Our main contributions are:

- We propose SFPNet-UIE, a deep framework that employs Dynamic Fusion Blocks (DFBs) to fuse spatial and frequency features and perform multi-scale aggregation, effectively preserving fine details for underwater image enhancement.
- We design a novel Spatial–Frequency Perception Module with a dual-path architecture. The spatial branch captures fine local details and textures, while the frequency branch models global structures, illumination, and color information, enabling a more comprehensive feature representation for underwater image enhancement.
- Extensive experiments show that SFPNet-UIE achieves state-of-the-art performance on multiple real-world underwater image datasets, with fair comparisons against other leading methods shown in Fig. 2.

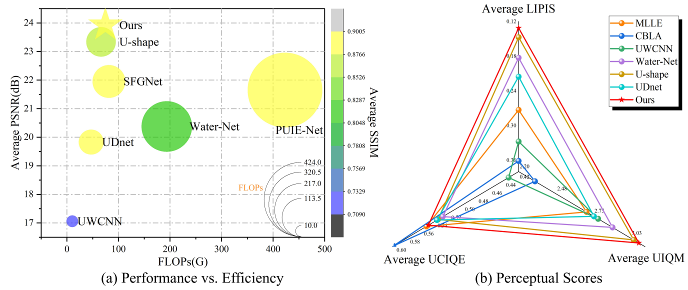

Fig. 2. Quantitative comparisons with other models. (a) Average PSNR and SSIM values between our method and others across four datasets. Regions closer to the top-left corner indicate superior performance. (b) Average scores for three perceptual metrics.

II. Related Work

A. Underwater Image Enhancement

Traditional Methods. Traditional UIE methods are primarily categorized into visual prior-based and methods physical model-based methods. Visual prior-based methods aim to restore degraded underwater images by modifying pixel values such as contrast [15] and retinex-based [16]. Hitam et al. [15] aimed to enhance contrast and suppress noise interference in underwater images by synergistically applying contrast adjustment and adaptive histogram equalization in both RGB and HSV color spaces. Physic model-based methods treat enhancement as an imaging inverse problem, establishing a degraded model and utilizing constraints such as the underwater dark channel prior [17], fuzzy prior [18], and minimum information prior [19]. However, the complex and dynamic nature of underwater environments imposes significant limitations on these methods, including insufficient precision regarding assumptions about underwater imaging conditions, difficulties in estimating key physical parameters, and limited generalization capabilities in scenarios requiring specific prior knowledge.

Deep learning-based UIE methods. In recent years, scholars have proposed deep learning-based underwater image enhancement methods, such as Qi et al. [8] signed a hybrid convolutional-axial attention module based on Retinex to fuse the strengths of CNNs and Transformers, enabling collaborative modeling of local features and global context in color correction and visibility enhancement. PUIE-Net [11] employs a generative model framework based on conditional variational autoencoders, addressing the inherent blurriness of reference images in UIE through probabilistic distribution modeling and a sampling consensus mechanism. However, most of these methods operate solely in the spatial domain, with limited exploration of frequency domain representations, despite the promising progress of frequency-based approaches.

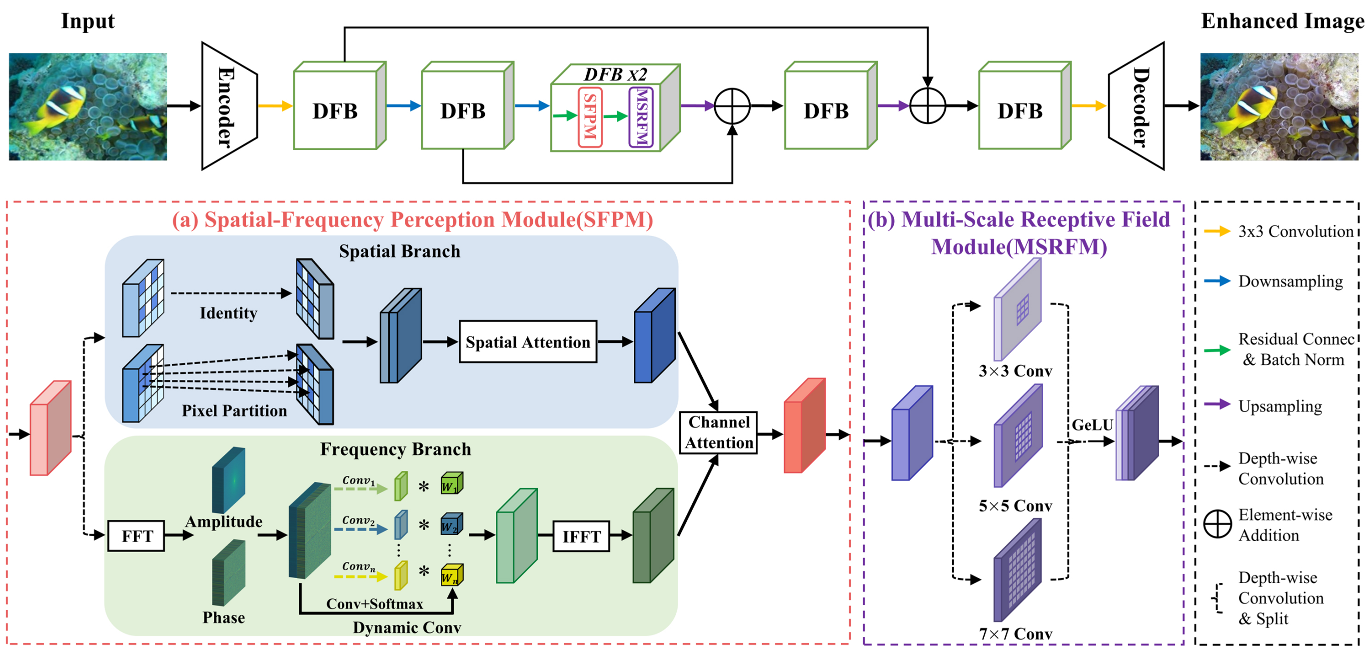

Fig. 3. The proposed SFPNet-UIE follows an encoder-decoder architecture composed of six Dynamic Fusion Blocks (DFBs) that extract and fuse spatial and frequency features. Each DFB includes a Spatial-Frequency Perception Module (SFPM), which integrates local spatial details with global frequency information, and a Multi-Scale Receptive Field Module (MSRFM), which captures multi-scale features using multi-kernel receptive fields to enhance texture restoration and correct color distortions..

B. Frequency Domain Features

Frequency domain learning has been applied in computer vision [20], [21] and UIE domains [14] to more effectively process specific features such as global illumination and local details by aggregating frequency domain features. For instance, Chi et al. [20] proposed directly extracting frequency domain features using fast Fourier convolutions, while Zhao et al. [14] proposed a spatial-frequency fusion network with gradient-aware corrector to enhance image details and structure. Yet, the cross-domain interaction merely combine spatial and frequency features without task-oriented feature fusion. Unlike existing frequency-domain restoration methods, our model directly extracts frequency features through learnable parameters using dynamic frequency convolution. By incorporating multi-scale representation learning to aggregate spatial and frequency domains, it enhances structural details and overall visual fidelity, effectively addressing the limitations of existing approaches in handling diverse underwater degradation scenarios.

III. PROPOSED METHOD

A. Overall Architecture of SFPNet-UIE

As illustrated in Fig. 3, the proposed SFPNet-UIE adopts an encoder–decoder architecture and introduces six Dynamic Fusion Blocks (DFB) to jointly capture both spatial and frequency features. Each DFB integrates two key modules: the Spatial-Frequency Perception Module (SFPM) and the Multi-Scale Receptive Field Module (MSRFM). SFPM adopts a spatial-frequency hybrid local-global modeling architecture that effectively leverages spatial and frequency domain information to enhance feature representations. MSRFM employs three parallel convolutional branches with different kernel sizes to capture features across multiple scales, enabling the network to preserve fine detail, reconstruct textures, and correct large-scale color distortions.

B. Spatial-Frequency Perception Module (SFPM)

The Spatial-Frequency Perception Module (SFPM) is the core of each DFB, designed to integrate local spatial information with global frequency domain representations. As shown in Fig. 3 (a), SFPM adopts a dual-branch architecture: the spatial branch focuses on local regional features, while the frequency branch captures global structure, illumination, and color cues.

The spatial branch extracts local features through two parallel paths: an identity path that preserves feature integrity and ensures stable gradient flow, and a pixel-partition path is constructed using depthwise convolution with a dilation factor of 2. By expanding the receptive field, contextual information is captured, and feature importance across spatial dimensions is learned. The contextually enriched features generated by this path provide structural prior knowledge for subsequent attention mechanisms, enabling the network to effectively distinguish salient features from redundant information. A lightweight spatial attention mechanism then highlights important regions and reduces background noise, improving feature representation with minimal computational cost.

The frequency branch focuses on modeling global information in the frequency domain. To capture both the amplitude (global color/illumination information) and phase (global structural information) under various degradation, the input feature map X is first transformed into the frequency domain via Fast Fourier Transform (FFT) [22], denoted as $F(X)$:

$$F(X)(u,v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X(h,w) e^{-j2\pi\left(\frac{hu}{H}+\frac{wv}{W}\right)}, \tag{1}$$

where (h,w) are spatial coordinates and (u,v) are frequency coordinates. The frequency domain representation consists of both real and imaginary components and can be written as:

$$F(X)(u,v) = \mathrm{R}_{X(u,v)} + \mathrm{I}_{X(u,v)} \cdot i, \tag{2}$$

where $R_{X(u,v)}$ and $I_{X(u,v)}$ represent the real and imaginary parts of the FFT, respectively. From these, the amplitude component $A(X(u,v))$ and phase component $P(X(u,v))$ are computed as:

$$A(X(u,v)) = \sqrt{\mathrm{R}^2_{X(u,v)} + \mathrm{I}^2_{X(u,v)}}, \tag{3}$$

$$P(X(u,v)) = \arctan \frac{\mathrm{I}_{X(u,v)}}{\mathrm{R}_{X(u,v)}}. \tag{4}$$

The amplitude and phase components are then aggregated to preserve the full frequency domain information. Subsequently, a series of convolutions $conv_1$, $conv_2$, ..., $conv_n$ are applied to the aggregated feature $F_j^{u,v}$ in the frequency domain, yielding the feature $F^{u,v}_{\mathrm{conv}_i}(i = 1, 2, \ldots, n)$. This operation explicitly decomposes $F_j^{u,v}$ into multiple parallel representations, each corresponding to a distinct frequency response pattern. Simultaneously, applying convolutions and softmax operations to $F_j^{u,v}$ generates dynamic weights $w_1^{u,v}, w_2^{u,v}, \ldots, w_n^{u,v}$ $(\sum_{i=1}^{n} w_i^{u,v} = 1)$. Dynamic feature fusion is achieved by combining the multiple feature representations generated through convolution with the learned attention weights, yielding the fused output feature $F^{u,v}_{out}$. This enables recalibration of different frequency components, allowing the network to autonomously emphasize information-rich spectral regions in underwater images while suppressing redundant frequency areas. This process can be represented as:

$$F^{u,v}_{out} = \sum_{i=1}^{n} w_i^{u,v} F^{u,v}_{conv_i}. \tag{5}$$

Finally, a channel attention operation aggregates the spatial and frequency branch outputs, achieving efficient fusion of multi-scale information.

C. Multi-Scale Receptive Field Module (MSRFM)

Inspired by [23], we employ the Multi-Scale Receptive Field Module (MSRFM) as another core component of the DFB, is designed to capture features at different spatial scales through parallel convolutional branches with varying kernel sizes (e.g., 3×3, 5×5, and 7×7), as shown in Fig. 3 (b). Specifically, the 3×3 kernel emphasizes fine-grained local textures and edge details, the 5×5 kernel captures mid-level structural information, and the 7×7 kernel models broader contextual features such as illumination variations and large-scale degradation patterns. Finally, all three-scale features from three branches are concatenated, and the GeLU activation function introduces nonlinearity to yield the output feature.

MSRFM extracts multi-scale features that capture both local details and global structures, which is essential for restoring fine textures and correcting large-scale color distortions. By learning from multiple receptive fields, it also generalizes well across different underwater environments, including shallow/deep waters and clear/turbid conditions.

D. Multi-scale Loss Function

To effectively guide the training process, a multi-scale loss function was designed for our model. I_x, I_y, and G_x represent the original input, ground truth image, and enhanced image from the generator output, respectively.

Firstly, we employed a content-pixel-aware loss function $L_{content}$, which combines L2 loss, SSIM loss, and VGG loss [24] to enhance image generation quality through multi-dimensional complementary supervision. The L2 ensures pixel-level accuracy, SSIM improves structural similarity, and VGG ensures perceptual quality.

Secondly, we designed a multi-color space loss function L_{color} that combines the LAB and LCH color spaces. This not only leverages the wider color gamut representation range of LAB and LCH, but also more accurately describes color saturation and brightness. It is expressed as follows:

$$\begin{aligned} L_{\mathrm{LAB}} = E_{I_x,I_y} \Bigg[& \left(L^{I_y} - L^{G_x}\right)^2 \\ & - \sum_{i=1}^{N} Q\left(A_i^{I_y}\right) \log\left(Q\left(A_i^{G_x}\right)\right) \\ & - \sum_{i=1}^{N} Q\left(B_i^{I_y}\right) \log\left(Q\left(B_i^{G_x}\right)\right) \Bigg], \end{aligned} \tag{6}$$

$$\begin{aligned} L_{\mathrm{LCH}} = E_{I_x,I_y} \Bigg[& - \sum_{i=1}^{N} Q\Big(L_i^{I_y}\Big) \log\Big(Q\big(L_i^{G_x}\big)\Big) \\ & + \left(C^{I_y} - C^{G_x}\right)^2 + \left(H^{I_y} - H^{G_x}\right)^2 \Bigg], \end{aligned} \tag{7}$$

$$L_{\text{color}} = \lambda_4 L_{\mathrm{LAB}} + \lambda_5 L_{\mathrm{LCH}}. \tag{8}$$

Among these, $L^{(\bullet)}$, $A^{(\bullet)}$, and $B^{(\bullet)}$ are used to convert the enhanced image G_x generated by the generator from an RGB space image to an LAB space image, resulting in the converted LAB space image. Similarly, $L^{(\bullet)}$, $C^{(\bullet)}$, and $H^{(\bullet)}$ are used to convert G_x to an LCH space image. Q represents the quantization operator, which is used to quantize a specific channel in different color spaces to calculate the cross-entropy loss between the enhanced image and the ground truth image on that channel. λ_4 and λ_5 represent the weights of the two components, set to 0.001 and 1, respectively.

The final loss function is expressed as:

$$L_{total} = L_{content} + L_{color}. \tag{9}$$

TABLE I
QUANTITATIVE AND EFFICIENCY COMPARISON OF PSNR/SSIM/LPIPS SCORES ON LSUI400 AND UIEB100 DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Datasets	Metrics	MLLE	CBLA	UWCNN	Water-Net	U-Shape	PUIE-Net	UDnet	SFGNet	Ours
LSUI400	PSNR↑	17.7340	13.8821	18.9816	20.2512	23.1687	21.2365	19.8700	22.7148	**24.0814**
	SSIM↑	0.7684	0.6215	0.7255	0.8127	0.8594	0.8778	0.8739	0.8775	**0.8814**
	LPIPS↓	0.2652	0.3580	0.3056	0.2123	0.1665	0.2028	0.2410	0.1789	**0.1424**
UIEB100	PSNR↑	19.5692	15.5308	15.1376	20.5016	23.5087	22.0661	19.7931	21.1895	**23.7673**
	SSIM↑	0.8515	0.6937	0.6924	0.8224	0.8780	0.9053	0.8993	0.9164	**0.9193**
	LPIPS↓	0.2814	0.3652	0.3504	0.1539	0.1287	**0.1201**	0.1905	0.1320	0.1207
Efficiency Evaluation	FLOPs↓	×	×	10.44G	193.7G	66.2G	423.05G	46.99G	81.48G	74.6G
	Params.↓	×	×	0.04M	24.8M	65.6M	1.4M	2.6M	1.3M	27M
	Times↓	0.77s	0.78s	1.43s	0.61s	0.07s	0.07s	0.03s	0.06s	0.06s

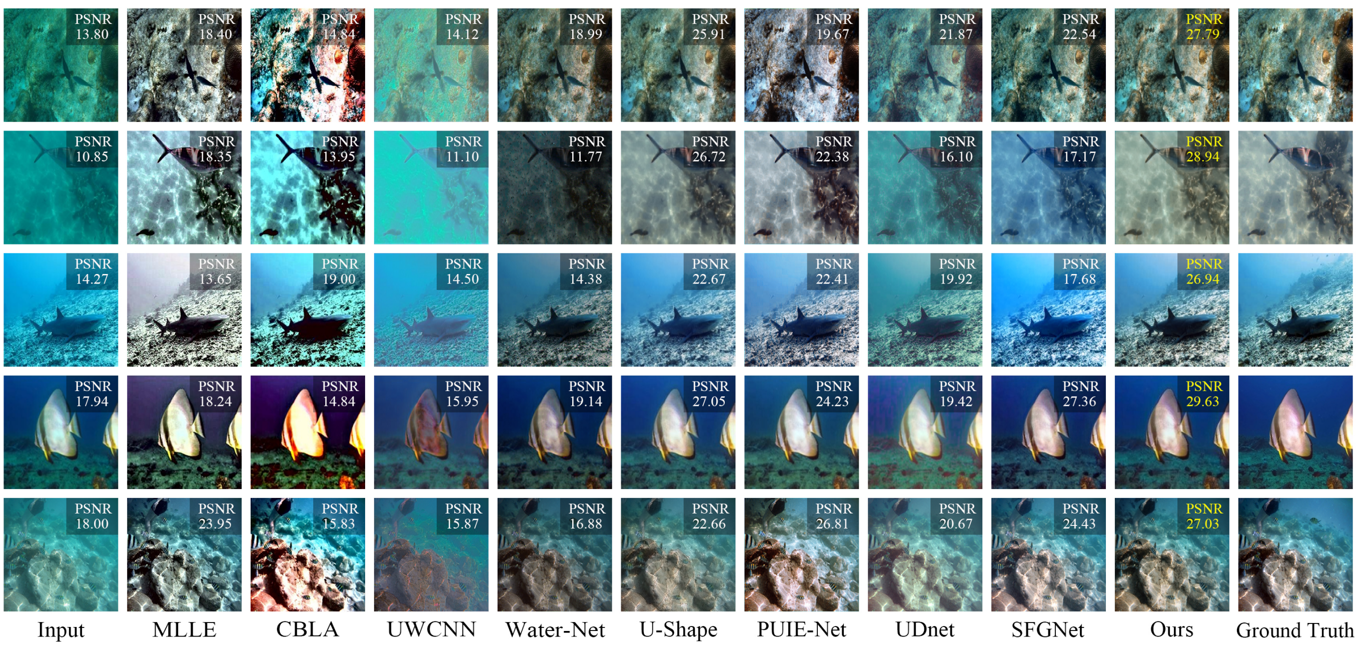

Fig. 4. Recovery results of different methods on full-reference datasets. The best PSNR scores are highlighted in yellow.

IV. EXPERIMENT

A. Datasets

To evaluate the effectiveness of the proposed SFPNet-UIE, we conducted experiments on two full-reference datasets and two non-reference datasets.

Full-reference Datasets. The LSUI dataset [25] comprises 4279 degraded underwater images and their corresponding ground truth. We randomly selected 3879 pairs of images from the LSUI for training, with the remaining 400 pairs serving as the test set (denoted as LSUI400). The UIEB dataset [26] consists of 890 paired images and 60 unpaired challenging images. From this dataset, 790 pairs were randomly selected for training, with remaining 100 pairs used for testing (denoted as UIEB100).

Non-reference Datasets. Non-reference datasets lack ground truth, facilitating evaluation of model generalization and robustness on real data. We evaluated our method on two challenging datasets: C60 [26] and UCCS [27]. The C60 consists of 60 images from the UIEB without ground truth (denoted as C60). UCCS is a subset of the real underwater image dataset RUIE, comprising 100 images each of blue, green, and blue-green color bias categories. We randomly selected 4 images from each category, totaling 12 images to form a test set (denoted as UCCS).

B. Evaluation Metrics

Full-reference Evaluation Metrics. For datasets with ground truth, we use PSNR [28], SSIM [29], and LPIPS [30] as full-reference evaluation metrics. PSNR measures fidelity to the reference image, SSIM assesses structural consistency in brightness, contrast, and structure. In contrast, LPIPS evaluates

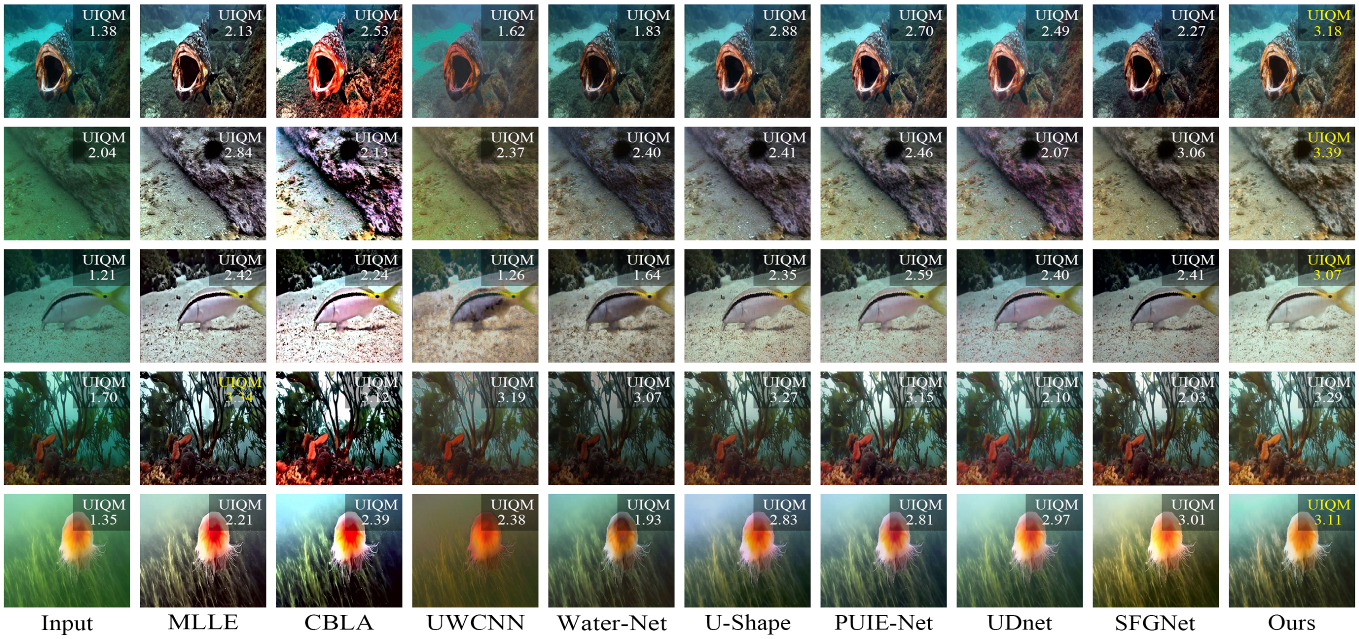

Fig. 5. Recovery results of different methods on non-reference datasets. The best UIQM scores are highlighted in yellow.

perceptual similarity to the ground truth, aligning with human visual judgment.

Non-reference Evaluation Metrics. We use UIQM [31] and UCIQE [32] as evaluation metrics. UIQM provides a comprehensive quality score across dimensions such as color, clarity, and contrast. UCIQE primarily measures color balance, contrast, and visual discernibility of images.

C. Experimental Setup

SFPNet-UIE was implemented in PyTorch and trained on a single NVIDIA GeForce RTX 3090 GPU. The ADAM optimizer is used with an initial learning rate of 5×10^{-4}. Additionally, the learning rate decreases by 20% every 40 epochs.

TABLE II
COMPARISON OF UIQM AND UCIQE SCORES ON C60 AND UCCS DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Datasets	C60		UCCS	
Metrics	UIQM↑	UCIQE↑	UIQM↑	UCIQE↑
MLLE	2.4446	0.5709	2.9533	0.5386
CBLA	2.1798	**0.6587**	2.4602	0.5473
UWCNN	2.4842	0.4510	3.0834	0.4183
Water-Net	2.5536	0.5178	3.2266	0.5439
U-Shape	2.8260	0.5343	3.2648	0.5411
PUIE-Net	2.7420	0.5519	3.1610	0.5277
UDNet	2.2452	0.5369	3.2618	0.5416
SFGNet	2.5763	0.5518	2.9739	0.5263
Ours	**2.9013**	0.5562	**3.2652**	**0.5478**

D. Performance Comparison with Existing Methods

To conduct a more comprehensive analysis, we compare the proposed SFPNet-UIE with two categories of current SOTA methods. These include traditional approaches such as MLLE [9] and CBLA [10], as well as deep learning-based methods including UWCNN [33], Water-Net [26], U-Shape [25], PUIE-Net [11], UDnet [34], and SFGNet [14].

Full-Reference Evaluation. As shown in Table I, evaluation on LSUI400 and UIEB100 shows that our SFPNet-UIE achieves the best performance in terms of PSNR and SSIM metrics, while also yielding relatively favorable results for the LPIPS metric. Notably, the SSIM value exceeds 0.9 on the UIEB100, indicating high similarity between the enhanced and reference images in terms of structure and brightness. We also present efficiency evaluations in Table I, where SFPNet-UIE achieves comparable performance to other SOTA methods with relatively reasonable parameters and inference time.

Fig. 4 compares the visual enhancement effects and PSNR metrics of the methods mentioned above. MLLE [9] and CBLA [10]) still exhibit color correction biases. UWCNN [33] tends toward oversaturation, and Water-Net [26] introduces color shifts when enhancing contrast. In contrast, our SFPNet-UIE effectively corrects color bias while preserving more natural scene perception and tonal continuity.

Non-Reference Evaluation. The results are presented in Table II, with visual comparisons shown in Fig. 5. Our SFPNet-UIE model achieved the highest UIQM scores across all three datasets and obtained the best UCIQE score on the UCCS dataset. As shown in Fig. 5, for common degraded underwater images, SFPNet-UIE demonstrates significant improvements in handling low-contrast and blurry scenes, effectively mitigating the impact of these factors. In contrast, UWCNN [33] may exacerbate color anomalies in certain scene images. Although it still has limitations in eliminating cyan/blue casts, its overall restoration quality and generalization capability are the best among the compared methods.

It is worth noting that existing non-reference metrics cannot fully reflect subjective perceived quality, a limitation also emphasized in [25]. As shown in Fig. 5, although MLLE and CBLA achieve high scores, their enhanced images exhibit unnatural reddish tints. These observations indicate that while such metrics provide useful quantitative references, high scores do not necessarily equate to superior perceived quality. When evaluating UIE models, non-reference metrics should be combined with full-reference metrics.

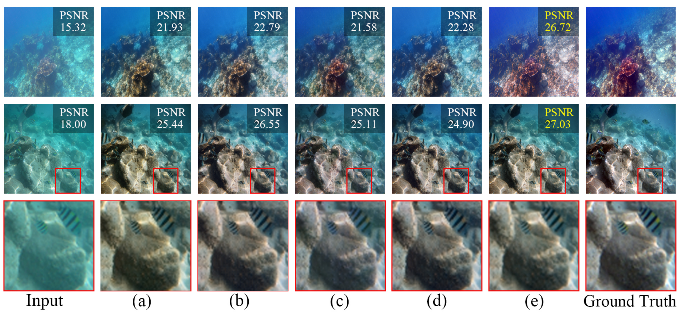

Fig. 6. Visual and detailed comparison of ablation studies for the key network components. (a) single-branch on spatial branch. (b) w/o spatial branch. (c) w/o frequency domain branch. (d) w/o MSRFM. (e) Full model.

TABLE III
THE ABLATION EXPERIMENT RESULTS OF THE KEY NETWORK COMPONENTS ON THE UIEB DATASET.

Method	PSNR↑	SSIM↑	LPIPS↓
single-branch on spatial branch	22.5506	0.9076	0.1386
w/o spatial branch	22.1357	0.8960	0.1383
w/o frequency branch	22.8873	0.9015	0.1259
w/o MSRFM	21.6582	0.8979	0.1442
Full model	**23.7673**	**0.9193**	**0.1207**

TABLE IV
THE ABLATION EXPERIMENTS RESULTS OF THE INDIVIDUAL COMPONENTS OF THE MULTI-SCALE LOSS FUNCTION.

Configuration		PSNR↑	SSIM↑	LPIPS↓
$L_{content}$	L_{color} (LAB+LCH)			
✓		21.1458	0.8652	0.1426
	✓	22.8975	0.9078	0.1217
✓	LAB only	22.3592	0.9039	0.1356
✓	LCH only	22.7845	0.8973	0.1287
✓	✓	**23.7673**	**0.9193**	**0.1207**

E. Ablation Study

Effectiveness of Key Network Components. As shown in Table III, we conducted an ablation study on four scenarios: single-branch on spatial branch, no spatial branch, no frequency branch, and no MSRFM. Evaluations were conducted on the UIEB dataset. Two conclusions emerged from the experiments: First, for the spatial branch, the PSNR and SSIM scores of the single-branch structure model were lower than those of the full model (i.e., dual-branch structure), indicating that the dual-branch structure outperforms the single-branch structure. Second, removing any one of the spatial branch, frequency branch, or MSRFM resulted in degraded model performance. As illustrated by the visual comparison in Fig. 6, the complete model achieves a balanced recovery of color and detail, outperforming variants lacking specific components. These results underscore the necessity of each branch.

Influence of the Multi-scale Loss Function. To evaluate the contribution of each component in the multi-scale loss function, we conduct ablation experiments on the UIEB dataset by progressively removing each loss. Results show that each loss provides complementary and irreplaceable supervision. Removing content loss causes a significant PSNR drop, highlighting its role in preserving contrast and pixel accuracy. Using only color loss leads to moderate degradation across metrics, confirming its independent contribution. Removing either LAB or LCH loss alone degrades performance, impairing local color details. Overall, each loss contributes to pixel accuracy, structural consistency, and color distribution, and their integration achieves superior restoration performance.

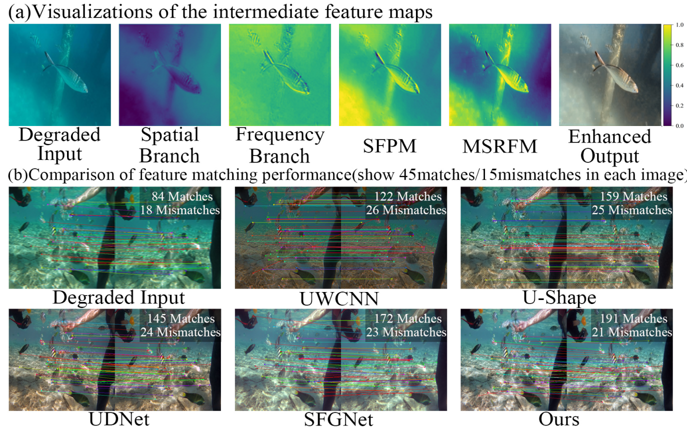

Fig. 7. (a) Visualizations of the intermediate feature maps. From left to right is the degraded input, the feature map from the spatial branch, the frequency branch, the SFPM, the MSRFM, and the enhanced output. (b) Comparison of feature matching among some SOTA methods.

F. Discussion and downstream task assessments

The experimental results clearly highlight the effectiveness of the proposed SFPNet-UIE framework in enhancing underwater images. As shown in Fig. 7(a), the feature map from spatial branch captures image texture and details, while the frequency branch captures illumination and color information, clearly preserving the structural details of the fish. The SFPM module effectively fuses information from both branches, and further enhances color fidelity, texture details, and overall image quality through MSRFM's multi-scale feature fusion.

Furthermore, to validate our advantage in restoring local structure and texture, we conducted a feature matching downstream experiment. As shown in Fig. 7(b), compared to methods such as U-Shape and SFGNet, our approach achieves the highest number of valid matches, demonstrating superior feature robustness and structural preservation.

V. CONCLUSION

In this work, we first re-examine the degradation process of underwater images from a frequency domain perspective and propose a novel Fourier-based enhancement framework for underwater image enhancement. The SFPM and MSRFM modules within this framework collaboratively process spatial and frequency domain features. By integrating dual-domain features, our approach achieves comprehensive enhancement of structural details and overall visual fidelity, effectively addressing the limitations of existing methods that neglect spatial-frequency domain fusion in degraded underwater scenarios. Experiments demonstrate the competitive performance of our SFPNet-UIE across four benchmark datasets.

REFERENCES

[1] X. Cong, Y. Zhao, J. Gui, J. Hou, and D. Tao, "A comprehensive survey on underwater image enhancement based on deep learning," *arXiv preprint arXiv:2405.19684*, 2024.

[2] S. Sun, T. Liu, Y. Wang, G. Zhang, K. Liu, and Y. Wang, "High-rate underwater acoustic localization based on the decision tree," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[3] T. Alexandri, Z. Z. Shamir, E. Bigal, A. Scheinin, D. Tchernov, and R. Diamant, "Localization of acoustically tagged marine animals in under-ranked conditions," *IEEE Transactions on Mobile Computing*, vol. 20, no. 3, pp. 1126–1137, 2019.

[4] J. Liu, S. Li, C. Zhou, X. Cao, Y. Gao, and B. Wang, "Sraf-net: A scene-relevant anchor-free object detection network in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[5] C. Liu, X. Shu, L. Pan, J. Shi, and B. Han, "Multiscale underwater image enhancement in rgb and hsv color spaces," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–14, 2023.

[6] J. Yuan, Z. Cai, and W. Cao, "Tebcf: Real-world underwater image texture enhancement model based on blurriness and color fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2021.

[7] J. Zhou, Q. Gai, D. Zhang, K.-M. Lam, W. Zhang, and X. Fu, "Iacc: Cross-illumination awareness and color correction for underwater images under mixed natural and artificial lighting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[8] H. Qi, H. Zhou, J. Dong, and X. Dong, "Deep color-corrected multiscale retinex network for underwater image enhancement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2023.

[9] W. Zhang, P. Zhuang, H.-H. Sun, G. Li, S. Kwong, and C. Li, "Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement," *IEEE Transactions on Image Processing*, vol. 31, pp. 3997–4010, 2022.

[10] M. Jha and A. K. Bhandari, "Cbla: Color-balanced locally adjustable underwater image enhancement," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–11, 2024.

[11] Z. Fu, W. Wang, Y. Huang, X. Ding, and K.-K. Ma, "Uncertainty inspired underwater image enhancement," in *European conference on computer vision*. Springer, 2022, pp. 465–482.

[12] Y. Tang, H. Kawasaki, and T. Iwaguchi, "Underwater image enhancement by transformer-based diffusion model with non-uniform sampling for skip strategy," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5419–5427.

[13] H. Wang, K. Köser, and P. Ren, "Large foundation model empowered discriminative underwater image enhancement," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[14] C. Zhao, W. Cai, C. Dong, and Z. Zeng, "Toward sufficient spatial-frequency interaction for gradient-aware underwater image enhancement," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 3220–3224.

[15] M. S. Hitam, E. A. Awalludin, W. N. J. H. W. Yussof, and Z. Bachok, "Mixture contrast limited adaptive histogram equalization for underwater image enhancement," in *2013 International conference on computer applications technology (ICCAT)*. IEEE, 2013, pp. 1–5.

[16] S. Zhang, T. Wang, J. Dong, and H. Yu, "Underwater image enhancement via extended multi-scale retinex," *Neurocomputing*, vol. 245, pp. 1–9, 2017.

[17] P. Drews, E. Nascimento, F. Moraes, S. Botelho, and M. Campos, "Transmission estimation in underwater single images," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 825–830.

[18] J. Y. Chiang and Y.-C. Chen, "Underwater image enhancement by wavelength compensation and dehazing," *IEEE transactions on image processing*, vol. 21, no. 4, pp. 1756–1769, 2011.

[19] C. Li, J. Guo, S. Chen, Y. Tang, Y. Pang, and J. Wang, "Underwater image restoration based on minimum information loss principle and optical properties of underwater imaging," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1993–1997.

[20] L. Chi, B. Jiang, and Y. Mu, "Fast fourier convolution," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4479–4488, 2020.

[21] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," *Advances in neural information processing systems*, vol. 34, pp. 980–993, 2021.

[22] E. O. Brigham, *The fast Fourier transform and its applications*. Prentice-Hall, Inc., 1988.

[23] Y. Li, Y. Chen, N. Wang, and Z.-X. Zhang, "Scale-aware trident networks for object detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6053–6062.

[24] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[25] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," *IEEE transactions on image processing*, vol. 32, pp. 3066–3079, 2023.

[26] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE transactions on image processing*, vol. 29, pp. 4376–4389, 2019.

[27] R. Liu, X. Fan, M. Zhu, M. Hou, and Z. Luo, "Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4861–4875, 2020.

[28] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?" in *2012 Fourth international workshop on quality of multimedia experience*. IEEE, 2012, pp. 37–38.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[30] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[31] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 541–551, 2015.

[32] M. Yang and A. Sowmya, "An underwater color image quality evaluation metric," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 6062–6071, 2015.

[33] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern recognition*, vol. 98, p. 107038, 2020.

[34] A. Saleh, M. Sheaves, D. Jerry, and M. R. Azghadi, "Adaptive deep learning framework for robust unsupervised underwater image enhancement," *Expert Systems with Applications*, vol. 268, p. 126314, 2025.