Single-Step Latent Diffusion for Underwater Image Restoration

Jiayi Wu[1,*], Tianfu Wang[1,*], Md Abu Bakr Siddique[2], Md Jahidul Islam[2],
Cornelia Fermuller[1], Yiannis Aloimonos[1], Christopher A. Metzler[1]
[1]University of Maryland [2]University of Florida

Abstract—Underwater image restoration algorithms seek to restore the color, contrast, and appearance of a scene that is imaged underwater. They are a critical tool in applications ranging from marine ecology and aquaculture to underwater construction and archaeology. While existing pixel-domain diffusion-based image restoration approaches are effective at restoring simple scenes with limited depth variation, they are computationally intensive and often generate unrealistic artifacts when applied to scenes with complex geometry and significant depth variation. In this work we overcome these limitations by combining a novel network architecture (SLURPP) with an accurate synthetic data generation pipeline. SLURPP combines pretrained latent diffusion models—which encode strong priors on the geometry and depth of scenes—with an explicit scene decomposition—which allows one to model and account for the effects of light attenuation and backscattering. To train SLURPP we design a physics-based underwater image synthesis pipeline that applies varied and realistic underwater degradation effects to existing terrestrial image datasets. This approach enables the generation of diverse training data with dense medium/degradation annotations. We evaluate our method extensively on both synthetic and real-world benchmarks and demonstrate state-of-the-art performance. Notably, SLURPP is over $200\times$ faster than existing diffusion-based methods while offering $\sim 3dB$ improvement in PSNR on synthetic benchmarks. It also offers compelling qualitative improvements on real-world data. Project website https://tianfwang.github.io/slurpp/.

Index Terms—Computational Imaging, Underwater Restoration, Denoising Diffusion, Foundational Models

1 Introduction

Underwater Image Restoration is a critical task due to the widespread degradation of visual quality in submerged environments caused by light absorption, scattering, and color distortion. These degradations significantly hinder visual perception, making it diffcult for computer vision systems to interpret underwater scenes accurately. Restoring underwater images is essential for a variety of applications, including marine biology research, underwater archaeology, environmental monitoring, autonomous underwater vehicle (AUV) navigation, and underwater robotics. However, underwater image restoration is inherently difficult due to the complex optical properties of water [1], which differ with depth, turbidity, and lighting conditions. Traditional model-based methods rely on physical priors [2], but often struggle with generalization across diverse underwater scenes. Recently, learning-based approaches have shown promising results by leveraging data-driven priors, yet they still face challenges such as a lack of ground truth data, domain shift, and poor interpretability. These limitations highlight the need for more robust and generalizable methods that can adapt to complex degradations of underwater scenes.

Modern text-to-image latent diffusion models [4], [5], trained on massive online datasets [6], have demonstrated remarkable generative capabilities. Crucially, the rich knowledge encoded within extends significantly beyond image synthesis. Recent works have highlighted this by successfully repurposing pretrained latent diffusion models for challenging computer vision tasks, including dense prediction problems such as monocular depth estimation and image intrinsic decomposition [7], [8], [9], [10]. Such successes strongly suggest that these models implicitly capture a sophisticated understanding of scene geometry and intrinsic properties [11], learned inherently from the structure present in their massive training corpora [12].

To this end, underwater image restoration presents a unique challenge, aiming to recover clear visual scene appearances distorted by wavelength-dependent scattering and absorption effects inherent to the water medium. Fundamentally, it requires a joint estimation of both the clear image and the physical medium parameters governing underwater visual degradation. Insights from underwater imaging [1], [13] suggest that these two components—scene content and water medium—can serve as mutual cues. We propose that the rich generative priors encoded in pretrained diffusion models offer a powerful framework to address both aspects of this problem. Specifically, the target clear images often depict natural scenes, aligning well with the distribution of content that such models are trained on. Moreover, backscattering and attenuation effects exhibit strong correlations with scene depth. Notably, recent advances in monocular depth estimation using pretrained latent diffusion [7], [8], [14], [15] reveal that these models inherently capture robust depth priors. This suggests a promising opportunity for modeling depth-dependent water medium parameters, offering a unified approach to underwater image restoration that is both data-efficient and physically grounded.

Building on this insight, we introduce **SLURPP**: **S**ingle-

* *Equal contribution*
Corresponding author: Tianfu Wang tianfuw@umd.edu

Fig. 1. **Real-world underwater restoration using our method.** We develop a single-step underwater restoration method that leverages pretrained latent diffusion priors. Given an underwater input image (top row), our method jointly predicts the clear image (middle row), and the per-pixel underwater medium parameters, specifically the backscattering (bottom row left) and transmission (bottom row right) parameters. In this figure, we present real-world results using images from the UIEB [3] underwater dataset. We show that our method can robustly restore underwater images in a variety of different scenes and water conditions.

step **L**atent **U**nderwater **R**estoration with **P**retrained **P**riors. Our method is a latent-diffusion-based restoration framework that offers a simple yet effective **single-step solution** for underwater image restoration. SLURPP is simple in that it performs direct, physically informed fine-tuning of the underlying latent diffusion model for the task of single-step underwater image restoration, without explicit auxiliary task training, such as dedicated depth prediction. We design a dual-branch architecture to jointly estimate the clear scene and the dense depth-dependent water medium. Crucially, we inject distinct diffusion priors into each branch, tailored to their respective tasks. Our method enables robust and data-efficient restoration across a wide range of underwater conditions (Fig. 1), overcoming challenges posed by the scarcity of real-world underwater datasets and the difficulty of obtaining paired ground truth data.

Our main contributions are summarized as follows:

1) We propose a novel underwater image restoration approach that leverages the foundational visual and geometric priors embedded in pretrained latent diffusion models. Our method jointly estimates the clear scene and water medium properties *in a single step*. By direct fine-tuning a dual-branch architecture tailored for disentangling image content from depth-dependent waterbody effects, our SLURPP method achieves efficient and high-quality restoration across diverse underwater scenes.
2) We develop a physically grounded and computationally efficient underwater image simulation pipeline, built upon the standard underwater image formation model and informed by real-world optical measurements of underwater environments. This pipeline enables the synthesis of high-quality, realistic paired training data by simulating diverse underwater conditions—including varying water types, depths, and lighting—on top of large-scale, easily accessible terrestrial image datasets.
3) By fine-tuning on our simulated dataset, our method effectively adapts the strong generative priors of pretrained latent diffusion models to the specific task of underwater image restoration. In contrast to prior approaches that rely on pixel-space diffusion, our framework operates in a more compact and expressive latent space, enabling fast single-step inference that is over $200\times$ faster than previous diffusion methods, while also supporting the restoration of higher-resolution images with greater visual fidelity. This efficiency, coupled with improved restoration quality, highlights the practical and technical benefits of leveraging latent generative priors for real-world underwater imaging applications.

2 Related Work

2.1 Underwater Image Restoration and Enhancement

Underwater image restoration and enhancement, although closely related, address different aspects of image quality improvement. Enhancement methods improve visual quality by adjusting contrast, color, and brightness without modeling physics [16], using techniques such as histogram equalization, Retinex, and local contrast adjustments [17]. While computationally efficient, these enhancement methods often produce visually appealing but physically implausible results. In contrast, restoration methods aim to recover true scene radiance by modeling underwater light propagation [18], accounting for absorption, scattering, and wavelength-dependent attenuation [13].

Traditional restoration methods rely on handcrafted priors and physical models to estimate and mitigate degradations [2], [3], [19], [20]. Deep learning has significantly advanced underwater image restoration and enhancement by offering adaptive solutions to learn complex mappings from data. [3] established the first CNN-based benchmark. [21] and [22] applied adversarial training for color and detail enhancement. [23] leverages unlabeled data pseudo-labeling and contrastive learning. Transformer approaches such as [24] and [25] introduced histogram and phase-based self-attention mechanisms. [26] uses wavelength-aware networks that enhance restoration using adaptive receptive fields and attentive skip connections. Most relevant to our work, [27] integrates RGBD diffusion priors with a physically-based sampling scheme.

2.2 Diffusion Models for Computer Vision Tasks

Recent large latent diffusion models (LDMs) [4], trained on massive online datasets of text-image pairs [6], can generate diverse and photorealistic images with a text prompt,

inspiring large attention in the computer vision community. The first extensions on LDMs focus on controllable image generation using additional conditions such as depth, inpainting, and segmentation maps [28], [29], [30], [31], [32], [33]. Since then, several works have repurposed LDMs for non-generative tasks, such as monocular depth and normal estimation [7], [8], [15], [34] and image intrinsic decomposition [10]. LDMs have also been used in the image restoration tasks such as deblurring [35], super-resolution [36], and flash removal [9]. On underwater image restoration, we note that Osmosis [27] is the first work to leverage diffusion priors. However, they only use a pixel space RGBD diffusion model trained from ImageNet [37], [38] with limited generation resolution and much less scale and generative capacity compared to current pretrained latent diffusion models [4]. Additionally, [27] uses a DDPM [5] sampling scheme that requires 1000 inference steps, while our method enables single-step inference.

3 Proposed Method

3.1 Preliminaries

3.1.1 Underwater Image Formation

The Jaffe-McGlamery (JM) model [39] serves as a fundamental framework in underwater imaging, providing a mathematical representation of the complex processes of light absorption and scattering in aquatic environments, which significantly influence the visual appearance of submerged objects. Subsequently, [1] introduced a revised underwater image formation model that incorporates variations in attenuation coefficients between direct transmission and backscatter to enhance the accuracy of underwater image correction techniques. A widely adopted formulation of the underwater image formation process is expressed by the following equation:

$$I_c = J_c \cdot e^{-\beta_c^D z} + B_c^{\infty} \cdot (1 - e^{-\beta_c^B z}), \quad (1)$$

where $c \in \{\text{R, G, B}\}$ represents the color channel; I represents the image captured underwater by the camera of a scene at distance z; J denotes the clear scene that would have been captured in the absence of water along the line of sight; and B^{∞} refers to the water color at infinity, commonly referred to as the background light. The two parameters β^D and β^B represent the attenuation and backscatter coefficients, respectively.

3.1.2 Latent Diffusion Model and Diffusion Fine-Tuning

Denoising Diffusion Probabilistic Models (DDPMs) [41], [42] are generative models that learn data distributions by reversing a Markovian process forward process, in which data is gradually corrupted by Gaussian noise over several steps. Early works on diffusion-based image generation are directly trained on RGB pixel space [38], [41], which imposes large computational and memory requirements for training and inference. Latent Diffusion Models (LDMs) such as Stable Diffusion (SD) [4] shift the diffusion process in a low dimensional latent space defined by a variational autoencoder (VAE) [43]. The VAE improves computational efficiency by contracting the image's spatial dimension, while expanding the feature dimension helps encode high-level features and creates a smoother sampling landscape for effective generation. The computational and modeling advantages of the LDMs lead to their wide adoption in image generation. However, LDMs are still slow due to their need for iterative denoising during inference, with work on few-step sampling [41], [44], [45], [46] trading inference time with generation quality.

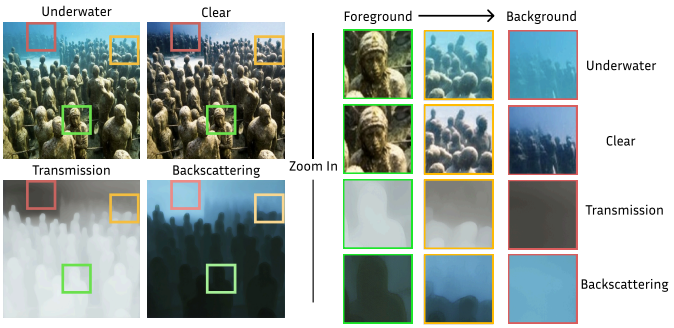

Fig. 2. **Our method captures the depth-varying change of water medium properties.** In this figure we demonstrate the depth-dependent nature of the underwater medium effects and show that our method can correctly capture this in our medium predictions. We can see in the zoomed-in regions of the input underwater image (right first row), that the water medium effects increase as we move from the foreground to the background. This illustrates that the water medium effects are strongly correlated with scene depth. Our model can recover both the clear image (right second row) while capturing the depth-correlated transmission and backscattering effects (right bottom rows). Our model predicts as the scene depth increases, the backscattering becomes stronger while the transmission becomes weaker. This prediction aligns with the observed medium phenomenon in the input underwater image.

Recent work repurpose LDMs for computer vision tasks [7], [8], [9], achieving impressive results. However, these works still model the estimation process as conditional generation based on additional image input. As such, they still need iterative denoising during inference. Inspired by recent theoretical and empirical progress in single-step diffusion [15], [34], [47], we hypothesize that the iterative denoising formulation is less crucial for underwater image restoration, where the distribution of the predicted restored image is narrow and peaks at the ground truth, compared to text-to-image generation, where there is a wide distribution of plausible images for a text prompt. As such, we choose a pipeline that removes stochasticity from the training and inference process and directly predicts restored properties in a single step. We show in our experiment that this single-step formulation does not degrade performance and could even outperform iterative denoising versions of our method, due to the additional advantage of the ability to directly supervise in the image space for single-step training.

3.2 Problem Setting

From the underwater image formation model Eq. (1), two key insights emerge: first, underwater light attenuation and scattering exhibit a strong dependence on both wavelength and propagation distance; second, the occluding backscatter layer, which degrades image clarity, is inherently independent of the scene content. Based on these insights, we formulate our restoration task as the joint estimation of both the restored clear image J, and medium-related parameters, including transmission T and backscattering B, of the input underwater image I. The output of our prediction should fit the underwater image formation model

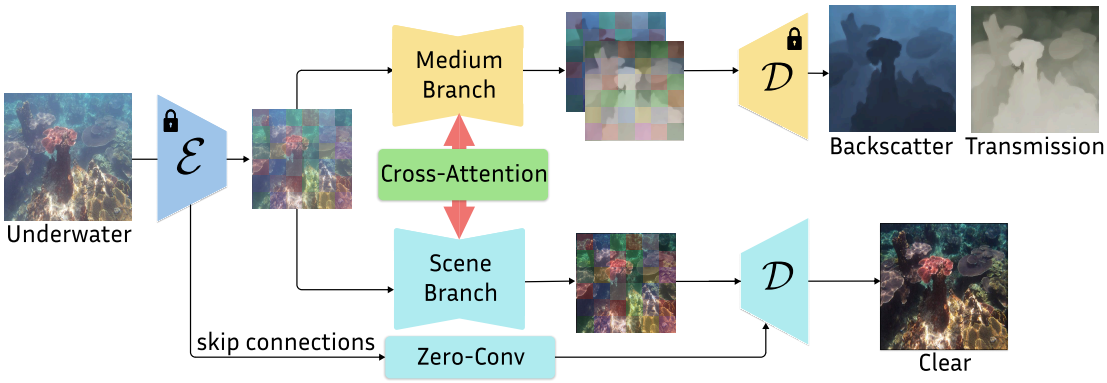

Fig. 3. **Pipeline overview of our single-step dual-branch underwater restoration method.** Our pipeline takes in an underwater image and aims to predict a clear image without water effects, along with the transmission and backscattering properties of the water medium in a single step. The input image is first encoded into latent space using the frozen VAE from pretrained Stable Diffusion (SD) [4]. This latent image is then fed into two UNet [40] branches: the scene branch to predict the clear scene, and the medium branch to predict the wavelength and depth-dependent medium effects. The two branches use different pretrained diffusion priors [4], [7] that fit their respective prediction modalities while exchanging mutual cues through a cross-attention mechanism. The UNets then predict the scene and medium latent images in a single step. To output the predictions, the attenuation and backscattering latent images are decoded using the standard SD decoder, while the clear image is decoded with a cross-latent decoder fine-tuned by incorporating high-frequency details passed from the input image through skip connection layers of the encoder.

specified in Eq. (1), where we now write as our dense scene-medium decomposition formulation:

$$I_c = J_c \cdot T_c + B_c \tag{2}$$

Compared to the imaging model of Eq. (1), we see that for each channel c, we have the relation $T_c = e^{-\beta_c^D z}$ and $B_c = B_c^\infty \cdot (1 - e^{-\beta_c^B z})$, showing that both medium predictions are highly correlated with the scene depth. We represent medium properties using two three-channel images, T and B, rather than separately estimating depth z and water parameters β^D, β^B, and B^∞ as in previous methods, for three main reasons.

First, to incorporate diffusion priors, our pipeline needs to preserve the architecture of pretrained latent diffusion models, which are naturally suited for high-dimensional dense signals such as images. Second, previous methods often assume medium homogeneity (i.e., a spatially uniform β) and even reduce the number of unknown parameters by setting $\beta^D = \beta^B$. Our approach avoids these simplifying assumptions, enabling more robust estimations. Finally, dense medium parameters are depth-related but represented as bounded image intensities, unlike raw depth values with infinite range. This makes them more robust for reconstructing scenes with large depth variations, as medium images are more robust to depth estimation errors in the distant background. We illustrate the effectiveness of our formulation in Fig. 2.

3.3 Our Core Idea

The core idea of our method can be broken down into three key points: foundational model prior, single-step task specific fine-tuning, and physically-accurate training data. Effective underwater restoration requires capturing two aspects: a clear image that resembles natural scenes and the water medium properties correlated with scene depth. With this in mind, we leverage current pretrained latent diffusion models to provide foundational natural image priors for clear image prediction and depth priors for medium prediction. We design a dual-branch architecture (Fig. 3) for joint scene and medium prediction, consisting of a scene branch for content restoration and a medium branch for estimating pixel-level medium parameters. The scene branch is initialized from a pretrained text-to-image diffusion model [4] containing strong priors on natural images, while the medium branch is initialized from a pretrained affine-invariant monocular depth diffusion model [7]. At the training level, our framework and fine-tuning strategy are designed to incorporate the prior knowledge of the underwater image formation model while enabling fast, high-quality single-step inference. We introduce inter-branch cross-attention to exploit the complementary relationship between the clear image and water medium, allowing them to serve as mutual cues during prediction. Additionally, our training objective includes a reconstruction loss that explicitly encourages the outputs to adhere to the dense scene–medium decomposition described in Eq. (2). At the data level, we train our model using physically accurate data, enabling it to learn the underwater image formation process for robust predictions. Addressing the lack of large-scale real paired underwater datasets, we synthesize training data by applying a physically accurate formation model to diverse terrestrial images as clean sources. Our physically accurate underwater image synthesis pipeline uses a carefully optimized medium-related parameter generation strategy to synthesize high-quality training data.

3.4 Physics-based Diverse Underwater Data Synthesis

Real-world underwater datasets are scarce and typically lack ground truth, hindering the development of generalizable image restoration models. 3D simulators, while an alternative, often suffer from high modeling costs, limited scene diversity, and a large domain gap compared to reality. Consequently, synthesizing physically plausible underwater images by applying imaging models to large-scale terrestrial data has become a standard approach in the field.

The underwater image formation model Eq. (1) shows that the degradation caused by the scattering medium is mainly governed by the depth z, the attenuation coefficient

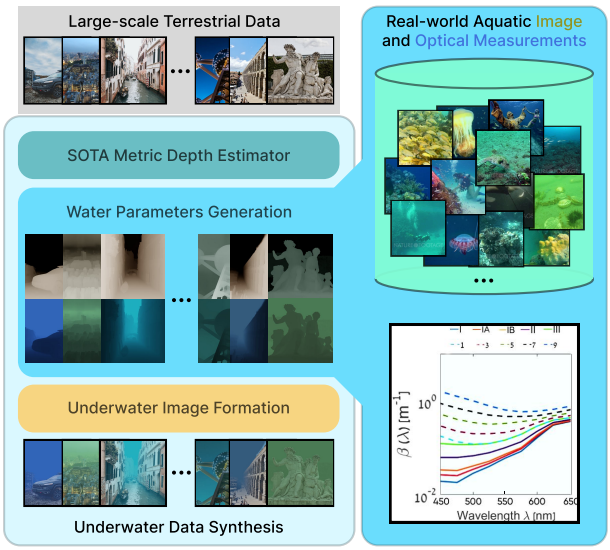

Fig. 4. **Physically accurate underwater image synthetics pipeline for diverse data generation.** Our model is trained on realistic underwater images synthesized from large-scale terrestrial data using precise modeling of depth, attenuation, and background light for physically accurate results. We generate accurate metric depth maps using a state-of-the-art metric depth estimator [48]. We sample attenuation values based on real-world water measurements [1], [49] (bottom-right, reproduced from [1]). We source diverse, realistic background light estimated from real-world underwater images [50] (top-right). These generation strategies enhance the realism and quality of our synthetic training data.

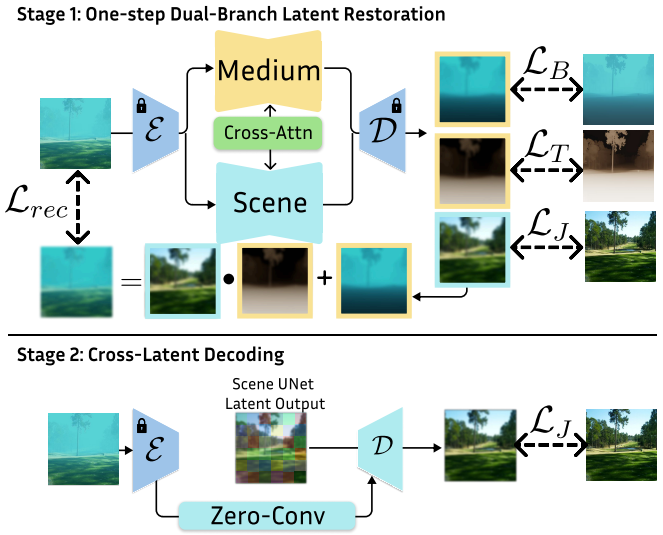

Fig. 5. **Two-stage training procedure of our method.** In our first stage, we train our dual-branch UNets with inter-branch cross-attention to directly predict the latent images of the clear scene J, as well as medium transmission T and backscattering B. The latent outputs are decoded and supervised with their respective ground truths using image losses. We also use the reconstruction loss to guide the predicted outputs to respect the underwater image formation model. For stage 2 cross-latent decoding, we fine-tune the decoder and additional zero convolution skip connections to transfer high-frequency details from the input underwater image to the restored image.

β, and the background light B^{∞}. Accurately and diversely generating these parameters is crucial for realistic synthetic data. Unlike prior methods that approximate the formation model, we optimize each parameter to achieve fine-grained rendering of light scattering and attenuation (Fig. 4).

Obtaining the depth z is challenging as the underwater image formation model Eq. (1) requires the absolute depth in meters. While RGBD datasets [51] provide metric depth, they are often sparse, lack scene diversity, and are costly to acquire. As a result, prior works [27], [50] often rely on monocular depth predictions with manual normalization. However, normalization without camera intrinsics introduces scale errors, limiting the reliability of depth and downstream water medium generation. To address this, we leverage Depth Pro [48], a recent advancement in metric monocular depth that directly predicts focal length from the input image, enabling accurate metric depth estimation. This allows us to generate dense, reliable, and diverse per-pixel metric depth maps from large-scale image collections.

While the values of the attenuation coefficient β and background light B^{∞} are theoretically arbitrary, they are intrinsically constrained on water type. To ensure consistency with real-world aquatic environments, our generation strategy is informed by global-scale underwater optical measurements and extensive real-world data priors. Specifically, we first randomly sample from the 10 Jerlov's water types [49], and then using the sample's measured coefficients at 600nm, 525nm, and 475nm to guide the RGB intensities of β respectively. To avoid unrealistic over or under degradation of the image, we bound the attenuation β, and resample when excessive information loss occurs in the generated image.

We source diverse, realistic background light B^{∞} values by extracting them from real underwater images using ULAP [50], where we swap its monocular depth component with DepthPro [48] for more precise background light estimation. The extracted background lights are then clustered (K=10 K-means in the Lab color space 'ab' channels) into perceptually distinct subsets representing different water types. During synthesis, we first randomly select a light cluster and then randomly sample within it to obtain the background light B^{∞}.

Using our generation strategies for depth z, attenuation β, and background light B^{∞}, and applying the underwater image formation model Eq. (1), we can efficiently generate high-quality paired data required to train our model.

3.5 Pipeline Architecture

Our proposed pipeline (Fig. 3) addresses underwater image restoration by jointly predicting the clear scene image, free from water effects, along with the transmission and backscattering properties characterizing the water medium. Initially, the input underwater image undergoes encoding into the latent space via the frozen pretrained Stable Diffusion (SD) VAE encoder [4], [43]. This latent representation serves as input to a dual-branch architecture comprising two UNets [40] connected with inter-branch cross-attention: a scene branch tasked with predicting the clear scene latent, and a medium branch predicting latent images of depth-dependent attenuation and backscattering. Both branches predict their respective latent images in a single step. Finally, the decoding process differs based on the output type: the attenuation and backscattering latent variables are decoded using the standard SD decoder. In contrast, the

Fig. 6. **Qualitative comparisons of restoration results in USOD10K [52] UIEB [3]**. We show extensive comparisons against previous methods [3], [22], [23], [24], [25], [26], [27], [53], [54]. As illustrated in the comparison, previous methods often struggle to achieve physically consistent restoration across the entire scene, and may even exhibit unnatural changes in water body color. In contrast, our method (second column to the left) achieves physically-consistent restoration across scenes of varying depth, with notably improved performance in severely degraded distant areas. Furthermore, our method accurately estimates per-pixel medium parameters and enables precise and faithful scene restoration across diverse water types and color profiles.

clear scene latent is decoded using the cross-latent decoder, which is fine-tuned to incorporate high-frequency details passed from the original input image via skip connections originating from the VAE encoder.

We now describe the training process illustrated in Fig. 5. We train the single-step restoration and cross-latent decoder in two stages. We note that during inference the frozen encoder can be trivially modified to introduce cross-latent decoding for unified single-step inference.

3.5.1 Stage 1: Single-step Restoration Fine-Tuning

Typical conditional diffusion fine-tuning [7], [9] first converts both the input and the output ground truth images to latent space and injects the output latent image with a random proportion of Gaussian noise. The latent diffusion UNet is then fine-tuned to predict the injected noise, given the input latent and noisy output latent as inputs. This training recipe is tailored for iterative denoising inference, but performs poorly for few-step inference. Additionally, the loss is applied to the noisy latent image that is uninterpretable and cannot leverage structural and perceptual image supervision. In our single-step fine-tuning, the diffusion UNet simply takes in the input latent concatenated with the zero image, which is the mean of the pure Gaussian noise distribution, and learns to directly predict the output latent image in one pass. The output latent image can then be decoded into RGB space where we can compare with the ground truth output using the pixel-space image loss:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{SSIM} + \mathcal{L}_{LPIPS}. \tag{3}$$

We use this training strategy to jointly train the two UNet branches (with cross-attention), with the encoder and decoder both frozen to their pretrained weights. We apply the image loss in Eq. (3) to all output modalities compared to

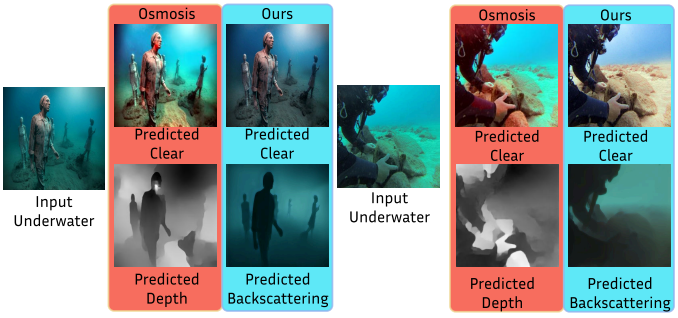

Fig. 7. **Comparison with Osmosis [27] on the UIEB dataset [3].** In this figure we show the predicted clear image and medium-related parameters for our method and Osmosis. In the medium visualization of both methods, objects in the foreground have lower depth/backscattering, while background objects have higher depth/backscattering. Osmosis highly depends on accurate depth estimation, incorrect depth (such as the diver's face region in the right image) leads to unrealistic restoration with spurious color artifacts. Our scene-medium separation formulation leverages depth priors indirectly through water medium prediction, and we obtain much better quality predictions for both clear restoration and depth-dependent medium parameters.

their respective ground truth. Additionally, we combine the predicted images using the dense scene-medium decomposition formulation in Eq. (2), and the apply the image loss in Eq. (3) to the input image as a self-supervised reconstruction loss $\mathcal{L}_{UIFM}$. Our final loss can be written as

$$\mathcal{L}_{total} = \lambda_J \mathcal{L}_J + \lambda_T \mathcal{L}_T + \lambda_B \mathcal{L}_B + \lambda_L \mathcal{L}_{UIFM}. \quad (4)$$

We use $\lambda_J = 1, \lambda_T = \lambda_B = 0.5, \lambda_L = 0.4$.

We note that L_{UIFM} plays a key role in mitigating the domain gap introduced by synthetic training data. While our improved data generation pipeline approximates real-world degradation, it remains limited by the lack of dense optical measurements in real underwater environments, leading to simplifications in light propagation and attenuation modeling. These assumptions introduce discrepancies between synthetic and real images, such as non-uniform media or mismatched scattering and attenuation coefficients. By guiding the model to learn from intrinsic data consistency rather than reply solely on synthetic labels, the reconstruciton loss L_{UIFM} improves robustness against synthetic data biases and enhances generalization to diverse real-world underwater environments.

3.5.2 Stage 2: High-frequency Preservation Decoding

Single-step latent restoration can already effectively restore the clean image. However, due to limitations of the vanilla SD decoder [4], we still observe blurriness and hallucinations in high-frequency details such as text. Following previous diffusion-based restoration methods [9], we use cross-latent decoding with additional zero convolution skip-connections to transfer high-frequency details from the underwater input to the clear image. Once the dual-branch diffusion is trained, we use pairs of underwater image and the latent image output of the scene branch UNet for the second stage corss-latent deocder training, where we fine-tune only the zero convolution and the decoder. This training is supervised using the same image loss in Eq. (3) between the decoded image and the ground truth image.

TABLE 1
Quantitative comparison on USOD10K [52] and UIEB [3] datasets using UIQM [55] and MUSIQ [56] reference-free metrics. Due to the lack of ground truth clear images for real-world underwater datasets, we use reference-free metrics that measure the clear image quality as an assessment to restoration effectiveness. For both datasets our method achieves the best reference-free metric performance.

Method	USOD10K [52]		UIEB [3]	
	UIQM↑	MUSIQ↑	UIQM↑	MUSIQ↑
WaterNet (TIP 2019) [3]	3.093	65.664	3.151	67.927
FUnIE-GAN (RA-L 2020) [22]	3.145	64.069	3.297	64.782
USUIR (AAAI 2022) [53]	3.129	64.955	3.231	67.231
MMLE (TIP 2022) [54]	2.019	67.550	2.229	69.711
Semi-UIR (CVPR 2023) [23]	2.850	66.720	2.961	68.844
DeepWaveNet (TOMM 2023) [26]	2.706	66.092	2.722	68.096
Histoformer (JOE 2024) [24]	3.024	65.278	3.026	67.918
Osmosis (ECCV 2024) [27]	2.821	63.294	2.997	64.089
Phaseformer (WACV 2025) [25]	2.200	66.810	2.319	68.824
Ours	**3.152**	**70.110**	**3.340**	**72.457**

TABLE 2
Qualitative evaluation on synthetic underwater dataset from [27]. We benchmark the clear scene restoration quality on synthetic underwater images curated by [27], which uses ground truth RGBD data from [57] to simulate underwater images. Our method achieves the best performance across all metrics.

	PSNR ↑	SSIM ↑	LPIPS ↓
WaterNet (TIP 2019) [3]	18.04	0.75	0.11
FUnIE-GAN (RA-L 2020) [22]	17.64	0.77	0.21
USUIR (AAAI 2022) [53]	16.76	0.80	0.18
Semi-UIR (CVPR 2023) [23]	17.82	0.83	0.12
MMLE (TIP 2022) [54]	17.00	0.74	0.17
DeepWaveNet (TOMM 2023) [26]	17.14	0.88	0.18
Histoformer (JOE 2024) [24]	16.15	0.82	0.28
Osmosis (ECCV 2024) [27]	22.74	0.89	0.06
Phaseformer (WACV 2025) [24]	16.19	0.78	0.26
Ours (latent loss following [7], [9])	24.62	0.93	0.05
Ours (w/o cross-latent decoder)	24.99	0.92	0.06
Ours (w/o cross-attention)	25.06	0.93	0.05
Ours	**25.66**	**0.95**	**0.05**

4 Experimental Results

4.1 Datasets and Experiment Setups

We use Stable Diffusion V2 (SDV2) [4] diffusion UNet to initialize the scene branch, and Marigold [7] monocular depth model to initialize the medium branch. We initialize the VAE weights from SDV2 pretrained weights [4]. For training data synthesis, we use high quality clean images from various sources, including natural images [58], [59], outdoor [60], indoor [61], night [62] images. We provide more details on training data in the Supplementary. We train the stage 1 single-step fine-tuning and stage 2 cross-latent decoder sequentially on a single NVIDIA A6000 GPU using the same learning rate of 10^{-5} and 512×512 image resolution. Stage 1 training took approximately 2 days and stage 2 took 1 day.

4.2 Real World and Synthetic Comparisons

We performed evaluations on both real-world and synthetic datasets to compare our method with existing approaches. A primary challenge in real-world data evaluation lies in the absence of ground truth clean images in available

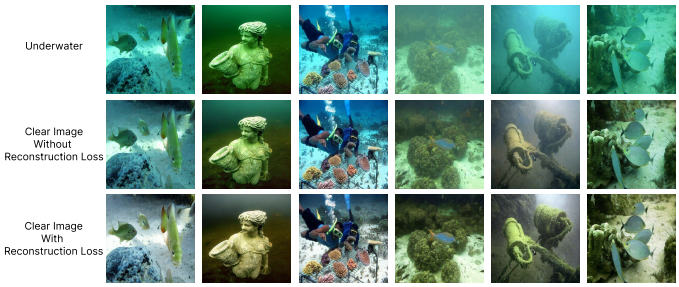

Fig. 8. **Reconstruction training objective improves restoration clarity.** Our single-step fine-tuning approach enables direct end-to-end supervision on images, aligning predicted clear and medium outputs with the dense scene-medium decomposition in Eq. (2). Real-world evaluation on [3] shows that this reconstruction loss leads to clearer restorations(bottom row), compared to outputs of the model that does not use reconstruction loss (middle row).

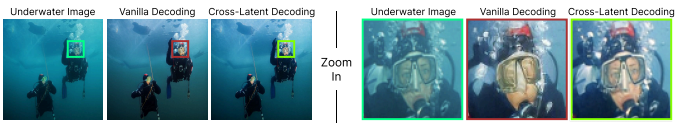

Fig. 9. **Cross-latent decoding enhances restoration details.** Even though our method uses image loss during training, the limitations of the vanilla SD [4] decoder could still hallucinate high-frequency details. The cross-latent decoding allows us to obtain restorations with better details, such as the eyes of the diver in this figure.

underwater datasets. We evaluated the quality of the restored image in two reference-free metrics following previous works: UIQM [55] is tailored for underwater image restoration and measures the colorfulness, sharpness, and contrast of the restored image; MUSIQ [56] is a multi-scale image quality assessment metric with a transformer-based architecture. We evaluated our method on two established real-world datasets in the underwater restoration literature: USOD10K [52] and UIEB [3] datasets. For a fair comparison on reference-free image quality metrics, we filtered out $\sim 10\%$ of images with apparent image artifacts unrelated to underwater effect, such as visible compression and pixelation artifacts. Our method achieves state-of-the-art performance across all metrics on both datasets, demonstrating its effectiveness in restoring degraded underwater images to high-quality clear images. Our extensive qualitative comparisons in Fig. 6 also show that our method achieves a more distinct separation between the underlying scene content and the water medium, whereas other methods often fail to completely remove the effects of water medium, such as backscattering, or estimate them incorrectly. We provide further examples and analysis of real-world underwater images in the Supplementary.

We additionally evaluated our method on synthetic benchmarks with ground truth clear images. In Tab. 2 we compared quantitative image metrics with baseline methods using the simulated underwater dataset in [27], achieving the best result across all evaluated metrics. We note that none of the images in this dataset is used in our training data. These results highlight the high color accuracy and structural fidelity of our predicted clear images.

We conducted an in-depth comparison with Osmosis [27], the previous state-of-the-art diffusion-based underwater reconstruction method. We show qualitative comparisons on real images in [3] in Fig. 7. Osmosis directly predicts depth and during iterative sampling enforces underwater image formation in Eq. (1). However, this method is vulnerable to incorrect depth predictions, which results in spurious color patches and red shifts in the restored images that are unrealistic. We observe that our water-medium predictions correctly capture scene depth relations even more than the direct depth predictions of [27], which also leads to more realistic clear image predictions. In terms of runtime, we ran both methods using the same A6000 GPU. Due to its RGBD diffusion prior and sampling scheme, Osmosis can only restore images up to 256×256 size, while taking more than 200 seconds to generate one image. In contrast, our method can restore up to $2K \times 2K$ images. Our single-step inference restores a 512×512 image in 0.75 seconds, marking a $> 200\times$ improvement over Osmosis. We provide further comparisons with Osmosis in the Supplementary, including an ablation on the training dataset, and quantitative comparison on water medium prediction.

4.3 Ablation Studies

Single Step Prediction and Training. We tailor our framework to train dual-branch diffusion for single-step latent inference. For comparison, we also train an iterative diffusion model following the fine-tuning protocol of [7], [9] where the diffusion UNets are supervised with a latent noise loss. For this model, we also include cross-latent decoder training. Synthetic results in Tab. 2 show that our single-step model performs better than the latent loss model with 50 inference steps. This improvement stems from our ability to directly supervise the RGB output in single-step training, rather than on uninterpretable latents. This includes the use of reconstruction loss that enforces the clear scene and medium output to respect the dense scene-medium decomposition formulation in Eq. (2). We show real-world examples in Fig. 8 that our training objective has better scene-medium separation and produces clearer restorations. We believe that this shows that the reconstruction loss improves our model's generalization to real-world underwater effects.

Cross-Latent Decoder. Even when guided by an image reconstruction loss during training, the standard SD decoder [4] can sometimes introduce hallucinations in detailed regions due to the inherent challenges of reconstructing detail from a compressed latent space. As shown in Fig. 9, the enhanced detail in the diver's eyes illustrates the effectiveness of cross-latent decoder in preserving critical high-frequency information, particularly in regions where the vanilla SD decoder might struggle.

5 Limitations

While our method effectively restores high-quality images and estimates water parameters from single underwater inputs, it has a few limitations. Although faster and more efficient than prior diffusion-based approaches, it still requires a consumer-grade GPU and does not yet achieve real-time performance. Additionally, as it operates on single images, temporal consistency is not enforced, which we demonstrate on the MKV underwater video dataset [63] in the Supplementary.

6 Conclusion

In conclusion, we propose a novel underwater image restoration framework that leverages the foundational natural image and geometric priors embedded in pretrained latent diffusion models. Our approach introduces a fast, single-step restoration pipeline capable of producing detailed and robust predictions of both the clear scene and the intervening water medium. To train our model, we develop a physically grounded underwater image synthesis pipeline that generates realistic and diverse synthetic training data at scale. Comprehensive experiments on both synthetic and real-world benchmarks demonstrate that our method achieves state-of-the-art restoration performance, significantly advancing the quality and efficiency of diffusion-based underwater image restoration.

Acknowledgments

J.W. and Y.A. were supported in part by USDA NIFA sustainable agriculture system program under award no. 20206801231805. T.W. and C.A.M. were supported in part by the UMD AIM Seed Grant Program, NSF CAREER grant no. 2339616, and ONR grant no. N00014-23-1-2752. M.A.S. and M.J.I. were supported in part by the NSF grant no. 2330416.

References

[1] D. Akkaynak and T. Treibitz, "A Revised Underwater Image Formation Model," in *CVPR*, 2018, pp. 6723–6732.

[2] M. Siddique, J. Wu, I. Rekleitis, and M. J. Islam, "AquaFuse: Waterbody Fusion for Physics Guided View Synthesis of Underwater Scenes," *IEEE Robotics and Automation Letters (RA-L)*, vol. 10, no. 5, 2025.

[3] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE transactions on image processing*, vol. 29, pp. 4376–4389, 2019.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[6] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, "LAION-5b: An open large-scale dataset for training next generation image-text models," in *NeurIPS Systems Datasets and Benchmarks Track*, 2022.

[7] B. Ke, K. Qu, T. Wang, N. Metzger, S. Huang, B. Li, A. Obukhov, and K. Schindler, "Marigold: Affordable adaptation of diffusion-based image generators for image analysis," *arXiv preprint arXiv:2505.09358*, 2025.

[8] X. Fu, W. Yin, M. Hu, K. Wang, Y. Ma, P. Tan, S. Shen, D. Lin, and X. Long, "Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image," in *ECCV*, 2024.

[9] T. Wang, M. Xie, H. Cai, S. Shah, and C. A. Metzler, "Flash-split: 2d reflection removal with flash cues and latent diffusion separation," *arXiv preprint arXiv:2501.00637*, 2024.

[10] Z. Chen, T. Xu, W. Ge, L. Wu, D. Yan, J. He, L. Wang, L. Zeng, S. Zhang, and Y.-C. Chen, "Uni-renderer: Unifying rendering and inverse rendering via dual stream diffusion," in *CVPR*, 2025.

[11] J. Wu, X. Lin, B. He, C. Fermuller, and Y. Aloimonos, "Viewactive: Active viewpoint optimization from a single image," *arXiv preprint arXiv:2409.09997*, 2024.

[12] T. Xiong, J. Wu, B. He, C. Fermuller, Y. Aloimonos, H. Huang, and C. Metzler, "Event3dgs: Event-based 3d gaussian splatting for high-speed robot egomotion," in *8th Annual Conference on Robot Learning*, 2024.

[13] D. Akkaynak and T. Treibitz, "Sea-Thru: A Method for Removing Water From Underwater Images," in *CVPR*, 2019, pp. 1682–1691.

[14] B. Yu, J. Wu, and M. J. Islam, "UDepth: Fast Monocular Depth Estimation for Visually-guided Underwater Robots," in *ICRA*, 2023.

[15] G. Martin Garcia, K. Abou Zeid, C. Schmidt, D. de Geus, A. Hermans, and B. Leibe, "Fine-tuning image-conditional diffusion models is easier than you think," in *WACV*, 2025.

[16] M. J. Islam, Y. Xia, and J. Sattar, "Fast Underwater Image Enhancement for Improved Visual Perception," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 2, pp. 3227–3234, 2020.

[17] M. J. Islam, P. Luo, and J. Sattar, "Simultaneous Enhancement and Super-Resolution of Underwater Imagery for Improved Visual Perception," in *Robotics: Science and Systems (RSS)*, Corvalis, Oregon, USA, July 2020.

[18] J. Wu, X. Lin, S. Negahdaripour, C. Fermüller, and Y. Aloimonos, "Marvis: Motion & geometry aware real and virtual image segmentation," in *IROS*, 2024.

[19] J. Wu, B. Yu, and M. J. Islam, "3d reconstruction of underwater scenes using nonlinear domain projection," in *IEEE Conference on Artificial Intelligence (CAI)*, 2023.

[20] J. Wu, "Low-cost depth estimation and 3d reconstruction in scattering medium," Ph.D. dissertation, University of Florida, 2023.

[21] C. Fabbri, M. J. Islam, and J. Sattar, "Enhancing underwater imagery using generative adversarial networks," in *ICRA*, 2018.

[22] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3227–3234, 2020.

[23] S. Huang, K. Wang, H. Liu, J. Chen, and Y. Li, "Contrastive semi-supervised learning for underwater image restoration via reliable bank," in *CVPR*, 2023.

[24] Y.-T. Peng, Y.-R. Chen, G.-R. Chen, and C.-J. Liao, "Histoformer: Histogram-based transformer for efficient underwater image enhancement," *IEEE Journal of Oceanic Engineering*, 2024.

[25] M. Khan, A. Negi, A. Kulkarni, S. S. Phutke, S. K. Vipparthi, and S. Murala, "Phaseformer: Phase-based attention mechanism for underwater image restoration and beyond," *arXiv preprint arXiv:2412.01456*, 2024.

[26] P. Sharma, I. Bisht, and A. Sur, "Wavelength-based attributed deep neural network for underwater image restoration," *ACM TOMM*, 2023.

[27] O. B. Nathan, D. Levy, T. Treibitz, and D. Rosenbaum, "Osmosis: Rgbd diffusion prior for underwater image restoration," in *ECCV*, 2024.

[28] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *ICCV*, 2023.

[29] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," *arXiv preprint arXiv:2108.01073*, 2021.

[30] T. Wang, M. Kanakis, K. Schindler, L. Van Gool, and A. Obukhov, "Breathing new life into 3d assets with generative repainting," in *BMVC*, 2023.

[31] Y. Jia, L. Hoyer, S. Huang, T. Wang, L. Van Gool, K. Schindler, and A. Obukhov, "Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control," in *ECCV*, 2024.

[32] T. Wang, A. Obukhov, and K. Schindler, "Consistencyˆ 2: Consistent and fast 3d painting with latent consistency models," *arXiv preprint arXiv:2406.11202*, 2024.

[33] H. Cai, T.-W. Huang, S. Gehlot, B. Y. Feng, S. Shah, G.-M. Su, and C. Metzler, "Parametric shadow control for portrait generation in text-to-image diffusion models," *arXiv preprint arXiv:2503.21943*, 2025.

[34] C. Ye, L. Qiu, X. Gu, Q. Zuo, Y. Wu, Z. Dong, L. Bo, Y. Xiu, and X. Han, "Stablenormal: Reducing diffusion variance for stable and sharp normal," *ACM Transactions on Graphics*, 2024.

[35] J. Whang, M. Delbracio, H. Talebi, C. Saharia, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in *CVPR*, 2022.

[36] K. Mei, H. Talebi, M. Ardakani, V. M. Patel, P. Milanfar, and M. Delbracio, "The power of context: How multimodality improves image super-resolution," in *CVPR*, 2025.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[38] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *NeurIPS*, 2021.

[39] J. S. Jaffe, "Computer modeling and the design of optimal underwater imaging systems," *IEEE journal of oceanic engineering*, vol. 15, no. 2, pp. 101–111, 1990.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., 2015.
[41] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *ICLR*, 2021.
[42] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *ICLR*, 2021.
[43] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.
[44] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," *arXiv preprint arXiv:2310.04378*, 2023.
[45] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in *NeurIPS*, 2023.
[46] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *ICLR*, 2022.
[47] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *NeurIPS*, 2022.
[48] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, "Depth pro: Sharp monocular metric depth in less than a second," *arXiv preprint arXiv:2410.02073*, 2024.
[49] M. G. Solonenko and C. D. Mobley, "Inherent optical properties of jerlov water types," *Appl. Opt.*, no. 17, pp. 5392–5401, 2015.
[50] W. Song, Y. Wang, D. Huang, and D. Tjondronegoro, "A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration," in *Advances in Multimedia Information Processing – PCM 2018*, 2018.
[51] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The international journal of robotics research*, 2013.
[52] L. Hong, X. Wang, G. Zhang, and M. Zhao, "Usod10k: a new benchmark dataset for underwater salient object detection," *IEEE transactions on image processing*, 2023.
[53] Z. Fu, H. Lin, Y. Yang, S. Chai, L. Sun, Y. Huang, and X. Ding, "Unsupervised underwater image restoration: From a homology perspective," in *AAAI*, 2022.
[54] W. Zhang, P. Zhuang, H.-H. Sun, G. Li, S. Kwong, and C. Li, "Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement," *IEEE Transactions on Image Processing*, vol. 31, pp. 3997–4010, 2022.
[55] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 541–551, 2015.
[56] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *ICCV*, 2021, pp. 5148–5157.
[57] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.
[58] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *CVPR Workshops*, 2017.
[59] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the association for computational linguistics*, vol. 2, pp. 67–78, 2014.
[60] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017.
[61] J. Zhu, F. Luan, Y. Huo, Z. Lin, Z. Zhong, D. Xi, R. Wang, H. Bao, J. Zheng, and R. Tang, "Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing," in *SIGGRAPH Asia*, 2022.
[62] C. Sakaridis, D. Dai, and L. Van Gool, "Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," *IEEE TPAMI*, 2020.
[63] Q.-T. Truong, T.-A. Vu, T.-S. Ha, J. Lokoč, Y. H. W. Tim, A. Joneja, and S.-K. Yeung, "Marine video kit: A new marine video dataset for content-based analysis and retrieval," in *MultiMedia Modeling*, 2023.

Jiayi Wu is a Ph.D. student at the Perception and Robotics Group of the University of Maryland, College Park. He received his M.Sc. (2023) in ECE from the University of Florida. His research focuses on 3D/4D generation, differentiable rendering, and active vision.

Tianfu Wang is a Ph.D. student at the Intelligent Sensing Lab of the University of Maryland, College Park. Tianfu completed his Master's degree in Computer Science at ETH Zurich, and his Bachelor's degree in Computer Science at Northwestern University. Tianfu is interested in computational imaging, generative models, and differentiable rendering.

Md Abu Bakr Siddique is a Ph.D. student at the dept of ECE, University of Florida. He completed an M.Sc. in ECE from Michigan Technological University (2024), and B.Sc. in EEE from IUT, Bangladesh. Post graduation, Abu worked as a faculty member at IUBAT, Bangladesh. His research interest lies in the intersection of Machine Learning and Robotics. He is exploring 3D reconstruction and mapping of underwater scenes. He is also exploring the visual servoing and image-guided exploration by autonomous mobile robots.

Md Jahidul Islam is an Assistant Professor at the Department of ECE of the University of Florida (UF). He received his Ph.D. (2021) in Robotics from the University of Minnesota. His research focuses on enabling active perception and navigation of autonomous underwater robots. He leads the RoboPI group toward developing next-generation robotics systems for subsea inspection, surveillance, and monitoring; his current projects are funded by multiple projects from the NSF, ONR, and TI.

Cornelia Fermüller is a research scientist at the Institute for Advanced Computer Studies (UMIACS) at the University of Maryland at College Park. She holds a Ph.D. from the Technical University of Vienna, Austria, and an M.S. from the University of Technology, Graz, Austria, both in Applied Mathematics. Her research is in Computer, Human, and Robot Vision. She studies and develops biologically inspired Computer Vision solutions for systems that interact with their environment. In recent years, her work has focused on interpreting human activities in the context of music education and on motion processing for fast active robots using bio-inspired event-based sensors as input.

Yiannis Aloimonos is Professor of Computational Vision and Intelligence at the Department of Computer Science, University of Maryland, College Park, and the Director of the Computer Vision Laboratory at the Institute for Advanced Computer Studies (UMIACS). He is also affiliated with the Institute for Systems Research and the Neural and Cognitive Science Program. He was born in Sparta, Greece and studied Mathematics in Athens and Computer Science at the University of Rochester, NY (PhD 1990). He is interested in Active Perception and the modeling of vision as an active, dynamic process for real time robotic systems. For the past five years he has been working on bridging signals and symbols, specifically on the relationship of vision to reasoning, action and language. He received the Presidential Young Investigator Award from President G. Bush. He is an IEEE Fellow.

Christopher A. Metzler is an Assistant Professor in the Department of Computer Science at the University of Maryland College Park, where he leads the UMD Intelligent Sensing Laboratory. He is a member of the University of Maryland Institute for Advanced Computer Studies (UMIACS) and has a courtesy appointment in the Electrical and Computer Engineering Department. His research develops new systems and algorithms for solving problems in computational imaging and sensing, machine learning, and wireless communications. His work has received multiple best paper awards; he recently received NSF CAREER, AFOSR Young Investigator Program, and ARO Early Career Program awards; and he was an Intelligence Community Postdoctoral Research Fellow, an NSF Graduate Research Fellow, a DoD NDSEG Fellow, and a NASA Texas Space Grant Consortium Fellow.

Supplementary Material for: "Single-Step Latent Diffusion for Underwater Image Restoration"

Fig. 10. **More restoration results of our method on real-world datasets [3], [52].** We showcase more real-world restoration image results of our method that we find visually appealing.

TABLE 3
Comparing our single-step method to iterative latent loss fine-tuning methods. We present quantitative comparisons on the synthetic dataset of [27] for our single-step diffusion restoration method, and iterative diffusion (50 steps) diffusion method fine-tuned following the latent loss of Marigold [7], [9]. Our results demonstrate that our single-step restoration model outperforms latent-loss-based models using 50 denoising steps, showing that the effectiveness of our training pipeline goes beyond merely accelerating inference speed.

	PSNR ↑	SSIM ↑	LPIPS ↓
Iterative Latent Loss (50 Steps)	23.46	0.89	0.08
Iterative Latent Loss (50 Steps) + CLD	24.62	0.93	0.05
Ours (Single-Step)	**25.66**	**0.95**	**0.05**

7 Comparison with Latent Loss Training Methods

Our single-step pipeline not only provides faster inference speeds, but also enables the use of image losses during training. Previous fine tuning methods for multi-step iterative inference, such as Marigold [7], [9], use a latent space loss, which is less interpretable. To show the advantage of our one-step architecture and training objectice, we present quantitative results for restoration models trained using the standard diffusion latent loss objective with 50 denoising steps used in Marigold [7], [9], with and without the cross-latent decoder (CLD). Our results in Tab. 3 show that our single-step model outperforms multi-step diffusion fine-tuned with latent loss. Finally, we highlight several key limitations regarding few-step diffusion distillation methods [44], [45], [46]. Distillation requires training an iterative diffusion teacher model before distilling it into a single-step model, while our method is trained for single-step prediction from the outset. Moreover, previous work [44], [45], [46] has consistently shown that the output quality of single-step distilled models is upper bounded by their iterative teacher models. In contrast, our single-step model already outperforms the multi-step diffusion.

TABLE 4
Quantitative evaluation on water medium prediction. We report PSNR and MAE for predicted transmission (T) and backscattering (B) against the ground truth on the simulated NYU [57] underwater dataset. Our method achieves higher accuracy for both components compared to Osmosis [27], which suffers from unreliable depth estimation.

	PSNR T. ↑	MAE T. ↓	PSNR B. ↑	MAE B. ↓
Osmosis	13.97	0.207	23.08	0.076
Ours	**24.69**	**0.060**	**32.37**	**0.024**

TABLE 5
Quantitative results for ablation studies in the main paper. We provide further quantitative results on the synthetic dataset of [27] for the ablation studies in Sec. 4.3 of the main paper.

	PSNR ↑	SSIM ↑	LPIPS ↓
Ours without Recon. Loss	24.80	0.93	0.05
Ours without CLD	24.99	0.92	0.06
Ours	**25.66**	**0.95**	**0.05**

8 Accuracy of Medium Prediction

Due to the lack of ground truth data on transmission and backscattering, we follow Osmosis [27] and evaluate our method on the unseen simulated NYU dataset [57]. In Tab. 4, we measure the PSNR and the MAE of the predicted transmission (T) and backscattering (B) compared to the ground truth. We achieve superior medium prediction accuracy for both predictions over Osmosis [27], which struggles due to its unreliable depth prediction similar to Fig. 7 in the main paper.

9 Additional Ablation Quantitative Comparisons

We report quantitative results on the simulated dataset from Osmosis [27], [57] for our ablation studies presented in Sec. 4.3 of the main paper in Tab. 5, specifically the use of reconstruction loss enabled by our single-step training, and the cross-latent decoder (CLD). Reconstruction loss improves the PSNR by 0.86 dB and the cross-latent decoder improves the PSNR by 0.67 dB.

TABLE 6
Ablation study on the effect of model architecture and data. We decouple our physics-based diverse underwater data synthesis pipeline and our single-step restoration network to further study how each component affects the performance of our method. Specifically, we train our method using the terrestrial dataset from Osmosis [27] instead of our terrestrial data, and we use randomized water medium parameters instead of our curated values from real-world measurements. We show quantitative results on the simulated underwater dataset using [57] from [27]. Our results show that using real-world water medium values during training data synthesis boosts the restoration accuracy of our model. On the other hand, even with randomized water parameters and using the same training data as Osmosis [27], our method still outperforms the baseline.

Method	Terrestrial Data	Water Medium Data	PSNR ↑	SSIM ↑	LPIPS ↓
Osmosis	Osmosis Data	-	22.74	0.89	0.06
Ours	Osmosis Data	Random Values	24.87	0.94	0.05
Ours	Osmosis Data	Sample From Real-World Measurements	**25.76**	0.95	0.05
Ours	Our Data	Sample From Real-World Measurements	25.66	**0.95**	**0.05**

10 Disentangling the Impact of Network Design and Training Data

Our work consists of two key components: a novel single-step diffusion underwater restoration network, and a physics-based diverse underwater training data synthesis pipeline. To further investigate the effect for each component on our method's final performance, we conduct an additional ablation study that disentangles data and network. Specifically, we train new versions of our network using the same RGBD terrestrial data that Osmosis [27] used for its RGBD diffusion prior. We also set the water parameters β^D, β^B, and B^∞ to random RGB values, instead of sampling from real-world water measurements. Our results in Tab. 6 show that using real-world water parameters boosts the performance of the trained restoration model, showing the strength to integrate domain-specific water medium knowledge to training. However, even after training with random water medium parameters, our method still outperforms Osmosis and other baselines (see Tab. 2 of the main paper), demonstrating the effectiveness of our single-step diffusion network and the underlying diffusion priors for underwater image restoration.

11 Robustness to Challenging Underwater Scenarios

Our simulation pipeline is based on the underwater image formation model in Eq. (1), which models scattering of light from the object surface without assuming co-location of the illumination source and the sensor. While Eq. (1) does not explicitly account for complex underwater phenomena such as turbidity, our prediction framework extends this formulation by leveraging the more flexible dense scene-medium decomposition in Eq. (2). To evaluate robustness of our formulation and our pipeline, we show real-world examples in Fig. 11 spanning a range of conditions including shallow water under solar illumination, deep water with non co-located light sources, and scenes with noticeable turbidity. Our method demonstrates strong performance across these diverse scenarios, although blurriness may appear in cases of severe turbidity. Incorporating a more explicit modeling of turbidity into our formulation presents a promising avenue for future research.

12 More Real-World Results

In our qualitative comparisons on real-world underwater datasets [3], [52], we observed a consistent trend in the color accuracy of our restored images. We show this effect in Fig. 12 on a wide range of examples. While evaluating samples across a wide range of scenes, our method consistently recovered more faithful color profiles for both foreground and background objects. In contrast, other approaches frequently introduced color distortions, such as unnatural red shifts, overcompensation for underwater effects, or incomplete removal of background light. We attribute this advantage to the strong natural image priors embedded in the pretrained latent diffusion backbone [4], our scene-medium decomposition formulation, as well as our physically informed fine-tuning objectives, which together enable more precise modeling of underwater image degradation and restoration. Finally, we show additional real-world scenes in Fig. 10 and comparison results in Fig. 13.

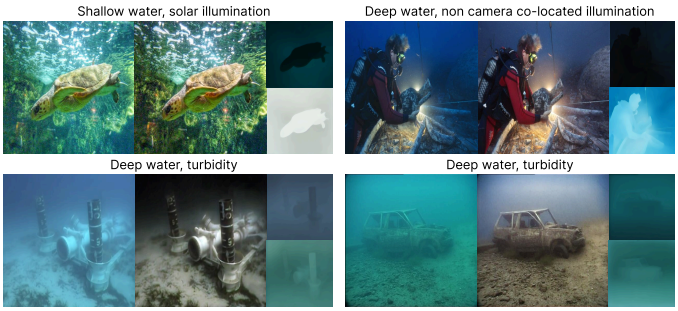

Fig. 11. **Visualizing our model's performance under challenging underwater scenarios.** We show method restoring real world examples from [3], [52] with challenging lighting scenarios and strong turbidity.

13 More Synthetic Training Data Examples

In Fig. 14 we present a diverse set of visualizations illustrating the synthetic underwater training data generated using our physically-accurate data synthesis pipeline detailed in Sec. 3.4 of the main paper. To ensure diversity and realism, we source clean images from a wide range of large-scale terrestrial datasets spanning both indoor and outdoor environments. These include ADE20K [60], an outdoor dataset originally designed for semantic segmentation; DIV2K [58], a high-resolution dataset containing diverse photographic scenes; and the Flickr dataset [59], which comprises a broad collection of crowd-sourced Internet images. We also incorporate Dark Zurich [62], which features urban street scenes captured in low-light, nighttime conditions, and InteriorVerse [61], a dataset focused on richly detailed indoor

environments. Using our synthesis pipeline, we simulate a variety of underwater conditions by varying medium parameters such as depth, attenuation, and background light, resulting in a rich and diverse training dataset that better captures the complexity of real-world underwater imaging scenarios. The visualizations in Fig. 14 also highlight the diversity of underwater medium profiles, showcasing that our data synthesis pipeline is able to capture a wide range of water types and lighting conditions to reflect the variability found in natural underwater environments.

14 Video Results

While our method is designed for single-image restoration and does not explicitly model temporal consistency, we evaluate it on underwater video sequences from the MVK dataset [63] to assess its performance across frames. We refer readers to the static HTML file provided in the supplementary files (supp_video.html) or directly in the "videos" folder to viewing of the video restoration results. The videos demonstrate restoration results in diverse underwater environments, from shallow reefs to deep-sea and wreck scenes. Our method yields stable, view-consistent outputs for foreground objects with minimal flickering, even for moving objects. However, in scenes with large depth variation or low light, flickering artifacts emerge between frames. A notable observation is that the model adapts well to focus changes in deep-sea footage, producing clearer predictions once objects come into focus. The failure case in the wreck video highlights limitations in our current approach, with significant flickering attributed to the absence of temporal modeling, noisy inputs, and ambiguous depth cues from particles in the scene. We believe that the flickering artifacts in our videos are temporal in nature and stems from the fact that our method does not model temporal consistency. Our foreground objects maintain strong view consistency across frames. On a per-frame basis, the restoration quality of the overall image is high with no apparent artifacts. Overall, these videos demonstrate the method's strong performance in restoring underwater visuals from single frames, while also motivating future work to incorporate temporal consistency for video-based applications.

Fig. 12. **Our method restores more accurate colors compared to other methods.** When evaluating restored outputs on real-world underwater datasets [3], [52], we observe that our method more accurately recovers the true color profiles of objects in both the foreground and background. In contrast, other methods often introduce artifacts such as spurious red shifts, overcompensation for medium effects, or incomplete removal of background light. We attribute our improved color fidelity to the strong natural image priors inherent in pretrained latent diffusion models [4], combined with our physically informed fine-tuning objectives.

Fig. 13. **Additional real-world comparisons on real-world underwater datasets [3], [52].** Our method (second column from the left) achieves physically consistent results across varying depths, with improved performance in degraded distant regions. It also accurately estimates per-pixel medium parameters across diverse water types, enabling faithful scene restoration under various underwater conditions.

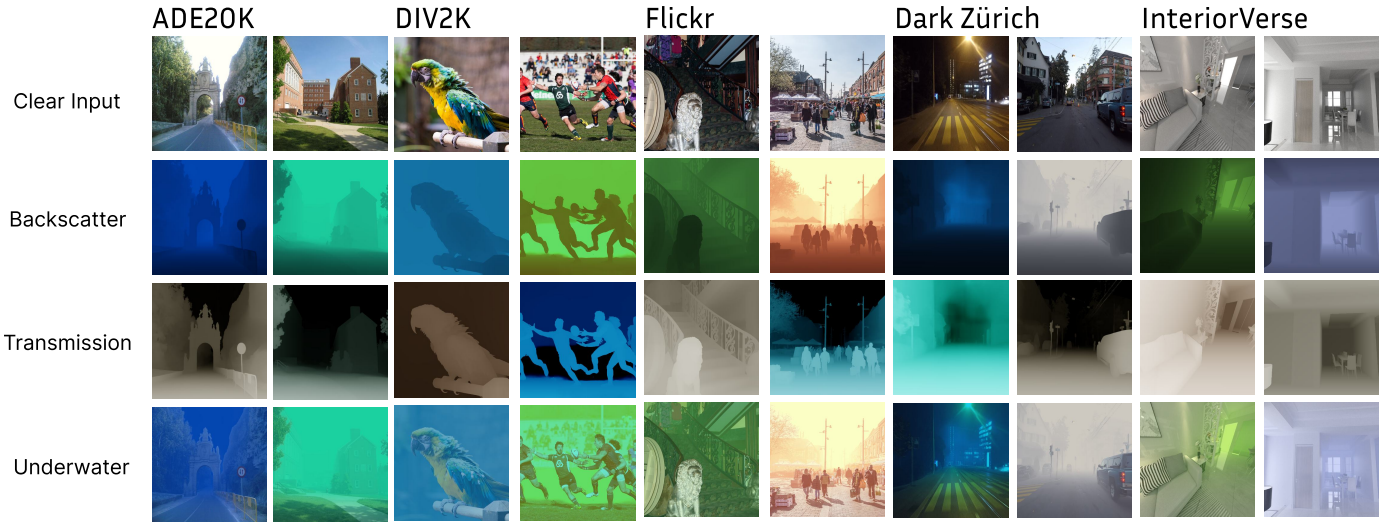

Fig. 14. **More visualizations of synthetic training data.** We show extensive visualizations of training data samples. We use a wide range of terrestrial image data sources as the clean image. Combined with our physically-accurate data synthesis pipeline, we generate diverse and realistic underwater images with various underwater medium profiles. ADE20K [60] is an outdoor dataset originally used for semanitic segmentation. DIV2K [58] is a high-resolution dataset with diverse scene content. Flickr dataset [59] is a large dataset consisting of crowd-sourced Internet images. Dark Zurich [62] is a dataset of city scenes in night environment. InteriorVerse [61] is an indoor dataset.