Wavelet-based Fourier Information Interaction with Frequency Diffusion Adjustment for Underwater Image Restoration

Chen Zhao, Weiling Cai*, Chenyu Dong, Chengwei Hu
School of Artificial Intelligence, Nanjing Normal University

Abstract

Underwater images are subject to intricate and diverse degradation, inevitably affecting the effectiveness of underwater visual tasks. However, most approaches primarily operate in the raw pixel space of images, which limits the exploration of the frequency characteristics of underwater images, leading to an inadequate utilization of deep models' representational capabilities in producing high-quality images. In this paper, we introduce a novel Underwater Image Enhancement (UIE) framework, named WF-Diff, designed to fully leverage the characteristics of frequency domain information and diffusion models. WF-Diff consists of two detachable networks: Wavelet-based Fourier information interaction network (WFI2-net) and Frequency Residual Diffusion Adjustment Module (FRDAM). With our full exploration of the frequency domain information, WFI2-net aims to achieve preliminary enhancement of frequency information in the wavelet space. Our proposed FRDAM can further refine the high- and low-frequency information of the initial enhanced images, which can be viewed as a plug-and-play universal module to adjust the detail of the underwater images. With the above techniques, our algorithm can show SOTA performance on real-world underwater image datasets, and achieves competitive performance in visual quality. The code is available at https://github.com/zhihefang/WF-Diff.

1. Introduction

Underwater image restoration is a practical but challenging technology in the field of underwater vision, widely used for tasks, such as underwater robotics[26] and underwater object tracking[6]. Due to light refraction, absorption, and scattering in underwater scenes, underwater images are usually severely distorted, with low contrast and blurriness [2]. Therefore, clear underwater images play a critical role in fields that need to interact with the underwater environment. The main goal of underwater image enhancement (UIE) is to obtain high-quality images by removing scattering and correcting color distortion in degraded images. UIE is crucial for vision-related underwater tasks.

*Corresponding Author. E-mail: caiwl@njnu.edu.cn

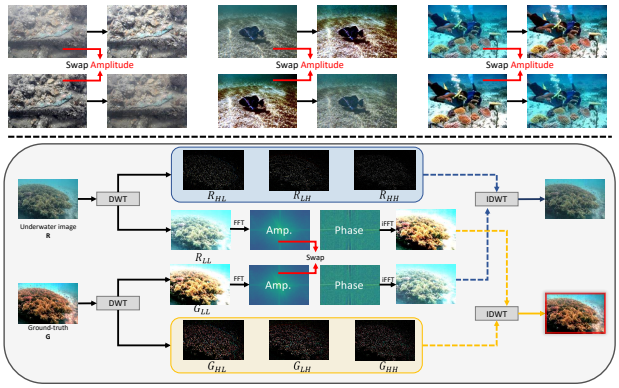

Figure 1. Our motivations. The amplitude and phase are produced by Fast Fourier Transform (FFT) and the recombined images are obtained by Inverse FFT (IFFT). We further explore the frequency properties for underwater images in Wavelet space.

Table 1. Evaluation of using different frequency domain transformation strategies on the UIEBD [28]. S1 refers to swapping the amplitude in original images, S2 refers to only swapping the amplitude of low-frequency sub-images in the wavelet space, and S3 refers to swapping the amplitude of low-frequency sub-images and high-frequency sub-images in the wavelet space.

Strategy	PSNR ↑	SSIM ↑	LPIPS↓	FID↓
S1	28.68	0.9027	0.0957	28.70
S2	27.10	0.8813	0.1023	33.04
S3	**29.97**	**0.9343**	**0.0820**	**23.55**

To address this problem, traditional UIE methods based on the physical properties of the underwater images were proposed [15, 17, 29–31]. These methods investigate the physical mechanism of the degradation caused by color cast or scattering and compensate for them to enhance the underwater images. However, these physics-based models with limited representation capacity cannot address all the complex physical and optical factors underlying the underwater scenes, which leads to poor enhancement results under highly complex and diverse underwater scenes. Recently, some learning-based methods [7, 18, 28, 36] for UIE can produce better results, since neural networks have powerful feature representation and nonlinear mapping capabili-

ties. It can learn the mapping of an image from degenerate to clear from a substantial quantity of paired training data. However, most previous methods are based on the raw pixel space of images, with limited exploration of the properties of the frequency space for underwater images, which results in an inability to effectively harness the representation power of the deep models for generating high-quality images.

Building on insights from previous Fourier-based works [12, 48], we explore the properties of the Fourier frequency information for the UIE task, as illustrated in Figure 1. Given two images (an underwater image and its corresponding ground-truth), we swap their amplitude components and combine them with corresponding phase components in the Fourier space. The recombined results show that the visual appearance is swapped following the amplitude swapping, which indicates the degradation information of underwater images is mainly contained in the amplitude component. We further explore the properties of the amplitude components in the Wavelet space. Specifically, the images can be decomposed into low-frequency sub-images and high-frequency sub-images using discrete wavelet transformations (DWT), and then we swap amplitude components of low-frequency sub-images. From visual results, we can find a similar phenomenon, which means the color degradation information is mainly contained in low-frequency sub-images, and the texture and detail degradation information is mainly contained in high-frequency sub-images. Table 1 shows the quantitative evaluation of the different frequency domain strategies, proving that our discovery is objective. Consequently, how to adequately exploit the properties of frequency domain information and effectively incorporate them into a unified image enhancement network is a crucial issue.

Recently, diffusion-based methods [10, 35] have garnered significant attention due to their outstanding performance in image synthesis [23, 24, 32, 34, 52] and restoration tasks [5, 40, 46, 51]. These methods rely on a hierarchical denoising autoencoder architecture, enabling them to iteratively reverse a diffusion process and achieve high-quality mapping from randomly sampled Gaussian noise to target images or latent distributions [10]. Tang et al. [36] present an image enhancement approach with diffusion model in underwater scenes. While standard diffusion models exhibit sufficient capability, unforeseen artifacts may arise as a result of the diversity introduced during the sampling process from randomly generated Gaussian noise to images [45]. Furthermore, the diffusion model needs to recover both the high and low-frequency information of images, which limits their ability to focus on fine-grained information, missing out on texture and details. Thereby, it is very crucial that the powerful representation capabilities of diffusion models can be fully utilized.

In this paper, we develop a novel UIE framework to fully exploit the properties of frequency domain information and diffusion models, called WF-Diff, which mainly consists of two stages: frequency preliminary enhancement and frequency diffusion adjustment. The first stage aims to preliminarily enhance the high-frequency and low-frequency components of underwater images by utilizing the frequency domain characteristics. Specifically, we first convert the input images into the wavelet space using discrete wavelet transformations (DWT), obtaining an average coefficient that represents the low-frequency content information of the input image and three high-frequency coefficients that represent sparse vertical, horizontal, and diagonal details of the input image. Then, we design a wavelet-based Fourier information interaction network (WFI2-net), fully integrating the characteristics of Transformer [22] and Fourier prior information to enhance high- and low-frequency content, respectively. Moreover, to achieve interaction of high- and low-frequency information, we propose a cross-frequency conditioner (CFC) to further improve the generation quality. The target of the second stage is to make adjustment to the initial enhanced coarse results in terms of details and textures via diffusion model. Consequently, we propose a frequency residual diffusion adjustment module (FRDAM). Unlike previous diffusion-based work, FRDAM learns the residual distribution of high- and low-frequency information between the ground-truth and the initial enhanced results using two diffusion models in the wavelet space, which can not only increase the model's focus on fine-grained information but also mitigate the adverse effects of the diversity of the sampling process.

In summary, the main contributions of our method are as follows:

- We explore in depth the properties of the frequency domain for underwater images. Based on the properties and diffusion model, we propose a novel UIE framework, named WF-Diff, with the goal of achieving frequency enhancement and diffusion adjustment.
- We propose a frequency residual diffusion adjustment module (FRDAM) to further refine the high- and low-frequency information of the initial enhanced images. FRDAM can be viewed as a plug-and-play universal module to adjust the detail of the underwater images.
- We propose a cross-frequency conditioner (CFC) to achieve the cross-frequency interaction of high- and low-frequency information.
- Experimental results compared with SOTAs considerably show that our developed WF-Diff performs the superiority against previous UIE approaches, and extensive ablation experiments can demonstrate the effectiveness of our contributions.

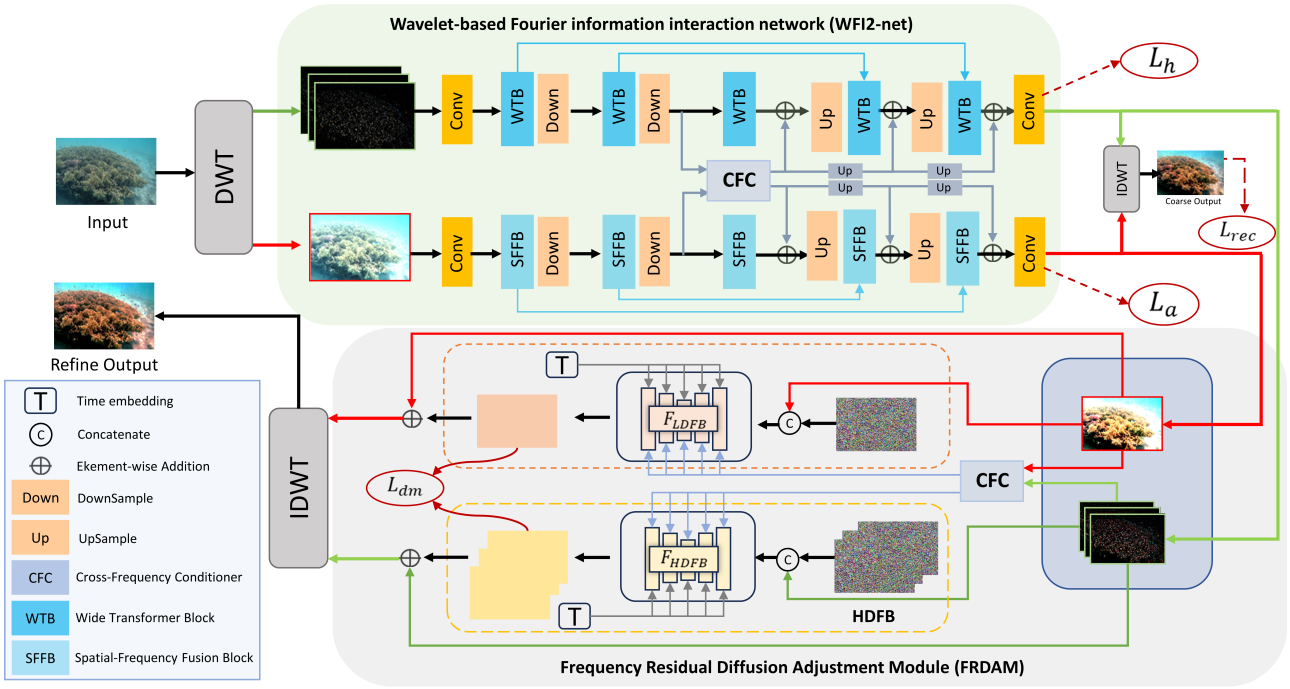

Figure 2. Overall framework of WF-Diff. It contains two detachable networks, Wavelet-based Fourier information interaction network (WFI2-net) and Frequency Residual Diffusion Adjustment Module (FRDAM). FRDAM consists of low-frequency diffusion branch (LDFB) and high-frequency diffusion branch (HDFB), which aims to further adjust the high- and low-frequency information of the initial enhanced images. Furthermore, the proposed cross-frequency conditioner (CFC) aims to achieve the cross-frequency interaction of high- and low-frequency information.

2. Related Works

2.1. Underwater Image Enhancement

Currently, existing UID methods can be briefly categorized into the physical and deep model-based approaches [15, 18, 28–30, 36]. Most UID methods based on the physical model utilize prior knowledge to establish models, such as water dark channel priors [29], attenuation curve priors [38], fuzzy priors [4]. In addition, Akkaynak and Treibitz [1] proposed a method based on the revised physical imaging model. However, the depth map of the underwater scene is difficult to obtain. This leads to unstable performance, which usually suffers from severe color cast and artifacts. Therefore, the manually established priors restrain the model's robustness and scalability under the complicated and varied circumstances. Recently, deep learning-based methods [18, 28, 36] have achieved acceptable performance. To alleviate the need for real-world underwater paired training data, many methods introduced GAN-based framework for UIE [7, 14, 21, 49], such as WaterGAN [21], UGAN [7] and UIE-DAL [37]. Recently, some complex frameworks were proposed and achieve the-state-of-the-art performance [15, 29]. Ucolor [19] combined the underwater physical imaging model in the raw space and designed a medium transmission guided model. Yang et al. [43] proposed a reflected light-aware multi-scale progressive restoration network to obtain images with both color equalization and rich texture in various underwater scenes. Huang et al. [13] proposed a mean teacher based semi-supervised network, which effectively leverages the knowledge from unlabeled data. However, most previous methods are based on the spatial domain, with limited exploration of the frequency space for underwater images, which results in an inability to effectively harness the representation power of the deep models.

2.2. Diffusion Model

Recently, Diffusion Probabilistic Models (DPMs) [10, 35] have been widely adopted for conditional image generation [5, 40, 42, 46, 50]. Saharia et al. [33] proposed Palette, which has demonstrated the excellent performance of diffusion models in the field of conditional image generation, including colorization, in-painting and JPEG restoration. Tang et al. [36] presented an image enhancement approach with diffusion model in underwater scenes. However, the reverse process starts from randomly sampled Gaussian noise to full images [45], which can lead to unexpected artifacts due to the diversity of the sampling process. Furthermore, the diffusion model needs to recover both the high and low-frequency information in images, which lim-

its their ability to focus on fine-grained information. Consequently, how to incorporate diffusion models into a unified underwater image enhancement network is a vital issue.

3. Methodology

3.1. Overall Framework

Given an underwater image as input, our goal is to learn a network to generate an output that eliminates the color cast from input while enhancing the image details. The overall framework of WF-Diff is shown in Figure 2. WF-Diff is designed to fully leverage the characteristics of frequency domain information and powerful ability of diffusion models. Specifically, WF-Diff consists of two detachable networks: Wavelet-based Fourier information interaction network (WFI2-net) and Frequency Residual Diffusion Adjustment Module (FRDAM). We first convert the input into the wavelet space using discrete wavelet transformations (DWT), obtaining a low-frequency coefficient and three high-frequency coefficients. WFI2-net is dedicated to achieving preliminary enhancement of frequency information. We fully integrate the characteristics of Transformer and Fourier prior Information, and design wide transformer block (WTB) and spatial-frequency fusion block (SFFB) to enhance high- and low-frequency content respectively. FRDAM consists of low-frequency diffusion branch (LDFB) and high-frequency diffusion branch (HDFB), which aims to further adjust the high- and low-frequency information of the initial enhanced images. Note that, our proposed FRDAM learns the residual distribution of high- and low-frequency information between the ground-truth and the initial enhanced results using two diffusion models, respectively. Additionally, the proposed cross-frequency conditioner (CFC) strives to achieve cross-frequency interaction between high- and low-frequency information.

3.2. Discrete Wavelet and Fourier Transform

Discrete Wavelet transform (DWT) has been widely applied to low-level vision tasks [11, 16]. We firstly use DWT to decompose an input into multiple frequency sub-bands so that we can achieve the color correction of low-frequency information and detail enhancement of high-frequency information, respectively. Given an underwater image as input $I \in \mathbb{R}^{H\times W\times c}$, we use DWT with Haar wavelets to decompose the input. Haar wavelets consist of the low-pass filter L, and the high-pass filter H, as follows:

$$L = \frac{1}{\sqrt{2}}[1,1]^T, H = \frac{1}{\sqrt{2}}[1,-1]^T. \tag{1}$$

We can obtain four sub-bands, which can be expressed as:

$$I_{LL}, \{I_{LH}, I_{HL}, I_{HH}\} = \text{DWT}(I), \tag{2}$$

where $I_{LL}, \{I_{LH}, I_{HL}, I_{HH}\} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times c}$ represent the low-frequency component of the input and high-frequency components in the vertical, horizontal, and diagonal directions, respectively. More specifically, the low-frequency component contains the content and color information of the input image, and the other three high-frequency coefficients contain details information of global structures and textures [31]. The sub-bands are downsampled to half-resolution of the input but do not result in information loss due to the biorthogonal property of DWT. For low-frequency component I_{LL}, we will explore its properties in Fourier space.

Then, we introduce the operation of the Fourier transform [48]. Given a image $x \in \mathbb{R}^{H\times W\times 1}$, whose shape is $H \times W$, the Fourier transform $\mathcal{F}$ which converts x to the Fourier space X can be expressed as:

$$\mathcal{F}(x)(u,v) = X(u,v) = \frac{1}{\sqrt{HW}}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} x(h,w)e^{-j2\pi(\frac{h}{H}u+\frac{w}{W}v)}, \tag{3}$$

where h, w are the coordinates in the spatial space and u, v are the coordinates in the Fourier space. $\mathcal{F}^{-1}$ denotes the inverse transform of $\mathcal{F}$. Complex component $X(u,v)$ in the Fourier space can be represented by a amplitude component $\mathcal{A}(X(u,v))$ and a phase component $\mathcal{P}(X(u,v))$ as follows:

$$\begin{aligned}\mathcal{A}(X(u,v)) &= \sqrt{R^2(X(u,v)) + I^2(X(u,v))},\\ \mathcal{P}(X(u,v)) &= arctan[\tfrac{I(X(u,v))}{R(X(u,v))}],\end{aligned} \tag{4}$$

where $R(x)$ and $I(x)$ represent the real and imaginary parts of $X(u,v)$, respectively. Note that, the Fourier operation can be computed alone in each channel for feature maps.

According to Figure 1 and Table 1 (our motivation), we conclude that the color degradation information of underwater images is mainly contained in the amplitude component of low-frequency sub-band, and the texture and detail degradation information is mainly contained in high-frequency sub-bands.

3.3. Frequency Preliminary Enhancement

Based on the above analysis, in frequency preliminary enhancement stage, we design a simple but effective WFI2-net with a parallel encoder-decoder (U-Net-like) format to restore the amplitude component of low-frequency information and high-frequency components, respectively. We also utilize skip connections to connect the features at the same level in the encoder and decoder. For high-frequency branch, we utilize advantage of transformer modeling global information to enhance high-frequency coefficients. We design wide transformer block (WTB) using multi-scale information, aiming to model long range dependencies. Our low-frequency branch aims to restore the amplitude component in Fourier space. In order to obtain rich frequency and spatial information, we design spatial-frequency fusion block (SFFB).

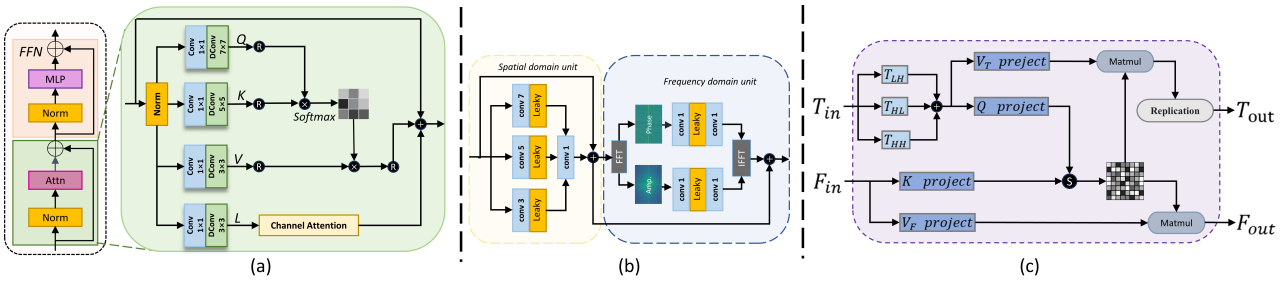

Figure 3. The architecture of (a) Wide Transformer Block, (b) Spatial-Frequency Fusion Block and (c) Cross-Frequency Conditioner.

Wide Transformer Block. Unlike low-frequency coefficients, high-frequency components contain global structure and texture details. Consequently, the high-frequency branch focuses on modeling global and local features. WTB is shown in Figure 3 (a). Given $I_{LH}, I_{HL}, I_{HH} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times c}$, WTB firstly obtains their embedding features $T_{in} \in \mathbb{R}^{3\times\frac{H}{2}\times\frac{W}{2}\times C}$ through convolution projection. To be specific, WTB are composed of an attention (Atten) module and a feed-forward network (FFN) module, and the computation can be denoted in the WTB as:

$$Q, K, V, L = \text{Split}(W_d W_p(Norm(T_{i-1}))), \tag{5}$$

$$\hat{T}_i = SA(Q, K, V) + CA(L) + T_{i-1}, \tag{6}$$

$$T_i = FFN(Norm(\hat{T}_i)) + \hat{T}_i, \tag{7}$$

where SA and CA refer to self-attention and channel attention, respectively. $Norm$ refers to normalization. T_{i-1} represents the input embeddings of the current WTB. W_d and W_p denote 1×1 point-wise convolution and multi-scale kernel depth-wise convolution, respectively; $Split$ refers to the split operation. L aims to focus on local information.

Spatial-Frequency Fusion Block. We show the structure of SFFB in Figure 3 (b), which has a spatial domain unit (SDU) and a frequency domain unit (FDU) for interaction of dual domain representations. In spatial domain unit, we employ multi-scale convolution kernels in order to enlarge the limited spatial receptive field. After obtaining the spatial embeddings F_s, we firstly utilize the FFT to obtain the amplitude $\mathcal{A}(F_s)$ and phase $\mathcal{P}(F_s)$ components. Then, the $\mathcal{A}(F_s)$ and $\mathcal{P}(F_s)$ are fed into two layers 1*1 conv to obtain $\mathcal{A}^{'}(F_s)$ and $\mathcal{P}^{'}(F_s)$. Finally, we use the IFFT algorithm to map $\mathcal{A}^{'}(F_s)$ and $\mathcal{P}^{'}(F_s)$ to image space and obtain the frequency embeddings F_f. The fusion embeddings of spatial domain and frequency domain can be expressed as:

$$F_{sf} = F_s + F_f. \tag{8}$$

Loss Function. We denote $I^{'}_{LL}$ as output of low-frequency branch, and $I^{'}_{LH}, I^{'}_{HL}$ and $I^{'}_{HH}$ as output of high-frequency branch. Ground-truth (G) can be decomposed $G_{LL}, G_{LH}, G_{HL}, G_{HH}$ by DWT. The high-frequency loss can be expressd as:

$$\mathcal{L}_h = \left\| I^{'}_{(i)} - G_{(i)} \right\|_2, \tag{9}$$

where $i \in \{LH, HL, HH\}$. For low-frequency information, we only constrain the amplitude components. Consequently, the low-frequency loss can be expressd as:

$$\mathcal{L}_a = ||\mathcal{A}(I^{'}_{LL})) - \mathcal{A}(G_{LL})||_1, \tag{10}$$

where $\mathcal{A}()$ refers to the amplitude component in Fourier transform. Finally, We further use an adversarial loss in Wasserstein GAN as reconstruction Loss $\mathcal{L}_{rec}$.

3.4. Cross-Frequency Conditioner

The detailed structure of CFC is shown in Figure 3 (c). The purpose of CFC is to facilitate information interaction between high and low frequency features, achieving cross-frequency mutual reinforcement between high and low frequency branches, thus enhancing overall representation power. We denote T_{in} and F_{in} as the input features of CFC, which represent high- and low-frequency embeddings. For the high-frequency embedding features $T_{in} \in \mathbb{R}^{3\times\frac{H}{2}\times\frac{W}{2}\times C}$, we can obtain $T_{LH}, T_{HL}, T_{HH} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times c}$ via split operation. By adding these extracted coefficients, we obtain the aggregated high-frequency embeddings. We use different linear projections to construct Q and K in CFC:

$$Q = Conv_{1\times1}(T_{LH} + T_{HL} + T_{HH}), \tag{11}$$

$$K = Conv_{1\times1}(F_{in}). \tag{12}$$

Similarly, V_T of high-frequency embeddings and V_F of low-frequency embeddings can be obtained:

$$V_T = Conv_{1\times1}(T_{LH} + T_{HL} + T_{HH}), \tag{13}$$

$$V_F = Conv_{1\times1}(F_{in}). \tag{14}$$

The output feature map T_{out} and F_{out} can then be obtained from the formula:

$$T_{out} = R(Softmax(\frac{QK^T}{\sqrt{d_k}})V_T), \tag{15}$$

$$F_{out} = Softmax(\frac{QK^T}{\sqrt{d_k}})V_F, \tag{16}$$

where R denotes a replication operation, and $\sqrt{d_k}$ is the number of columns of matrix Q.

3.5. Frequency Diffusion Adjustment

FRDAM aims to further adjust the high- and low-frequency information using the powerful representation of diffusion model. Generally, FRDAM can be divided into two branch, namely low-frequency diffusion branch (LDFB) and high-frequency diffusion branch (HDFB). We adopt the diffusion process proposed in DDPM [10] to construct the residual distribution of high- and low-frequency information for each branch, which can be described as a forward diffusion process and a reverse diffusion process.

Forward Diffusion Process. The forward diffusion process can be viewed as a Markov chain progressively adding Gaussian noise to the data. Given the initial enhanced frequency components $I_i^{'}$ and its ground truth G_i, $i \in \{LL, LH, HL, HH\}$, we calculate their residual distribution $x_0 = G_i - I_i^{'}$, then introduce Gaussian noise based on the time step, as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \quad (17)$$

where β_t is a variable controlling the variance of the noise. Introducing $\alpha_t = 1 - \beta_t$, this process can be described as:

$$x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1}, \quad \epsilon_{t-1} \sim \mathcal{N}(0, \mathcal{Z}). \quad (18)$$

With Gaussian distributions are merged, We can obtain :

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I). \quad (19)$$

Reverse Diffusion Process. The reverse diffusion process aims to restore the residual distribution from the Gaussian noise. The reverse diffusion can be expressed as:

$$p_\theta(x_{t-1}|x_t, x_c^{(l)}) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, x_c^{(l)}, t), \sigma_t^2 \mathcal{Z}), \quad (20)$$

where we take the LDFB as an example, and $x_c^{(l)}$ refers to the conditional image $I_{LL}^{'}$. $\boldsymbol{\mu}_\theta(x_t, x_c^{(l)}, t)$ and σ_t^2 are the mean and variance from the estimate of step t, respectively. In LDFB and HDFB, we follow the setup of [35], they can be expressed as:

$$\mu_\theta(x_t, x_c^{(l)}, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{(1-\bar{\alpha}_t)}\epsilon_\theta(x_t, x_c^{(l)}, t)), \quad (21)$$

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \quad (22)$$

where $\epsilon_\theta(x_t, x_c^{(l)}, t)$ is the estimated value with a Unet.

We optimize an objective function for the noise estimated by the network and the noise $\epsilon^{(l)}$ actually added in LDFB. Therefore, the diffusion loss process is:

$$L_{dm}(\theta) = \|\epsilon^{(l)} - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon^{(l)}, x_c^{(l)}, t)\|. \quad (23)$$

Generally, the frequency diffusion adjustment process is to refine the high- and low-frequency component of the initial enhancement. The whole diffusion process can be formulated:

$$\hat{I}_{(i)} = \mathcal{F}_{HDFB}(\epsilon_s^{(h)}, I_{(i)}^{'}), i \in \{LH, HL, HH\}, \quad (24)$$

$$\hat{I}_{LL} = \mathcal{F}_{LDFB}(\epsilon_s^{(l)}, I_{LL}^{'}), \quad (25)$$

where $\epsilon_s^{(h)} \in \mathbb{R}^{3\times\frac{H}{2}\times\frac{W}{2}\times 3}$ and $\epsilon_s^{(l)} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 3}$ are Gaussian noise.

Ultimately, the refined frequency components are obtained as the addition of the diffusion generative residual distribution and the initial enhanced frequency components. Then, we employ IDWT to obtain the final generated image:

$$I_{final} = IDWT(I_{(i)}^{'} + \hat{I}_{(i)}, I_{LL}^{'} + \hat{I}_{LL}), i \in \{LH, HL, HH\} \quad (26)$$

4. Experiments

4.1. Setup

Implementation details. Our network, implemented using PyTorch 1.7, underwent training and testing on an NVIDIA GeForce RTX 3090 GPU. We employed the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The patch size was configured as 256 × 256, and the batch size was set to 2. The diffusion model's total time steps, denoted as T, were set to 1000, and the number of training iterations reached one million. The initial learning rate was established at 0.0001.

Datasets. We utilize the real-world UIEBD dataset [28] and the LSUI dataset [18] for training and evaluating our model. The UIEBD dataset comprises 890 underwater images with corresponding labels. Out of these, 700 images are allocated for training, and the remaining 190 are designated for testing. The LSUI dataset is randomly partitioned into 4500 images for training and 504 images for testing. In addition, to verify the generalization of WF-Diff, we use non-reference benchmarks U45 [20], which contains 45 underwater images for testing.

Comparison methods. We conduct a comparative analysis between WF-Diff and eight state-of-the-art (SOTA) UIE methods, namely UIEC^2-Net [39], Water-Net [28], UWCNN [3], SCNet [8],UIEWD [25], U-color [19], U-shape [18], and DM-water [36]. To ensure a fair and rigorous comparison, we utilize the provided source codes from the respective authors and adhere strictly to the identical experimental settings across all evaluations.

Evaluation Metrics. We primarily utilize well-established full-reference image quality assessment metrics: PSNR and SSIM [41]. PSNR and SSIM offer quantitative comparisons of our method with other approaches at both pixel and structural levels. Higher PSNR and SSIM values signify superior quality of the generated images. Additionally, we incorporate the LPIPS and FID metrics for full-reference image evaluation. LPIPS [47] is a deep neural network-based image quality metric that assesses the perceptual similarity between an image and a reference image. FID [9] measures the distance between the distributions of real and generated images. A lower LPIPS and FID score indicates a more effective UIE approach. For non-reference benchmarks U45,

Table 2. Quantitative comparison of different UIE methods on the UIEBD, LSUI and U45 datasets. The best results are highlighted in bold and the second best results are underlined.

	Methods	UIEWD	UWCNN	UIEC^2-Net	Water-Net	SCNet	U-color	U-shape	DM-water	Ours
UIEBD	FID↓	85.12	94.44	35.06	37.48	33.66	38.25	46.11	<u>31.07</u>	**27.85**
	LPIPS↓	0.3956	0.3525	0.2033	0.2116	0.2497	0.2337	0.2264	<u>0.1436</u>	**0.1248**
	PSNR↑	14.65	15.40	20.14	19.35	20.41	20.71	21.25	<u>21.88</u>	**23.86**
	SSIM↑	0.7265	0.7749	0.8215	0.8321	0.8235	0.8411	<u>0.8453</u>	0.8194	**0.8730**
LSUI	FID↓	98.49	100.5	34.51	38.90	158.99	45.06	28.56	<u>27.91</u>	**26.75**
	LPIPS↓	0.3962	0.3450	0.1432	0.1678	0.283	0.123	**0.1028**	0.1138	<u>0.1096</u>
	PSNR↑	15.43	18.24	20.86	19.73	22.63	22.91	24.16	**27.65**	<u>27.26</u>
	SSIM↑	0.7802	0.8465	0.8867	0.8226	0.9176	0.8902	<u>0.9322</u>	0.8867	**0.9437**
U45	UIQM↑	2.458	2.379	2.780	2.957	2.856	3.104	<u>3.151</u>	3.086	**3.181**
	UCIQE↑	0.583	0.567	0.591	0.601	0.594	0.586	0.592	**0.634**	<u>0.619</u>

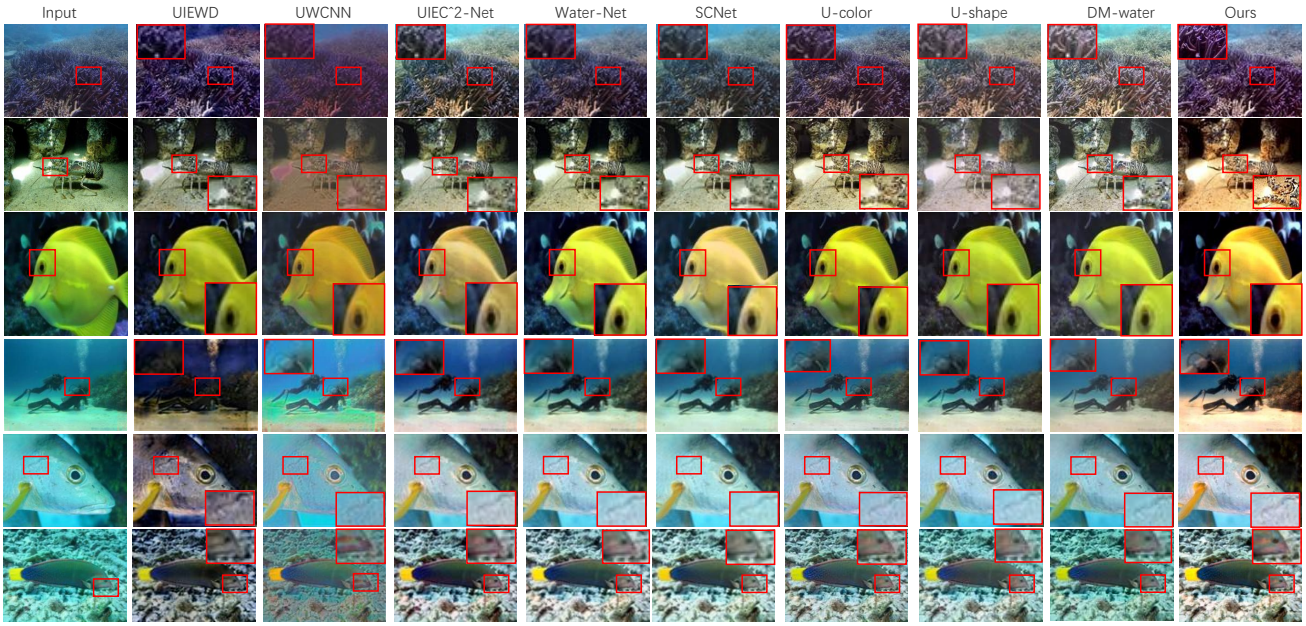

Figure 4. Qualitative comparison with other SOTA methods on real underwater images.

we introduce UIQM [27] and UCIQE [44] to evaluate our method.

4.2. Results and Comparisons

Table 2 shows the quantitative results compared with different baselines on UIEBD, LSUI and U45 datasets, including with UIEC^2-Net, Water-Net, UIEWD, UWCNN, SCNet, U-color, U-shape, and DM-water. We mainly use PSNR, SSIM, LPIPS and FID as our quantitative indices for UIEBD and LSUI datasets, and UIQM and UCIQE for non-reference dataset U45. The results in Table 2 show that our algorithm outperforms state-of-the art methods obviously, and achieve state-of-the-art performance in terms of image quality evaluation metrics on UIE task, which verifies the robustness of the proposed WF-Diff. To better validate the superiority of our methods, in Figure 4, we show the visual results comparison with state-of-the-art methods on real underwater images. The six examples are randomly selected on the UIEBD and LSUI datasets. Our methods consistently generate natural and better visual results on testing images, strongly proving that WF-Diff has good generalization performance for real-world applications.

4.3. Ablation Study

Ablation study with WFI2-net. In order to evaluate the effectiveness of each part in WFI2-net, we conduct two ab-

Table 3. Ablation study with nerwork structure of WFI2-net on UIEBD dataset.

WTB		SFFB		UIEBD	
SA	CA	SDU	FDU	PSNR ↑	SSIM↑
×	✓	✓	✓	20.94	0.8541
✓	×	✓	✓	20.82	0.8473
✓	✓	×	✓	21.11	0.8586
✓	✓	✓	×	20.23	0.8346
✓	✓	✓	✓	**21.87**	**0.8622**

Table 4. Ablation study with loss of WFI2-net on UIEBD dataset.

Methods				UIEBD	
$\mathcal{L}_h$	$\mathcal{L}_{rec}$	$\mathcal{L}_a$	CFC	PSNR ↑	SSIM↑
×	✓	✓	✓	20.46	0.8274
✓	×	✓	✓	20.65	0.8399
✓	✓	×	✓	19.81	0.8213
✓	✓	✓	×	20.97	0.8425
✓	✓	✓	✓	**21.87**	**0.8622**

Table 5. Ablation study with FRDAM on UIEBD dataset. DM refers to diffusion model (DM), RDM refers to residual DM. D-L refers to refining low-frequency component in wavelet space, and D-H refers to refining high-frequency components in wavelet space.

Method	DM	RDM	D-L	D-H	CFC	PSNR↑
A	✓	×	×	×	×	20.86
B	×	✓	×	×	×	22.37
C	×	✓	✓	×	×	22.32
D	×	✓	×	✓	×	22.58
E	×	✓	✓	✓	×	23.44
F	×	✓	✓	✓	✓	**23.86**

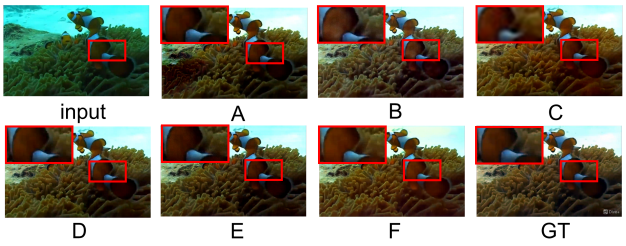

Figure 5. Visual results of ablation study with FRDAM.

lation experiments with WFI2-net on the UIEBD dataset in Table 3 and 4. CFC is cross-frequency conditioner. Note that, we do not discuss the effect of our proposed FRDAM here. We regularly remove one component to each configuration at one time, and our strategy achieves the best performance by using all loss functions and blocks, proving that each part of WFI2-net is useful for UIE task.

Table 6. Comparison results of inference speed.

Method	SCNet	Ushape	DM-water	WFI2-net	WF-Diff
Inference time (s)	0.0149	0.0353	0.1441	0.0739	0.2816

Table 7. Comparison results of FLOPs.

Method	SCNet	WaterNet	Ucolor	Ushape	WFI2-net
FLOPs (G)	5.88	193.7	443.8	66.2	94.1

Ablation study with FRDAM. In this section, we will discuss the effectiveness of FRDAM. Table 5 shows the quantitative results on the UIEBD dataset, and Figure 5 shows the visual results. Note that, Model A and B achieve diffusion process in the pixel level, and model C, D, E and E achieve diffusion process in wavelet space. Model A obtains relatively poor results in PSNR and generates images with color distortion or artifacts in Figure 5 due to the diversity of the sampling process. Model B does not yield fully satisfactory results, because it needs to adjust both the high and low-frequency information in images, which limits their ability to focus on fine-grained information. Compared to model C, model D achieves better results, suggesting that the degradation information is mainly in the high-frequency information in frequency diffusion adjustment stage. Model F achieves the best performance, proving that our designed FRDAM is best for UIE task.

5. Conclusion

In this paper, we develop a novel UIE framework, namely WF-Diff. With fully utilizing the frequency domain characteristics and diffusion model, WFI2-net can achieve enhancement and adjustment of frequency information. Our proposed FRDAM is a plug-and-play universal module to adjust the details of the underwater images. WF-Diff shows SOTA performance on UIE task, and extensive ablation experiments prove that each of our contributions is effective.

Limitations. As a result of employing two diffusion models, our approach doesn't confer an advantage in terms of inference speed. Table 6 presents the actual values of inference speed for WF-Diff, indicating that our model does not outperform recent approaches in terms of inference time. Furthermore, Table 7 provides a comparison of FLOPs for WFI2-net. It should be mentioned that the implicit sampling step is set to 10 for WF-Diff. Hence, in the future, we will delve into methods to expedite the sampling process.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Grant No. 62276138 and 62371232).

References

[1] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1682–1691. Computer Vision Foundation / IEEE, 2019. 3

[2] Derya Akkaynak, Tali Treibitz, Tom Shlesinger, Yossi Loya, Raz Tamir, and David Iluz. What is the space of attenuation coefficients in underwater computer vision? In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 568–577, 2017. 1

[3] Saeed Anwar, Chongyi Li, and Fatih Porikli. Deep underwater image enhancement. *CoRR*, abs/1807.03528, 2018. 6

[4] John Yi-Wu Chiang and Ying-Ching Chen. Underwater image enhancement by wavelength compensation and dehazing. *IEEE Trans. Image Process.*, 21(4):1756–1769, 2012. 3

[5] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: conditioning method for denoising diffusion probabilistic models. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14347–14356. IEEE, 2021. 2, 3

[6] Karin de Langis and Junaed Sattar. Realtime multi-diver tracking and re-identification for underwater human-robot collaboration. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pages 11140–11146, 2020. 1

[7] Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar. Enhancing underwater imagery using generative adversarial networks. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 7159–7165. IEEE, 2018. 1, 3

[8] Zhenqi Fu, Xiaopeng Lin, Wu Wang, Yue Huang, and Xinghao Ding. Underwater image enhancement via learning water type desensitized representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 2764–2768. IEEE, 2022. 6

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. 6

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 3, 6

[11] Huaibo Huang, Ran He, Zhenan Sun, and Tieniu Tan. Wavelet-srnet: A wavelet-based CNN for multi-scale face super resolution. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1698–1706, 2017. 4

[12] Jie Huang, Yajing Liu, Feng Zhao, Keyu Yan, Jinghao Zhang, Yukun Huang, Man Zhou, and Zhiwei Xiong. Deep fourier-based exposure correction network with spatial-frequency interaction. pages 163–180, 2022. 2

[13] Shirui Huang, Keyan Wang, Huan Liu, Jun Chen, and Yunsong Li. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18145–18155. IEEE, 2023. 3

[14] Md Jahidul Islam, Youya Xia, and Junaed Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robotics Autom. Lett.*, 5(2):3227–3234, 2020. 3

[15] Paulo Drews Jr., Erickson Rangel do Nascimento, F. Moraes, Silvia S. C. Botelho, and Mario F. M. Campos. Transmission estimation in underwater single images. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 825–830, 2013. 1, 3

[16] Eunhee Kang, Won Chang, Jae Jun Yoo, and Jong Chul Ye. Deep convolutional framelet denosing for low-dose CT via wavelet residual network. *IEEE Trans. Medical Imaging*, 37 (6):1358–1369, 2018. 4

[17] Chongyi Li, Jichang Guo, Runmin Cong, Yanwei Pang, and Bo Wang. Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior. *IEEE Trans. Image Process.*, 25(12):5664–5677, 2016. 1

[18] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Trans. Image Process.*, 29:4376–4389, 2020. 1, 3, 6

[19] Chongyi Li, Saeed Anwar, Junhui Hou, Runmin Cong, Chunle Guo, and Wenqi Ren. Underwater image enhancement via medium transmission-guided multi-color space embedding. *IEEE Trans. Image Process.*, 30:4985–5000, 2021. 3, 6

[20] Hanyu Li, Jingjing Li, and Wei Wang. A fusion adversarial underwater image enhancement network with a public test dataset. *arXiv preprint arXiv:1906.06819*, 2019. 6

[21] Jie Li, Katherine A. Skinner, Ryan M. Eustice, and Matthew Johnson-Roberson. Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robotics Autom. Lett.*, 3(1):387–394, 2018. 3

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 2

[23] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 2

[24] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. *arXiv preprint arXiv:2403.06135*, 2024. 2

[25] Ziyin Ma and Changjae Oh. A wavelet-based dual-stream network for underwater image enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 2769–2773. IEEE, 2022. 6

[26] James McMahon and Erion Plaku. Autonomous data collection with timed communication constraints for unmanned underwater vehicles. *IEEE Robotics Autom. Lett.*, 6(2): 1832–1839, 2021. 1

[27] Karen Panetta, Chen Gao, and Sos Agaian. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, 41(3):541–551, 2015. 7

[28] Lintao Peng, Chunli Zhu, and Liheng Bian. U-shape transformer for underwater image enhancement. *IEEE Trans. Image Process.*, 32:3066–3079, 2023. 1, 3, 6

[29] Yan-Tsung Peng and Pamela C. Cosman. Underwater image restoration based on image blurriness and light absorption. *IEEE Trans. Image Process.*, 26(4):1579–1594, 2017. 1, 3

[30] Yan-Tsung Peng, Keming Cao, and Pamela C. Cosman. Generalization of the dark channel prior for single image restoration. *IEEE Trans. Image Process.*, 27(6):2856–2868, 2018. 3

[31] Priyadharsini Ravisankar, T. Sree Sharmila, and V. Rajendran. A wavelet transform based contrast enhancement method for underwater acoustic images. *Multidimens. Syst. Signal Process.*, 29(4):1845–1859, 2018. 1, 4

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 2

[33] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH '22: Special Interest Group on Computer Graphics and Interactive Techniques Conference, Vancouver, BC, Canada, August 7 - 11, 2022*, pages 15:1–15:10. ACM, 2022. 3

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 2, 3, 6

[36] Yi Tang, Hiroshi Kawasaki, and Takafumi 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, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 5419–5427. ACM, 2023. 1, 2, 3, 6

[37] Pritish M. Uplavikar, Zhenyu Wu, and Zhangyang Wang. All-in-one underwater image enhancement using domain-adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1–8, 2019. 3

[38] Yi Wang, Hui Liu, and Lap-Pui Chau. Single underwater image restoration using adaptive attenuation-curve prior. *IEEE Trans. Circuits Syst. I Regul. Pap.*, 65-I(3):992–1002, 2018. 3

[39] Yudong Wang, Jichang Guo, Huan Gao, and HuiHui Yue. Uiecˆ2-net: Cnn-based underwater image enhancement using two color space. *Signal Process. Image Commun.*, 96: 116250, 2021. 6

[40] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023. 2, 3

[41] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612, 2004. 6

[42] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G. Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16272–16282. IEEE, 2022. 3

[43] Jian Yang, Chen Li, and Xuelong Li. Underwater image restoration with light-aware progressive network. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE, 2023. 3

[44] Miao Yang and Arcot Sowmya. An underwater color image quality evaluation metric. *IEEE Transactions on Image Processing*, 24(12):6062–6071, 2015. 7

[45] Zongyuan Yang, Baolin Liu, Yongping Xiong, Lan Yi, Guibin Wu, Xiaojun Tang, Ziqi Liu, Junjie Zhou, and Xing Zhang. Docdiff: Document enhancement via residual diffusion models. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 2795–2806. ACM, 2023. 2, 3

[46] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with A generative diffusion model. *CoRR*, abs/2308.13164, 2023. 2, 3

[47] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595, 2018. 6

[48] Chen Zhao, Wei-Ling Cai, Chenyu Dong, and Ziqi Zeng. Toward sufficient spatial-frequency interaction for gradient-aware underwater image enhancement. *arXiv preprint arXiv:2202.08537*, 2023. 2, 4

[49] Chen Zhao, Wei-Ling Cai, and Zheng Yuan. Spectral normalization and dual contrastive regularization for image-to-

image translation. *The Visual Computer*, pages 1–12, 2024. 3

[50] Chen Zhao, Chenyu Dong, and Weiling Cai. Learning a physical-aware diffusion model based on transformer for underwater image enhancement. *arXiv preprint arXiv:2403.01497*, 2024. 3

[51] Dewei Zhou, Zongxin Yang, and Yi Yang. Pyramid diffusion models for low-light image enhancement. *arXiv preprint arXiv:2305.10028*, 2023. 2

[52] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. *arXiv preprint arXiv:2402.05408*, 2024. 2