Learning Underwater Image Enhancement Iteratively without Reference Images

Yi Tang[1] Hiroshi Kawasaki[2], Takafumi Iwaguchi[2], Yuhang Zhang[3], Hiroshi Masui[1]

[1]School of Regional Innovation and Social Design Engineering, Kitami Institute of Technology
[2]Graduate School and Faculty of Information Science and Electrical Engineering
[3]School of Computer Science and Cyber Engineering, Guangzhou University
tangyi@mail.kitami-it.ac.jp, {kawasaki,iwaguchi}@ait.kyushu-u.ac.jp
yuhang.zhang@gzhu.edu.cn, hgmasui@mail.kitami-it.ac.jp

Abstract

Since high-fidelity reference images are difficult to obtain in real underwater scenes, most deep models trained by synthetic paired data cannot match real-world data exactly. In this paper, we propose an unsupervised training framework for underwater image enhancement (UIE) by leveraging an iterative training strategy and quantification of specific neural units. Specifically, to eliminate the heavy color cast and distortion in the underwater images, we decompose the unsupervised image enhancement as two targeted sub-tasks, namely colorization and color compensation. First, a diffusion model is introduced for colorization to correct the green and blue color casts. Then, to intensify the learning ability of balanced color information, we introduce an extra network branch and propose a quantification mechanism for color compensation. The extra branch encodes style information from normal images into the generative model, while the quantification mechanism identifies and adjusts neural units relevant to warm colors, improving the model's ability to learn balanced color feature representations for robust generation. In the end, through iterative training, color cast and distortion are progressively reduced, leading to a gradual improvement in the quality of the generated images. Experimental results on various widely used underwater datasets demonstrate that our approach achieves excellent performance, even when compared to recent supervised methods.

Code — https://github.com/piggy2009/DM-noreference

Introduction

Underwater images are essential for ocean exploration, marine ecological protection, and ocean biology research (Kimball et al. 2018). However, these images often suffer from significant degradation, such as color cast and distortion, due to wavelength-dependent light absorption and scattering in water. These issues make it challenging to use underwater images for advanced visual tasks like object tracking (Wang et al. 2022), autonomous underwater vehicle (AUV) navigation (Aladem, Baek, and Rawashdeh 2019), and underwater biological detection (Guo, Lu, and Wu 2021). Therefore, it is a critical research topic to improve images to high-quality and clear ones in underwater scenes.

Traditional approaches, based on physical models such as the Retinex model (Land 1977) or Koschmieder's model (McGlamery 1980), can reduce some noise in some simple underwater images but struggle with the complexity of modern scenes. Current underwater image datasets, like the LSUI (Peng, Zhu, and Bian 2021) and RUIE (Liu et al. 2020), include diverse underwater scenes and extremely severe color casts. Physical model-based methods can address specific types of noise in some simple scenes without reference images, but their performance is limited in complex ones with diverse noise sources. Recently, deep learning-based methods (Li et al. 2019; Fabbri, Islam, and Sattar 2018; Islam, Luo, and Sattar 2020) have employed generative adversarial networks (GAN) (Islam, Xia, and Sattar 2020; Li et al. 2017; Fabbri, Islam, and Sattar 2018), attention modules (Li et al. 2021; Kim et al. 2021), and other techniques for underwater image enhancement. As we cannot remove the water and capture high-quality images as reference images in real underwater scenes, according to human visual senses, researchers try to create some synthetic datasets (Li et al. 2019; Peng, Zhu, and Bian 2021), which are widely used in recent frameworks for their model training in supervising ways. However, the annotation process for these datasets, such as UIEB (Li et al. 2019) and LSUI (Peng, Zhu, and Bian 2021), is still labor intensive, resulting in limited dataset sizes.

In order to solve the issue of insufficient training data, several approaches propose tailored deep models with unsupervised training strategies. USUIR (Fu et al. 2022a) proposes an unsupervised deep framework by using the homology property and physical model. DIP (Ulyanov, Vedaldi, and Lempitsky 2018) and DDIP (Gandelsman, Shocher, and Irani 2019) introduce self-supervising losses or regularizers for network training. These unsupervised approaches mainly rely on Koschmieder's model and then exploit the enhancement and reconstruction framework to complete the network training with a reconstruction loss function, whose brief framework is shown in Fig. 1(b). Similarly, these unsupervised approaches also struggle with heavy or mixed noise in complex underwater scenes.

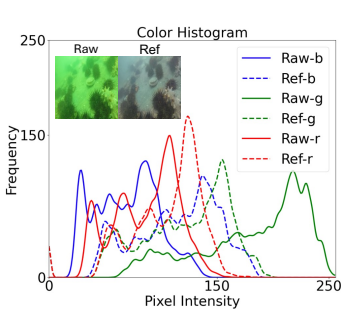
Statistics from (Xie et al. 2024) indicate that most of the underwater images in existing datasets suffer from green or blue color casts. Consequently, a primary goal of underwater image enhancement is to eliminate color degradation and

recover the original, natural colors. Here, we present a comparison between an underwater image and its corresponding reference image. As shown in Fig. 1(a), the color histogram reveals an unbalanced color distribution, with an excess of blue and green pixels compared to red. Additionally, the reference image typically exhibits warmer tones, which align more closely with human visual preferences. This observation also underpins the design of many synthetic datasets, where reference images are produced to reflect perceptually favorable color characteristics.

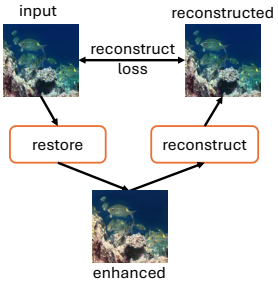
Based on these observations, we propose a diffusion model-based unsupervised training framework in this paper. Specifically, we decompose the enhancement into colorization and color compensation. To address the issue of color cast, we propose a conditional diffusion model to colorize the gray-scale underwater image. Due to a shortage of supervising color information from reference images, the initial generative image contains blur noise and lacks warm colors. To mitigate this and improve the color contrast, we introduce an extra style encoding branch to extract balanced color information from normal images for color compensation. Additionally, inspired by the analytic framework of deep networks in (Bau et al. 2018), we propose a tailored quantification mechanism to detect the neural units relevant to the warm colors and then adjust the values of the corresponding feature maps by a Gaussian-like weight, enabling the generative network to activate warm color information effectively. Furthermore, as shown in Fig. 1(c), we propose an iterative training strategy in an unsupervised manner. The trained diffusion model by the original underwater images is used as a generator to produce the enhanced images, which serve as pseudo-reference images for retraining the network in the next round. This iterative process can gradually improve the generator's ability to produce higher-quality images.

In summary, our contributions can be concluded:

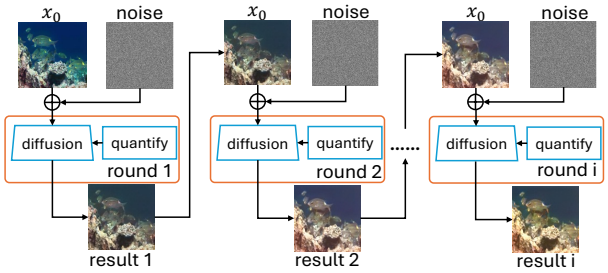
- We introduce a novel diffusion model-based unsupervised deep learning framework that decomposes underwater image enhancement into colorization and color compensation, and propose an iterative training framework to enhance image quality progressively.
- We design an extra style branch in the denoising network and a tailored quantification mechanism to encode style information from normal images into the network and intensify the learning of relevant color information, effectively eliminating color cast and compensating for balanced color information in the generative results.
- Our diffusion model can generate sufficient clean images for training. Besides, the proposed method is visually valid against recent unsupervised and supervised approaches by experiments from the widely used datasets.



(a) Comparison of the color curves between an underwater image and its corresponding reference image.



(b) Previous unsupervised training framework by an enhancement-reconstruction pipeline.



(c) Proposed iterative training process via diffusion model and quantification mechanism

Figure 1: Insight of our motivation and comparison between the previous unsupervised methods and the proposed framework.

Related Works

Underwater Image Enhancement

Underwater image enhancement has progressed from early physics-based techniques to modern deep learning methods. Classical approaches rely on models such as Retinex (Fu et al. 2014), fusion-based strategies (Ancuti et al. 2012, 2017; Zhang et al. 2022), pixel-range stretching (Iqbal et al. 2010), and distribution adjustment (Ghani and Isa 2015), but they struggle in complex scenes and are often computationally heavy. Deep learning has since enabled more effective solutions (Li et al. 2019; Islam, Xia, and Sattar 2020; Fabbri, Islam, and Sattar 2018; Uplavikar, Wu, and Wang 2019; Li et al. 2021; Kim et al. 2021). WaterNet (Li et al. 2019) employs gated fusion; LaffNet (Yang, Huang, and Chen 2021) uses a lightweight U-Net with adaptive fusion; and Ucolor (Li et al. 2021) incorporates attention with multi–color-space guidance. The synthetic LSUI dataset (Peng, Zhu, and Bian 2021) further supports stronger models, including transformer-based Ushape (Peng, Zhu, and Bian 2021) and the CVAE-driven PUIENet (Fu et al. 2022b). Unsupervised approaches have also emerged, leveraging GANs (Li et al. 2017; Li, Guo, and Guo 2018; Yang et al. 2020) and physical priors.

Diffusion Model

To simplify the training process of the GAN, the denoising diffusion probabilistic model (DDPM) (Ho, Jain, and Abbeel 2020) is proposed as a new generative model. Its framework includes two processes: forward and backward process. First, based on the Markov chain, a normal image x_0 is transferred into a noisy image x_t as follows:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon \quad (1)$$

where $\boldsymbol{\epsilon}$ denotes a gaussian noise, α_t is a hyper-parameter and t is timestep. Second, the backward process is a denoising process, whose purpose is to remove the noise step by step with the following equation:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\Big(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\Big) + \sigma_t \boldsymbol{z} \qquad (2)$$

where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \cdot \beta_t$, $\bar{\alpha}_t, \beta_t$ are hyper-parameter and $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ is a trainable denoising neural network. By iteratively using this equation, we can obtain the predicted x_0 in the end.

The original diffusion model generates images in an unconditional manner. To enable targeted generation, a conditional diffusion model was introduced in (Saharia et al. 2022), where a condition image c is added to the noise prediction network: $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, c)$. This approach has shown strong performance in various low-level vision tasks, including super-resolution (Saharia et al. 2022), low-light enhancement (Wang et al. 2023), and rain removal (Li, Liu, and Ma 2023). In the field of underwater image enhancement, TDM (Tang, Kawasaki, and Iwaguchi 2023) incorporates transformers within a supervised conditional diffusion framework, while Diff-Retinex (Yi et al. 2023) fuses physical priors and diffusion models for low-light enhancement.

The Proposed Method

The proposed framework adopts an iterative training strategy in which the network progressively learns balanced color information and produces improved pseudo-reference images for the next round. As shown in Fig. 2, training involves two tasks: colorization and color compensation. For colorization, a conditional grayscale image and a noisy image are fed into a denoising network to generate a colorful output. To address insufficient color cues, we enhance the network using two sources: semantic features from the CLIP text encoder (Radford et al. 2021) and style features extracted from an additional normal-scene image (Zhang, Rao, and Agrawala 2023). For color compensation, we introduce a quantification mechanism that identifies and enhances relevant neural units, improving output fidelity. After diffusion-based generation, a post-processing module (Lu et al. 2024) is applied to reduce haze and refine contrast.

Colorization with Diffusion Model

In the iterative training process, the denoising network from the previous round, $\boldsymbol{\epsilon}_\theta^{r-1}$, is used to generate pseudo-reference images x_0^r for the current round r. As illustrated in Fig. 2(a), the denoising network takes a quintuple input: a noisy image x_t^r, a conditional grayscale image c, a prompt text p, a style image s, and the timestep t.

Unlike prior methods (Saharia et al. 2022; Tang, Kawasaki, and Iwaguchi 2023), which directly use low-quality underwater images as conditional inputs, we introduce a color filter to retain only warm hues. The RGB image is first converted to HSV space, and a hue filter retains values within $(0, 40)$ and $(125, 180)$. The resulting conditional image c is concatenated with x_t^r to form a 6-channel input for the denoising network. The noisy image x_t^r is obtained from x_0^r using Eq. 1. We also incorporate two auxiliary inputs: a prompt text p and a style image s. Prompts are derived from semantic labels in (Peng, Zhu, and Bian 2021) and enhanced with descriptive vocabulary (e.g., "*natural coral on land and make it clearer*"). The text features are extracted by using the pre-trained CLIP model (Radford et al. 2021) and embedded via transformer modules in the network. These features can improve color contrast and alleviate other degradation types like blurriness or low-light to some extent. The style image provides global color cues and is processed by a dedicated branch that mirrors the encoder structure, with a modified first layer to accept three channels.

During training, the full input is passed through the denoising network to predict the noisy image, which is supervised using an L1 loss against the target noise distribution $\boldsymbol{\epsilon}_t^r$. The overall process is defined as:

$$L_s = \|\boldsymbol{\epsilon}_t^r - \boldsymbol{\epsilon}_\theta^r(x_t^r, c, p, s, t)\| \qquad (3)$$

where $\boldsymbol{\epsilon}_\theta^r(\cdot)$ is the trainable denoising network, θ represents the network parameters.

The proposed denoising network adopts a U-Net with eight trainable blocks: four in the encoder and four in the decoder. Each block incorporates a modified transformer-based design inspired by (Tang, Kawasaki, and Iwaguchi 2023). As shown in Fig. 2(a), each block first uses a convolution layer to adjust channel dimensions, followed by down- or up-sampling to resize feature maps. A lightweight ECA-transformer module (Tang, Kawasaki, and Iwaguchi 2023) then encodes visual features along with the timestep. Finally, text features are integrated through a spatial transformer module (Rombach et al. 2022), which applies self-attention to refine visual features and cross-attention to fuse visual and textual information. The operations within each block are summarized as follows:

$$\begin{aligned} F_1^r &= RE_m\big(Conv(f_i^r)\big), \quad m \in \{up, do\} \\ F_2^r &= T_e(F_1^r, t), \quad f_o^r = T_c(F_2^r, p) \end{aligned} \qquad (4)$$

where f_i^r and f_o^r denote the input and output features in the r-th training round. $Conv(\cdot)$ is the convolution layer. $RE_m(\cdot)$ represents the resize operation, where up and do denotes the up-sampling the down-sampling, respectively. $T_e(\cdot)$ is the ECA-transformer module. T_c is the spatial transformer. F_1^r and F_2^r are the intermediate features.

Style Encoding and Quantification Mechanism for Color Compensation

While the proposed diffusion model achieves colorization and produces preliminary results, it is insufficient to generate a more color-balanced image by only using the semantic information from the text embedding module. Inspired by (Zhang, Rao, and Agrawala 2023), we introduce an additional branch in the denoising network and use a normal image from the ImageNet dataset as a style reference. In the network, the secondary encoder is exploited to extract multi-level color features s_l^r, and then embed them into the

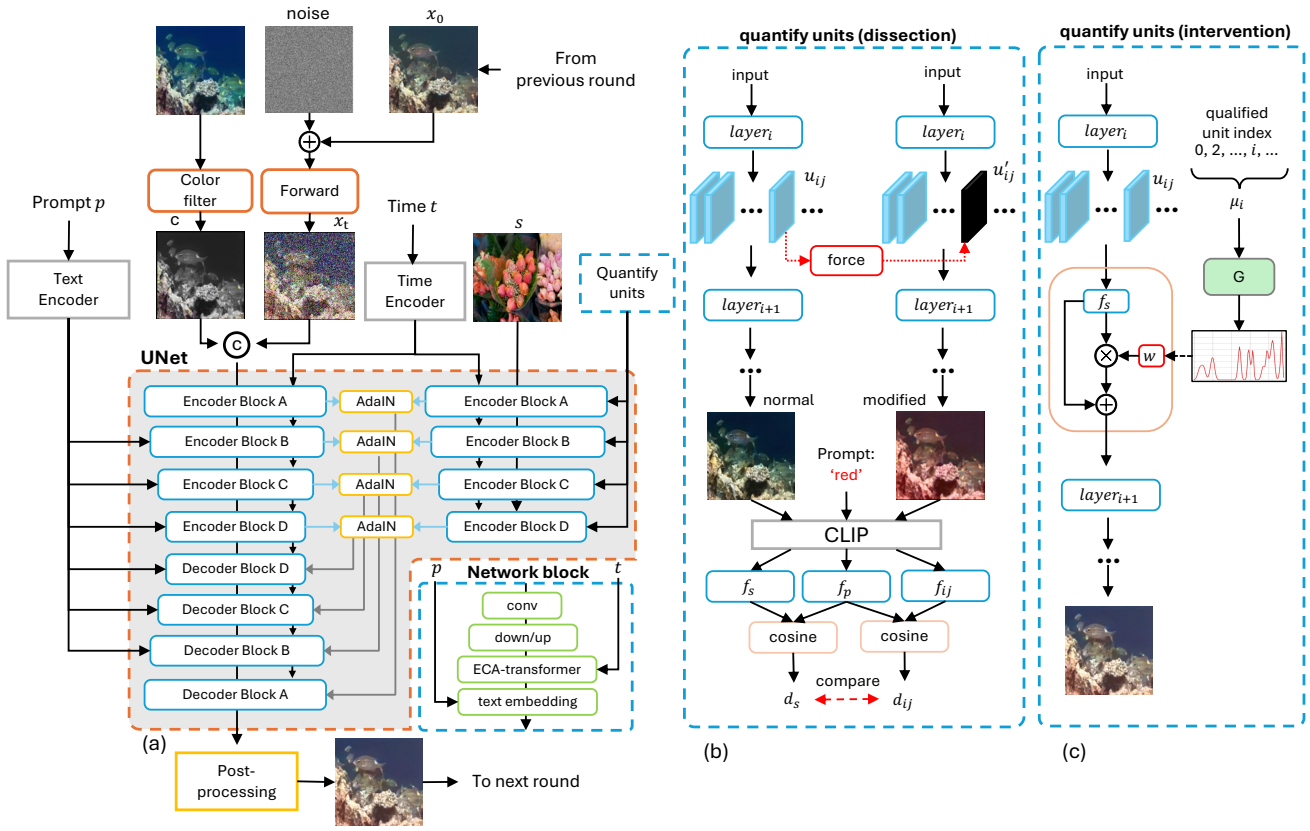


Figure 2: Denoising network and quantification mechanism in one training round. (a) The quintuple input is processed by a U-Net with an encoder–decoder backbone and a style branch, whose color features are injected via AdaIN. (b) Quantification mechanism for identifying relevant neural units in the style branch. (c) Intervention step, where feature maps of the selected units are weighted using a Gaussian-based vector.

backbone by adaptive instance normalization (AdaIN) layers (Huang and Belongie 2017). The specific process can be written as follows:

$$AdaIN(f_l^r, s_l^r) = \sigma(s_l^r)\left(\frac{f_l^r - \mu(f_l^r)}{\sigma(f_l^r)}\right) + \mu(s_l^r) \qquad (5)$$

where f_l^r and s_l^r denote the features from noisy image and style image at l-th block in the r-th round training. $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation of the feature maps. By using this branch and AdaIN layer, the network can transfer some color information into the generative results and keep color balance to some extent.

However, it is still difficult to enable the network to learn the remarkably robust color information against the supervising training with the reference images by indirectly using the semantic text and style images. Inspired by (Bau et al. 2018), we try to detect relevant neural units, which control the learning of colors and slightly adjust their values of the corresponding feature maps so that the balanced color information can be fully learned by the network. Consequently, we present the color compensation with a quantification mechanism of the neural units. As illustrated in Fig. 2(a), the quantification mechanism is activated on the style branch in the network. The reasons for these operations are two-fold. First, compared with the input in the backbone of U-Net, the style image contains more balanced color information, which needs to be transferred into the denoising network. Second, the backbone of the denoising network is used to remove the noise. Modifying the neural units in the backbone could potentially result in incomplete noise removal in the generated images, negatively impacting denoising performance. Therefore, the specific pipeline of the proposed quantification mechanism is shown in Fig. 2(b) and (c), which includes two steps: dissection and intervention.

Dissection. To identify neural units relevant to color semantics that are typically missing in the original underwater image, Fig. 2(b) shows the specific process. Given a trained generator $g(\cdot)$, we examine neural units u_{ij} in the i-th layer and j-th unit sequentially. For clarity, the quintuple input of the denoising network is denoted as k, the output of the i-th layer is represented as o_i and the final generative result is denoted as re. First, we generate a standard result re_s without any modification. Then, we test whether neural unit u_{ij} causes the generation of re by constraining the unit:

$$\begin{aligned} re_s &= g(k, o_i), \\ re_{ij} &= g(k, o_{ij}), \quad \text{where } o_{ij} = \tau \end{aligned} \qquad (6)$$

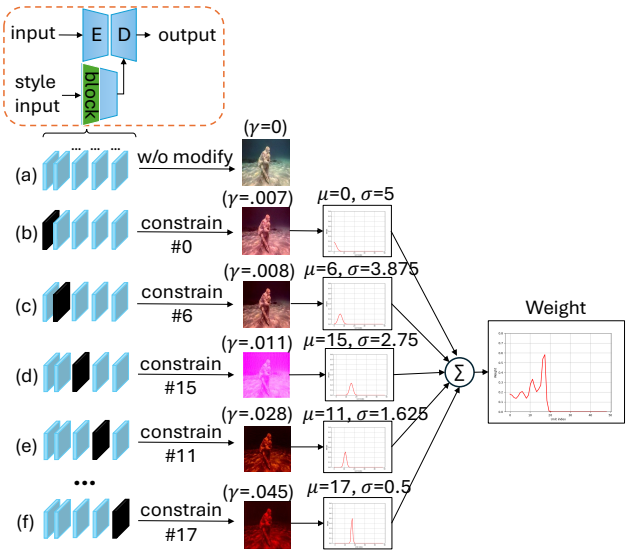


Figure 3: A use case of the proposed quantification mechanism. (a) normal results without any modification. (b)-(f) the different outputs by constraining a unit, where the units are denoted as $\#j$.

Here, we constrain the corresponding j-th feature map o_{ij} of neural unit u_{ij} in the i-th layer as a constant τ, that is set to -5 in our experiments to force the unit off, and generate new generative result re_{ij}.

After that, the two different generative results are fed into the pre-trained CLIP model to extract the visual features. Meanwhile, we also extract the text features of the color semantics by the CLIP model. Due to the shortage of warm colors in the underwater scenes, we choose the *'red'* color as the color semantic. Therefore, we will obtain the three feature vectors f_s, f_p, f_{ij} from the standard result, color semantics and modified generative results, respectively. Then, we use cosine distance to compute the similarity (d_s and d_{ij}) between the visual features and text features. In our experiments, we measure these two distances, namely $\gamma = d_s - d_{ij}$. If $\gamma > \gamma_0$, the unit u_{ij} is qualified. Here, the value of γ_0 is set to 0.005, which is chosen by the experiments. The γ denotes a similar degree with color semantics. The higher value indicates that the corresponding unit is more capable of controlling the representations of targeted color semantics.

Intervention. Even if we obtain the qualified units, their values cannot be modified arbitrarily, because the direct modification can affect the generative process and introduce extra noise in the results. Therefore, we introduce a weighting method to generate a channel-wise weight for the feature maps in the i-th layer, so that the relevant color features can be intensified, thus improving the quality of the generated images. As shown in Fig. 2(c), given the index of qualified units j and its similar degree γ_j, we format the sets of qualified units in the i-th layer as $U_i^q = \{< j, \gamma_j > | j \in C_i \ \& \ \gamma_j \in (\gamma_0, 1)\}$, where C_i denotes the channel number in the i-th layer. Then, a Gaussian distribution function $G_j(\cdot)$ is used to generate the channel-wise weight, where the index j is regarded as the mean and σ_j is sampled by an inverse uniform distribution $\hat{U}_{\sigma_j} \sim (b, a)$, where $b > a$. The process can be written as follows:

$$\begin{aligned} \sigma_j(\gamma_j) &= b + (a - b) * R(\gamma_j), \\ G_j(x) &= \frac{1}{\sigma_j \sqrt{2\pi}} e^{\frac{(x-j)^2}{2\sigma_j}} \end{aligned} \tag{7}$$

where $R(\cdot)$ represents the ranking index of the γ_j. The σ_j sampling principle is that the neural unit is more capable of affecting the color semantics, and the corresponding weight is greater. Therefore, if the γ_j is bigger, the sampled σ_j is smaller, so the corresponding Gaussian curve is narrower and sharper, which can highlight and intensify the neural unit that is more relevant to the necessary color semantics. Next, we sum all of the $G_j(x)$ and compute their sum to generate the channel-wise weight w_i for the i-th layer:

$$w_i = \sum_j^n G_j(x), \quad \hat{f}_i = f_i * w_i + f_i \tag{8}$$

where n is the number of the qualified units in the i-th layer. After that, the weight will multiply the original feature maps f_i and combine with a skip connection to obtain the weighted feature maps $\hat{f}_i$.

As shown in Fig. 3, we present a specific example of our method. Here, we quantify block 1 in the style branch as an example. The dimension of output features in block 1 is $48 * 256 * 256$, namely 48 units. During the dissection, we sequentially test the 48 units by constraining the unit. Then, we obtain the 5 qualified units. Fig. 3(a) presents the generation without any modification. From (b) to (f), we sequentially constrain the units, namely their corresponding feature maps are set to -5, and then check the values of γ. Finally, we obtain five qualified units and their respective γ_j. During the intervention process, the index of the qualified unit is regarded as the mean of the Gaussian function. The standard deviation is sampled by an inverse uniform distribution $U \sim (5, 0.5)$. As for the unit with the lowest γ, namely unit 0, the mean is set to 0 and the standard deviation is set to 5. Their means are their indices and the standard deviations are sampled from 3.875 to 0.5, respectively. After that, all of the Gaussian probability distribution functions are summed together to generate the weight w_i of the corresponding feature maps with $48 * 1 * 1$ dimensions. Then, the weighting feature maps can be obtained by Eq.8 to balance the color distribution of the final generative results.

Experiments

This section describes the experimental setup, including datasets, evaluation metrics, and implementation details. Then, we present a comparison and ablation study.

Experimental Setting

Training and testing datasets. The LSUI dataset (Peng, Zhu, and Bian 2021) contains 5,004 underwater images across various scenes (e.g., marine life, diving, wreckage).

Method	LSUI		C60			RUIE		
	PSNR↑	SSIM↑	Musiq↑	UIQM↑	URanker↑	Musiq↑	UIQM↑	URanker↑
Sea-thru	15.78	0.7307	39.15	2.214	1.278	30.16	2.759	1.035
UDCP	13.61	0.5784	39.86	1.211	-0.101	30.10	1.865	-0.227
HLRP	12.64	0.1929	40.93	1.292	1.354	32.91	2.447	0.896
FUnIE	18.78	0.6196	41.56	3.008	0.424	31.49	3.083	0.082
DGD-cGAN	16.57	0.7031	40.85	2.866	0.480	31.23	2.399	-0.786
UIEDAL	21.12	0.7231	44.28	3.044	0.858	34.27	3.279	1.241
Ushape	24.16	0.9322	40.71	2.168	1.034	25.64	2.850	1.111
HAAM	20.44	0.8243	43.42	3.009	1.262	29.63	3.021	1.238
WF-Diff	27.26	0.9437	39.71	2.009	-0.091	28.05	2.901	0.970
Lite	20.02	0.8239	42.94	2.612	1.214	31.56	2.914	1.013
Osmosis	17.18	0.7352	46.01	2.720	1.151	28.32	2.912	1.283
UUUIR†	15.64	0.7255	42.24	2.548	1.005	30.07	3.188	0.503
UDNet†	19.51	0.8045	13.67	2.616	0.681	28.36	2.977	0.389
USUIR†	19.46	0.8050	27.59	2.682	0.945	31.11	3.077	0.882
Ours†	18.42	0.8293	44.67	3.251	1.536	34.98	3.413	1.326

Table 1: Quantitative comparison on the LSUI, C60 and RUIE datasets. The top three results are successively marked in red, blue and green, respectively. † represents the unsupervised methods.

We only use the 4,500 original underwater images for training and the remaining 504 for testing. The UIEB dataset (Li et al. 2019) follows a similar annotation process. For evaluation, we use its challenging subset (C60), which includes 60 images with severe color cast or distortion. Lastly, we adopt the RUIE dataset (Liu et al. 2020), using its 630-image testing set for evaluation. Since C60 and RUIE lack reference images, we apply no-reference metrics for these datasets.

Evaluation metrics. For datasets with reference images (e.g., LSUI), we use PSNR and SSIM as evaluation metrics. PSNR measures pixel-level fidelity, while SSIM evaluates structural and textural similarity. For datasets without reference images, such as C60 and RUIE, we employ three no-reference quality metrics. UIQM (Panetta, Gao, and Agaian 2015) assesses underwater images based on colorfulness, sharpness, and contrast. Musiq (Ke et al. 2021) is a transformer-based image quality method. Additionally, we adopt URanker (Guo et al. 2023), a learning-based metric built on a transformer architecture to rank underwater image quality in a reference-free setting.

Implementation. We train the network using Adam with an initial learning rate of 1.0×10^{-4}. Inputs are resized to 256×256, normalized to $[-1, 1]$, and trained with a batch size of 6. The diffusion model uses 2000 timesteps with $\beta_t \in [10^{-6}, 10^{-2}]$. For inference, we adopt DDIM (Song, Meng, and Ermon 2020) with 10 sampling steps. All experiments run on an NVIDIA RTX 6000 Ada GPU.

Our iterative training follows 5 rounds. In each round, the base denoising network is first trained for 1000 epochs, then fine-tuned for 200 epochs with the style branch using ImageNet style images and text prompts (randomly sampled for better generalization; details in the supplementary). During fine-tuning, the quantification mechanism selects relevant units and applies weights to guide feature learning. The model is optimized solely with an L1 loss. After each round, generated images are refined with the post-processing method in (Lu et al. 2024) and fed into the next round.

Comparisons with the State-of-the-Art

We compare our method with 14 representative UIE approaches, including three traditional (non-deep) methods: Sea-thru (Akkaynak and Treibitz 2019), UDCP (Drews et al. 2013), and HLRP (Zhuang et al. 2022); eight supervised deep learning methods: FUnIE (Islam, Xia, and Sattar 2020), DGD-cGAN (Gonzalez-Sabbagh, Robles-Kelly, and Gao 2024), UIEDAL (Uplavikar, Wu, and Wang 2019), Ushape (Peng, Zhu, and Bian 2021), HAAM (Zhang et al. 2023), WF-Diff (Zhao et al. 2024), Lite (Zhang et al. 2024), and Osmosis (Nathan et al. 2024), and three unsupervised methods: UUUIR (Chai et al. 2022), UDNet (Saleh et al. 2022), and USUIR (Fu et al. 2022a). Among them, WF-Diff and Osmosis are diffusion-based models.

Table 1 summarizes quantitative results on the LSUI, C60, and RUIE datasets. On LSUI, where reference images are available, supervised models such as WF-Diff and Ushape achieve strong PSNR and SSIM scores, surpassing 24 dB and 0.93, respectively. In contrast, unsupervised methods generally show lower performance on reference-based metrics. Nevertheless, our method achieves excellent performance among unsupervised models. For C60 and RUIE, where no reference images are available, we evaluate them by using no-reference metrics: UIQM, Musiq, and URanker. As shown in Table 1, our method can achieve strong performance. For example, it achieves a UIQM score of 3.251 on C60, exceeding the second-best by 0.207. Our URanker results on C60 and RUIE also show notable improvement.

Fig. 4 provides visual comparisons. Existing methods such as HAAM, UUUIR, and USUIR struggle with residual color noise, while HLRP often overexposes images and introduces artifacts. Unsupervised approaches like UDNet cannot fully eliminate the bluish cast. In contrast, our method corrects the color cast and improves overall

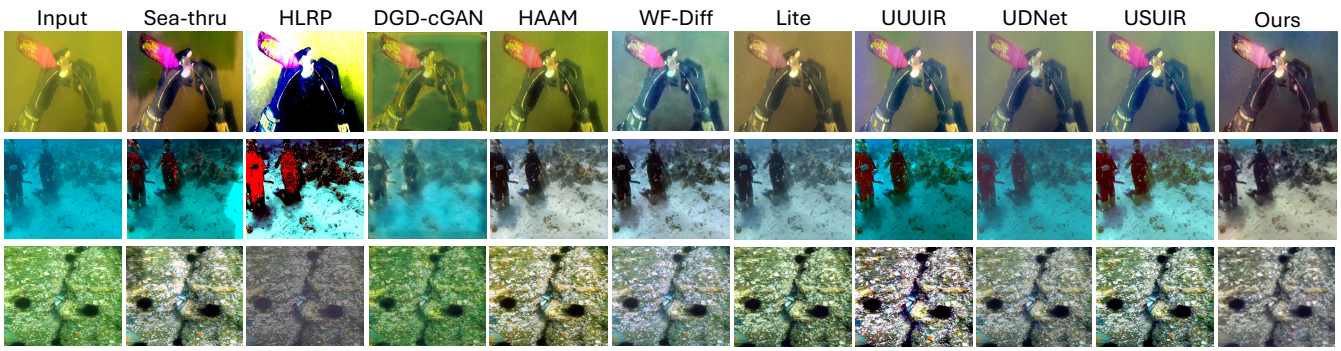


Figure 4: Visual comparison. The first column is the underwater images. The remaining are the results of Sea-thru, HLRP, DGD-cGAN, HAAM, WF-Diff, Lite, Osmosis, UUUIR, UDNet, USUIR and the proposed methods.

Rounds	round 1	round 2	round 3	round 4	round 5	round 6
UIQM↑	3.115	3.302	3.417	3.450	3.508	3.479
URanker↑	1.839	1.986	2.115	2.132	2.249	2.235

Table 2: Quantitative comparison of the quality of x_0 in the different training rounds on LSUI dataset.

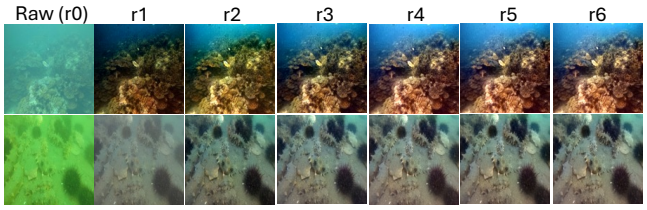


Figure 5: Visual comparison of results from r0 to r6, where the model refines x_0 progressively to improve image quality.

module	w/o SB	w/o FW	w/o CF	w/o PP	w/o TP	Full
UIQM↑	3.121	3.256	3.297	3.318	3.322	3.508
URanker↑	1.964	2.025	1.904	2.108	2.122	2.249

Table 3: Quantitative comparison of different modules on the LSUI dataset. **SB**: style branch in the denoising network. **FW**: features weights generated by the proposed quantification mechanism. **CF**: color filter function. **PP**: post-processing module. **TP**: text prompts module.

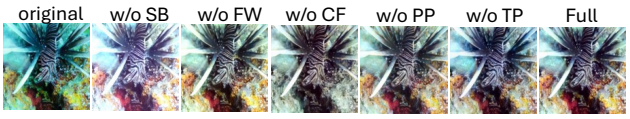


Figure 6: Visual comparison of the generative results by ablating different modules in our framework.

color balance. Although minor issues remain (e.g., undersaturation in fine details like the red bottle in the first row), our results are visually excellent with other methods.

Ablation Study

We begin by evaluating the iterative training process. Unlike conventional unsupervised frameworks, our method incrementally improves generative quality over multiple rounds. Initially, the original underwater images are treated as x_0 in the diffusion model. The trained model then generates new x_0 samples for subsequent rounds, repeating this process until performance stabilizes. As shown in Table 2 and Fig. 5, performance improves rapidly in the early rounds, especially from round 1 to 2. Color cast and haze effects are effectively removed early on, while color contrast continues to improve in later rounds. Peak performance is observed at round 5, with UIQM and URanker scores of 3.508 and 2.249, respectively. Beyond this point, improvements plateau, and a slight decline begins in round 6, suggesting that five training rounds is an optimal hyperparameter.

We conduct an ablation study to evaluate five key components of our unsupervised framework: the style branch, quantification mechanism, color filter, post-processing module, and text prompts. As shown in Table 3 and Fig. 6, each module contributes positively to overall performance. The style branch injects balanced color information from natural images. As for the quantification mechanism, we first fully identify the relevant neural units and introduce a weight to refine the color features, thus enhancing warm color cues and improving the color contrast of the generative images. The color filter is introduced to retain the warm color from the original inputs and to further guide the network training. The post-processing module, along with text prompts, helps reduce haze, blurriness, and other artifacts.

Conclusions

We propose a novel unsupervised framework for UIE based on an iterative diffusion model. By analyzing color imbalance in underwater images, we decompose the task into two subtasks: colorization to correct cast and distortion, and color compensation to restore balanced color information. To support compensation, we introduce a quantification mechanism that detects and enhances warm-color-related neural units. Experimental results show that our iterative framework effectively improves image quality, while the proposed mechanism enables robust unsupervised color learning.

References

Akkaynak, D.; and Treibitz, T. 2019. Sea-thru: A method for removing water from underwater images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1682–1691.

Aladem, M.; Baek, S.; and Rawashdeh, S. A. 2019. Evaluation of Image Enhancement Techniques for Vision-Based Navigation under Low Illumination. *Journal of Robotics*, 2019(1): 5015741.

Ancuti, C.; Ancuti, C. O.; Haber, T.; and Bekaert, P. 2012. Enhancing underwater images and videos by fusion. In *2012 IEEE conference on computer vision and pattern recognition*, 81–88. IEEE.

Ancuti, C. O.; Ancuti, C.; De Vleeschouwer, C.; and Bekaert, P. 2017. Color balance and fusion for underwater image enhancement. *IEEE Transactions on image processing*, 27(1): 379–393.

Bau, D.; Zhu, J.-Y.; Strobelt, H.; Zhou, B.; Tenenbaum, J. B.; Freeman, W. T.; and Torralba, A. 2018. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*.

Chai, S.; Fu, Z.; Huang, Y.; Tu, X.; and Ding, X. 2022. Unsupervised and Untrained Underwater Image Restoration Based on Physical Image Formation Model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2774–2778. IEEE.

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

Fabbri, C.; Islam, M. J.; and Sattar, J. 2018. Enhancing underwater imagery using generative adversarial networks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 7159–7165. IEEE.

Fu, X.; Zhuang, P.; Huang, Y.; Liao, Y.; Zhang, X.-P.; and Ding, X. 2014. A retinex-based enhancing approach for single underwater image. In *2014 IEEE International Conference on Image Processing (ICIP)*, 4572–4576. IEEE.

Fu, Z.; Lin, H.; Yang, Y.; Chai, S.; Sun, L.; Huang, Y.; and Ding, X. 2022a. Unsupervised Underwater Image Restoration: From a Homology Perspective. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 643–651.

Fu, Z.; Wang, W.; Huang, Y.; Ding, X.; and Ma, K.-K. 2022b. Uncertainty Inspired Underwater Image Enhancement. In *European Conference on Computer Vision (ECCV)*, 465–482.

Gandelsman, Y.; Shocher, A.; and Irani, M. 2019. " Double-DIP": unsupervised image decomposition via coupled deep-image-priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11026–11035.

Ghani, A. S. A.; and Isa, N. A. M. 2015. Underwater image quality enhancement through integrated color model with Rayleigh distribution. *Applied soft computing*, 27: 219–230.

Gonzalez-Sabbagh, S.; Robles-Kelly, A.; and Gao, S. 2024. DGD-cGAN: A dual generator for image dewatering and restoration. *Pattern Recognition*, 148: 110159.

Guo, C.; Wu, R.; Jin, X.; Han, L.; Zhang, W.; Chai, Z.; and Li, C. 2023. Underwater ranker: Learn which is better and how to be better. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 702–709.

Guo, H.; Lu, T.; and Wu, Y. 2021. Dynamic low-light image enhancement for object detection via end-to-end training. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 5611–5618. IEEE.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Iqbal, K.; Odetayo, M.; James, A.; Salam, R. A.; and Talib, A. Z. H. 2010. Enhancing the low quality images using unsupervised colour correction method. In *2010 IEEE International Conference on Systems, Man and Cybernetics*, 1703–1709. IEEE.

Islam, M. J.; Luo, P.; and Sattar, J. 2020. Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception. *arXiv preprint arXiv:2002.01155*.

Islam, M. J.; Xia, Y.; and Sattar, J. 2020. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5(2): 3227–3234.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5148–5157.

Kim, H.; Choi, S.-M.; Kim, C.-S.; and Koh, Y. J. 2021. Representative Color Transform for Image Enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4459–4468.

Kimball, P. W.; Clark, E. B.; Scully, M.; Richmond, K.; Flesher, C.; Lindzey, L. E.; Harman, J.; Huffstutler, K.; Lawrence, J.; Lelievre, S.; et al. 2018. The ARTEMIS under-ice AUV docking system. *Journal of field robotics*, 35(2): 299–308.

Land, E. H. 1977. The retinex theory of color vision. *Scientific american*, 237(6): 108–129.

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 Transactions on Image Processing*, 30: 4985–5000.

Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; and Tao, D. 2019. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29: 4376–4389.

Li, C.; Guo, J.; and Guo, C. 2018. Emerging from water: Underwater image color correction based on weakly supervised color transfer. *IEEE Signal processing letters*, 25(3): 323–327.

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 Robotics and Automation letters*, 3(1): 387–394.

Li, Y.; Liu, L.; and Ma, B. 2023. Image rain removal algorithm based on conditional diffusion model. In *2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 68–71. IEEE.

Liu, R.; Fan, X.; Zhu, M.; Hou, M.; and Luo, Z. 2020. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE transactions on circuits and systems for video technology*, 30(12): 4861–4875.

Lu, L.; Xiong, Q.; Xu, B.; and Chu, D. 2024. Mixdehazenet: Mix structure block for image dehazing network. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–10. IEEE.

McGlamery, B. 1980. A computer model for underwater camera systems. In *Ocean Optics VI*, volume 208, 221–231. SPIE.

Nathan, O. B.; Levy, D.; Treibitz, T.; and Rosenbaum, D. 2024. Osmosis: Rgbd diffusion prior for underwater image restoration. In *European Conference on Computer Vision*, 302–319. Springer.

Panetta, K.; Gao, C.; and Agaian, S. 2015. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, 41(3): 541–551.

Peng, L.; Zhu, C.; and Bian, L. 2021. U-shape Transformer for Underwater Image Enhancement. *arXiv preprint arXiv:2111.11843*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Saleh, A.; Sheaves, M.; Jerry, D.; and Azghadi, M. R. 2022. Adaptive uncertainty distribution in deep learning for unsupervised underwater image enhancement. *arXiv preprint arXiv:2212.08983*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tang, Y.; Kawasaki, H.; and Iwaguchi, T. 2023. 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*, 5419–5427.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9446–9454.

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

Wang, H.; Lu, Y.; Chen, Z.; Shen, J.; and Zhang, M. 2022. Underwater object tracking by image enhancement and feature fusion. In *2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*, 448–450. IEEE.

Wang, Y.; Yu, Y.; Yang, W.; Guo, L.; Chau, L.-P.; Kot, A. C.; and Wen, B. 2023. Exposurediffusion: Learning to expose for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12438–12448.

Xie, Y.; Kong, L.; Chen, K.; Zheng, Z.; Yu, X.; Yu, Z.; and Zheng, B. 2024. UVEB: A Large-scale Benchmark and Baseline Towards Real-World Underwater Video Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22358–22367.

Yang, H.-H.; Huang, K.-C.; and Chen, W.-T. 2021. Laffnet: A lightweight adaptive feature fusion network for underwater image enhancement. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 685–692. IEEE.

Yang, M.; Hu, K.; Du, Y.; Wei, Z.; Sheng, Z.; and Hu, J. 2020. Underwater image enhancement based on conditional generative adversarial network. *Signal Processing: Image Communication*, 81: 115723.

Yi, X.; Xu, H.; Zhang, H.; Tang, L.; and Ma, J. 2023. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12302–12311.

Zhang, D.; Wu, C.; Zhou, J.; Zhang, W.; Li, C.; and Lin, Z. 2023. Hierarchical attention aggregation with multi-resolution feature learning for GAN-based underwater image enhancement. *Engineering Applications of Artificial Intelligence*, 125: 106743.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, S.; Zhao, S.; An, D.; Li, D.; and Zhao, R. 2024. LiteEnhanceNet: A lightweight network for real-time single underwater image enhancement. *Expert Systems with Applications*, 240: 122546.

Zhang, W.; Zhuang, P.; Sun, H.; Li, G.; Kwong, S.; and Li, C. 2022. Underwater Image Enhancement via Minimal Color Loss and Locally Adaptive Contrast Enhancement. *IEEE Transactions on Image Processing*.

Zhao, C.; Cai, W.; Dong, C.; and Hu, C. 2024. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8281–8291.

Zhuang, P.; Wu, J.; Porikli, F.; and Li, C. 2022. Underwater image enhancement with hyper-laplacian reflectance priors. *IEEE Transactions on Image Processing*, 31: 5442–5455.