WaterDiffusion: Learning a Prior-involved Unrolling Diffusion for Joint Underwater Saliency Detection and Visual Restoration

Laibin Chang[1], Yunke Wang[2], Longxiang Deng[1], Bo Du[1*], Chang Xu[2*]

[1] School of Computer Science, National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, China.

[2] School of Computer Science, The University of Sydney, Sydney, Australia.

{changlb666, denglx, dubo}@whu.edu.cn, {yunke.wang, c.xu}@sydney.edu.au

Abstract

Underwater salient object detection (USOD) plays a pivotal role in various vision-based marine exploration tasks. However, existing USOD techniques face the dilemma of object mislocalization and imprecise boundaries due to the complex underwater environment. The quality degradation of raw underwater images (caused by selective absorption and medium scattering) makes it challenging to perform instance detection directly. One conceivable approach involves initially removing visual disturbances through underwater image enhancement (UIE), followed by saliency detection. However, this two-stage approach neglects the potential positive impact of the restoration procedure on saliency detection due to it executes in a cascade. Based on this insight, we propose a generalized prior-involved diffusion model, called WaterDiffusion for collaborative underwater saliency detection and visual restoration. Specifically, we first propose a revised self-attention joint diffusion, which embeds dynamic saliency masks into the diffusive network as latent features. By extending the underwater degradation prior into the multi-scale decoder, we innovatively exploit optical transmission maps to aid in localizing underwater salient objects. Then, we further design a gate-guided binary indicator to select either normalized or raw channels for improving feature generalization. Finally, the Half-quadratic Splitting is introduced into the unfolding sampling to refine saliency masks iteratively. Comprehensive experiments demonstrate the superior performance of WaterDiffusion over state-of-the-art methods in both quantitative and qualitative evaluations.

Introduction

Underwater Salient Object Detection (USOD) aims to rapidly identify the visually salient objects within observed underwater scenes, which has received increasing attention due to various vision-based marine exploration requirements (Hong et al. 2023). However, the raw images captured directly by underwater vehicles tend to lose visual saliency, presenting various types of degradation such as color bias, low contrast, and blurred details (Song et al. 2023; Jiang et al. 2024). Degraded images with these defects often lack visual appeal and hinder the localization of salient objects in the forward view.

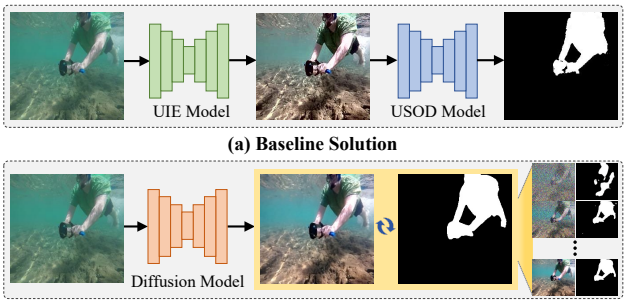


Figure 1: Baseline solution leverages two separate models for UIE and USOD, while our waterDiffusion can jointly accomplish both tasks with a collaborative diffusion.

Benefiting from the available large-scale datasets and advanced deep network designs, many saliency detection methods (Wang et al. 2019; Zhao et al. 2019) related to the terrestrial domains have made remarkable progress. However, USOD mainly faces three limitations: 1) Poor visibility of underwater images interferes with the localization of scene saliency. 2) Underwater environments exhibit specific object categories (*e.g.*, fish, corals, shipwrecks) and water patterns, presenting a unique diversity compared to that on land. 3) Fine-grained annotation leads to existing USOD datasets are relatively limited in scale. Several USOD methods (Islam et al. 2020; Fu et al. 2024; Shen, Zhou, and Liu 2024) have been proposed to stack multiple convolutional sequences with limited-expression ability into a deep network to extract deep feature representations, but the results still leave much room for improvement.

One receivable solution is to first acquire clear images using underwater image enhancement (UIE) techniques and then perform salient object detection, as shown in Figure 1 (a). Many UIE methods have been proposed to perform the pre-processing of degraded underwater images, ranging from physics-based (Chang et al. 2023; Zhou et al. 2023a) to deep learning-based methods (Zhao et al. 2024; Zhang et al. 2023). Despite their success, these methods only consider enhancing the visual perception of degraded images from a global perspective, without highlighting objects that lose saliency caused by underwater light attenuation. In other

*Corresponding author.

words, they ignore the potential beneficial promotion that visual restoration may bring to saliency detection.

Diffusion models have attracted increasing attention due to their impressive performance in image restoration, especially in terms of noise robustness. Some works (Tang, Kawasaki, and Iwaguchi 2023; Lu et al. 2023; Zhao et al. 2024) try to utilize diffusion-based algorithms for enhancing underwater images, achieving more satisfactory results than GAN-based methods which may introduce artifacts due to mode collapse. However, they primarily concentrate on noise removal during the diffusion process, rather than detecting the salient objects distributed within local perception. Up to now, there is no relevant research on underwater saliency detection based on diffusion models. To this end, UIE aims to improve the visual quality of salient objects, and USOD relies on underwater clear images as well, it is beneficial to explore a collaborative approach that integrates saliency detection with visual restoration.

Motivated by the aforementioned analysis, we innovatively propose a generalized prior-involved diffusion model, called WaterDiffusion, which can collaboratively perform underwater saliency detection and visual restoration with a single generative network. As depicted in Figure 2, WaterDiffusion integrates the generative diffusion with underwater degradation prior, and mainly includes three modules: self-attention joint diffusion (SAJD), medium transmission prior (MTP), and gate-guided feature selection (GFS). Specifically, we introduce a dynamic saliency-aware diffusion model that formulates the underwater saliency detection problem as a joint pursuit of visual restoration and refined mask generation. Within this baseline network, the SAJD module cascades preceding intermediate features to ensure high-fidelity content, while the MTP module aims to assist saliency localization by integrating optical transmission map into the decoder. Given the heterogeneity of underwater degraded characteristics, the GFS module designed with gate-guided instance normalization is adopted for selecting generalized features. Finally, we employ an unrolling approach based on the half-quadratic splitting, which refines the mask and estimates the noise via iterative sampling.

Our key contributions are summarized as follows:

- We propose WaterDiffusion, the novel prior-involved diffusion model that joints generative diffusion and underwater transmission prior for collaborative underwater saliency detection and visual restoration.
- We design a gate-guided binary indicator to select either normalized or original channels for improving feature generalization. We propose the deep unrolling-inspired sampling to explicitly integrate the mask and map constraints into the intrinsic iterative of WaterDiffusion.
- Comprehensive experiments on the public datasets validate that our WaterDiffusion surpasses existing state-of-the-art UIE and USOD solutions in both qualitative and quantitative outcomes.

Related Work

Underwater Image Enhancement. UIE is a practical yet challenging technique within the field of visual restoration, which mainly includes the physics-based and deep learning-based methods (Wang et al. 2024). The former methods focused on improving visual perception by directly manipulating image pixel values with well-designed techniques, including multi-scale fusion (Song et al. 2022; Chang et al. 2025), Retinex-based (Fu et al. 2014; Zhuang et al. 2022), and histogram equalization (Huang et al. 2018; Zhou et al. 2023b), *etc*. Deep learning-based UIE methods concentrated on enhancing degraded images by autonomously learning non-linear restoration mappings from paired underwater image datasets. Specifically, many generative adversarial networks (GANs)-based UIE methods have been proposed, such as UGAN (Fabbri, Islam, and Sattar 2018), UWGAN (Guo, Li, and Zhuang 2020), and TwinGAN (Liu et al. 2022). Inspired by the diffusion model, Lu *et al.* (Lu et al. 2023) proposed a UIE method called UW-DDPM, which utilizes two U-Net networks to perform denoising and distribution transforms. Then, Tang *et al.* (Tang, Kawasaki, and Iwaguchi 2023) introduced a transformer-based diffusion network for UIE. Although the images restored by the diffusion model are superior to the GAN-based generative methods, they are only conditioned on degraded images in diffusion, without caring about saliency generation.

Underwater Saliency Detection. Existing feature-based USOD methods (Kumar, Sardana, and Shome 2019; Kanwal, Riaz, and Ghafoor 2024) attempt to encode low-level image features (*e.g.*, color, texture, and contour) into superpixel descriptors and then infer visual saliency by quantifying their global relative sharpness. Jian *et al.* (Jian et al. 2018) proposed a framework that combines quaternionic distance-based Weber descriptor, pattern distinctness, and local contrast to highlight salient objects and suppress background regions. Instead of linearly stacking multiple convolutional layers in the network, several deep learning-based USOD methods (Deng et al. 2023; Hong et al. 2023) have incorporated visual transformers with wider receptive fields into their deep architectures. This way relieves the computational burden imposed by convolution and improves the saliency detection performance to some extent. However, these methods do not consider the underwater degradation prior when designing the encoder architecture.

Methodology

We first demonstrate the overall framework of WaterDiffusion by introducing the dynamic saliency-aware joint diffusion. Then, we integrate the underwater medium transmission map into the network's decoder to assist in saliency localization. Next, we introduce the gate-guided module designed to improve feature generalization. Finally, we outline a deep unrolling optimization based on half-quadratic splitting for iterative sampling.

Dynamic Saliency-Aware Joint Diffusion

Instead of the two-stage approach involving restoration followed by detection, we propose a dynamic saliency-aware diffusion network that redefines saliency detection as a joint task of restoring the clear image and refining the saliency mask. We revise the previous diffusion model (Xia et al.

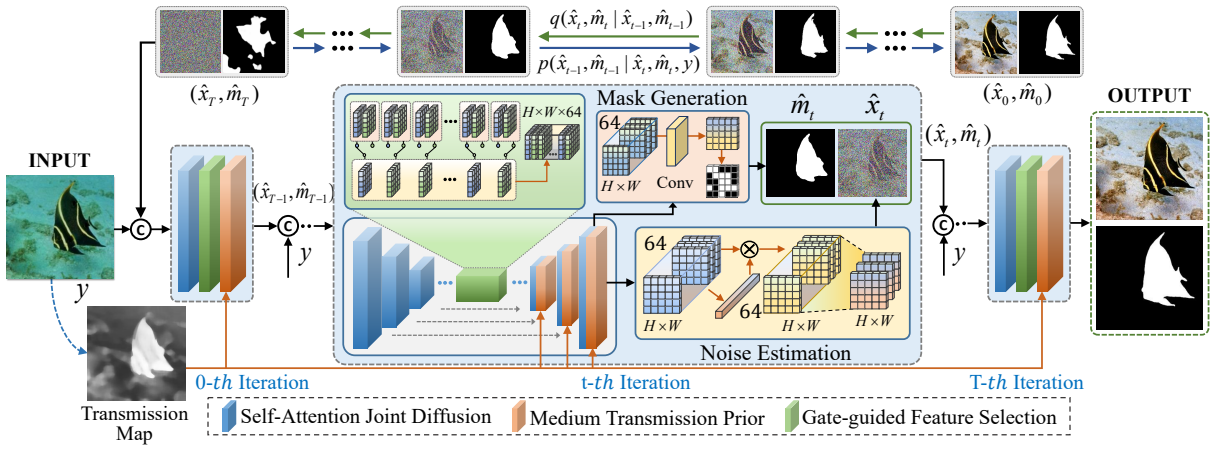


Figure 2: Illustration of the proposed WaterDiffusion, whose baseline framework primarily consists of the self-attention joint diffusion (SAJD) module, medium transmission prior (MTP) module, and gate-guided feature selection (GFS) module. Instead of the single-task generation manner, it utilizes prior-involved unrolling diffusion and iterative sampling to jointly accomplish saliency detection and visual restoration with the raw underwater image as input.

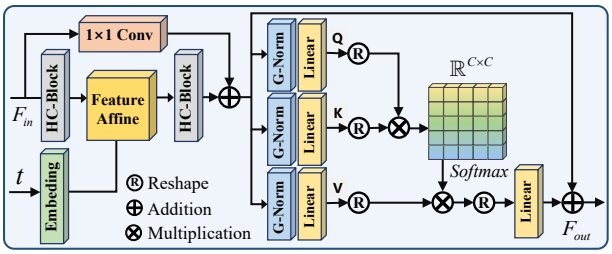


Figure 3: Architecture of the self-attention joint diffusion (SAJD) module. It cascades preceding features F_{in} and time steps t to ensure high-fidelity of the generated content.

2023), which learns a conditional reverse process $p_\theta(\mathrm{x}_{0:T}|\mathrm{y})$ without modifying the forward diffusion $q(\mathrm{x}_{1:T}|\mathrm{x}_0)$, ensuring the sampled x_t is faithful to the data distribution. Unlike them take degraded image y as an invariant condition, we sample $(\mathrm{x}_0, \mathrm{y}, \mathrm{m}_t)$ from a triplet data distribution $q(\mathrm{x}_t, \mathrm{y}, \mathrm{m}_t)$, in which the raw degraded image y and dynamic saliency mask m_t are joined as conditions. Mathematically, the reverse process of learning dynamic mask-aware is expressed as follows:

$$p_\theta(\mathrm{x}_{0:T}|\mathrm{y}, \mathrm{m}_{0:T}) = p(\mathrm{x}_T) \prod_{t=1}^{T} p_\theta(\mathrm{x}_{t-1}|\mathrm{x}_t, \mathrm{y}, \mathrm{m}_t), \quad (1)$$

where $p_\theta(\mathrm{x}_{t-1}|\mathrm{x}_t, \mathrm{y}, \mathrm{m}_t)$ is defined as:

$$p_\theta(\mathrm{x}_{t-1}|\mathrm{x}_t, \mathrm{y}, \mathrm{m}_t) = \mathcal{N}\left(\mathrm{x}_{t-1}; \mu_\theta(\mathrm{x}_t, \mathrm{y}, \mathrm{m}_t, t), \sigma_t^2 \boldsymbol{I}\right), \quad (2)$$

where $\mathcal{N}(\mathrm{x}_t; \mu, \sigma_t^2)$ represents a standard normal distribution with standard deviation $\sigma_t^2 = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$, and $\boldsymbol{I}$ is the identity matrix having the same dimensions as x_t. The mean μ_θ can be written as the following form:

$$\mu_\theta(\mathrm{x}_t, \mathrm{y}, \mathrm{m}_t, t) = \frac{1}{\sqrt{\alpha_t}}(\mathrm{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathrm{x}_t, \mathrm{y}, \mathrm{m}_t, t)), \quad (3)$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and ϵ belongs to the normal distribution $\mathcal{N}(0, \boldsymbol{I})$.

During the diffusion, we design a noise and mask predictor akin to U-Net, which incorporates numerous self-attention joint diffusion (SAJD) modules as the baseline network. As illustrated in Figure 3, the SAJD module cascades embedded time steps t and the preceding features F_{in} to ensure the reliable generation of high-fidelity content. However, for severely degraded underwater images with blurred details, accurately localizing salient objects based on these intermediate features poses a significant challenge.

Underwater Prior Aids Saliency Localization

The transmission map of underwater light characterizes the degradation differences between salient and non-salient regions across the entire image (Hou et al. 2024; Li et al. 2021), as depicted in Figure 2. Inspired by this, we adopt a medium transmission prior (MTP) module, which incorporates the optical transmission map into the decoder network, aiding in localizing underwater salient objects.

The well-known underwater imaging model primarily includes the degraded image $I(x)$ captured by the device and the desired clear image $J(x)$, which is defined as:

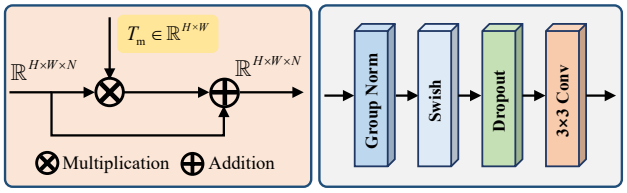
$$I_c(x) = J_c(x)T_{\mathrm{m}}(x) + B_c\left(1 - T_{\mathrm{m}}(x)\right), c \in \{r, g, b\}, \quad (4)$$

where B_c and $T_{\mathrm{m}}(x)$ represent the background light and transmission map, respectively. We employ the dark channel prior-based scheme to calculate the map of $T_{\mathrm{m}}(x)$, expressed as follows:

$$T_m(x) = \max_{y \in \Omega(x)} \left(\frac{B_c - I_c(y)}{\max(B_c, 1 - B_c)} \right), c \in \{r, g, b\}, \quad (5)$$

where $\Omega(x)$ denotes a local filter of size 15 × 15 centered on x, and c denotes the color channel of each pixel.

After that, we utilize the calculated $T_{\mathrm{m}}(x)$ as a feature selector to weight the prominence of different regions, as



(a) MTP Module (b) HC-Block

Figure 4: (a) denotes the medium transmission prior (MTP) module that assists in underwater saliency localization, and (b) is a hierarchical block in the SAJD module.

depicted in Figure 4 (a). In other words, pixels in salient regions are assigned higher weights, expressed as:

$$F_{out}(x) = F_{in}(x) + F_{in}(x) \times T_{\mathrm{m}}(x), \tag{6}$$

where $F_{in}, F_{out} \in \mathbb{R}^{H\times W\times N}$ represent the input and output features after the map weighting, respectively.

Feature Generalization via Gate-Guided Selection

There are two concerns that require attention: (1) Underwater images exhibit obvious inter-domain gaps, showcasing diverse degradation types; (2) The SAJD module cascades multiple variables, and effectively utilizing their features is meaningful. Taking both into account, we employ the instance normalization technique to improve generalization by extracting their invariant representations.

Given an intermediate feature $\mathcal{F}_{in}(x)$ with the shape of $H \times W \times N$, we adopt the IN algorithm (Ulyanov, Vedaldi, and Lempitsky 2016) to normalize it by subtracting the mean value $\mu(\mathcal{F}_{in}(x))$ followed by dividing the standard deviation $\sigma(\mathcal{F}_{in}(x))$, expressed as follows:

$$\mathcal{F}_{nor}(x) = \gamma \frac{\mathcal{F}_{in}(x) - \mu(\mathcal{F}_{in}(x))}{\sigma(\mathcal{F}_{in}(x))} + \rho, \tag{7}$$

where $\gamma, \rho \in \mathbb{R}^C$ are scalable parameters learned from the training data, respectively.

Although normalized features have a robust representation that is not affected by similar degradation, it inevitably leads to the sacrifice of feature information when calculating the mean and variance. To address this issue, we propose a gate-guided feature selection (GFS) module to adaptively select normalized or original features. As depicted in Figure 5, the GFS module initially reduces the given feature $\mathcal{F}_{in}(x)$ to a set of one-dimensional values using Global Average Pooling (GAP), followed by several fully connected layers and ReLU activations to acquire the probability vector $\mathbf{\Gamma} \in 1 \times 1 \times N$. Then, the mean of $\mathbf{\Gamma}$ is calculated and represented as $\mathbf{\Gamma}_{avg}$. For each value $\mathbf{\Gamma}(i)$, if $\mathbf{\Gamma}(i) \geq \mathbf{\Gamma}_{avg}$, it is set to 1, otherwise set to 0, thereby obtaining a series of binary switch indicators $\mathbf{\Theta}$. The obtained $\mathbf{\Theta}$ is used to guide the adaptive selection of normalized or original feature channel, which is expressed as follows:

$$\mathcal{F}_{out}(x) = [(1 - \mathbf{\Theta}) \odot \mathcal{F}_{nor}(x) + \mathbf{\Theta} \odot \mathcal{F}_{in}(x)] \otimes \mathbf{\Gamma}, \tag{8}$$

where $\odot$ and $\otimes$ represent channel-wise and pixel-wise multiplications, respectively.

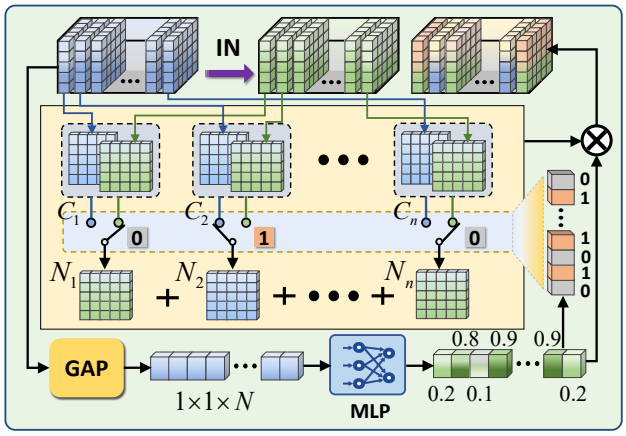


Figure 5: Architecture of the gate-guided feature selection (GFS) module. It improves the model's generalizability by adaptively selecting raw or normalized features.

Unrolling Sampling with Half-Quadratic Splitting

During the joint sampling, it heavily relies on refined mask information and light transmission properties. Conditioned on the estimated map T_{m}, we strive to combine the refined mask m_t with the generated intermediate sample x_t to highlight salient regions. That is, the two tasks of underwater saliency detection and visual restoration promote each other with synergy. Specifically, we first define the mask-guided modification function $\Psi(\mathrm{x}_t)$ as follows:

$$\Psi(\mathrm{x}_t) = T_{\mathrm{m}} \times \mathrm{m}_t \times \mathrm{x}_t + (1 - \mathrm{m}_t) \times \mathrm{x}_t, \tag{9}$$

where $\Psi(\mathrm{x}_t)$ represents the weighted sampling version.

After that, we iteratively refine the desired mask $\hat{\mathrm{m}}_t$ together with the clear image x_t in an unrolling iterative manner. It is achieved by minimizing the following energy function with an image-mask regularizer:

$$\begin{aligned}\min_{\mathrm{x},\mathrm{m},\varphi,\phi} \|\Psi(\varphi) - \Psi(\mathrm{x}_0)\|^2 + \alpha \|\phi - \mathrm{m}_{gt}\|^2 + \frac{\beta}{2}\mathcal{R}(\langle \mathrm{x}|\mathrm{m}\rangle) \\ + \frac{\lambda}{2}(\|\mathrm{x} - \varphi\| + \|\mathrm{m} - \phi\|)\end{aligned} \tag{10}$$

where α, β, λ are the weighted parameters, and $\mathcal{R}(\cdot)$ represents the regularizer for jointing image and mask prior. φ and ϕ are two defined auxiliary variables that transform the unrolling sampling into a constrainable problem. By adopting the half-quadratic splitting algorithm (Afonso, Bioucas-Dias, and Figueiredo 2010), the optimization of Eq. (10) is addressed by calculating the following sub-problems:

$$\begin{aligned}\varphi_{t-1} &= arg\min_{\varphi_t} \|\Psi(\varphi_t) - \Psi(\mathrm{x}_0)\|^2 + \frac{\lambda}{2}\|\mathrm{x}_{t-1} - \varphi_t\|^2 \\ \phi_{t-1} &= arg\min_{\phi_t} \alpha\|\phi_t - \mathrm{m}_{gt}\|^2 + \frac{\lambda}{2}\|\mathrm{m}_{t-1} - \phi_t\|^2 \\ \langle \mathrm{x}|\mathrm{m}\rangle_{t-1} &= arg\min_{\mathrm{x},\mathrm{m}} \frac{\beta}{2}\mathcal{R}(\langle \mathrm{x}|\mathrm{m}\rangle) + \frac{\lambda}{2}\|\langle \mathrm{x}|\mathrm{m}\rangle - \langle \varphi_t|\phi_t\rangle\|^2\end{aligned} \tag{11}$$

Both of them are least-squares problems with quadratic penalty, and we learn from (Schlemper et al. 2017) that updates φ_{t-1} and ϕ_{t-1} from the time T-step to 0-stpe by sharing inputs and reconstructing information between variables.

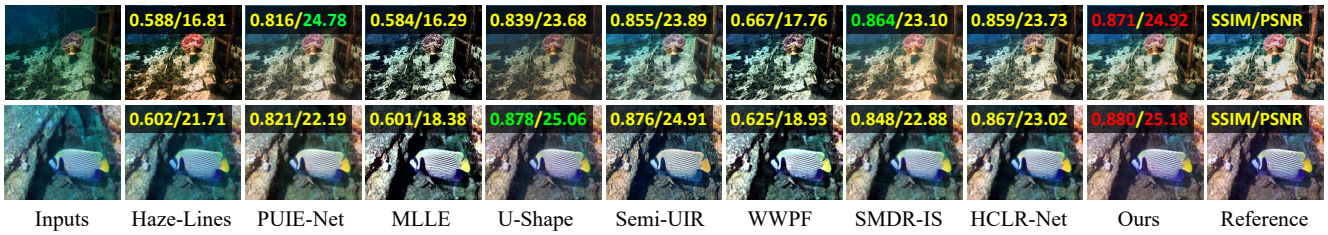


Figure 6: Qualitative results of visual restoration for each method.

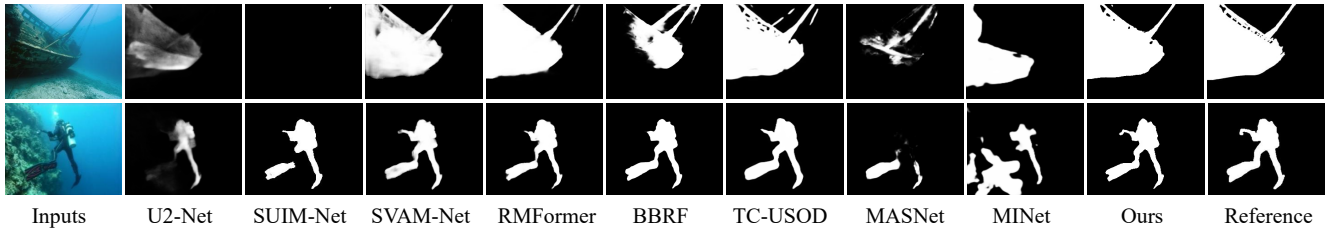


Figure 7: Qualitative results of saliency detection for each method.

While the update of x and m is achieved by reverse a sampling of the WaterDiffusion model, expressed as:

$$\langle \mathrm{x}|\mathrm{m}\rangle_{t-1} = \mathcal{S}_\theta(\varphi_t, \mathrm{y}, \phi_t, t), \tag{12}$$

where $\mathcal{S}_\theta$ represents the sampling procedure.

Loss function. We take the underwater degraded image y, the intermediate variable $(\mathrm{x}_t, \mathrm{m}_t)$, and the time step t as input to predict the noise $\hat{\epsilon}_t$ and the saliency mask $\hat{\mathrm{m}}_t$, expressed as follows:

$$\hat{\epsilon}_t, \hat{\mathrm{m}}_t = \epsilon_\theta((\sqrt{\bar{\alpha}_t}\mathrm{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon, \mathrm{y}, \mathrm{m}_t), t). \tag{13}$$

The training constraint for the conditional diffusion model is to predict the noise vector $\hat{\epsilon}_t$ by optimizing the denoising network parameters. The diffusive objective function is simplified as,

$$\mathcal{L}_{noise} = \mathbb{E}_{\mathrm{x}_0,\epsilon_t,t}\|\hat{\epsilon}_t - \epsilon_t\|^2. \tag{14}$$

To obtain the dynamic refined mask, we incorporate a feature reduction block comprising convolution layers and Sigmoid function after the final self-attention diffusion module. Specifically, we utilize the saliency mask corresponding to the real-world underwater scene as a reference to constrain the rationality of the refined mask, as expressed below:

$$\mathcal{L}_{mask} = \mathbb{E}_{t\sim[1,T]}\|\hat{\mathrm{m}}_t - \mathrm{m}_{gt}\|^2. \tag{15}$$

Based on the noise estimation term $\mathcal{L}_{noise}$ and saliency mask refinement term $\mathcal{L}_{mask}$, the hybrid objective function $\mathcal{L}_{total}$ is defined by combining them as follows:

$$\mathcal{L}_{Total} = \mathcal{L}_{noise} + \zeta\mathcal{L}_{mask}, \tag{16}$$

where $\zeta = 0.1$ is weighted to coordinate the significance of each term in the experiment.

Experiments

Experimental Setups

Implementation details. The proposed WaterDiffusion is trained using the Pytorch framework on two NVIDIA GeForce RTX 4090 GPUs for 5×10^5 iterations. During the training phase, the batch size and patch size are set to 16 and 256×256, respectively. The Adam optimizer comes with an initial learning rate of 1×10^{-4}. The diffusion steps are set to $T = 1000$ with a noise schedule β_t that increases linearly from 0.0001 to 0.02, and the sampling steps are set to $S = 5$ for efficient restoration.

Benchmark datasets. The visual restoration evaluation is conducted on three public UIE datasets (UIEB (Li et al. 2020), UWScene (Islam, Xia, and Sattar 2020), and LSUI (Peng, Zhu, and Bian 2023)), while the saliency detection evaluation is worked on three USOD datasets (UFO-120 (Islam, Luo, and Sattar 2020), SUIM (Islam et al. 2020), and USOD10K (Hong et al. 2023)), all of which are real-world underwater images with reference. We use 3019 images with saliency masks from SUIM and USOD10K for training and the remaining images for testing.

Evaluation metrics. To comprehensively evaluate the compared UIE/USOD methods, we employed popular image evaluation metrics including UIF (Zheng et al. 2022), SSIM, and PSNR for visual restoration evaluation, and S-measure (Fan et al. 2017), E-measure (Fan et al. 2018), F-measure (Achanta et al. 2009), and MAE (Perazzi et al. 2012) for saliency detection evaluation.

Comparison with State-of-the-Arts

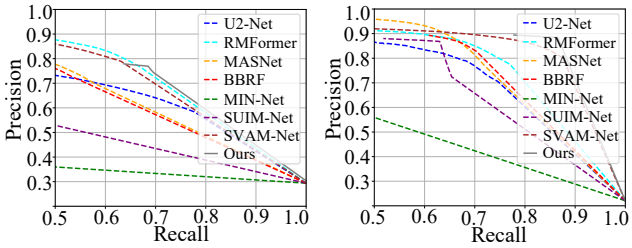
We compare the proposed WaterDiffusion model with state-of-the-art UIE methods, including three physical-based types (*i.e.*, Haze-Lines (Berman et al. 2021), MLLE (Zhang et al. 2022), and WWPF (Zhang et al. 2024b)) and five deep learning-based types (*i.e.*, PUIE-Net (Fu et al. 2022), U-Shape (Peng, Zhu, and Bian 2023), Sem-UIR (Huang et al. 2023), SMDR-IS (Zhang et al. 2024a), and HCLR-Net (Zhou et al. 2024)). For the underwater saliency detection evaluation, we compare the proposed model with eight well-known USOD methods, including U2-Net (Qin et al. 2020),

Config.	*UIEB*			*UWScene*			*LSUI*			Average		
	UIF↑	SSIM↑	PSNR↑	UIF↑	SSIM↑	PSNR↑	UIF↑	SSIM↓	PSNR↓	UIF↑	SSIM↓	PSNR↓
Haze-Lines	0.426	0.607	19.578	0.421	0.623	18.917	0.416	0.571	19.877	0.419	0.591	19.556
PUIE-Net	0.476	0.826	25.453	0.437	0.832	23.778	**0.464**	0.818	24.002	0.457	0.823	24.111
MLLE	0.408	0.658	19.481	0.347	0.584	17.268	0.366	0.631	19.059	0.365	0.620	18.578
U-Shape	0.473	0.831	24.556	0.420	0.839	25.242	0.445	0.848	24.986	0.441	0.843	25.010
Semi-UIR	0.480	0.852	25.463	0.434	0.830	23.106	0.453	0.835	23.943	0.451	0.836	23.878
WWPF	0.449	0.736	20.096	0.361	0.642	18.038	0.391	0.680	19.412	0.389	0.675	19.087
SMDR-IS	**0.486**	0.878	**26.947**	0.441	0.856	23.747	0.457	0.845	23.443	0.456	0.852	23.957
HCLR-Net	0.479	**0.881**	25.872	0.456	0.847	22.647	0.458	**0.858**	23.983	0.460	0.858	23.815
Ours	0.481	0.865	24.938	**0.468**	**0.868**	**25.751**	0.463	0.855	**25.027**	**0.467**	**0.860**	**25.231**

Table 1: Quantitative evaluation of UIE's visual restoration on three public underwater datasets (*UIEB*, *UWScene*, and *LSUI*). The best and second-best results are highlighted with bold and underlined, respectively.

Config.	*UFO-120*				*T-SUIM*				USOD10K			
	$F^{mean}\uparrow$	$E_\xi^{mean}\uparrow$	$S_m\uparrow$	MAE↓	$F^{mean}\uparrow$	$E_\xi^{mean}\uparrow$	$S_m\uparrow$	MAE↓	$F^{mean}\uparrow$	$E_\xi^{mean}\uparrow$	$S_m\uparrow$	MAE↓
U2-Net	0.495	0.593	0.598	0.219	0.657	0.722	0.708	0.130	0.650	0.751	0.753	0.091
SUIM-Net	0.393	0.521	0.520	0.248	0.684	0.758	0.740	0.124	0.506	0.661	0.637	0.130
SVAM-Net	0.640	0.680	0.675	0.180	0.801	0.852	0.827	0.087	0.619	0.735	0.722	0.101
RMFormer	0.615	0.676	0.665	**0.177**	0.743	0.790	0.757	0.118	**0.788**	0.857	**0.829**	0.083
BBRF	0.560	0.645	0.641	0.203	0.740	0.794	0.760	0.111	0.776	0.859	0.823	0.074
TC-USOD	0.632	0.712	0.673	0.185	0.840	0.895	0.844	0.072	0.758	0.853	0.816	0.086
MASNet	0.543	0.638	0.630	0.210	0.710	0.776	0.755	0.112	0.720	0.801	0.781	0.067
MINet	0.299	0.487	0.346	0.381	0.484	0.632	0.526	0.210	0.315	0.573	0.502	0.236
Ours	**0.654**	**0.726**	**0.678**	0.179	**0.862**	**0.915**	**0.860**	**0.056**	0.767	**0.875**	0.827	**0.053**

Table 2: Quantitative evaluation of USOD's saliency detection on three public underwater datasets (*UFO-120*, *T-SUIM*, and *USOD10K*). The best and second-best results are highlighted with bold and underlined, respectively.



(a) UFO-120 Dataset (b) T-SUIM Dataset

Figure 8: Comparison of PR curves for each saliency detection model on *UFO-120* and *T-SUIM* datasets.

SUIM-Net (Islam et al. 2020), SVAM-Net (Islam, Wang, and Sattar 2021), RMFormer (Deng et al. 2023), BBRF (Ma et al. 2023), TC-USOD (Hong et al. 2023), MASNet (Fu et al. 2024), and MINet (Shen, Zhou, and Liu 2024), all of which are deep learning-based methods.

Qualitative evaluation. Figure 6 presents the example results of each UIE method for visual restoration on the benchmark datasets. In the visual comparison, even though most methods have achieved satisfactory enhancement results, our WaterDiffusion demonstrates the closest perception to the reference in terms of color saturation, content similarity, and global contrast. Figure 7 shows the saliency detection results of each method on the public datasets. It can be easily observed that our WaterDiffusion shows superior robustness in object localization and detail segmentation compared to other methods.

Quantitative evaluation. The results of quantitative evaluation regarding visual restoration and saliency detection on several public datasets are presented in Table 1 and Table 2, respectively. Our WaterDiffusion achieves the best or second-best performance in the full-reference metrics. The reason is that the designed diffusion model has strong generation ability, coupled with the GFS and MTP modules that focus on exploiting feature channel attention and underwater transmission prior. As shown in Figure 8, the Precision-Recell (PR) curves of each method on two datasets further validate the superior performance of our WaterDiffusion through the area-under-the-curve (AUC)-based analysis.

Evaluation of Model Efficiency

Inference time. We first compare the inference time of WaterDiffusion with each UIE and USOD method. In Figure 9, the vertical rows with white background indicate the inference time for visual restoration using the UIE methods, while the horizontal rows indicate the time for saliency detection using the USOD methods. We present the summation of any two linear cascades of UIE and USOD into the center region. In Figure 9, the time of 0.196s (ours) is less than the minimum of 0.201s (MLLE and TC-USOD), thus validating that our model requires less inference time to achieve both visual restoration and saliency detection tasks.

Parameters and FLOPs. We supplement the evaluation of WaterDiffusion with deep learning-based UIE and USOD methods in terms of Parameters and FLOPs. As shown in Table 3, we observe that our model achieves relatively fewer

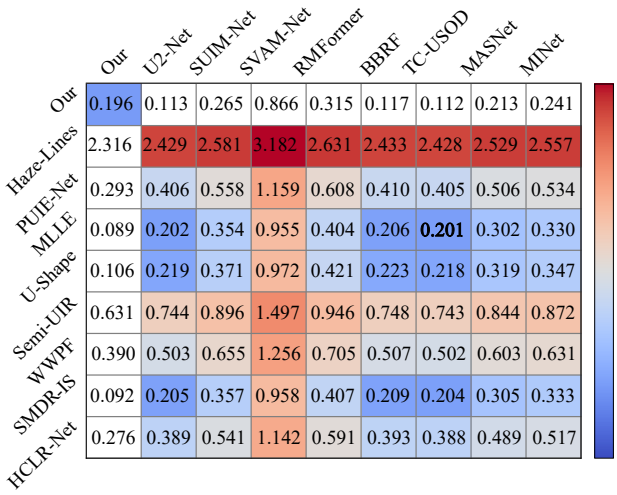


Figure 9: Efficiency evaluation of each compared method in terms of inference time. The center region indicates the summed time of two tasks at the corresponding position.

UIE Methods			USOD Methods		
Config.	Params.	FLOPs	Config.	Params.	FLOPs
Haze-Lines	*Null*	*Null*	U2-Net	44.01	37.65
PUIE-Net	10.69	30.09	SUIM-Net	12.22	71.46
MLLE	*Null*	*Null*	SVAM-Net	29.11	356.27
U-Shape	65.60	66.20	RMFormer	174.19	563.14
Semi-UIR	12.78	36.46	BBRF	74.01	31.13
WWPF	*Null*	*Null*	TC-USOD	117.64	29.64
SMDR-IS	12.26	46.59	MASNet	26.20	17.20
HCLR-Net	4.87	401.97	MINet	47.00	137.00
Ours	55.49	169.56	Ours	55.49	169.56

Table 3: Efficiency evaluation of each compared method in terms of Parameters (M) and FLOPs (G).

parameters and faster FLOPs compared to the partial methods, even though it processes both UIE and USOD tasks.

Ablation Study

The effect of iterative mask refinement. We evaluate the necessity of iterative mask refinement (IMR) in performing both UIE and USOD tasks. Table 4 illustrates the quantitative results with three metrics on the UIEB dataset, showcasing the distinctions with/without mask refinement ("-w/" IMR and "-w/o" IMR) as input. The decreased scores validate the necessity of IMR in the UIE task. However, without the iterative refinement of the saliency mask, the USOD task ceases to function.

The effect of medium transmission prior. Table 5 illustrates the ablation results with and without the medium transmission prior ("-w/" MTP and "-w/o" MTP), encompassing three metrics for quantitative comparison in the UIE task. Table 6 shows the quantitative results of the MTP module for USOD task. Both results confirm that incorporating underwater priors into diffusion is beneficial for two tasks.

The effect of gate-guided feature selection. Figure 10 compares the learning stability with and without gate-guided feature selection ("-w/" GFS and "-w/o" GFS) by recording the SSIM and PSNR metrics during the training procedure. Although the convergence speed is slow and exhibited oscillations within the first 100 epochs after the introduction of GFS module, the final training scores ultimately surpass that of the "-w/o" GFS instance.

Config.	UIF↑	SSIM↑	PSNR↑
"-w/o" IMR	0.442	0.828	22.864
"-w/" IMR	0.481	0.865	24.938

Table 4: Ablation studies of iterative mask refinement (IMR) in terms of visual restoration.

Config.	UIF↑	SSIM↑	PSNR↑
"-w/o" MTP	0.425	0.785	20.338
"-w/" MTP	0.481	0.865	24.938

Table 5: Ablation studies of medium transmission prior (MTP) in terms of visual restoration.

Config.	$F^{mean}\uparrow$	$E_{\xi}^{mean}\uparrow$	$S_m\uparrow$	MAE ↓
"-w/o" MTP	0.610	0.692	0.631	0.203
"-w/" MTP	0.654	0.726	0.678	0.179

Table 6: Ablation studies of medium transmission prior (MTP) in terms of saliency detection.

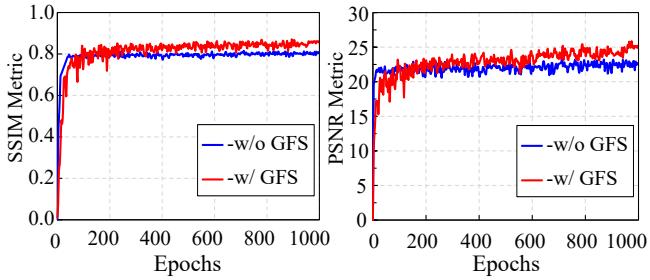


Figure 10: Ablation studies of gate-guided feature selection (GFS) with SSIM and PSNR metrics during the training.

Conclusion

In this paper, we propose a generalized prior-involved diffusion model named WaterDiffusion for joint underwater saliency detection and visual restoration. Firstly, we introduce a dynamic saliency-aware diffusion that embeds the refined saliency mask and underwater image priors into the well-designed baseline network to achieve high-fidelity content. Given the heterogeneity of underwater degraded characteristics, a gate-guided binary indicator is designed to select the normalized or raw channels to improve feature generalization. Then, the half-quadratic splitting is introduced into the sampling to iteratively optimize the generated results. Finally, comprehensive experiments demonstrate the superiority of our WaterDiffusion, which outperforms the state-of-the-art USOD and UIE methods.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62225113, the Innovative Research Group Project of Hubei Province under Grant 2024AFA017, and the Australian Research Council under Project DP210101859.

References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S. 2009. Frequency-tuned salient region detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1597–1604.

Afonso, M. V.; Bioucas-Dias, J. M.; and Figueiredo, M. A. 2010. Fast image recovery using variable splitting and constrained optimization. *IEEE Transactions on Image Processing*, 19(9): 2345–2356.

Berman, D.; Levy, D.; Avidan, S.; and Treibitz, T. 2021. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2822–2837.

Chang, L.; Song, H.; Li, M.; and Xiang, M. 2023. UIDEF: A real-world underwater image dataset and a color-contrast complementary image enhancement framework. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196: 415–428.

Chang, L.; Wang, Y.; Du, B.; and Xu, C. 2025. Rectangling and enhancing underwater stitched image via content-aware warping and perception balancing. *Neural Networks*, 181: 106809.

Deng, X.; Zhang, P.; Liu, W.; and Lu, H. 2023. Recurrent multi-scale transformer for high-resolution salient object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7413–7423.

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*, 7159–7165.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4548–4557.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment measure for binary foreground map evaluation. *arXiv:1805.10421*.

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 *IEEE International Conference on Image Processing*, 4572–4576.

Fu, Z.; Chen, R.; Huang, Y.; Cheng, E.; Ding, X.; and Ma, K.-K. 2024. Masnet: A robust deep marine animal segmentation network. *IEEE Journal of Oceanic Engineering*, 49(3): 1104–1115.

Fu, Z.; Wang, W.; Huang, Y.; Ding, X.; and Ma, K.-K. 2022. Uncertainty inspired underwater image enhancement. In *European Conference on Computer Vision*, 465–482.

Guo, Y.; Li, H.; and Zhuang, P. 2020. Underwater image enhancement using a multiscale dense generative adversarial network. *IEEE Journal of Oceanic Engineering*, 45(3): 862–870.

Hong, L.; Wang, X.; Zhang, G.; and Zhao, M. 2023. USOD10K: a new benchmark dataset for underwater salient object detection. *IEEE Transactions on Image Processing*, 1–1.

Hou, G.; Li, N.; Zhuang, P.; Li, K.; Sun, H.; and Li, C. 2024. Non-uniform illumination underwater image restoration via illumination channel sparsity prior. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2): 799–814.

Huang, D.; Wang, Y.; Song, W.; Sequeira, J.; and Mavromatis, S. 2018. Shallow-water image enhancement using relative global histogram stretching based on adaptive parameter acquisition. In *2018 International Conference on MultiMedia Modeling*, 453–465.

Huang, S.; Wang, K.; Liu, H.; Chen, J.; and Li, Y. 2023. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18145–18155.

Islam, M. J.; Edge, C.; Xiao, Y.; Luo, P.; Mehtaz, M.; Morse, C.; Enan, S. S.; and Sattar, J. 2020. Semantic segmentation of underwater imagery: Dataset and benchmark. In *IEEE International Conference on Intelligent Robots and Systems*, 1769–1776. IEEE.

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

Islam, M. J.; Wang, R.; and Sattar, J. 2021. SVAM: Saliency-guided visual attention modeling by autonomous underwater robots. *arXiv:2011.06252*.

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.

Jian, M.; Qi, Q.; Dong, J.; Yin, Y.; and Lam, K.-M. 2018. Integrating QDWD with pattern distinctness and local contrast for underwater saliency detection. *Journal of Visual Communication and Image Representation*, 53: 31–41.

Jiang, Q.; Kang, Y.; Wang, Z.; Ren, W.; and Li, C. 2024. Perception-driven deep underwater image enhancement without paired supervision. *IEEE Transactions on Multimedia*, 26: 4884–4897.

Kanwal, M.; Riaz, M. M.; and Ghafoor, A. 2024. Unveiling underwater structures: pyramid saliency detection via homomorphic filtering. *Multimedia Tools and Applications*, 1–18.

Kumar, N.; Sardana, H. K.; and Shome, S. 2019. Saliency based shape extraction of objects in unconstrained underwater environment. *Multimedia Tools and Applications*, 78: 15121–15139.

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. 2020. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29: 4376–4389.

Liu, R.; Jiang, Z.; Yang, S.; and Fan, X. 2022. Twin adversarial contrastive learning for underwater image enhancement and beyond. *IEEE Transactions on Image Processing*, 31: 4922–4936.

Lu, S.; Guan, F.; Zhang, H.; and Lai, H. 2023. Underwater Image Enhancement Method Based on Denoising Diffusion Probabilistic Model. *Journal of Visual Communication and Image Representation*, 96: 103926.

Ma, M.; Xia, C.; Xie, C.; Chen, X.; and Li, J. 2023. Boosting broader receptive fields for salient object detection. *IEEE Transactions on Image Processing*, 32: 1026–1038.

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

Perazzi, F.; Krähenbühl, P.; Pritch, Y.; and Hornung, A. 2012. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 733–740.

Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O. R.; and Jagersand, M. 2020. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognition*, 106: 107404.

Schlemper, J.; Caballero, J.; Hajnal, J. V.; Price, A. N.; and Rueckert, D. 2017. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Transactions on Medical Imaging*, 37(2): 491–503.

Shen, K.; Zhou, X.; and Liu, Z. 2024. MINet: Multiscale Interactive Network for Real-Time Salient Object Detection of Strip Steel Surface Defects. *IEEE Transactions on Industrial Informatics*, 20(5): 7842–7852.

Song, H.; Chang, L.; Chen, Z.; and Ren, P. 2022. Enhancement-registration-homogenization (ERH): A comprehensive underwater visual reconstruction paradigm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6953–6967.

Song, H.; Chang, L.; Wang, H.; and Ren, P. 2023. Dual-model: Revised imaging network and visual perception correction for underwater image enhancement. *Engineering Applications of Artificial Intelligence*, 125: 106731.

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. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*.

Wang, H.; Sun, S.; Chang, L.; Li, H.; Zhang, W.; Frery, A. C.; and Ren, P. 2024. INSPIRATION: A reinforcement learning-based human visual perception-driven image enhancement paradigm for underwater scenes. *Engineering Applications of Artificial Intelligence*, 133: 108411.

Wang, W.; Zhao, S.; Shen, J.; Hoi, S. C.; and Borji, A. 2019. Salient object detection with pyramid attention and salient edges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1448–1457.

Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; and Van Gool, L. 2023. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13095–13105.

Zhang, D.; Zhou, J.; Guo, C.; Zhang, W.; and Li, C. 2024a. Synergistic Multiscale Detail Refinement via Intrinsic Supervision for Underwater Image Enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7033–7041.

Zhang, W.; Zhou, L.; Zhuang, P.; Li, G.; Pan, X.; Zhao, W.; and Li, C. 2024b. Underwater Image Enhancement via Weighted Wavelet Visual Perception Fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4): 2469–2483.

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 Transactions on Image Processing*, 31: 3997–4010.

Zhang, Z.; Jiang, Z.; Liu, J.; Fan, X.; and Liu, R. 2023. Waterflow: Heuristic normalizing flow for underwater image enhancement and beyond. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7314–7323.

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.

Zhao, J.-X.; Liu, J.-J.; Fan, D.-P.; Cao, Y.; Yang, J.; and Cheng, M.-M. 2019. EGNet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8779–8788.

Zheng, Y.; Chen, W.; Lin, R.; Zhao, T.; and Le Callet, P. 2022. UIF: An objective quality assessment for underwater image enhancement. *IEEE Transactions on Image Processing*, 31: 5456–5468.

Zhou, J.; Liu, Q.; Jiang, Q.; Ren, W.; Lam, K.-M.; and Zhang, W. 2023a. Underwater camera: Improving visual perception via adaptive dark pixel prior and color correction. *International Journal of Computer Vision*, 1–19.

Zhou, J.; Pang, L.; Zhang, D.; and Zhang, W. 2023b. Underwater image enhancement method via multi-interval subhistogram perspective equalization. *IEEE Journal of Oceanic Engineering*, 48(2): 474–488.

Zhou, J.; Sun, J.; Li, C.; Jiang, Q.; Zhou, M.; Lam, K.-M.; Zhang, W.; and Fu, X. 2024. HCLR-Net: Hybrid Contrastive Learning Regularization with Locally Randomized Perturbation for Underwater Image Enhancement. *International Journal of Computer Vision*, 1–25.

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.