Adaptive Dual-domain Learning for Underwater Image Enhancement

Lintao Peng, Liheng Bian[*]

Beijing Institute of Technology, Beijing & Zhuhai, China
{penglintao, bian}@bit.edu.cn

Abstract

Recently, learning-based Underwater Image Enhancement (UIE) methods have demonstrated promising performance. However, existing learning-based methods still face two challenges. 1) They rarely consider the inconsistent degradation levels in different spatial regions and spectral bands simultaneously. 2) They treat all regions equally, ignoring that the regions with high-frequency details are more difficult to reconstruct. To address these challenges, we propose a novel UIE method based on spatial-spectral dual-domain adaptive learning, termed SS-UIE. Specifically, we first introduce a spatial-wise Multi-scale Cycle Selective Scan (MCSS) module and a Spectral-Wise Self-Attention (SWSA) module, both with linear complexity, and combine them in parallel to form a basic Spatial-Spectral block (SS-block). Benefiting from the global receptive field of MCSS and SWSA, SS-block can effectively model the degradation levels of different spatial regions and spectral bands, thereby enabling degradation level-based dual-domain adaptive UIE. By stacking multiple SS-blocks, we build our SS-UIE network. Additionally, a Frequency-Wise Loss (FWL) is introduced to narrow the frequency-wise discrepancy and reinforce the model's attention on the regions with high-frequency details. Extensive experiments validate that the SS-UIE technique outperforms state-of-the-art UIE methods while requiring cheaper computational and memory costs.

Code — https://github.com/LintaoPeng/SS-UIE

Introduction

Underwater imaging technology (Sahu, Gupta, and Sharma 2014; Yang et al. 2019) is essential for obtaining underwater images and investigating the underwater environment. However, due to the strong absorption and scattering effects on light caused by dissolved impurities and suspended matter in the medium (water), underwater images often suffer from inconsistent levels of degradation in different spatial regions and spectral bands, which ultimately leads to problematic issues, such as color casts, artifacts and blurred details (Schettini and Corchs 2010). Therefore, Underwater Image Enhancement (UIE) related innovations are of great significance for improving the visual quality and merit of images in accurately understanding the underwater world.
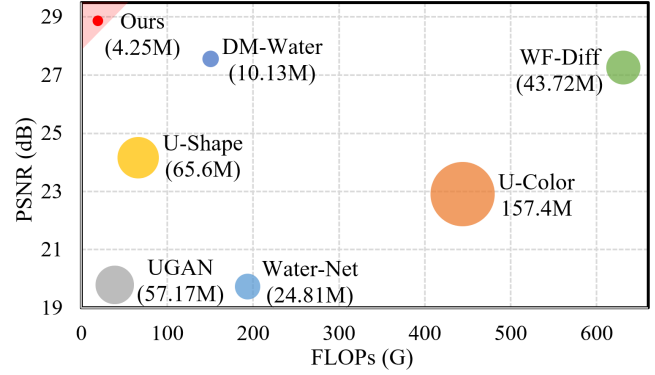
Figure 1: PSNR-Parameters-FLOPs comparisons with existing UIE methods. The vert axis is PSNR (dB), the horizontal axis is FLOPs (computational cost), and the circle radius is the parameter number (memory cost) of each UIE model. Our SS-UIE outperforms state-of-the-art (SOTA) methods while requiring fewer FLOPs and Params.

The purpose of UIE is to enhance image quality by removing scattering and correcting color distortions in degraded images. To achieve this goal, traditional physical model-based methods (Ancuti et al. 2012; Drews Jr et al. 2013; Li et al. 2016b) first investigate the physical mechanisms of underwater image degradation and then reverse the degradation process using various hand-crafted physical priors. However, these methods rely on manually tuned physical parameters and cannot account for all the complex physical and optical factors in underwater scenes, leading to poor generalizability. In contrast, learning-based UIE methods eliminate the need for manual parameter tuning. Researchers initially applied CNNs (Zeiler and Fergus 2014) to learning-based UIE methods to improve generalization capability. However, CNN-based methods (Li et al. 2017, 2020, 2021; Islam, Xia, and Sattar 2020) show limitations in capturing non-local self-similarity and long-range dependencies in images, resulting in suboptimal UIE performance.

Recently, Transformer-based UIE methods (Peng, Zhu, and Bian 2023; Tang, Kawasaki, and Iwaguchi 2023; Zhao et al. 2024) have demonstrated promising performance. The self-attention mechanism in Transformer (Vaswani et al. 2017) can model long-range dependencies and non-local

similarities, which offers the possibility to address the short-comings of CNN-based methods. However, the existing Transformer-based methods still have the following issues. **First**, in the global attention Transformer (Dosovitskiy et al. 2020), the computational complexity is quadratic to the spatial size, this burden is non-trivial and sometimes unaffordable. **Second**, existing Transformer-based UIE methods ignore the physical prior that underwater images have inconsistent degradation levels in different spatial regions and spectral bands (Sahu, Gupta, and Sharma 2014; Yang et al. 2019), and still treat all spatial regions and spectral bands equally, resulting in unsatisfactory enhancement results.

Naturally, a question arises: *How to achieve spatial-spectral dual domain adaptive UIE with linear complexity?* The recently proposed state space model, i.e., Mamba (Gu and Dao 2023) inspired us, as it can obtain global receptive field and content-based reasoning ability with linear complexity (Han et al. 2024). Based on this, we propose a spatial-wise Multi-scale Cycle Selective Scan (MCSS) module to obtain the global receptive field with linear complexity, thereby modeling the degradation levels in different spatial regions. We then combine it with the proposed Spectral-Wise Self-Attention (SWSA) module in parallel to create a basic Spatial-Spectral block (SS-block). SWSA uses a fast FFT (Nussbaumer and Nussbaumer 1982) to transform the image into the frequency domain and applies a learnable global filter to mix frequency-wise tokens. This frequency-domain processing method can create global correlations for each spectral band and capture the inter-spectral long-distance dependencies, thereby modeling the degradation levels in different spectral bands. Thanks to the parallel design, SS-block can model the degradation levels in different spatial regions and spectral bands simultaneously with linear complexity, thereby achieving degradation level-based dual-domain adaptive UIE. By stacking multiple SS-blocks, we build our spatial-spectral dual-domain adaptive learning UIE network, termed SS-UIE. In addition, considering that the spatial regions with high-frequency details are more difficult to reconstruct, we introduce a Frequency-Wise Loss (FWL) function to reinforce the network's attention in the regions with rich detail. Equipped with the proposed techniques, our SS-UIE achieves SOTA UIE performance, as shown in Fig. 1. Our contributions can be summarized as follows,

- Our proposed MCSS and SWSA module can obtain the spatial-wise and spectral-wise global receptive fields with linear complexity, respectively, thereby modeling the degradation levels in different spatial regions and spectral bands.

- We combined MCSS and SWSA in parallel to form an SS-block, which can reinforce the network's attention to the spatial regions and spectral bands with serious attenuation, and achieve degradation level-based adaptive UIE.

- The proposed FWL function can narrow the frequency-wise discrepancy, and force the model to restore high-frequency details adaptively without additional memory and computational costs.

Related Work

Underwater Image Enhancemnet

Physical model-based methods. Traditional physical model-based UIE methods address underwater image degradation by employing hand-crafted priors and estimated parameters. Prior is the basis of the physical model-based UIE, in which existing work includes underwater dark channel priors (Drews Jr et al. 2013), attenuation curve priors (Wang, Liu, and Chau 2018), fuzzy priors (Chiang and Chen 2012) and minimum information priors (Li et al. 2016a), etc. For example, Akkaynak et. al (2019) proposed a method based on the revised physical imaging model. However, the depth map of the underwater image is difficult to obtain, which leads to unstable performance. Consequently, hand-crafted priors limit the model's robustness and scalability in complex scenarios.

CNN-based methods. Existing CNN-based UIE methods can be divided into two main technical routes, (1) *designing an end-to-end UIE model;* (2) *utilizing deep-learning models to estimate physical parameters, and then restoring the clean image based on the degradation model.* To alleviate the need for real-world underwater paired training data, many end-to-end methods introduced GAN-based (Goodfellow et al. 2020) framework for UIE, such as Water GAN (Li et al. 2017), UGAN (Fabbri, Islam, and Sattar 2018) and UIE-DAL (Uplavikar, Wu, and Wang 2019). However, the training datasets used by the above methods are not matched real underwater images, which leads to limited enhancement effects in diverse real-world underwater scenes. Recently, Li et al. (2021) combined the underwater physical imaging model and designed a medium transmission guided model to reinforce the network's response to areas with more severe quality degradation. Nevertheless, physical models sometimes fail with varied underwater environments. Compared with physical model-based methods, CNN-based methods have improved generality and speed. However, CNN-based methods still have limitations in capturing long-distance dependencies and non-local similarities.

Transformer-based methods. Recently, the Global Vision Transformer has achieved great success in image classification (Dosovitskiy et al. 2020). However, for dense image processing tasks such as UIE, the Transformer's computational complexity scales quadratically with image size, this burden is non-trivial and sometimes unaffordable. Local-window self-attention (Liu et al. 2021) can effectively reduce the computational complexity, but its receptive field is quite limited. Moreover, existing Transformer-based UIE methods (Peng, Zhu, and Bian 2023; Tang, Kawasaki, and Iwaguchi 2023; Khan et al. 2024) often ignore the inconsistent degradation level of underwater images in different spatial regions and spectral bands, and still treat all spatial regions and spectral bands equally, resulting in unsatisfactory enhancement results. In this work, we introduce the MCSS and SWSA modules with linear complexity and combine them in parallel to form the SS-block, which can simultaneously model the degradation levels of different spatial regions and spectral bands, and achieve degradation level-based adaptive UIE.

State Space Models

Originating from control theory, State Space Models (SSMs)(Gu, Goel, and Ré 2021) have garnered increasing attention due to their efficacy in long-term language modeling. Unlike self-attention based transformers, most SSMs capture long-range token interactions through linear recurrent processes, entailing $\mathcal{O}(N)$ complexity theoretically. Mamba (Gu and Dao 2023; Xu et al. 2024) improves the expressiveness of SSMs by introducing a selective mechanism, with its structural parameters adaptively learned from inputs. Motivated by its potential for modeling high-resolution images, many researchers try to apply Mamba to vision tasks. For instance, VMamba (Liu et al. 2024) introduces a cross-scan module to enable 1D selective scanning in 2D image space. LocalMamba (Huang et al. 2024) utilizes local windows to enhance local modeling capability. EfficientV-Mamba (Pei, Huang, and Xu 2024) designs an atrous-based selective scan approach to enhance efficiency. In this work, inspired by the physical prior that underwater images have different degradation levels in different spatial regions, we designed the MCSS module based on Mamba. Its global receptive field enables it to model the degradation levels in different regions and reinforce the network's attention on severely degraded regions, achieving adaptive UIE.

Method

The Overall Architecture of SS-UIE Network

We designed the SS-UIE network based on the SS-block. As shown in Fig. 2, the network consists of three modules, including the shallow feature extraction module, the deep feature fusion module, and the image reconstruction module.

Shallow Feature Extraction. Given a low-quality underwater image $\mathbf{I}_{LQ} \in \mathbb{R}^{H*W*C}$ (H, W and C are the image's height, width and channel number), we use the shallow feature extractor $\mathrm{H}_{\mathrm{FE}}(*)$ to explore its shallow features $\mathbf{F}_0 \in \mathbb{R}^{\frac{H}{2^i}*\frac{W}{2^i}*2^i C}$ as,

$$\mathbf{F}_0 = \mathrm{H}_{\mathrm{FE}}(\mathbf{I}_{LQ}), \tag{1}$$

where i represents the number of downsampling convolution blocks. After each convolution block, the resolution of feature map $\mathbf{F}$ is halved and the number of channels is doubled.

Deep Feature Fusion. Next, we use 6 Densely Connected SS-Blocks (DCSSB) to extract different levels of high dimensional features $\mathbf{F}_j \in \mathbb{R}^{\frac{H}{2^i}*\frac{W}{2^i}*2^i C}$ from $\mathbf{F}_0$ to $\mathbf{F}_{j-1}$, denoted as,

$$\mathbf{F}_j = \mathrm{H}^{\mathrm{j}}_{\mathrm{DCSSB}}(\mathbf{F}_0, \mathbf{F}_1, \ldots, \mathbf{F}_{j-1}), \tag{2}$$

where $\mathbf{H}^{\mathrm{j}}_{\mathrm{DCSSB}}$ represents the j_{th} DCSSB block, which consists of 4 SS-blocks and a convolution layer. The last block of the deep feature fusion module is the gated fusion block, which fuses the outputs of different DCSSB blocks with self-adaptive weights w_i,

$$\mathbf{F}_{DF} = w_1\mathbf{F}_1 + w_2\mathbf{F}_2 + \ldots + w_n\mathbf{F}_n, \tag{3}$$

where $\mathbf{F}_{DF}$ represents the multi-level deep fusion features output by the gated fusion block. The self-adaptive weights w_i control the weighted fusion of each DCSSB's output,

which are adaptively adjusted through backpropagation during network training. Such a module structure is conducive to deep mining of different levels of high dimensional features, which prevents losing long-term memory as the network deepens and enhances local details.

Image Reconstruction. The image reconstruction module consists of several upsampling convolutional blocks. We retrieve high-quality underwater images by aggregating shallow features and multi-level deep fusion features. The operation is described as,

$$\mathbf{I}_{pred} = \mathrm{H}_{\mathrm{REC}}(\mathbf{F}_0 + \mathbf{F}_{DF}). \tag{4}$$

The shallow features $\mathbf{F}_0$ are mainly low-frequency image features, while the multi-level deep fusion features $\mathbf{F}_{DF}$ focus on recovering lost medium-frequency and high-frequency features.

Spatial-Spectral Block

The detailed structure of SS-block is shown in Fig. 2. Assuming that the inputs of SS-block are feature maps $\mathbf{X}_{in} \in \mathbb{R}^{\frac{H}{2^i}*\frac{W}{2^i}*2^i C}$ at different scales. To be specific, for an input feature map $\mathbf{X}_{in}$, it is first passed through a 1×1 convolution and split evenly into two feature maps $\mathbf{X}_1$ and $\mathbf{X}_2$,

Next, $\mathbf{X}_1 \in \mathbb{R}^{\frac{H}{2^i}*\frac{W}{2^i}*2^i C}$ is fed into the SWSA module for further processing. As shown in Fig. 2, in the SWSA, we propose a global filter $\mathbf{K}$ as an alternative to the self-attention layer which can mix tokens representing different spectrum information. For input feature $\mathbf{X}_1$, we first perform layer normalization (LN). Then, we use 2D FFT to convert $\mathbf{X}_1$ to the frequency domain,

$$\mathbf{X}_F = \mathcal{F}(\mathrm{LN}(\mathbf{X}_1)), \tag{5}$$

where $\mathcal{F}(\cdot)$ denotes the 2D FFT. The output feature $\mathbf{X}_F$ represents the spectrum of $\mathbf{X}_1$. We can then modulate the spectrum by multiplying a learnable filter $\mathbf{K} \in \mathbb{R}^{\frac{H}{2^i}*\frac{W}{2^i}*2^i C}$ to the $\mathbf{X}_F$,

$$\mathbf{Y}_F = \mathbf{K} \odot \mathbf{X}_F, \tag{6}$$

where $\odot$ is the element-wise multiplication (Hadamard product). The filter K has the same dimension with $\mathbf{X}_F$, which can represent an arbitrary filter in the spectral domain and enables global spectral information exchange in the SWSA module. Finally, we adopt the IFFT to transform the modulated spectrum $\mathbf{Y}_F$ back to the spatial domain and add with a residual value controlled by the learnable scale factor α_1 to get the $\mathbf{Y}_1$,

$$\mathbf{Y}_1 = \mathrm{MLP}(\mathrm{LN}(\mathcal{F}^{-1}(\mathbf{Y}_F) + \alpha_1 * \mathbf{X}_1)), \tag{7}$$

where $\mathcal{F}^{-1}(\cdot)$ denotes the 2D IFFT. The global filter $\mathbf{K}$ is motivated by the frequency filters in the digital image processing (Pitas 2000), which can be regarded as a set of learnable frequency filters for different hidden dimensions.

Similarly, $\mathbf{X}_2 \in \mathbb{R}^{\frac{H}{2^i}*\frac{W}{2^i}*2^i C}$ is sent to the MCSS module. Our MCSS module is built on the Select-Scan Structured State Space for Sequences (S6) block (Gu and Dao 2023). The S6 block is the core of Mamba and can achieve global receptive field, dynamic weight, and linear complexity at the same time. While the sequential nature of the scanning operation in S6 aligns well with NLP tasks involving
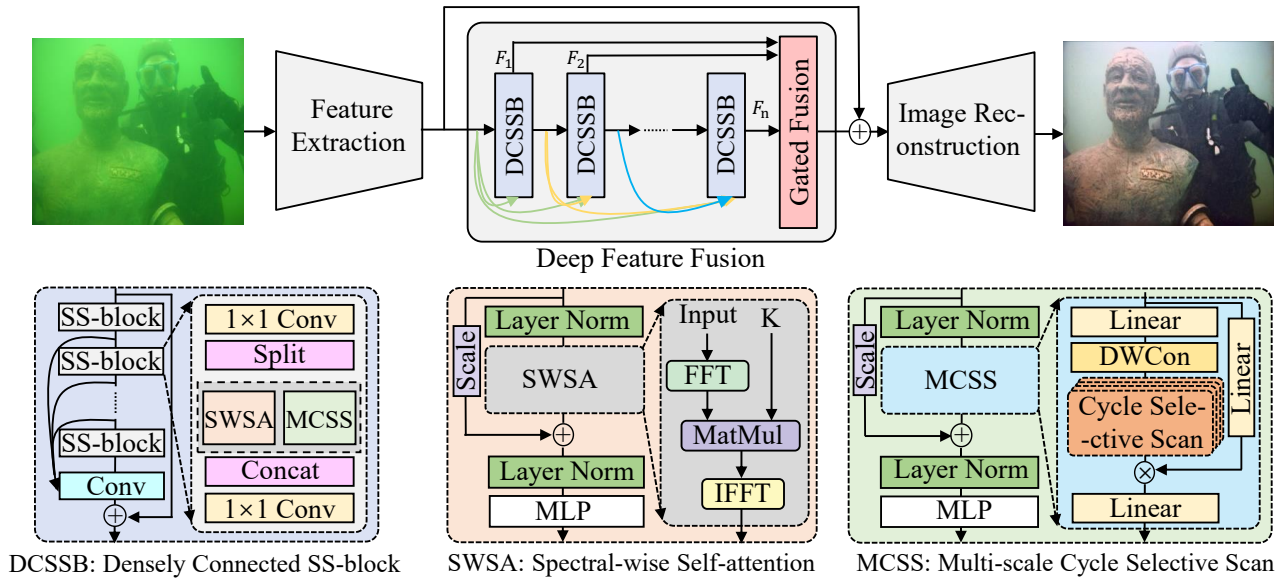
Figure 2: The overall structure of SS-UIE. We combine the spatial-wise Multi-scale Cycle Selective Scan (MCSS) module with the Spectral-Wise Self-Attention (FWSA) module in parallel to form the Spatial-Spectral block (SS-block). MCSS and SWSA can obtain the spatial-wise and spectral-wise global receptive field with linear complexity, respectively. The parallel design facilitates complementary interactions between spatial and spectral features, enabling the SS-block to capture the degradation levels in different spatial regions and spectral bands, thereby achieving degradation level-based dual-domain adaptive UIE.

temporal data, it poses a significant challenge when applied to vision data, which is inherently nonsequential and encompasses spatial information (e.g., local texture and global structure). In order to simultaneously perceive local textures and global features in visual data, we propose the multi-scale cycle selective-scan strategy (as shown in Fig. 3) to adapt S6 to vision data without compromising its advantages.

As shown in Fig. 2, the input feature $\mathbf{X}_2$ will go through two parallel branches. In the first branch, the feature's channel number is expanded to $\lambda 2^i C$ by a linear layer, where λ is a pre-defined channel expansion factor. Subsequently, the feature is processed by a depth-wise convolution (DW-Conv) (Han et al. 2023) and an MCSS layer (as shown in Fig. 3). In the second branch, the features channel is also expanded to $\lambda 2^i C$ with a linear layer. The linear layer consists of a 1×1 convolution and a SiLu activation function. After that, features from the two branches are aggregated with the Hadamard product. Finally, the channel number is projected back to $2^i C$ to generate output $\mathbf{Y}_2'$ with the same shape as input:

$$\mathbf{X}_2' = \text{MCSS}(\text{DWConv}(\text{Linear}(\text{LN}(\mathbf{X}_2)))),$$
$$\mathbf{Y}_2' = \text{Linear}(\mathbf{X}_2' \odot (\text{Linear}(\mathbf{X}_2))). \quad (8)$$

Then, $\mathbf{Y}_2'$ needs to added with a residual value controlled by the learnable scale factor α_2 to get the output of MCSS module,

$$\mathbf{Y}_2 = \text{MLP}(\text{LN}(\mathbf{Y}_2' + \alpha_2 * \mathbf{X}_2)). \quad (9)$$

Finally, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are concatenated as the input of a 1×1 convolution, which has a residual connection with the input

$\mathbf{X}_{in}$. As such, the final output of SS-block is given by,

$$\mathbf{X}_{out} = \text{Conv1} \times 1(\text{Concat}(\mathbf{Y}_1, \mathbf{Y}_2)) + \mathbf{X}_{in}. \quad (10)$$

Frequency-wise Loss

The F-Principle (Xu, Zhang, and Xiao 2019) proves that deep learning networks tend to prefer low frequencies to fit the objective, which will result in the frequency domain gap. Recent studies (Wang et al. 2020; Zhang, Karaman, and Chang 2019) indicate that the periodic pattern shown in the frequency spectrum may be consistent with the artifacts in the spatial domain. In UIE task, the model overfitting at low frequencies brings smooth textures and blurry structures. So exploring adaptive constraints on specific frequencies is essential for the refined reconstruction.

In this work, we use the Discrete Fourier Transform (DFT) to convert the underwater images to the frequency domain and to supervise the frequency distance between truth and predicted images adaptively. We calculate the frequency spectrum for each channel. In a specific channel k, the conversion relationship between spatial coordinates (h, w, k) and frequency domain coordinates (u, v) is expressed as,

$$\mathbf{F}^k(u, v) = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \mathbf{I}(h, w, k) e^{-j2\pi(\frac{uh}{H} + \frac{vw}{W})}, \quad (11)$$

where $\mathbf{F}$ is the frequency spectra of all channels corresponding to the underwater image $\mathbf{I}$. Then we can calculate the dynamic weights based on the frequency distance to make the model concentrate on high-frequency details that are difficult to reconstruct. Specifically, In each channel k, the frequency distance between ground truth and predicted image
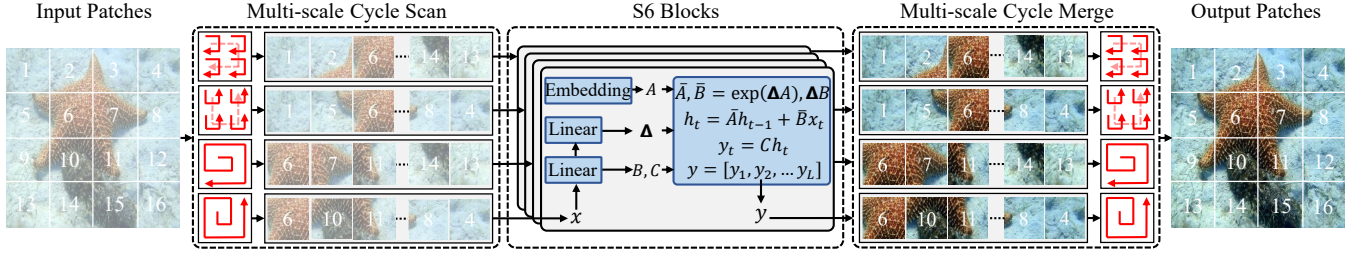
Figure 3: The framework of MCSS module. Given the input data **X**, MCSS first unfolds input patches into sequences along multiple distinct traversal paths (i.e., multi-scale cycle scan), processes each patch sequence using a separate S6 block in parallel, and subsequently reshapes and merges the resultant sequences to form the output map (i.e., feature-merge).

	Methods	Fusion	MLLE	Water-Net	UGAN	U-Color	U-Shape	DM-Water	WF-Diff	Ours
UIEB	FID↓	38.28	37.89	37.48	42.98	38.25	46.11	31.07	<u>27.85</u>	**27.31**
	LPIPS↓	0.239	0.217	0.212	0.204	0.234	0.226	0.144	<u>0.125</u>	**0.117**
	PSNR↑	19.04	20.47	19.35	20.68	20.71	21.25	21.88	<u>23.86</u>	**24.19**
	SSIM↑	0.821	0.862	0.832	0.842	0.841	0.845	0.819	<u>0.873</u>	**0.897**
LSUI	FID↓	39.25	30.78	38.90	32.66	45.06	28.56	27.91	<u>26.75</u>	**25.98**
	LPIPS↓	0.244	0.153	0.168	0.203	0.123	**0.103**	0.114	0.110	<u>0.106</u>
	PSNR↑	17.48	21.55	19.73	19.79	22.91	24.16	<u>27.65</u>	27.26	**28.87**
	SSIM↑	0.791	0.892	0.823	0.782	0.891	0.932	0.887	<u>0.944</u>	**0.952**
U45	UIQM↑	2.623	2.683	2.951	3.033	3.104	3.151	3.086	<u>3.181</u>	**3.246**
	UCIQE↑	0.612	0.456	0.601	0.573	0.586	0.592	<u>0.634</u>	0.619	**0.703**

Table 1: Quantitative comparison of different UIE methods on the UIEB (Li et al. 2020), LSUI (Peng, Zhu, and Bian 2023) and U45 (Li, Li, and Wang 2019) datasets. The best results are highlighted in bold and the second best results are underlined.

is equivalent to the power distance between their spectrum $\mathbf{F}_{gt}^k$ and $\mathbf{F}_{pred}^k$, which is defined as,

$$d^k(u,v) = \left\| \mathbf{F}_{gt}^k(u,v) - \mathbf{F}_{pred}^k(u,v) \right\|^2 . \quad (12)$$

Next, we define a dynamic weight factor $\theta(u,v)$ linearly related to the distance $d(u,v)$ to make the model pay more attention to the frequencies that are hard to synthesize. Then the distance between the ground truth and the predicted image in a single channel k is formulated as,

$$d(\mathbf{F}_{gt}^k, \mathbf{F}_{pred}^k) = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} \theta^k(u,v) d^k(u,v), \quad (13)$$

where the $\theta^k(u,v)$ changes linearly with the absolute value of the k_{th} channel frequency distance $\sqrt{(|d^k(u,v)|)}$. We traverse $k = \{0,1,2...C-1\}$ and sum each distance to calculate the frequency domain loss L_{FWL} as,

$$\mathrm{L}_{FWL}(\mathbf{F}_{gt}, \mathbf{F}_{pred}) = \sum_{k=0}^{C-1} d(\mathbf{F}_{gt}^k, \mathbf{F}_{pred}^k). \quad (14)$$

In addition, we also used the spatial-wise L_1 loss, and combined it with the L_{FWL} to form a dual domian loss,

$$\mathrm{L}_{total} = \mathrm{L}_1(\mathbf{I}_{gt}, \mathbf{I}_{pred}) + \lambda \mathrm{L}_{FWL}(\mathbf{F}_{gt}, \mathbf{F}_{pred}), \quad (15)$$

where λ is a weight factor, which makes the values of L_1 and L_{FWL} be of the same order of magnitude.

Experiments

Experiment Setup

Datasets. We utilize the datasets UIEB (Li et al. 2020) and LSUI (Peng, Zhu, and Bian 2023) to evaluate our model. The UIEB dataset comprises 890 images with corresponding labels. Out of these, 800 images are allocated for training, and the remaining 90 are designated for testing. The LSUI dataset is randomly partitioned into 3879 images for training and 400 images for testing. To ensure fairness, we only divide the LSUI dataset once, and then use it for the training and testing of all UIE methods involved in the comparison. In addition, to verify the generalization of SS-UIE, we use non-reference benchmarks U45 (Li, Li, and Wang 2019), which contains 45 underwater images for testing.

Evaluation Metrics. For the full-reference experiment, we primarily utilize two metrics: PSNR and SSIM (Wang et al. 2004). 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 (Zhang et al. 2018) is a deep neural network-based image quality metric that assesses the perceptual similarity between two images. FID (Heusel et al. 2017) measures the distance between the distributions of two images. A lower LPIPS and FID score indicates a more effective UIE approach. For non-reference benchmark U45, we introduce UIQM (Panetta, Gao, and Agaian 2016) and
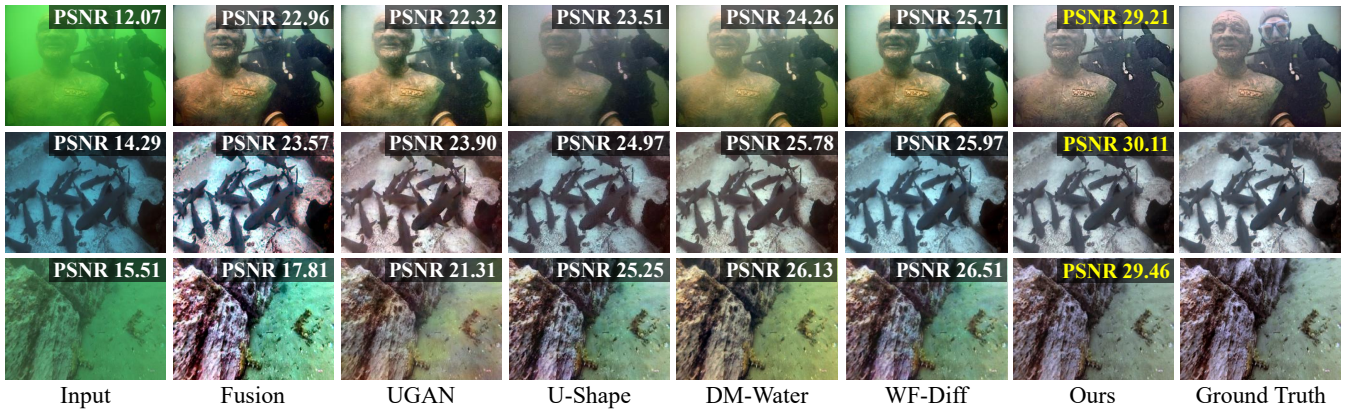
Figure 4: Visual comparison of enhancement results sampled from the test set of LSUI and UIEB dataset. The highest PSNR scores are marked in yellow. It can be seen that the enhancement results of our method are the closest to the ground truth.

UCIQE (Yang and Sowmya 2015) to evaluate our method.

Comparison Methods. We conduct a comparative analysis between our SS-UIE and eight state-of-the-art (SOTA) UIE methods, namely Fusion (Ancuti et al. 2012), MLLE (Zhang et al. 2022), Water-Net (Li et al. 2020), UGAN (Fabbri, Islam, and Sattar 2018), U-Color (Li et al. 2021), U-Shape (Peng, Zhu, and Bian 2023), DM-Water (Tang, Kawasaki, and Iwaguchi 2023), and WF-Diff (Zhao et al. 2024). To ensure a fair comparison, we utilize the provided source codes from the respective authors and adhere strictly to the identical experimental settings across all evaluations.

Experiment Results

Full-Reference Evaluation. The test sets of UIEB and LSUI datasets were used for evaluation. The statistical results and visual comparisons are presented in Tab. 1 and Fig. 4. As shown in Tab. 1, our SS-UIE demonstrates the best UIE performance on almost all full-reference metrics while requiring fewer Params and FLOPs (as shown in Fig. 1). The visual comparisons shown in Fig. 4 reveal that the enhancement results of our method are the closest to the reference image, which has high-fidelity local details and fewer color casts and artifacts. This demonstrates the effectiveness of using SS-block to capture the degradation degrees in different spatial regions and spectral bands, and performing degradation level-based adaptive UIE. It also proves that the parallel design can help the SS-block to achieve the interaction of spatial and spectral features while maintaining linear complexity, thereby obtaining both the spatial-wise and spectral-wise global receptive field.

Non-Reference Evaluation. The U45 test set was utilized for the non-reference evaluation, in which statistical results and visual comparisons are shown in Tab. 1 and Fig. 5. As shown in Tab. 1, our method achieved the highest scores on UIQM and UCIQE metrics, demonstrating the superior generalization performance of SS-UIE. Moreover, as shown in Fig. 5, the UIE performance of previous methods in high-frequency local details is unsatisfactory. These methods either produce overly smooth results, sacrificing fine-grained textural details, or introduce undesirable speckled textures.

BL	SS2D	MCSS	SWSA	Serial	Parallel	FWL	PSNR	SSIM
✓							22.67	0.859
✓	✓						23.88	0.873
✓		✓					24.47	0.892
✓			✓				24.03	0.879
✓		✓	✓	✓			26.87	0.927
✓		✓	✓		✓		28.13	0.938
✓		✓	✓		✓	✓	**28.87**	**0.952**

Table 2: Break-down ablation study. Models are trained and tested on the LSUI dataset.

In contrast, our SS-UIE can accurately reconstruct the high-frequency local details. This is because the FWL can reinforce the network's attention in the regions with rich detail and improve the UIE quality in those regions.

Ablation Study

To demonstrate the effectiveness of each component, we conducted a series of ablation studies on the LSUI dataset. We consider several factors including the SWSA and MCSS module in the SS-block, and the FWL function. The baseline (BL) model is derived by replacing our SS-block with the Swin transformer block (Liu et al. 2021), and removing the FWL from SS-UIE. Serial and Parallel represent combining the SWSA and MCSS modules using serial and parallel design, respectively. Additionally, to validate the effectiveness of the multi-scale cycle selective scanning strategy in MCSS, we compared it with the original 2D Selective Scan (SS2D) in Vision-Mamba (Liu et al. 2024).

Effectiveness of SS-Blcok. From Tab. 2, we observe that the removal of any component from the SS-block leads to a degradation in performance, which demonstrates the effectiveness of each component in the SS-block and the effectiveness of their combination. Moreover, the UIE performance of BL+MCSS is better than that of BL+ SS2D,
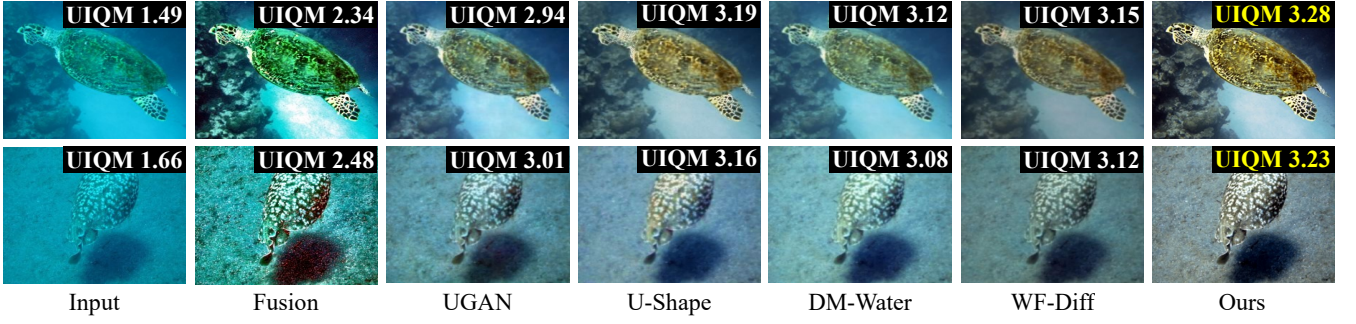
Figure 5: Visual comparison of the non-reference evaluation sampled from the U45 dataset. Compared with existing methods, our method exhibits fewer color casts and artifacts, and recovers high-frequency local details better.
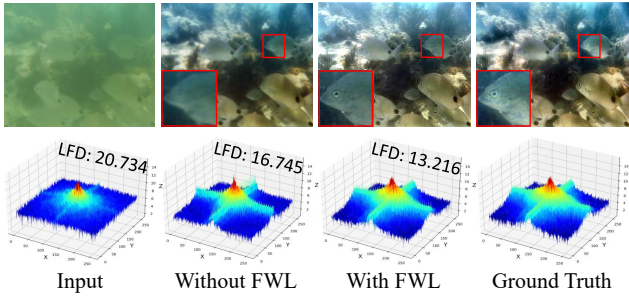

Figure 6: Frequency spectrum visualization with or without FWL. The metric LFD is used to measure the frequency similarity. The smaller the LFD, the closer the reconstructed image is to the ground truth.

which proves the effectiveness of the multi-scale cyclic selective scanning strategy. However, only using MCSS cannot capture the degradation levels between different spectral bands, which is why the reconstruction performance of BL+MCSS is not as good as BL+MCSS+SWSA. Furthermore, we can see that using a parallel design to combine MCSS and SWSA results in better UIE performance than using a serial design. Moreover, the computational and memory costs of parallel design (4.25M, 19.37G) are also fewer than serial design (5.78M, 28.95G). This is because in a serial design, the network processes spatial and spectral features in successive steps, which makes these two types of features less interweaved. Conversely, the parallel design facilitates the interaction of spatial and spectral features while maintaining linear complexity, thereby enabling the SS-block to simultaneously capture spatial-wise and spectral-wise global receptive fields, model the degradation levels in different spatial regions and spectral bands, and perform degradation level-based adaptive UIE.

Effectiveness of FWL. We calculate the log frequency distance (LFD) as a frequency-level metric to evaluate the spectrum difference.

$$\mathrm{F}_{LFD} = log\left[\frac{1}{HW}\left(\sum_{u=0}^{H-1}\sum_{v=0}^{W-1}|d(u,v)|\right) + 1\right]. \quad (16)$$

The LFD has a logarithmic relationship with the frequency

distance $d(u,v)$ in Eq. 12. As shown in Fig. 6, we visualize the 3D-spectra reconstructed with or without the FWL and provide the corresponding LFD. It can be seen that the 3D-spectra optimized with our proposed FWL exhibits more accurate frequency reconstruction and lower LFD, aligning the frequency statistics more closely with the ground truth. On the contrary, the frequency 3D-spectra of the reconstructed image without FWL supervision loses a lot of frequency information, resulting in amplitude and phase distortions that manifest as blurred details and artifacts in the predicted image. Additionally, as shown in Tab. 2, the inclusion of FWL leads to an increase in the PSNR and SSIM scores of SS-UIE. This is because fine-grained spectrum supervision can preserve more high-frequency information that is difficult to synthesize, thereby achieving higher UIE performance. Moreover, since FWL does not change the network structure, the improvements achieved by FWL will not bring any additional memory and computational costs during testing.

Conclusion

In this paper, we explore how to achieve adaptive UIE based on the degradation levels in different spatial regions and spectral bands with linear complexity. To this end, we propose a novel adaptive dual-domain learning method, termed SS-UIE. Specifically, we first introduce the SWSA and MCSS modules and combine them in parallel to form a basic SS-block. MCSS and SWSA can obtain the spatial-wise and spectral-wise global receptive field with linear complexity , respectively. Parallel design allows the interaction of spatial and spectral features, and helps the SS-block to capture the degradation levels in different spatial regions and spectral bands, thereby achieving degradation level-based adaptive UIE. In addition, FWL is introduced for UIE to narrow the frequency-wise discrepancy. Dynamic frequency-level supervision forces the model to reconstruct fine-grained frequencies and improves the imaging quality in regions with high-frequency details. Extensive quantitative and qualitative experiments demonstrate that our SS-UIE outperforms other UIE methods while requiring cheaper computational and memory costs.

Acknowledgments

References

Akkaynak, D.; and Treibitz, T. 2019. Sea-Thru: A Method for Removing Water From Underwater Images. In *CVPR*, 1682–1691.

Ancuti, C.; Ancuti, C. O.; Haber, T.; and Bekaert, P. 2012. Enhancing Underwater Images and Videos by Fusion. In *CVPR*, 81–88.

Chiang, J. Y.; and Chen, Y.-C. 2012. Underwater Image Enhancement by Wavelength Compensation and Dehazing. *IEEE T. Image Process.*, 21(4): 1756–1769.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Drews Jr, P.; do Nascimento, E.; Moraes, F.; Botelho, S.; and Campos, M. 2013. Transmission Estimation in Underwater Single Images. In *ICCV workshops*, 825–830.

Fabbri, C.; Islam, M. J.; and Sattar, J. 2018. Enhancing Underwater Imagery Using Generative Adversarial Networks. *ICRA*, 7159–7165.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Commun. ACM*, 63(11): 139–144.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Han, D.; Wang, Z.; Xia, Z.; Han, Y.; Pu, Y.; Ge, C.; Song, J.; Song, S.; Zheng, B.; and Huang, G. 2024. Demystify Mamba in Vision: A Linear Attention Perspective. *arXiv preprint arXiv:2405.16605*.

Han, Q.; Fan, Z.; Dai, Q.; Liu, J.; and Wang, J. 2023. ON THE CONNECTION BETWEEN LOCAL ATTENTION AND DYNAMIC DEPTH-WISE CONVOLUTION. *IEEE TPAMI*, 45(6): 29.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 6629–6640. ISBN 9781510860964.

Huang, T.; Pei, X.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2024. LocalMamba: Visual State Space Model with Windowed Selective Scan. *arXiv preprint arXiv:2403.09338*.

Islam, M. J.; Xia, Y.; and Sattar, J. 2020. Fast Underwater Image Enhancement for Improved Visual Perception. *IEEE Robot. Autom. Lett.*, 5(2): 3227–3234.

Khan, M. R.; Mishra, P.; Mehta, N.; Phutke, S. S.; Vipparthi, S. K.; Nandi, S.; and Murala, S. 2024. Spectroformer: Multi-Domain Query Cascaded Transformer Network For Underwater Image Enhancement. In *WACV*, 1443–1452.

Li, C.; Anwar, S.; Hou, J.; Cong, R.; Guo, C.; and Ren, W. 2021. Underwater Image Enhancement via Medium Transmission-Guided Multi-Color Space Embedding. *IEEE T. Image Process.*, 30: 4985–5000.

Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; and Tao, D. 2020. An Underwater Image Enhancement Benchmark Dataset and Beyond. *IEEE T. Image Process.*, 29: 4376–4389.

Li, C.; Guo, J.; Chen, S.; Tang, Y.; Pang, Y.; and Wang, J. 2016a. Underwater Image Restoration Based on Minimum Information Loss Principle and Optical Properties of Underwater Imaging. In *ICIP*, 1993–1997.

Li, C.-Y.; Guo, J.-C.; Cong, R.-M.; Pang, Y.-W.; and Wang, B. 2016b. Underwater Image Enhancement by Dehazing With Minimum Information Loss and Histogram Distribution Prior. *IEEE T. Image Process.*, 25(12): 5664–5677.

Li, H.; Li, J.; and Wang, W. 2019. A fusion adversarial underwater image enhancement network with a public test dataset. *arXiv preprint arXiv:1906.06819*.

Li, J.; Skinner, K. A.; Eustice, R. M.; and Johnson-Roberson, M. 2017. WaterGAN: Unsupervised Generative Network to Enable Real-Time Color Correction of Monocular Underwater Images. *IEEE Robot. Autom. Lett.*, 3(1): 387–394.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. VMamba: Visual State Space Model. *arXiv preprint arXiv:2401.10166*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 10012–10022.

Nussbaumer, H. J.; and Nussbaumer, H. J. 1982. *The fast Fourier transform*. Springer.

Panetta, K.; Gao, C.; and Agaian, S. 2016. Human-Visual-System-Inspired Underwater Image Quality Measures. *IEEE J. Ocean. Eng.*, 41(3): 541–551.

Pei, X.; Huang, T.; and Xu, C. 2024. EfficientVMamba: Atrous Selective Scan for Light Weight Visual Mamba. *arXiv preprint arXiv:2403.09977*.

Peng, L.; Zhu, C.; and Bian, L. 2023. U-Shape Transformer for Underwater Image Enhancement. *IEEE T. Image Process.*, 32: 3066–3079.

Pitas, I. 2000. *Digital image processing algorithms and applications*. John Wiley & Sons.

Sahu, P.; Gupta, N.; and Sharma, N. 2014. A survey on underwater image enhancement techniques. *IJCA*, 87(13).

Schettini, R.; and Corchs, S. 2010. Underwater Image Processing: State of the Art of Restoration and Image Enhancement Methods. *EURASIP. J. Adv. Signal Process.*, 2010: 1–14.

Tang, Y.; Kawasaki, H.; and Iwaguchi, T. 2023. Underwater image enhancement by transformer-based diffusion model

with non-uniform sampling for skip strategy. In *ACM MM*, 5419–5427.

Uplavikar, P. M.; Wu, Z.; and Wang, Z. 2019. All-in-One Underwater Image Enhancement Using Domain-Adversarial Learning. In *CVPR Workshops*, 1–8.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. In *NIPS*, 5998–6008.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-generated images are surprisingly easy to spot...for now. In *CVPR*.

Wang, Y.; Liu, H.; and Chau, L.-P. 2018. Single Underwater Image Restoration Using Adaptive Attenuation-Curve Prior. *IEEE Trans. Circuits. Syst. I. Regul. Pap.*, 65(3): 992–1002.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4): 600–612.

Xu, R.; Yang, S.; Wang, Y.; Du, B.; and Chen, H. 2024. A survey on vision mamba: Models, applications and challenges. *arXiv preprint arXiv:2404.18861*.

Xu, Z.-Q. J.; Zhang, Y.; and Xiao, Y. 2019. Training behavior of deep neural network in frequency domain. In *ICONIP*, 264–274. Springer.

Yang, M.; Hu, J.; Li, C.; Rohde, G.; Du, Y.; and Hu, K. 2019. An In-Depth survey of underwater image enhancement and restoration. *IEEE Access.*, 7: 123638–123657.

Yang, M.; and Sowmya, A. 2015. An Underwater Color Image Quality Evaluation Metric. *IEEE T. Image Process.*, 24(12): 6062–6071.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *ECCV*, 818–833. Springer.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 586–595.

Zhang, W.; Zhuang, P.; Sun, H.-H.; Li, G.; Kwong, S.; and Li, C. 2022. Underwater Image Enhancement via Minimal Color Loss and Locally Adaptive Contrast Enhancement. *IEEE T. Image Process.*, 31: 3997–4010.

Zhang, X.; Karaman, S.; and Chang, S.-F. 2019. Detecting and Simulating Artifacts in GAN Fake Images. In *WIFS*, 1–6.

Zhao, C.; Cai, W.; Dong, C.; and Hu, C. 2024. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *CVPR*, 8281–8291.