Journal of *Marine Science and Engineering*

MDPI

Article

DSFNet: Directional Spatial–Frequency Collaboration Network for Underwater Optical Image Enhancement

Weibin Li [1], Jiahui Li [2], Jingchun Zhou [2,3,4,5,*], Dehuan Zhang [2], Guoqing Zhang [6] and Qiuping Jiang [7]

[1] College of Marine Engineering, Dalian Maritime University, Dalian 116026, China; cloud9@dlmu.edu.cn
[2] School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China; a2012255201@gmail.com (J.L.); zhangdehuan@dlmu.edu.cn (D.Z.)
[3] State Key Laboratory of Ocean Sensing, ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou 311215, China
[4] School of Information and Communication Engineering, North University of China, Taiyuan 030051, China
[5] Shanxi Province Key Laboratory of Intelligent Detection and Equipment Technology and Equipment, Taiyuan 030051, China
[6] Navigation College, Dalian Maritime University, Dalian 116026, China; zgq_dlmu@163.com
[7] School of Information Science and Engineering, Ningbo University, Ningbo 315211, China; jiangqiuping@nbu.edu.cn
* Correspondence: zhoujingchun03@gmail.com

Abstract

Underwater optical images frequently suffer from wavelength-dependent attenuation and scattering, resulting in color distortion, contrast reduction, and structural-detail loss. Recent methods have already explored directional modeling and spatial–frequency fusion; coordinating these operations through explicitly specified feature interactions remains a design question. To address this issue, we propose the Directional Spatial–Frequency Collaboration Network (DSFNet) for underwater optical image enhancement. Building on LU2Net's encoder–decoder and axial depthwise convolution, DSFNet adds separate sparse main- and anti-diagonal branches to form OctaRayDW. A spatially conditioned Fourier module updates amplitude-derived coefficients and phase responses, while two reciprocal cross-domain attention operations exchange spatial and frequency-enhanced features before gated fusion. A parallel convolutional path is adaptively combined with histogram attention for local and contextual modeling. These components are integrated into a three-stage encoder–decoder containing 5.58 M parameters. On the Underwater Image Enhancement Benchmark (UIEB) dataset, DSFNet achieves 25.352 dB PSNR and 0.929 SSIM, exceeding the retrained original two-stage LU2Net by 0.704 dB and 0.004, respectively. Additional evaluations on C60, U45, MABLs, and UFO120 indicate favorable perceived quality under the adopted no-reference metrics across diverse underwater conditions. The results confirm the effectiveness of directional spatial–frequency collaboration for underwater optical image enhancement.

Keywords: underwater image enhancement; underwater optical imaging; spatial-frequency interaction; multidirectional structural modeling; local-global adaptive fusion

Check for updates

Academic Editor: Weicheng Cui

Received: 16 September 2026
Revised: 5 October 2026
Accepted: 7 October 2026
Published: 10 October 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) license.

1. Introduction

Underwater optical imaging is an important sensing modality for acquiring visual information in submerged environments, supporting applications such as ecological monitoring, seabed exploration, underwater infrastructure inspection, and autonomous robotic

operations [1]. However, the quality and reliability of acquired imagery are strongly affected by complex underwater light propagation [2]. Wavelength-dependent absorption selectively attenuates different color components, while scattering introduces veiling effects and weakens image contrast and structural details [3]. Because these effects vary with water type, imaging depth, illumination conditions, and camera-to-object distance, underwater images commonly exhibit spatially nonuniform color casts, reduced visibility, and texture degradation [4]. Such distortions impede human interpretation and subsequent automated visual analysis. Islam et al. [5] investigated underwater enhancement as a preprocessing step for visual perception tasks. Therefore, improving the visual fidelity and structural integrity of underwater optical images is essential for reliable information extraction in complex marine environments.

Recent deep learning-based underwater image enhancement methods have achieved remarkable progress by learning nonlinear mappings from degraded images to enhanced results [6–8], including diffusion-based methods that decouple chromatic degradation from structural content [9]. Existing approaches mainly improve restoration quality from several perspectives, including color correction, spatial feature refinement, frequency-domain representation, and attention-based contextual modeling. Recent methods such as Cross-Domain Fusion for Underwater Image Enhancement (CDF-UIE) [10] and FSIF-Net [11] already combine spatial and frequency information for underwater enhancement. The proposed Directional Spatial–Frequency Collaboration Network (DSFNet) investigates a specific coupling of sparse orientation-dependent spatial aggregation, conditioned Fourier refinement, reciprocal cross-domain attention, and gated local–histogram fusion. Such feature coordination is particularly relevant to complex underwater environments, where spatially varying degradation increases the need to preserve both visual fidelity and structural information. Adaptation of in-air models to underwater enhancement and detection has also been investigated [12].

Specifically, three challenges remain unresolved in enhancing images acquired under complex underwater conditions. First, underwater targets and structures often contain directional patterns, such as object contours, elongated textures, and artificial facilities, which require effective directional feature modeling beyond conventional isotropic convolution. Second, underwater degradation alters spatial appearance and is also reflected in the Fourier-domain representations of captured images, where amplitude statistics and phase-related structural cues provide complementary information, motivating explicit spatial-frequency interaction. Third, underwater scenes usually contain spatially distributed regions with similar degradation properties, making it necessary to simultaneously preserve local details and capture global contextual dependencies.

To address these challenges, we propose DSFNet, a degradation-aware spatial-frequency collaborative network for underwater optical image enhancement. The network extends LU2Net's two-stage encoder–decoder to three stages and couples the added diagonal depthwise branches with an amplitude–phase refinement module, reciprocal cross-domain attention, and gated local–histogram fusion. Figure 1 compares restoration performance and model parameter count on the Underwater Image Enhancement Benchmark (UIEB) validation set, while Figure 2 illustrates its conceptual integration as an image enhancement front end in an underwater robotic vision system.

The main contributions of this work are summarized as follows:

1. We extend the two-stage LU2Net backbone to three stages and integrate orientation-specific spatial aggregation, Fourier refinement, cross-domain interaction, and response-guided context within an underwater enhancement network.
2. We extend axial depthwise convolution with separate sparse main-diagonal and anti-diagonal depthwise branches in OctaRayDW, explicitly representing four orientation axes.

3. We couple spatially conditioned residual refinement of Fourier amplitude-derived coefficients and phase responses with two independently parameterized reciprocal cross-domain attention operations and gated fusion.
4. We extend histogram attention with a separate local convolutional path and a content-dependent gate to combine local features and sorted-response context in Local-Global Adaptive Fusion (LGAF).

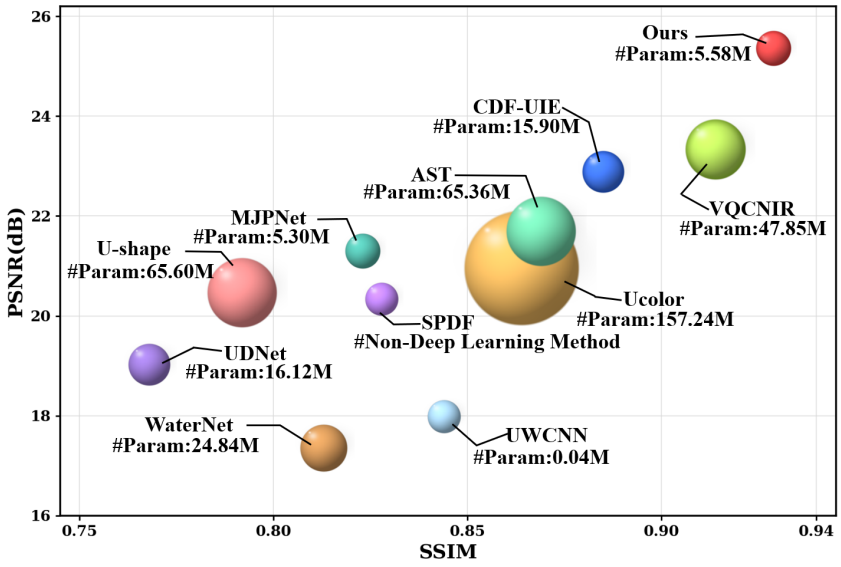


Figure 1. UIEB restoration performance and model parameter count. The horizontal and vertical axes represent SSIM and PSNR, respectively, while the bubble size indicates the number of model parameters.

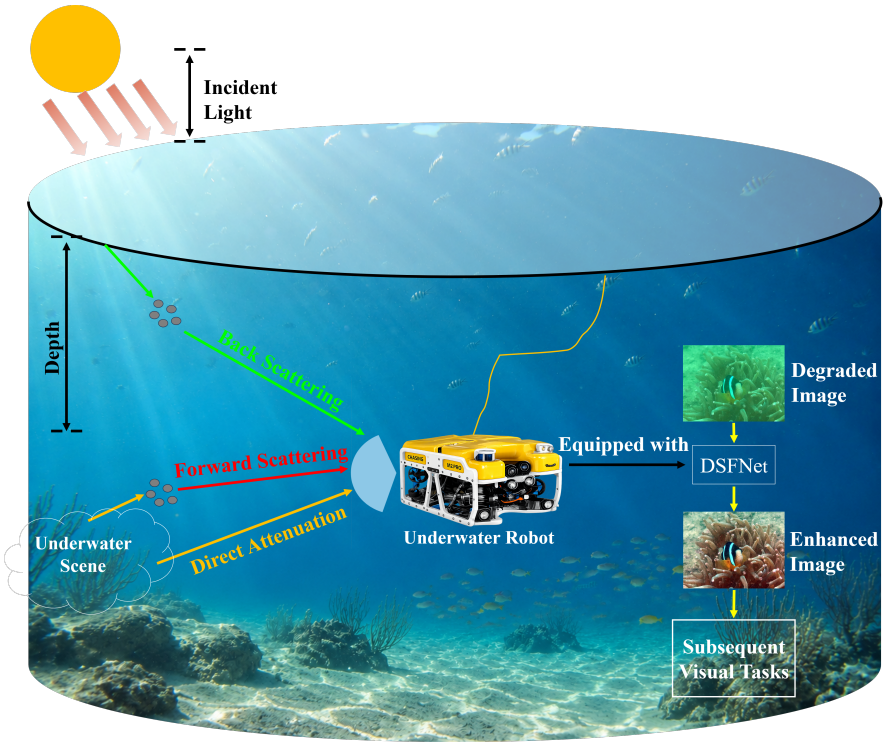


Figure 2. Conceptual integration of the proposed method into an underwater robotic vision system, where DSFNet serves as an image enhancement front end prior to subsequent visual perception tasks.

2. Related Work

In this section, we review methods closely related to the proposed approach. Anwar and Li [13] provide a broader survey of underwater image enhancement methods.

2.1. Multidirectional Structural Representation

Existing image restoration methods enhance structural representation and detail recovery through residual learning, cross-layer feature propagation, multi-scale feature fusion, and attention mechanisms. Such feature reuse and aggregation approaches often represent spatial structure implicitly, motivating explicit modeling of edges and elongated patterns in underwater scenes. In contrast, direction-aware structural modeling can aggregate contextual information along specific spatial directions, thereby capturing geometric structures in images more directly. Recent studies have explored directional modeling from different perspectives. For instance, Liu et al. [14] proposed VMamba, whose two-dimensional selective scan module traverses visual features along four scanning routes to aggregate contextual information from different spatial directions. Guo et al. [15] introduced MambaIR for image restoration, combining two-dimensional selective scanning with local enhancement to balance long-range dependency modeling and local detail preservation. For underwater enhancement, Kang et al. [16] used structural patch decomposition to fuse the intensity, contrast, and structure of complementary pre-processed images in SPDF. Yang et al. [17] combined axial depthwise convolution with channel attention in LU2Net. Li et al. [11] incorporated directional convolutions and local detail enhancement into FSIF-Net. Together, these underwater methods illustrate complementary ways to represent structural information, from patch decomposition to directional feature extraction.

Efficient representation of complementary directional structures thus remains a relevant design question for underwater image enhancement.

2.2. Frequency-Domain Representation and Spatial-Frequency Interaction

Frequency-domain representations have been widely studied in image restoration. Huang et al. [18] propose the Degradation Frequency Curve (DFC) as an explicit spectral representation to characterize degradation responses in the frequency domain. Li et al. [19] separately process the Fourier amplitude and phase components and integrate frequency-domain processing with spatial feature enhancement for low-light image restoration, while Ren et al. [20] introduce spatial-frequency parallel fusion to improve restoration performance through frequency-aware feature interaction. These studies demonstrate that frequency representations can provide effective global priors and degradation-aware information for image restoration.

Spatial–frequency interaction has also been explored for underwater image enhancement. Huo et al. [21] used wavelet-based multi-stage refinement to process low- and high-frequency features. SFGNet [22] exchanges Fourier and spatial features through cross-connections between fusion blocks. TSNet-SFFI [23] combines Fourier-derived global features with local spatial features in a two-stage framework. CDF-UIE [10] processes amplitude and phase information and fuses spatial and frequency representations. FSIF-Net [11] combines directional spatial features with wavelet and Fourier representations. How to coordinate spectral refinement with spatial feature interaction remains worth further investigation in underwater image enhancement.

2.3. Local-Global Context Aggregation

Image restoration methods have explored local and global feature interactions from different perspectives. Pei et al. [24] integrated local feature extraction with broader contextual information through an efficient visual state space model, achieving competitive

performance for lightweight image restoration with linear complexity. Guo et al. [15] introduced restoration-specific local enhancement into a visual state-space model to mitigate local pixel forgetting while retaining efficient long-range dependency modeling. Beyond state space models, transformer-based approaches have investigated response-aware attention mechanisms to capture long-range dependencies. For underwater enhancement, the U-shape Transformer [25] combines channel-wise multi-scale fusion with spatial global feature modeling. Ucolor [26] uses attention to integrate features from multiple color spaces and transmission guidance to focus on degraded regions. Histoformer [27], developed for general image restoration, combines histogram self-attention and dynamic-range convolution for contextual and local feature modeling. These studies motivate further exploration of how local details and broader contextual information can be balanced across varying underwater image content.

3. Proposed Method

This section provides a detailed introduction to the proposed underwater image enhancement network, DSFNet. The overall network architecture of DSFNet is presented, including the basic workflow of its encoder–bottleneck–decoder structure. The OctaRayDW for multidirectional spatial structure modeling is introduced. This module enhances the network's ability to represent direction-sensitive structures through depthwise convolutions along horizontal, vertical, main-diagonal, and anti-diagonal directions. The Explicit Frequency Modeling (EFM) and Explicit Cross-Domain Fusion (ECDF) modules are described. They are designed to characterize the response variations of underwater image degradation in the frequency domain amplitude and phase spectra and to enable bidirectional interaction between spatial and frequency-domain features. The Local-Global Adaptive Fusion (LGAF) module is introduced. Built upon histogram-domain attention, LGAF explicitly coordinates convolutional local refinement with response-domain contextual aggregation through content-adaptive fusion, thereby balancing fine-detail restoration and global visual consistency. The overall architecture of DSFNet and the motivation for its multidirectional structural design are illustrated in Figures 3 and 4, respectively. In Figure 4, the directional contours in the underwater scene motivate the comparison between the square-kernel sampling pattern and the horizontal, vertical, and diagonal branches of OctaRayDW.

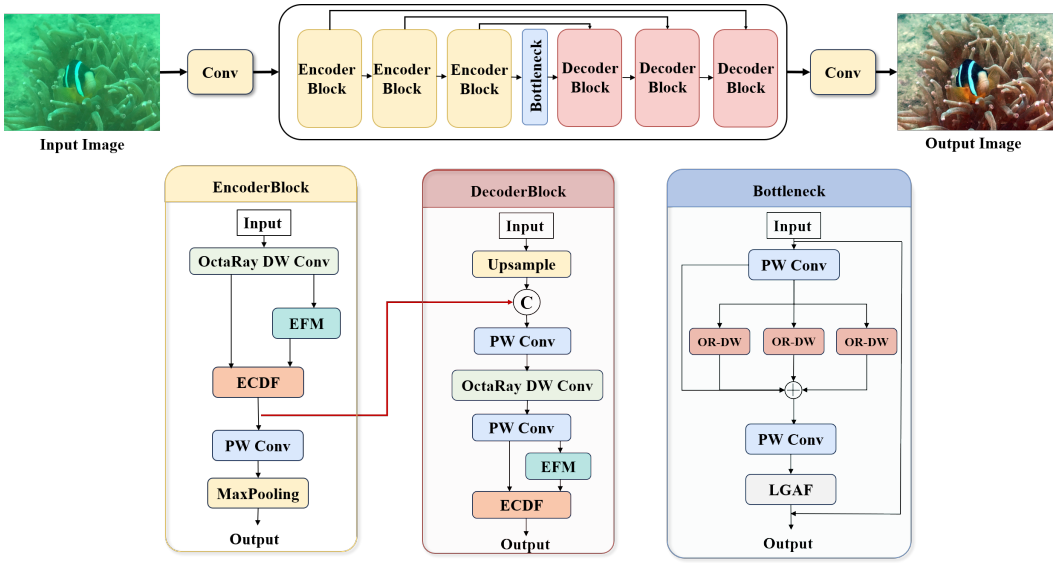


Figure 3. Overall architecture of DSFNet, showing the three-stage encoder–decoder, EFM–ECDF blocks, and LGAF bottleneck.

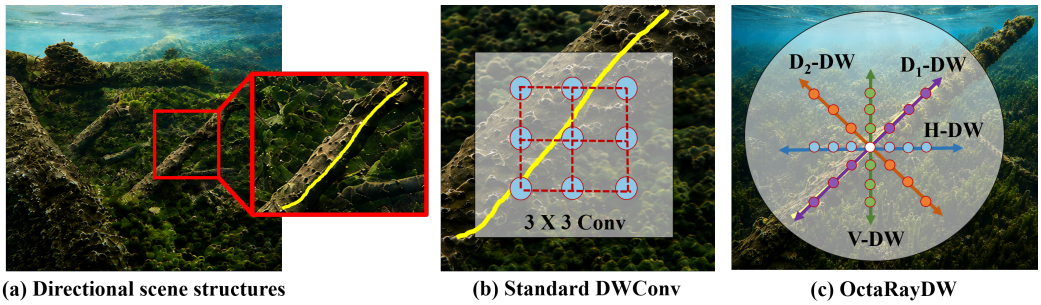


Figure 4. Motivation for OctaRayDW. (**a**) Underwater scene with directional structures. (**b**) Standard 3×3 depthwise convolution. (**c**) Horizontal, vertical, and diagonal branches of OctaRayDW.

3.1. Overall Architecture of DSFNet

The backbone of DSFNet is constructed by extending the two-stage encoder–decoder architecture of LU2Net [17] to three stages, enabling hierarchical feature extraction at three spatial scales. LU2Net also provides the axial depthwise branches and channel-attention operation used as the starting point for OctaRayDW. The added diagonal branches, EFM–ECDF interaction, and LGAF are described below as modifications and extensions of this starting point. Given an input underwater image $I \in \mathbb{R}^{3\times H\times W}$, DSFNet learns a nonlinear mapping from the degraded image to the clear image and outputs the enhanced image $\hat{J}$. As shown in Figure 3, the overall network consists of an input convolution, a three-stage encoder, a bottleneck block, a three-stage decoder, and an output convolution. The overall process can be briefly written as

$$\hat{J} = F_\theta(I) \tag{1}$$

where F_θ denotes the proposed DSFNet, and θ represents the learnable parameters of the network. The input image is first mapped to a shallow feature space through a 7×7 convolution:

$$X_0 = \mathrm{Conv}_{7\times 7}(I) \tag{2}$$

Then, the network sequentially passes through three encoder blocks for feature extraction and downsampling. Each encoder block consists of OctaRayDW, EFM, and ECDF, which jointly extract spatial directional structures and frequency-domain degradation responses. The encoding process can be formulated as

$$X_i, S_i = \mathrm{Enc}_i(X_{i-1}) \quad i = 1, 2, 3 \tag{3}$$

where S_i denotes the skip connection feature at the i-th scale, and X_i denotes the downsampled deep feature. The input convolution first maps the image to 16-channel shallow features, after which the three encoder stages progressively expand the feature dimensions to 64, 256, and 512 channels. The bottleneck layer performs multi-scale contextual modeling on the lowest resolution features. Specifically, this layer first enlarges the effective receptive field through OctaRayDW with different dilation rates and then employs LGAF to adaptively integrate convolutionally refined local features with histogram-domain contextual representations:

$$X_b = \mathrm{BottleNeck}(X_3) \tag{4}$$

The three parallel OctaRayDW branches in the bottleneck use a kernel length of 3 with dilation rates of 1, 2, and 3, respectively. The decoder progressively restores the spatial resolution, starting from the bottleneck feature. Specifically, the first decoder stage takes

the bottleneck representation as input, and subsequent decoder stages recursively utilize the output of the previous stage together with the corresponding skip features:

$$\begin{cases} Y_3 = \mathrm{Dec}_3(X_b, S_3) \\ Y_i = \mathrm{Dec}_i(Y_{i+1}, S_i), \quad i = 2, 1 \end{cases} \tag{5}$$

Finally, the network maps the features back to the RGB image space through a 1×1 convolution:

$$\hat{J} = \mathrm{Conv}_{1\times1}(Y_1) \tag{6}$$

3.2. OctaRay Depthwise Convolution

Conventional convolution typically employs square kernels with uniform spatial support for local feature aggregation. Although learnable kernel weights can implicitly capture diverse local patterns, their sampling geometry does not explicitly emphasize structures along specific orientations. Asymmetric convolution has shown that horizontal and vertical one-dimensional branches can strengthen the directional skeletons of square convolution kernels [28]. In underwater image enhancement, horizontal, vertical, and diagonal convolutions have also been employed to improve the representation of anisotropic structures [11]. Nevertheless, axial convolution mainly aggregates information along the horizontal and vertical directions and therefore provides limited explicit responses to oblique textures and diagonal structures. To address this issue, we design the OctaRayDW, whose structure is shown in Figure 5. This module consists of horizontal depthwise convolution, vertical depthwise convolution, main-diagonal depthwise convolution, and anti-diagonal depthwise convolution. These are four learnable orientation branches; each line-shaped support contains samples on the two opposite rays of its orientation. The name OctaRay refers to these eight rays, rather than eight independently parameterized convolutions. Given an input feature X, the multidirectional response of OctaRayDW can be formulated as

$$\begin{aligned} X_o = CAB\big(X + \mathrm{DW}_h(X) + \mathrm{DW}_w(X) \\ + \mathrm{DW}_{d1}(X) + \mathrm{DW}_{d2}(X)\big) \end{aligned} \tag{7}$$

where DW_h and DW_w denote the horizontal and vertical depthwise convolutions, respectively. DW_{d1} and DW_{d2} denote the main-diagonal and anti-diagonal depthwise convolutions, respectively. $\mathrm{CAB}(\cdot)$ denotes the channel attention block. The two diagonal branches are initialized with zero weights and biases; their responses can subsequently be learned during training.

The horizontal and vertical directions employ $1 \times k$ and $k \times 1$ depthwise convolutions, respectively, to model axial structures with a low parameter cost. For diagonal directions, instead of directly using a full $k \times k$ depthwise convolution, we only retain the learnable weights located on the main-diagonal or anti-diagonal of the convolution kernel. In all three encoder and decoder blocks, $k = 7$ and the dilation rate is 1. Each bottleneck branch uses $k = 3$, with the branch-specific dilation rates specified in Section 3.1. In each diagonal branch, only k positions of the corresponding $k \times k$ support have learnable weights. Taking the main-diagonal convolution as an example, its computation can be written as

$$\mathrm{DW}_{d1}(X)_{n,c,p,q} = \sum_{t=0}^{k-1} w_{c,t} \widetilde{X}_{n,c,p+(t-r)d,\, q+(t-r)d} + b_c \tag{8}$$

where n, c, p, and q denote the sample, channel, row, and column indices, respectively. k is the kernel size, $r = \lfloor k/2 \rfloor$, and d denote the dilation rate. $\widetilde{X}$ is the padded feature map of X, $w_{c,t}$ is the weight of the c-th channel at the t-th diagonal position, and b_c is the bias

term. The anti-diagonal convolution adopts the opposite offset direction along the column dimension, i.e., $q-(t-r)d$. Compared with standard $k \times k$ depthwise convolution, the diagonal depthwise convolution requires only k parameters for each channel, allowing it to capture inclined edges and diagonal textures with a sparse learnable kernel. After multidirectional feature aggregation, OctaRayDW uses channel attention to recalibrate the channels of the combined feature:

$$\begin{cases} \mathrm{CAB}(X) = X \odot \sigma(M_c(X)), \\ M_c(X) = \mathrm{Conv}_{1\times1}(\mathrm{ReLU}(\mathrm{Conv}_{1\times1}(\mathrm{GAP}(X)))) \end{cases} \tag{9}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, $M_c(\cdot)$ denotes the channel-attention mapping, $\mathrm{ReLU}(\cdot)$ denotes the ReLU activation function, $\sigma(\cdot)$ denotes the Sigmoid function, and $\odot$ denotes element-wise multiplication. For a feature map $X \in \mathbb{R}^{C\times H\times W}$, $[\mathrm{GAP}(X)]_c = \frac{1}{HW}\sum_{p=1}^{H}\sum_{q=1}^{W} X_{c,p,q}$ and $\mathrm{ReLU}(z) = \max(0, z)$. The Sigmoid output assigns one weight to each channel, and $\odot$ multiplies that weight with every spatial position in the corresponding channel. As shown in Equation (7), CAB receives the sum of the identity and four directional branches. Its weights therefore rescale the combined feature channels rather than the individual directional branches. Figure 6 conceptually compares the directional information propagation of axial depthwise convolution and OctaRayDW. The illustrated diagonal paths add aggregation along oblique contours to the horizontal and vertical paths.

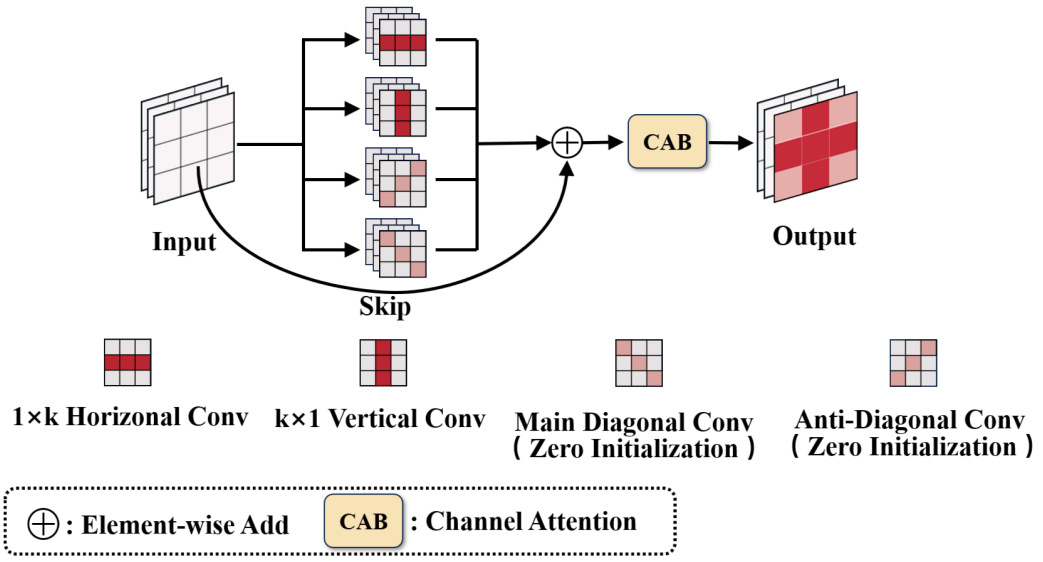


Figure 5. Architecture of OctaRayDW with four directional depthwise branches, identity addition, and channel attention.

3.3. Explicit Frequency Modeling

Underwater image degradation is not only manifested as color shifts, brightness attenuation, and texture blurring in the spatial domain, but it also affects amplitude distribution anomalies and phase structure perturbations in the frequency domain [23]. The amplitude spectrum typically reflects image energy, brightness, and contrast distribution, while the phase spectrum is closely related to edges, contours, and structural positions [29]. In the captured image, wavelength-dependent absorption produces imbalanced responses across color channels, while scattering and blurring attenuate fine structures and high-frequency content [30]. These effects are reflected in the Fourier representation through changes in amplitude statistics and phase-related structural cues. Consequently, spatial-

domain convolution alone may be insufficient to fully characterize these complementary frequency-domain characteristics [11].

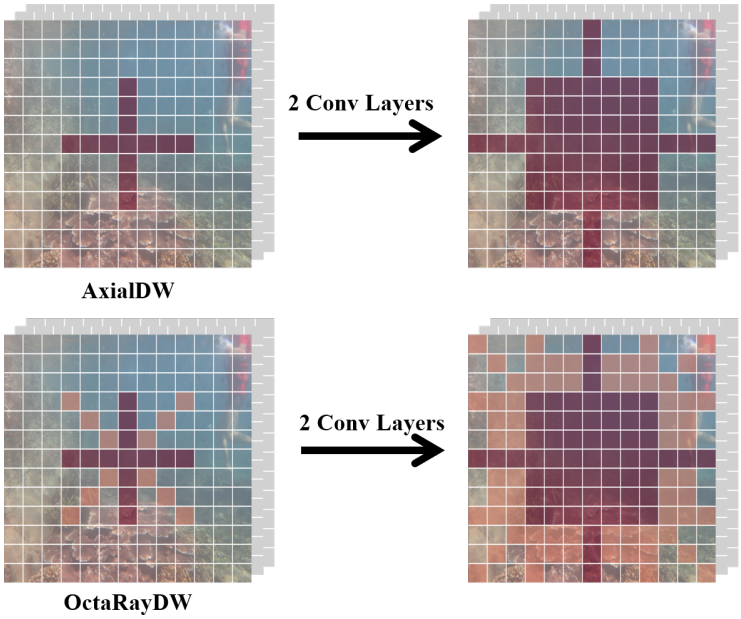


Figure 6. Conceptual comparison of directional aggregation by axial depthwise convolution and OctaRayDW.

Fourier amplitude–phase decomposition has already been used in underwater enhancement [10]. Our EFM specifies a spatially conditioned residual update to the amplitude-derived coefficient and an independent residual update to the phase, followed by recomposition and inverse Fourier transformation. The EFM structure is shown in Figure 7. The diagram shows that the resized spatial feature gates the amplitude update, whereas phase follows a separate branch until the two responses are recomposed.

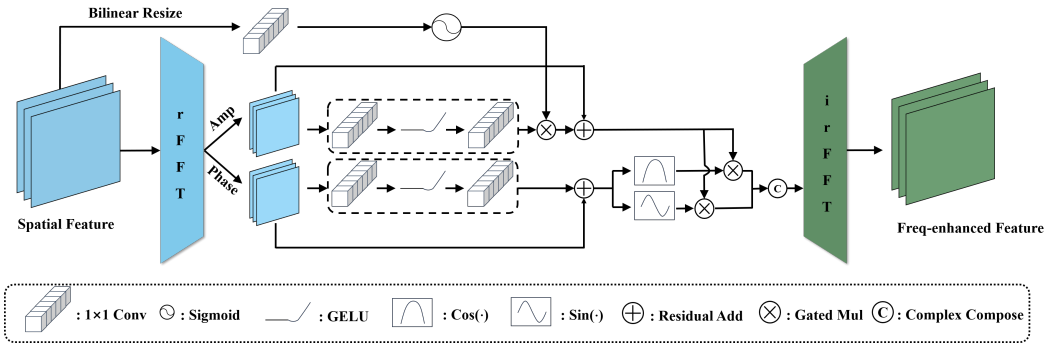


Figure 7. Architecture of EFM with separate amplitude and phase updates and inverse Fourier reconstruction.

Given an input spatial feature $X \in \mathbb{R}^{C\times H\times W}$, let $W_r = \lfloor W/2 \rfloor + 1$. The two-dimensional real fast Fourier transform (rFFT) is applied to obtain the one-sided complex spectrum $\mathcal{F}_r(X) \in \mathbb{C}^{C\times H\times W_r}$, which is decomposed into its amplitude and phase components:

$$\mathcal{F}_r(X) = A \odot e^{jP} \tag{10}$$

where $A = |\mathcal{F}_r(X)|$ and $P = \angle\mathcal{F}_r(X)$ denote the amplitude and phase spectra, respectively, and $A, P \in \mathbb{R}^{C\times H\times W_r}$, j denotes the imaginary unit satisfying $j^2 = -1$.

Since the spatial conditional feature and the one-sided spectral representation have different widths, the spatial feature is first resized through bilinear interpolation to match the

width of the one-sided spectral representation: $\bar{X} = \mathcal{R}_{\text{bil}}(X; H, W_r)$, where $\bar{X} \in \mathbb{R}^{C\times H\times W_r}$. Resizing only matches tensor dimensions and does not align spatial coordinates with Fourier indices. The gate is thus a learned modulation on the spectral grid. The amplitude-derived coefficient and phase response are then updated as

$$\begin{cases} A_m = A + \phi_A(A) \odot \sigma\big(\phi_g(\bar{X})\big) \\ \widetilde{P} = P + \phi_P(P) \end{cases} \tag{11}$$

where $\phi_A(\cdot)$ and $\phi_P(\cdot)$ denote the amplitude and phase transformation branches, respectively. Each branch consists of a 1×1 convolution, a GELU activation, and another 1×1 convolution. $\phi_g(\cdot)$ denotes the 1×1 convolution used to generate the amplitude gate, and $\sigma(\cdot)$ denotes the Sigmoid function. A_m is a real-valued spectral coefficient derived from the original amplitude response, while $\widetilde{P}$ denotes the refined phase response.

The two responses are subsequently recomposed into a one-sided complex spectrum and transformed back to the spatial domain:

$$X_f = \mathcal{F}^{-1}_{r,H,W}\Big(A_m \odot \Big[\cos(\widetilde{P}) + j\sin(\widetilde{P})\Big]\Big) \tag{12}$$

where $\mathcal{F}^{-1}_{r,H,W}(\cdot)$ denotes the two-dimensional irFFT with the output size specified as $H \times W$, and X_f denotes the reconstructed frequency-enhanced spatial feature. The cosine–sine recomposition is 2π-periodic, but the phase branch does not explicitly address wrapping or unstable phase values at low-magnitude frequencies. Specifying (H, W) in irFFT produces a real-valued feature, although the learned spectrum is not explicitly constrained to satisfy all Hermitian boundary relations. Circular shifts may alter spectral phase and the resized gate, and changes in image size alter the spectral grid, so translation equivariance and resolution invariance are not guaranteed.

3.4. Explicit Cross-Domain Fusion

Although EFM can explicitly characterize underwater degradation responses in the frequency domain and reconstruct frequency-enhanced spatial representations, the original spatial features and frequency-enhanced features emphasize complementary structural and spectral information. Cross-domain fusion has previously been studied in CDF-UIE [10], while FSIF-Net [11] interactively fuses spatial and frequency features. ECDF differs in its specified reciprocal construction: separate channel-wise attention operations take spatial queries with frequency-enhanced keys and values, and frequency-enhanced queries with spatial keys and values, respectively; a learned gate then combines their outputs. Figure 8 shows the two attention paths feeding the fusion gate before the residual refinement block. Given the spatial structural feature X_s and the frequency-enhanced spatial feature X_f obtained by EFM, ECDF first performs bidirectional cross-domain attention to exchange information:

$$\begin{cases} \widehat{X}_s = X_s + \text{CDA}_{s\leftarrow f}(X_s, X_f) \\ \widehat{X}_f = X_f + \text{CDA}_{f\leftarrow s}(X_f, X_s) \end{cases} \tag{13}$$

where $\text{CDA}_{s\leftarrow f}(\cdot,\cdot)$ and $\text{CDA}_{f\leftarrow s}(\cdot,\cdot)$ denote two cross-domain attention operations with independent learnable parameters. The arrow indicates the direction of information aggregation. In each operation, the query is generated from the first input, whereas the key and value are generated from the second input. Then, multi-head normalized attention is adopted for information aggregation:

$$\mathrm{CDA}(X_{\mathrm{src}}, X_{\mathrm{tgt}}) = \mathrm{Proj}_{\mathrm{CDA}}\Big(\mathrm{Softmax}\Big(\tau \odot (\overline{Q}_{\mathrm{src}}\overline{K}_{\mathrm{tgt}}^{T})\Big)V_{\mathrm{tgt}}\Big) \tag{14}$$

where $\overline{Q}_{\mathrm{src}}$ and $\overline{K}_{\mathrm{tgt}}$ denote the query and key normalized along the spatial dimension, respectively. V_{tgt} denotes the value generated from the target feature. τ is a learnable per-head scaling parameter, and $\mathrm{Proj}_{\mathrm{CDA}}(\cdot)$ denotes the 1×1 output projection of the cross-domain attention operation. The Softmax operation is performed along the last channel dimension. This attention mechanism establishes channel-wise response relationships between spatial features and frequency-enhanced representations, enabling the spatial branch to exploit degradation-aware spectral information while avoiding the quadratic complexity of spatial-wise attention. ECDF employs a gating mechanism to adaptively fuse the bidirectionally enhanced features:

$$\begin{cases} \alpha = \sigma\Big(\mathrm{Conv}_{1\times1}\Big([\widehat{X}_s, \widehat{X}_f]\Big)\Big) \\ X_c = \alpha \odot \widehat{X}_s + (1-\alpha) \odot \widehat{X}_f \end{cases} \tag{15}$$

where $[\cdot,\cdot]$ denotes channel-wise concatenation. The concatenated feature is used only to predict the content-adaptive fusion weight $\alpha \in \mathbb{R}^{C\times H\times W}$. The spatial and frequency-enhanced representations are then fused using the complementary weights α and $1-\alpha$, respectively. ECDF further enhances the fused feature using a lightweight refinement branch composed of depthwise convolution, GELU activation, and 1×1 convolution:

$$Y = X_c + \psi(X_c) \tag{16}$$

where $\psi(\cdot)$ denotes the local refinement operation.

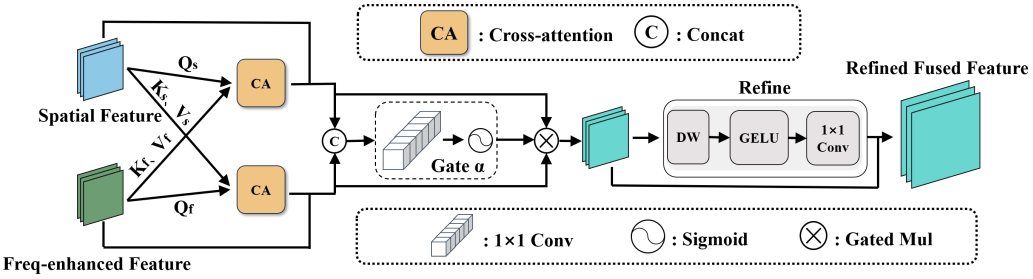


Figure 8. Architecture of ECDF with bidirectional attention, gated fusion, and residual refinement.

3.5. Local-Global Adaptive Fusion

Underwater image enhancement requires both local texture restoration and global visual consistency. Convolutional operations effectively characterize edges, textures, and other fine-grained structures, but their restricted receptive fields limit the association of spatially distant regions with similar degradation responses. Histoformer [27] already combines histogram self-attention and dynamic-range convolution for context and detail modeling. In LGAF, our specific modification is a separate spatially local convolutional path evaluated alongside the adopted histogram attention, with a content-dependent gate that explicitly combines their outputs.

To achieve a better balance between these two types of information, we develop the Local-Global Adaptive Fusion module, as illustrated in Figure 9. LGAF places a conventional local convolutional refinement path in parallel with the histogram path inherited from Histoformer. A content-adaptive gate regulates the contributions of the two outputs before the residual connection. Given an input feature X, LGAF first creates a

residual copy X_{res} to preserve the original feature for the final residual connection. Only the first half of channels in the copied feature are sorted along the height and width dimensions. Let $C_s = \{1, \ldots, \lfloor C/2 \rfloor\}$ denote the selected channel indices. This partial H/W sorting process can be written as

$$\begin{cases} X_{\text{res}} = \text{Clone}(X) \\ X_h, \Pi_h = \text{Sort}_H((X_{\text{res}})_{C_s}) \\ X_{hw}, \Pi_w = \text{Sort}_W(X_h) \\ X_{\text{sort}} = \text{Replace}_{C_s}(X_{\text{res}}, X_{hw}) \end{cases} \tag{17}$$

where Π_h and Π_w denote the sorting indices along the height and width dimensions, respectively. The rearranged feature is subsequently projected into the shared query, key, value, and local refinement features:

$$Q, K, V, L = \text{DWConv}_{3\times3}(\text{Conv}_{1\times1}(X_{sort})) \tag{18}$$

where X_{sort} denotes the spatially sorted feature. Q, K, and V are used for global attention modeling, while L is used for the local branch. The local enhancement branch adopts a lightweight convolutional structure to extract fine-grained textures:

$$X_l = \text{Conv}_{1\times1}(\text{GELU}(\text{DWConv}_{3\times3}(L))) \tag{19}$$

where $\text{GELU}(\cdot)$ denotes the GELU activation function. The local path employs a lightweight depthwise convolutional mixer. For the contextual path, we adopt the value-based sorting and complementary histogram reshaping operations of [27]. Specifically, V is sorted according to its response values, and the resulting indices are used to rearrange Q and K consistently. The rearranged features are processed by contiguous and interleaved attention paths, after which the original spatial order is restored:

$$\begin{aligned} X_g = \text{Restore}\big(&\text{Attn}_{box}(Q_s, K_s, V_s) \\ &\odot \text{Attn}_{inter}(Q_s, K_s, V_s)\big) \end{aligned} \tag{20}$$

where Q_s, K_s, and V_s denote the features rearranged according to the sorting indices of V. Attn_{box} and Attn_{inter} denote block-wise sorted attention and interleaved sorted attention, respectively. $\text{Restore}(\cdot)$ denotes restoring the attention result to the original spatial order according to the sorting indices. The attention computation adopts normalized queries and keys, and introduces learnable temperature parameters to improve training stability. Directly using the histogram-domain feature X_g treats contextual aggregation as an isolated transformation and does not explicitly regulate its interaction with convolutional refinement. LGAF therefore predicts a content-dependent fusion weight from the joint responses of X_l and X_g:

$$\begin{cases} Y_{\text{LGAF}} = X + \text{Proj}_{\text{LGAF}}(\gamma \odot X_l + (1-\gamma) \odot X_g) \\ \gamma = \sigma(\text{Conv}_{1\times1}([X_l, X_g])) \end{cases} \tag{21}$$

where $\gamma \in [0,1]^{C\times H\times W}$ denotes the content-dependent fusion weight, and $\text{Proj}_{\text{LGAF}}(\cdot)$ represents the 1×1 output projection of LGAF.

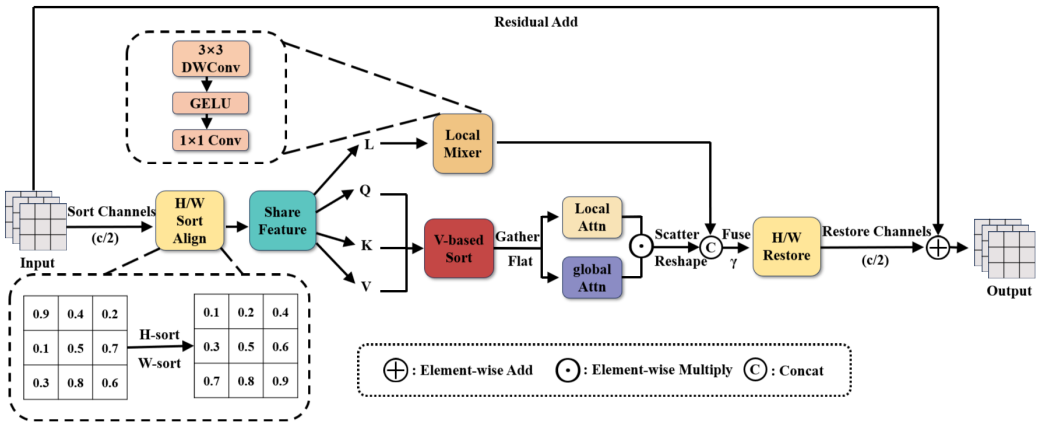


Figure 9. Architecture of LGAF with local convolution, histogram-domain attention, and adaptive fusion.

3.6. Loss Function

We train DSFNet in a supervised manner. Given a degraded underwater image I and its corresponding clear reference image J, the network outputs the enhanced result $\hat{J} = F_\theta(I)$. To simultaneously constrain pixel-level reconstruction accuracy and structural similarity, we jointly optimize the network using the L_1 loss and the SSIM loss:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{SSIM} \tag{22}$$

The L_1 loss is used to constrain the pixel-wise difference between the enhanced result and the reference image:

$$\mathcal{L}_1 = \left\| \hat{J} - J \right\|_1 \tag{23}$$

The SSIM loss is used to constrain luminance, contrast, and structural consistency:

$$\mathcal{L}_{SSIM} = 1 - SSIM(\hat{J}, J) \tag{24}$$

The L_1 loss helps improve pixel-level restoration accuracy, while the SSIM loss enhances the structure-preserving capability. By combining these two losses, the model can achieve more stable performance in color correction, contrast recovery, and texture detail enhancement.

4. Experiments

This section presents a systematic experimental evaluation of the proposed DSFNet to assess full-reference restoration on the paired UIEB validation set and perceived image quality on the other evaluation sets. Experiments are conducted on several publicly available underwater image enhancement datasets, including UIEB [31], U45 [32], MABLs [33], and UFO120 [34]. In the quantitative assessment, DSFNet is compared with eleven representative image enhancement methods, namely the original two-stage LU2Net [17], WaterNet [31], UWCNN [35], Ucolor [26], U-shape [25], SPDF [16], AST [36], MJPNet [37], VQCNIR [38], UDNet [39], and CDF-UIE [10]. The evaluation is carried out from both subjective visual quality and objective assessment metrics, providing a comprehensive analysis of the enhancement performance of the proposed method. Furthermore, to investigate the contribution of each component in DSFNet, a series of ablation experiments is designed to analyze the effects of the three core components, including OctaRayDW, the coupled EFM–ECDF branch, and LGAF on the overall network performance, thereby validating the effectiveness and complementarity of these core modules.

4.1. Implementation Details

We used the 800 paired UIEB training images to train DSFNet. The model was trained for 600 epochs with a batch size of 8. AdamW was adopted as the optimizer, with an initial learning rate of 2×10^{-4} and a weight decay of 1e-6. The exponential decay rates β_1 and β_2 in AdamW were set to 0.9 and 0.999, respectively. A cosine annealing learning rate scheduler with warm restarts was employed, where the initial restart period was set to 5 and the period multiplication factor was set to 3. To improve the generalization ability and robustness of the model, data augmentation strategies including random cropping, random horizontal flipping, random vertical flipping, and random rotations in multiples of 90° were applied during training. Training image–reference pairs were resized bilinearly so that the shorter side was 256 pixels, followed by synchronized random 256×256 crops and augmentations. The 90 paired UIEB images outside the training set were resized to 256×256 for validation after each epoch. We saved the checkpoint with the highest mean validation PSNR and used that checkpoint for the reported results. These same 90 images were also used for the reported full-reference evaluation. Experiments were conducted using NVIDIA GPUs, and the random seed was fixed to 42 to improve the reproducibility of the experimental results. For the LU2Net baseline, we used the authors' original two-stage architecture rather than the three-stage DSFNet ablation backbone. We retrained it on the same 800 paired UIEB training images with the same 600-epoch schedule, batch size, optimizer settings, L_1–SSIM objective, and best-validation-PSNR checkpoint selection as DSFNet. Both models were evaluated on the same 90 paired images. At inference, each image was processed at its original resolution after reflection padding to a multiple of 16, and the output was cropped back to the original size. For the reported metrics, enhanced images and their reference images, when available, were resized to 256×256. For epoch-wise validation, fixed 256×256 inputs provided a consistent and computationally manageable basis for checkpoint selection. For final evaluation, processing the images at their original resolution avoided discarding spatial detail before enhancement. Only the outputs and references were resized to 256×256 for metric calculation, so checkpoint selection and reported evaluation used different inference resolutions.

4.2. Datasets and Metrics

Datasets: We conduct experiments on the paired UIEB dataset and its Challenging-60 (C60) subset, U45, MABLs, and UFO120. UIEB is a real-world underwater image enhancement dataset containing various complex degradation scenarios, such as low illumination, turbid water, and color cast. It consists of 950 images, among which 890 are paired images. We use 800 paired images for training and the other 90 paired images for validation, checkpoint selection, and quantitative evaluation. The Challenging-60 (C60) subset is a separate subset of UIEB containing 60 challenging underwater images with severe degradation and no reference images, and is used to further evaluate enhancement performance under difficult conditions. U45 contains typical underwater degradations, including low contrast, color deviation, and haze-like effects. MABLs and UFO120 are further used to assess perceived image quality in different underwater scenarios using the adopted no-reference metrics. For consistency with the other no-reference evaluations, UFO120 is evaluated using no-reference metrics despite the availability of reference images.

Full Reference (FR): We adopt three full-reference metrics, namely PSNR, SSIM, and RMSE, to evaluate the similarity between the enhanced results and the reference images. PSNR measures the pixel-level reconstruction quality, SSIM evaluates image similarity in terms of luminance, contrast, and structure, and RMSE calculates the root mean squared error between the enhanced image and the reference image. Higher PSNR and SSIM values, together with a lower RMSE value, indicate better enhancement performance.

No Reference (NR): Considering that reference images are often unavailable for real-world underwater images, we employ UCIQE [40] and URanker [41] for no-reference quality assessment. URanker estimates perceived underwater image quality, including the appearance of color and detail. UCIQE summarizes chroma, saturation, and luminance contrast. Higher URanker and UCIQE values indicate better perceptual quality.

4.3. Performance Assessment

This section provides a comprehensive comparative analysis between the proposed method and existing advanced UIE methods, covering both qualitative and quantitative aspects. The qualitative analysis demonstrates the performance of different methods in underwater image enhancement through visual comparisons, while the quantitative analysis objectively evaluates their performance using multiple evaluation metrics introduced in Section 4.2.

Qualitative Comparison: Figure 10 presents a visual comparison between our method and ten representative underwater image enhancement methods. As can be observed, although WaterNet alleviates the original color cast to some extent, its results generally appear grayish and dark, with limited recovery of details in shadow regions. UWCNN still exhibits noticeable cool color casts in certain scenes, while UDNet tends to introduce greenish tones, leading to less natural enhanced results. U-shape and AST significantly improve image brightness; however, they are prone to overenhancement in bright regions, such as sandy areas and cave entrances. SPDF produces purple or pink color casts in some results, whereas MJPNet tends to generate locally warm tones and overly strong contrast. Ucolor, VQCNIR, and CDF-UIE improve the overall visual quality to a certain extent, but still show limitations in local color consistency and natural detail restoration. In contrast, our method more consistently corrects underwater color casts across different scenes, appropriately improves brightness and contrast, and preserves more texture and structural details, ultimately producing enhanced images with more natural colors and clearer visual quality.

To further evaluate the visual enhancement performance of the proposed method, Figure 11 presents local zoom-in comparisons of different methods on the same underwater scene. The red and blue boxes highlight regions with prominent texture and edge details, respectively. It can be observed that some competing methods suffer from insufficient color correction, local blurring, or overenhancement. In contrast, the proposed method restores the colors of underwater scenes more naturally while better preserving detailed textures in regions such as ropes, rocks, and attached objects. Figure 12 shows the RGB channel ratios of the enhanced results. In the displayed examples, the channel-ratio plots show reduced blue-green dominance relative to the raw images. Figure 13 presents the pixel distribution comparison between different methods and the reference clear image, where the red curve represents the enhanced image and the blue curve represents the reference image. For the displayed example, the proposed result exhibits a pixel distribution closer to the reference, reflecting similarity in global intensity statistics.

Figure 14 further provides the three-dimensional L-channel error surfaces between the enhanced results and the reference image. The blue-to-red color scale indicates increasing error, and a lower, flatter surface indicates less luminance mismatch in the illustrated scene. For this example, the DSFNet error surface shows less mismatch with the reference than several competing methods.



Figure 10. Qualitative comparison on the UIEB validation set. (**a**) Raw, (**b**) WaterNet [31], (**c**) UWCNN [35], (**d**) Ucolor [26], (**e**) U-shape [25], (**f**) SPDF [16], (**g**) AST [36], (**h**) MJPNet [37], (**i**) VQCNIR [38], (**j**) UDNet [39], (**k**) CDF-UIE [10], (**l**) ours.

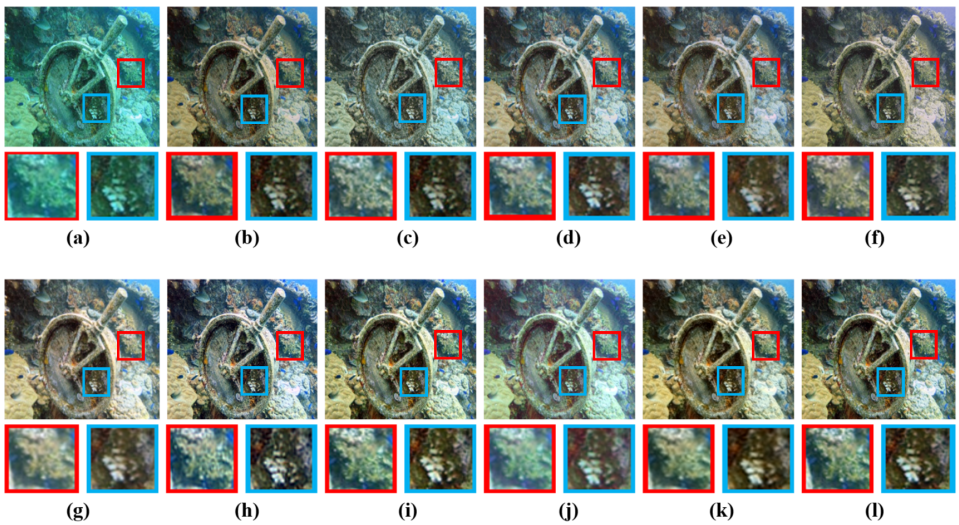


Figure 11. Local-detail comparison on the UIEB validation set. (**a**) Raw, (**b**) WaterNet [31], (**c**) UWCNN [35], (**d**) Ucolor [26], (**e**) U-shape [25], (**f**) SPDF [16], (**g**) AST [36], (**h**) MJPNet [37], (**i**) VQCNIR [38], (**j**) UDNet [39], (**k**) CDF-UIE [10], (**l**) ours.

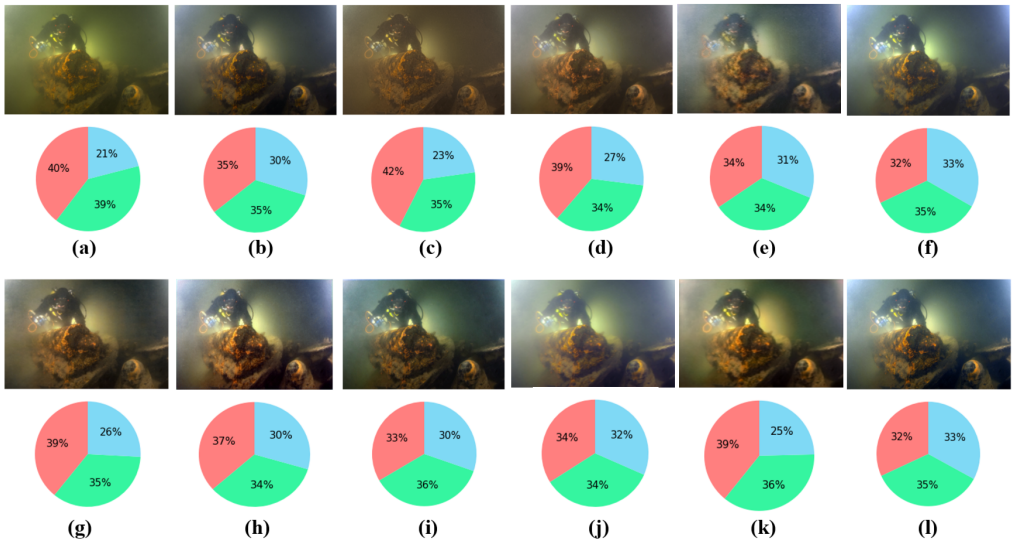


Figure 12. Analyzing the color ratio for the red, green, and blue channels. (**a**) Raw, (**b**) WaterNet [31], (**c**) UWCNN [35], (**d**) Ucolor [26], (**e**) U-shape [25], (**f**) SPDF [16], (**g**) AST [36], (**h**) MJPNet [37], (**i**) VQCNIR [38], (**j**) UDNet [39], (**k**) CDF-UIE [10], (**l**) ours.

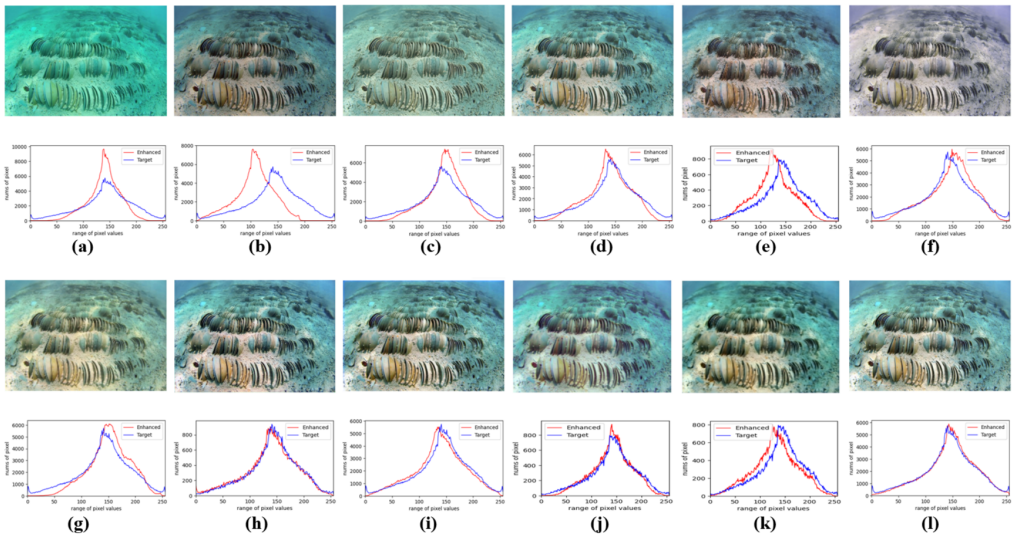


Figure 13. Pixel distributions of enhanced results and the reference image. (**a**) Raw, (**b**) WaterNet [31], (**c**) UWCNN [35], (**d**) Ucolor [26], (**e**) U-shape [25], (**f**) SPDF [16], (**g**) AST [36], (**h**) MJPNet [37], (**i**) VQCNIR [38], (**j**) UDNet [39], (**k**) CDF-UIE [10], (**l**) ours.

Quantitative Comparison: FSIF-Net is discussed as a related architecture but is not assigned a numerical result: at the time of this revision, we could not identify verifiable author-released implementation code or pretrained weights for evaluation in the same protocol. The quantitative evaluation results are presented in Table 1. For the UIEB dataset with reference images, PSNR, SSIM, and RMSE are adopted as evaluation metrics. For C60, U45, MABLs, and UFO120, we employ URanker and UCIQE as no-reference image quality assessment metrics. As shown in Table 1, the proposed method achieves the best performance on the UIEB dataset, with a PSNR of 25.352 dB, an SSIM of 0.929, and an RMSE of 0.062, indicating that our method can effectively improve image reconstruction quality while preserving structural consistency. The retrained original two-stage LU2Net obtains 24.648 dB PSNR, 0.925 SSIM, and 0.066 RMSE under the same evaluation protocol, so DSFNet improves PSNR by 0.704 dB and SSIM by 0.004 relative to this direct baseline. For C60, U45, MABLs, and UFO120, DSFNet is evaluated in terms of perceived image

quality using no-reference metrics. It obtains the highest URanker and UCIQE scores on the C60 dataset, and achieves the best or near-best results on the U45, MABLs, and UFO120 datasets. These no-reference results indicate favorable perceptual quality under the adopted no-reference evaluation protocol. Because the same 90 paired UIEB images were used for checkpoint selection and full-reference evaluation, the reported PSNR, SSIM, and RMSE are validation-set results and may be optimistic estimates of performance on independent paired images. The comparisons in Table 1 should therefore be interpreted within this validation split.

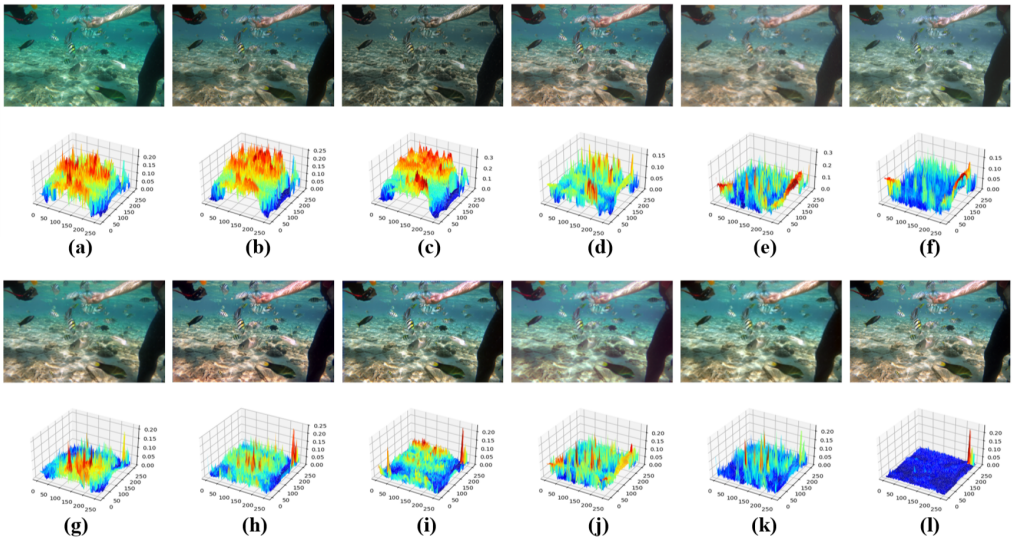


Figure 14. Error surfaces for a natural-light scene in the L-channel of the HLS space. (**a**) Raw, (**b**) WaterNet [31], (**c**) UWCNN [35], (**d**) Ucolor [26], (**e**) U-shape [25], (**f**) SPDF [16], (**g**) AST [36], (**h**) MJPNet [37], (**i**) VQCNIR [38], (**j**) UDNet [39], (**k**) CDF-UIE [10], (**l**) ours.

Table 1. Quantitative comparisons of different methods on the UIEB, C60, U45, MABLs, and UFO120 datasets. The best and second-best results are highlighted in bold and underlined, respectively. Ties at the reported precision share the same highlighting.

Method	UIEB					C60		U45		MABLs		UFO120	
	PSNR ↑	SSIM ↑	RMSE ↓	URanker ↑	UCIQE ↑	URanker ↑	UCIQE ↑	URanker ↑	UCIQE ↑	URanker ↑	UCIQE ↑	URanker ↑	UCIQE ↑
WaterNet	17.349	0.813	0.144	1.471	0.606	0.931	0.591	1.330	0.594	1.448	**0.600**	1.748	0.615
UWCNN	17.985	0.844	0.134	1.309	0.554	0.499	0.530	1.513	0.554	1.030	0.532	1.893	0.579
Ucolor	20.962	0.864	0.097	1.534	0.591	0.841	0.565	1.471	0.586	1.255	0.573	1.754	0.612
U-shape	20.462	0.792	0.100	1.665	0.576	1.059	0.553	1.502	0.572	1.514	0.561	1.530	0.572
SPDF	20.331	0.828	0.103	1.433	0.570	1.034	0.569	1.284	0.565	1.413	0.560	1.678	0.592
AST	21.695	0.869	0.089	1.791	0.585	1.042	0.558	1.776	0.589	1.363	0.564	1.938	0.608
MJPNet	21.288	0.823	0.093	2.062	0.608	1.431	0.585	1.913	0.607	1.735	0.588	**2.144**	0.629
VQCNIR	23.330	0.914	0.074	2.069	0.615	1.446	0.588	**2.067**	**0.616**	1.812	0.592	2.122	**0.635**
UDNet	19.017	0.768	0.118	1.216	0.585	0.618	0.568	0.858	0.583	0.997	0.569	1.506	0.602
CDF-UIE	22.886	0.885	0.077	2.042	0.594	1.387	0.569	1.915	0.596	1.798	0.578	2.134	0.620
LU2Net (two-stage)	24.648	0.925	0.066	1.537	0.611	0.809	0.592	1.917	**0.616**	1.414	0.592	2.008	0.631
Ours	**25.352**	**0.929**	**0.062**	**2.133**	**0.617**	**1.516**	**0.597**	1.938	0.612	**1.869**	0.598	**2.144**	0.633

4.4. Ablation Study

To verify the effectiveness of each module in DSFNet, we conducted quantitative and visual ablation experiments on the UIEB validation set, with the visual comparisons presented in Figure 15. In Tables 2–5, Base denotes the internal three-stage encoder–decoder with LU2Net's axial depthwise convolution and channel attention, without EFM, ECDF, or LGAF. This internal ablation baseline is distinct from the original two-stage LU2Net evaluated in Table 1. In these runs, the three-stage Base reaches 23.730 dB PSNR and 0.906 SSIM, below the retrained two-stage LU2Net at 24.648 dB and 0.925; the added stage alone does not account for the full-model result. Table 2 compares different spatial structure modeling strategies. Compared with SquareDW and AxialDW, OctaRayDW

achieves the best performance, indicating that relying solely on local neighborhoods or horizontal and vertical directional modeling is insufficient to fully characterize the complex directional structures in underwater images. By employing multidirectional depthwise convolutions, OctaRayDW enhances the representation of edges, textures, and diagonal structures, thereby improving structural restoration capability.

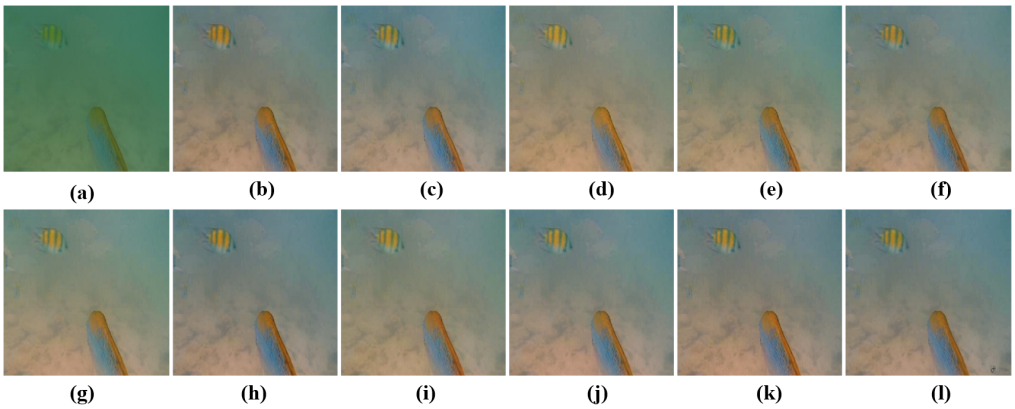


Figure 15. Visual comparison for ablation studies on the UIEB dataset. (**a**) Raw, (**b**) OctaRayDW, (**c**) EFM ECDF, (**d**) LGAF, (**e**) OctaRayDW EFM ECDF, (**f**) OctaRayDW LGAF, (**g**) EFM ECDF LGAF, (**h**) AxialDW EFM ECDF LGAF, (**i**) OctaRayDW EFM LGAF, (**j**) OctaRayDW EFM ECDF DHSA, (**k**) OctaRayDW EFM ECDF LGAF, (**l**) GT.

Table 2. Quantitative analysis of ablation experiments. The best performance is highlighted in bold.

Operation	PSNR	SSIM
Base	23.730	0.906
AxialDW EFM ECDF LGAF	24.870	0.926
SquareDW EFM ECDF LGAF	25.280	0.926
OctaRayDW EFM ECDF LGAF	**25.352**	**0.929**

Table 3 presents the ablation results of EFM and ECDF. Without ECDF, the EFM output is added directly to the spatial feature in each encoder and decoder block. This fixed fusion reduces PSNR from 24.677 to 23.910 dB relative to the variant without EFM, suggesting that simply adding the frequency-derived response can interfere with the spatial representation in this configuration. Concatenation and either one-way interaction perform better than direct addition. The stronger one-way variant, in which frequency features query spatial features, reaches 24.997 dB and 0.924 SSIM. Bidirectional ECDF reaches 25.352 dB and 0.929 SSIM, an improvement of 0.355 dB and 0.005 SSIM over that variant. All variants share the same backbone and training protocol, while the fusion modules have different parameter counts.

Table 3. Ablation of spatial–frequency fusion on UIEB. The best performance is highlighted in bold.

Operation	PSNR	SSIM
Base	23.730	0.906
OctaRayDW LGAF	24.677	0.926
OctaRayDW EFM LGAF (addition)	23.910	0.914
OctaRayDW EFM LGAF (concatenation)	24.739	0.924
OctaRayDW EFM LGAF (spatial from frequency)	24.724	0.915
OctaRayDW EFM LGAF (frequency from spatial)	24.997	0.924
OctaRayDW EFM ECDF LGAF (bidirectional)	**25.352**	**0.929**

Table 4 evaluates LGAF within the DSFNet backbone. Under this configuration, replacing LGAF with the Dynamic-range Histogram Self-Attention (DHSA) module of Histoformer [27] yields 24.426 dB PSNR and 0.913 SSIM, compared with 25.352 dB and 0.929 for LGAF. This comparison supports the added local path and its gate in the evaluated architecture; it does not imply that the original Histoformer lacks convolutional or local feature modeling.

Table 4. Quantitative analysis of ablation experiments. The best performance is highlighted in bold.

Operation	PSNR	SSIM
Base	23.730	0.906
OctaRayDW EFM ECDF	25.263	0.927
OctaRayDW EFM ECDF DHSA	24.426	0.913
OctaRayDW EFM ECDF LGAF	**25.352**	**0.929**

Table 5 further verifies the complementarity among the modules. Introducing OctaRayDW, LGAF, or the coupled EFM–ECDF branch individually brings performance improvements, indicating that they enhance the model representation from the perspectives of directional structure, local-global adaptive fusion, and frequency-domain degradation, respectively. When all three components are used simultaneously, DSFNet achieves the best results, demonstrating that although some pairwise combinations do not yield monotonic improvements, the complete configuration achieves the best overall performance, indicating that the three components are most effective when jointly integrated in the full architecture.

Table 5. Quantitative analysis of ablation experiments. The best performance is highlighted in bold.

Operation	PSNR	SSIM
Base	23.730	0.906
OctaRayDW	24.379	0.924
EFM ECDF	25.070	0.927
LGAF	24.500	0.922
OctaRayDW EFM ECDF	25.263	0.927
OctaRayDW LGAF	24.677	0.926
EFM ECDF LGAF	24.870	0.926
OctaRayDW EFM ECDF LGAF	**25.352**	**0.929**

5. Conclusions

In this work, we proposed DSFNet, a degradation-aware spatial-frequency collaborative network for underwater optical image enhancement. By jointly modeling directional structures, frequency-domain degradation characteristics, and local-global contextual dependencies, DSFNet effectively restored underwater visual information under complex degradation conditions. On the UIEB validation split, DSFNet achieved the highest PSNR and SSIM and the lowest RMSE among the compared methods. The no-reference scores indicated favorable perceived quality in the evaluated scenarios. Ablation studies further verified the effectiveness and complementarity of the proposed components. The model contained 5.58 M parameters. Its deployment on underwater robotic hardware, including inference latency, memory use, and downstream vision performance, was not evaluated in this study.

Author Contributions: Conceptualization, W.L., J.L., J.Z., D.Z. and Q.J.; Methodology, W.L., J.L., J.Z. and D.Z.; Software, J.Z.; Validation, W.L. and J.Z.; Formal analysis, J.Z.; Investigation, J.Z.; Resources, J.Z.; Data curation, W.L. and J.Z.; Writing—original draft, W.L., J.L., J.Z., D.Z., G.Z. and Q.J.; Writing—review & editing, W.L., J.L., J.Z., D.Z., G.Z. and Q.J.; Visualization, W.L., J.L. and J.Z.; Supervision, J.Z. and G.Z.; Project administration, J.L., J.Z. and Q.J.; Funding acquisition, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Open Project of the State Key Laboratory of Ocean Sensing (No. 2026003), in part by the Open Project of the State Key Laboratory of Ocean Sensing (No. OSKF-2025M04), in part by the Natural Science Foundation of Zhejiang (No. LR22F020002), in part by the Liaoning Provincial Science and Technology Plan Joint Program (No. 2025-MSLH-113), in part by the Fundamental Research Funds for the Central Universities (No. 3132026240), and in part by the Fundamental Research Funds of the Liaoning Provincial Department of Education (No. LJ212510151021).

Data Availability Statement: The datasets used in this study are publicly available from their original sources. The DSFNet implementation code is publicly available at https://github.com/a2012255201/DSFNet.

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

References

1. Zhang, D.; Cao, W.; Zhou, J.; Peng, Y.T.; Zhang, W.; Lin, Z. Two-Branch Underwater Image Enhancement and Original Resolution Information Optimization Strategy in Ocean Observation. *J. Mar. Sci. Eng.* **2023**, *11*, 1285. [CrossRef]
2. Wang, R.; Zhang, Y.; Zhang, Y. Agent-Guided Non-Local Network for Underwater Image Enhancement and Super-Resolution Using Multi-Color Space. *J. Mar. Sci. Eng.* **2024**, *12*, 358. [CrossRef]
3. Akkaynak, D.; Treibitz, T. A Revised Underwater Image Formation Model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6723–6732.
4. Peng, Y.T.; Cosman, P.C. Underwater Image Restoration Based on Image Blurriness and Light Absorption. *IEEE Trans. Image Process.* **2017**, *26*, 1579–1594. [CrossRef] [PubMed]
5. Islam, M.J.; Xia, Y.; Sattar, J. Fast Underwater Image Enhancement for Improved Visual Perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. [CrossRef]
6. Wu, J.; Zhang, G.; Fan, Y. MambaRA-GAN: Underwater Image Enhancement via Mamba and Intra-Domain Reconstruction Autoencoder. *J. Mar. Sci. Eng.* **2025**, *13*, 1745. [CrossRef]
7. Lu, C.; Hong, L.; Fan, Y.; Shu, X. A Multi-Scale Contextual Fusion Residual Network for Underwater Image Enhancement. *J. Mar. Sci. Eng.* **2025**, *13*, 1531. [CrossRef]
8. Yeh, C.H.; Lai, Y.W.; Lin, Y.Y.; Chen, M.J.; Wang, C.C. Underwater Image Enhancement Based on Light Field-Guided Rendering Network. *J. Mar. Sci. Eng.* **2024**, *12*, 1217. [CrossRef]
9. Fan, G.; Zhou, Y.; Zhou, J.; Ju, Y.; Chen, G.Y.; Li, J.; Kot, A.C. DCD-UIE: Decoupled Chromatic Diffusion Model for Underwater Image Enhancement. *IEEE Trans. Image Process.* **2026**, *35*, 449–464. [CrossRef] [PubMed]
10. Zhang, H.; Xu, H.; Yu, X.; Zhang, X.; Gao, X.; Wu, C. CDF-UIE: Leveraging Cross-Domain Fusion for Underwater Image Enhancement. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4203715. [CrossRef]
11. Li, G.; Sun, S.; Zhang, C.; Jiao, G. A frequency and spatial domain interactive fusion network for underwater image enhancement. *Knowl.-Based Syst.* **2026**, *338*, 115499. [CrossRef]
12. Zhou, J.; Zhang, D.; He, Z.; Gai, Q.; Jiang, Q. Multi-Prior Fusion Transfer Plugin for Adapting In-Air Models to Underwater Image Enhancement and Detection. *IEEE Trans. Image Process.* **2025**, *34*, 7773–7785. [CrossRef] [PubMed]
13. Anwar, S.; Li, C. Diving Deeper into Underwater Image Enhancement: A Survey. *Signal Process. Image Commun.* **2020**, *89*, 115978. [CrossRef]
14. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. Vmamba: Visual State Space Model. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024; Volume 37.
15. Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; Xia, S.T. MambaIR: A Simple Baseline for Image Restoration with State-Space Model. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 222–241. [CrossRef]

16. Kang, Y.; Jiang, Q.; Li, C.; Ren, W.; Liu, H.; Wang, P. A Perception-Aware Decomposition and Fusion Framework for Underwater Image Enhancement. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 988–1002. [CrossRef]
17. Yang, H.; Xu, J.; Lin, Z.; He, J. LU2Net: A lightweight network for real-time underwater image enhancement. *arXiv* **2024**, arXiv:2406.14973.
18. Huang, X.; Yang, Z.; Wu, C.; Li, S.; Zhi, S.; Zhang, Y.; Hou, Q.; Deng, X.; Xia, J. Degradation Frequency Curve: An Explicit Frequency-Quantified Representation for All-in-One Image Restoration. *arXiv* **2026**, arXiv:2605.17506.
19. Li, C.; Guo, C.L.; Zhou, M.; Liang, Z.; Zhou, S.; Feng, R.; Loy, C.C. Embedding Fourier for Ultra-High-Definition Low-Light Image Enhancement. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
20. Ren, B.; Zamfir, E.; Wu, Z.; Li, Y.; Li, Y.; Paudel, D.P.; Timofte, R.; Yang, M.H.; Van Gool, L.; Sebe, N. Any Image Restoration via Efficient Spatial-Frequency Degradation Adaptation. *Trans. Mach. Learn. Res.* **2026**. [CrossRef]
21. Huo, F.; Li, B.; Zhu, X. Efficient Wavelet Boost Learning-Based Multi-Stage Progressive Refinement Network for Underwater Image Enhancement. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 17 October 2021.
22. Zhao, C.; Cai, W.; Dong, C.; Zeng, Z. Toward sufficient spatial-frequency interaction for gradient-aware underwater image enhancement. In *Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing*; IEEE: New York, NY, USA, 2024; pp. 3220–3224.
23. Lin, P.; Wang, Y.; Wang, Z.; Cao, N.; Yáng, Y.R.; Fu, X. Underwater image enhancement based on spatial-frequency feature interaction. *Opt. Lasers Eng.* **2026**, *198*, 109492. [CrossRef]
24. Pei, X.; Huang, T.; Xu, C. EfficientVMamba: Atrous Selective Scan for Light Weight Visual Mamba. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 6443–6451. [CrossRef]
25. Peng, L.; Zhu, C.; Bian, L. U-Shape Transformer for Underwater Image Enhancement. *IEEE Trans. Image Process.* **2023**, *32*, 3066–3079. [CrossRef] [PubMed]
26. Li, C.; Anwar, S.; Hou, J.; Cong, R.; Guo, C.; Ren, W. Underwater Image Enhancement via Medium Transmission-Guided Multi-Color Space Embedding. *IEEE Trans. Image Process.* **2021**, *30*, 4985–5000. [CrossRef] [PubMed]
27. Sun, S.; Ren, W.; Gao, X.; Wang, R.; Cao, X. Restoring images in adverse weather conditions via histogram transformer. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 111–129.
28. Ding, X.; Guo, Y.; Ding, G.; Han, J. ACNet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1911–1920.
29. Zeng, R.; Zhu, X.; Peng, W.; Cao, J.; Guo, Y.; Wang, Z.; Jiang, Q. Amplitude exchanging network for unsupervised underwater image enhancement. *Pattern Recognit.* **2026**, *175*, 113119. [CrossRef]
30. Chen, C.; Zhou, Z.; Zhao, X.; Yang, Y.; Zhang, X.; Tao, R.; Su, B. Dual-Domain Fractional Fourier Transformer for Underwater Image Degradation Removal. *IEEE Signal Process. Lett.* **2026**, *33*, 1121–1125. [CrossRef]
31. Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; Tao, D. An Underwater Image Enhancement Benchmark Dataset and Beyond. *IEEE Trans. Image Process.* **2020**, *29*, 4376–4389. [CrossRef] [PubMed]
32. Li, H.; Li, J.; Wang, W. A Fusion Adversarial Underwater Image Enhancement Network with a Public Test Dataset. *arXiv* **2019**, arXiv:1906.06819.
33. Song, W.; Wang, Y.; Huang, D.; Liotta, A.; Perra, C. Enhancement of Underwater Images With Statistical Model of Background Light and Optimization of Transmission Map. *IEEE Trans. Broadcast.* **2020**, *66*, 153–169. [CrossRef]
34. Islam, M.J.; Luo, P.; Sattar, J. Simultaneous Enhancement and Super-Resolution of Underwater Imagery for Improved Visual Perception. In Proceedings of the Robotics: Science and Systems, Virtual, 12–16 July 2020. [CrossRef]
35. Li, C.; Anwar, S.; Porikli, F. Underwater Scene Prior Inspired Deep Underwater Image and Video Enhancement. *Pattern Recognit.* **2020**, *98*, 107038. [CrossRef]
36. Zhou, S.; Chen, D.; Pan, J.; Shi, J.; Yang, J. Adapt or Perish: Adaptive Sparse Transformer with Attentive Feature Refinement for Image Restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, DC, USA, 16–22 June 2024; pp. 2952–2963.
37. Fan, J.; Xu, J.; Zhou, J.; Meng, D.; Lin, Y. See Through Water: Heuristic Modeling Toward Color Correction for Underwater Image Enhancement. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 4039–4054. [CrossRef]
38. Zou, W.; Gao, H.; Ye, T.; Chen, L.; Yang, W.; Huang, S.; Chen, H.; Chen, S. VQCNIR: Clearer Night Image Restoration with Vector-Quantized Codebook. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 7873–7881. [CrossRef]
39. Saleh, A.; Sheaves, M.; Jerry, D.; Rahimi Azghadi, M. Adaptive deep learning framework for robust unsupervised underwater image enhancement. *Expert Syst. Appl.* **2025**, *268*, 126314. [CrossRef]

40. Yang, M.; Sowmya, A. An underwater color image quality evaluation metric. *IEEE Trans. Image Process.* **2015**, *24*, 6062–6071. [CrossRef] [PubMed]
41. Guo, C.; Wu, R.; Jin, X.; Han, L.; Zhang, W.; Chai, Z.; Li, C. Underwater ranker: Learn which is better and how to be better. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 702–709. [CrossRef]

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.