WaterGen: Decoupling Scene and Medium in Underwater Image Generation

Jiayi Wu[*,1], Tianfu Wang[*,1], Tianyi Xiong[1], Dehao Yuan[1], Xiaomin Lin[2],
Md Jahidul Islam[3], Cornelia Fermuller[1], Christopher Metzler[1], Yiannis Aloimonos[1]
[1]University of Maryland, [2]University of South Florida, [3]University of Florida

Abstract

Underwater computer vision tasks, such as detection, restoration, and segmentation, are limited by the scarcity of large-scale and diverse training data. We introduce WaterGen, a method for generating large-scale realistic and diverse underwater images that provides independent control of the scene and water medium conditions. Our approach treats underwater image generation as the decoupled control of two factors: realistic and diverse scene content (what is in the image), and accurate and controllable water medium effects (what the water does to the image). Existing methods generally achieve only part of this objective: they either provide controllability with limited realism or diversity, or generate realistic scenes without accurately and independently modeling water-medium effects. Our key insight, that allows us to avoid this compromise, is that scene generation and medium modeling can be decoupled within a latent diffusion framework, enabling diverse scene generation together with accurate and controllable underwater appearance. To do this, we decompose underwater image synthesis into two stages. First, we fine-tune the latent diffusion U-Net using degradation-free underwater images so that it learns to generate diverse and realistic latent embeddings of underwater scene content without medium-induced degradation. Second, we formulate the physically accurate medium degradation synthesis as a conditional decoding process applied to these latent embeddings. This decoupled design allows our model to generate diverse scenes with full control of underwater appearance. We leverage WaterGen to build large-scale synthetic underwater datasets that are diverse in scene structures and accurate in water effects and pseudo-labels. We demonstrate that our synthetic data consistently improve downstream performance in underwater restoration and segmentation.

1. Introduction

Exploring underwater environments is crucial for marine biology, archaeological preservation, and offshore industrial inspection. However, the efficacy of computer vision algorithms in these domains—such as object detection, image restoration, and segmentation—is severely bottlenecked by the scarcity of high-quality, diverse labeled data. Unlike terrestrial imaging, underwater photography is plagued by complex physical degradation, including wavelength-dependent attenuation, scattering, and low contrast, making large-scale dataset collection both expensive and logistically challenging. To make matters worse, collecting underwater data at scale is challenging, requiring specialized platforms and equipment (e.g. divers, ROVs/AUVs, calibrated lighting) and operating under limited visibility, harsh conditions, and safety/regulatory constraints. Even when imagery is captured, obtaining reliable ground truth, especially for depth, geometry, and detailed annotations, is costly, slow, and often impractical.

Current approaches to underwater data augmentation and synthesis generally fall into three categories: physics-based rendering, style transfer networks, and image generation models. Physics-based methods use the underwater image formation model (UIFM) to degrade in-air images, but because real underwater scenes rarely have scattering-free ground truth, they are typically built on terrestrial datasets, creating a semantic gap and limiting underwater-native scene diversity. Data-driven style transfer methods (e.g. GAN-based translation), while popular, typically rely on a fixed scene content source domain. Because they are designed to preserve the semantic structure of the input image, they cannot increase the diversity of scene content. Furthermore, most of these methods neglect the depth-dependent nature of underwater optical physics, reducing the simulation to a superficial color style transfer rather than an accurate media transition. Large-scale generative models, particularly diffusion models, demonstrate impressive capability in synthesizing diverse scene content. However, these models lack explicit knowledge of underwater image formation. Consequently, they often hallucinate results with low physical fidelity, exhibiting a significant domain gap from real underwater environments. In addition, text prompts alone do not enable precise control of medium parameters, and the entanglement between scene content and

[*]Equal Contribution

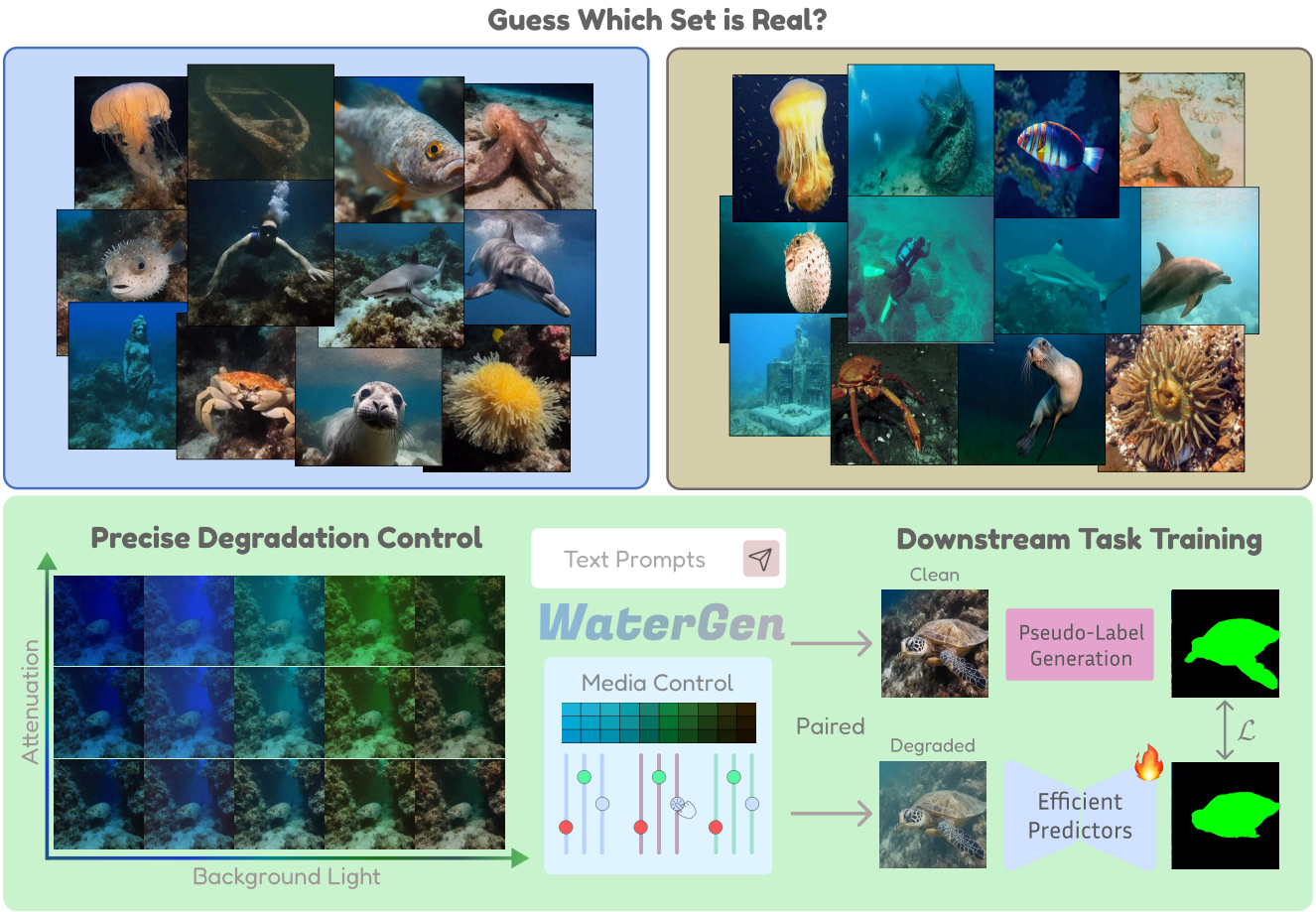

Figure 1. **We introduce WaterGen, an underwater image generation method that enables independent and precise control over medium degradation.** Taking text descriptions and physical water parameters as inputs, WaterGen synthesizes diverse, high-fidelity underwater scenes (**top-left**) with accurate medium degradation, achieving a striking resemblance to real underwater images (**top-right**). Our scene-medium decoupled design allows a single latent scene embedding to be rendered across various water types. This capability positions WaterGen as a powerful synthetic data engine for diverse downstream underwater tasks, where perfectly paired degradation-free scenes drastically improve annotation convenience and quality.

scattering effects prevents independent adjustment of water conditions without altering scene geometry.

To overcome these limitations, we build on pretrained diffusion models and propose WaterGen (Fig. 1), a scene–medium decoupled framework for underwater image generation that bridges the gap between semantic diversity and physical fidelity. Given a natural-language prompt and water parameters (e.g. transmission and backscattering), our method synthesizes high-quality underwater images with realistic, underwater-native structures and physically consistent degradation effects. Our key idea is to decouple underwater image synthesis into two controllable factors: scene content and water-medium effects. Scene content generation governs image content and structure and thus relies on diverse global semantics. By contrast, water-medium effects are local in nature and require accurate modeling of pixel-wise geometry and physical parameters. Based on this view, we decompose generation into two sequential stages. First, **Semantic Scene Latent Diffusion** generates latent representations that capture diverse and plausible underwater scene geometry and layout without imposing medium degradation. Second, **Multi-scale Medium-conditioned Decoding** applies detailed, controllable, and physically grounded attenuation and scattering effects during decoding. This decoupled design lets each stage be optimized with independent datasets and objective-specific losses: we fine-tune the diffusion U-Net on restored clear underwater images to adapt scene structure, while training spatially-aware decoding using physically simulated data to enforce accurate medium control. WaterGen serves as a scalable data engine for producing effectively unlimited amounts of diverse, physically grounded under-

water imagery, helping alleviate the data bottleneck in underwater computer vision.

Overall, our main contributions are summarized as follows:

1. **We propose a novel scene-medium decoupled underwater image generation framework.** By decoupling semantic scene synthesis from media degradation and adopting a tailored isolated training strategy, our method preserves the diversity of modern generative models while substantially improving the realism and controllability of underwater medium effects. To the best of our knowledge, this work is the first to introduce accurate medium control into diffusion-based underwater image generation.
2. **We design a physics-aware decoding mechanism that achieves fine-grained medium control.** Based on the insight that degradation is fundamentally a spatial phenomenon, we inject multi-scale media conditions exclusively into the latent decoding process. This ensures that the generated attenuation and scattering effects adhere to the underwater image formation rule without compromising the semantic integrity of the scene.
3. **We show that medium-controlled underwater image generation directly improves downstream vision tasks.** By generating paired water-free ground truth and physically degraded underwater images at scale, our framework provides effective training data for task-specific learning. Experiments demonstrate clear gains in downstream applications, including underwater restoration and segmentation, validating the utility of accurate and controllable synthetic data generation.

2. Background and Related Work

Underwater Image Synthesis and Style Transfer. Underwater image synthesis aims to simulate or generate realistic underwater images that conform to the optical physics of light propagation in scattering media [12]. Model-based image synthesis [2, 5, 6, 17, 31, 41, 44] simulates image degradation using underwater imaging equations that capture attenuation and backscatter during light propagation in water. Because these methods are interpretable and controllable, they remain widely used for synthetic data generation. Typically, they take a clean image and a depth map as input and then apply degradation under assumed water parameters. Examples include RUIG [6], RSUIGM [5], and AquaFuse [41], all of which extend physical modeling to produce more realistic underwater imagery. However, such approaches depend on prior clean images, offer limited scene diversity, inherit domain bias from terrestrial data, and rely on simplified assumptions about underwater environments.

On the other hand, data-driven methods [25, 45, 50, 51, 54, 55] learn to synthesize underwater imagery from real underwater data examples. FunieGAN [13], WaterGAN [25], and image-to-image translation models such as Pix2Pix [14] and CycleGAN [56] synthesize underwater appearance by transferring color cast and haze styles from underwater images. These methods increase dataset diversity but often capture only the visual appearance rather than the underlying physics, leading to physically inconsistent results. Depth-guided approaches such as UStyle [40] improve realism by linking degradation to depth, but medium effects remain entangled and difficult to control. More fundamentally, because style-transfer methods can only transform existing images, they remain limited in scene diversity.

Latent Diffusion Models for Underwater Image Synthesis Denoising Diffusion Probabilistic Models (DDPMs) [8, 43] generate data by reversing a noise corruption process. To reduce the high computational cost of pixel-space diffusion [7, 38], Latent Diffusion Models (LDMs) such as Stable Diffusion [37] perform diffusion in a compressed latent space encoded by a pretrained Variational Autoencoder (VAE) [21]. This latent formulation enables efficient high-quality image synthesis while preserving semantic structure [4, 16, 18, 46, 47]. Controllable frameworks such as ControlNet [53] and T2I-Adapter [32] further improve generation by incorporating conditions such as depth, edges, and segmentation, enabling scene-level control by combining semantic prompts with structural guidance. In the underwater domain, Atlantis [52] extends Stable Diffusion with a Depth2Underwater ControlNet, while TIDE [29] introduces a unified framework for generating underwater images together with dense annotations such as depth, segmentation, and edge maps. However, these methods remain limited by domain bias and the lack of explicit underwater image formation modeling, often producing stylized *clear water* scenes instead of physically accurate scattering, attenuation, and turbidity, while offering limited control over water optical properties.

3. Method

3.1. Problem Setting

The Jaffe-McGlamery (JM) model [15] provides a foundational description of underwater image formation by modeling light absorption and scattering in water. Subsequently, [1] refined this formulation by distinguishing the attenuation coefficients of direct transmission and backscatter, improving physical accuracy for underwater imaging. A common formulation is:

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

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 d; J is the corresponding clear scene radiance without water; B^∞ is the background light or water

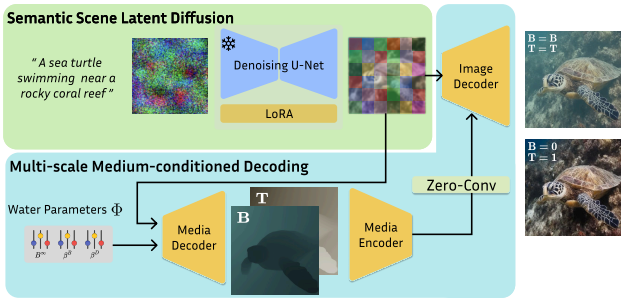

Figure 2. **Overview of the WaterGen pipeline.** Given a text prompt and water parameters Φ, WaterGen decouples scene generation from medium control to synthesize realistic underwater images. A LoRA-adapted denoising U-Net first produces clean scene latents from text. A medium-conditioned decoder then applies attenuation and scattering consistent with Φ. This design natively supports one scene, multiple waters: the same scene latent can be decoded under different water conditions by varying Φ.

color at infinity; and the two parameters β^D and β^B represent the attenuation and backscatter coefficients, respectively.

Drawing from the underwater image formation model in Eq. (1), two key insights emerge. First, wavelength- and depth-dependent attenuation and scattering are mainly determined by a few intrinsic water parameters ($\Phi = \{B^\infty, \beta^D, \beta^B\}$). Second, this media-induced degradation is an extrinsic physical effect on the radiance of the scene, not an intrinsic semantic property of the scene itself. Therefore, we train a latent diffusion model $\mathcal{G}$ that generates underwater images from a text prompt P while independently controlling the medium with physically meaningful parameters Φ, $I_{gen} = \mathcal{G}(P, \ \Phi)$, such that the generated image matches the prompt semantics while exhibiting physically consistent degradation under the specified water conditions.

3.2. Our Core Idea: Decoupling Scene Generation & Media Control

Our core idea is to decouple scene generation from media control by treating underwater imagery as the combination of two components: semantic scene content (what is in the image) and participating-medium effects (what the water does to the image). Existing text-to-image models often entangle these factors into a single learned "underwater style," which makes medium control weak and biased towards dominant training appearances (e.g. clear blue/green water), even when the prompt specifies challenging turbidity or color conditions. A natural attempt is to inject medium conditions into the diffusion backbone using established controllable generation mechanisms such as ControlNet [53], T2I-Adapter [32], or by concatenating condition maps with diffusion noise [18]. These tools are primarily designed for structural guidance (edges, depth, segmentation) and are ill-suited for underwater degradation, which requires radiometrically accurate, spatially varying, pixel-level modulation. In our experiments, direct medium injection led to inconsistent medium control and sometimes reduced generation quality.

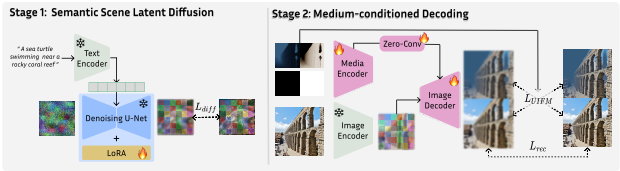

Figure 3. **Training pipeline of WaterGen.** We adopt a two-stage isolated training strategy to decouple scene generation from medium degradation. Stage 1 fine-tunes a latent diffusion backbone with LoRA on restored, water-free underwater images to learn scene geometry and layout. Stage 2 independently trains the decoder on on physically accurate degraded terrestrial data to learn to conditionally inject water medium effects into the clean image latent.

To resolve this mismatch, we adopt a principled separation within the latent diffusion framework: semantic scene synthesis is performed in the latent diffusion backbone, while medium modulation is executed in the decoding module. This design follows the distinct nature of the two tasks. Diffusion is the right place for scene synthesis because it is a semantic generative task requiring diversity, compositional reasoning, global context, and strong prompt alignment. In contrast, medium simulation is best treated as a reconstruction-time radiometric process requiring per-pixel precision and direct supervision from physics. Consequently, we keep the diffusion stage responsible for producing a clean, water-free semantic latent image that preserves the diversity and text conditioning, and we relocate physical injection to the decoder, where spatial fidelity is crucial and physically meaningful maps derived from input water parameters can modulate pixel reconstruction. This decoupled design enables precise, independent control over water conditions, enforces physics-consistent water degradation, and preserves semantic quality by preventing the semantic manifold of the diffusion latent space from being distorted by learning optically degraded domain statistics. Our decoupled framework naturally supports "one scene, many waters": the same semantic latent can be decoded under different water conditions by varying the input parameters. This enables efficient generation of perfectly aligned clean/degraded pairs, providing scalable paired data for downstream tasks while keeping scene content fixed.

3.3. Pipeline Architecture

Our proposed computational pipeline is outlined in Fig. 2; it addresses physics-informed underwater image generation with precise media control by strictly decoupling semantic

scene generation from physical media degradation simulation. Given a text prompt and a set of water parameters, the model first performs **Semantic Scene Latent Diffusion** to generate a clean latent representation of the underlying scene, without underwater effects. This stage is built on an SDXL backbone [35] fine-tuned to model *pristine, water-free* underwater scene content, allowing it to focus on semantic structure and diversity rather than medium degradation. The clean scene latent is then passed to our proposed **Multi-scale Medium-Conditioned Decoder**. Conditioned on the input water parameters Φ and the clean image latent, A media decoder, which contains a standard diffusion decoder and depth estimation module [3], produces the corresponding backscattering and transmission maps, denoted by (B, T). These physically grounded pixel-level media descriptors are subsequently encoded and introduced into the image decoder as explicit spatial conditions for reconstructing the final image. Specifically, based on the definitions in the Underwater Image Formation Model (Eq. (1)), the decoder uses the input water parameters Φ to formulate pixel-level media descriptions—namely the backscattering map and transmission map. These physically-derived maps serve as explicit spatial conditions that modulate the decoding of the clean latent. The final decoded output preserves the semantic structure of the generated scene while accurately matching the specified water conditions.

3.4. Two Stage Isolated Training Process

We now describe the isolated two stage training process illustrated in Fig. 3.

3.4.1. Tailored Data Curation and Synthesis

The isolated training strategy empowers us to employ the most optimal data sources tailored to the different training objectives of the diffusion and decoding components. For fine-tuning the semantic scene diffusion stage, we aggregate a large-scale collection of pristine underwater scene representations by applying the state-of-the-art restoration model, SLURPP [48], to multiple publicly available real-world underwater datasets. The corresponding textual captions are generated using the BLIP [26] vision-language model. Conversely, for the media-conditioned decoding stage, the isolation from the semantic backbone grants us the flexibility to leverage large-scale terrestrial datasets, which offer the "water-free" ground truth to synthesize precise underwater optical degradation. To ensure physical precision, we synthesize diverse and realistic degradation effects by projecting these terrestrial scenes through the underwater image formation model, utilizing water parameters randomly sampled from empirical distributions derived from extensive real-world measurements (following SLURPP [48]). Crucially, this synthetic generation pipeline also facilitates the implementation of the stochastic media noise injection mechanism detailed below.

3.4.2. Stage 1: Clean-Target Diffusion Fine-Tuning

The primary objective of training **Semantic Scene Latent Diffusion** is to adapt the diffusion backbone to generate clear and degradation-free underwater scene semantics conditioned on text prompts, without compromising the generative quality of the original pretrained model. To achieve this, we construct a large-scale training dataset by applying a state-of-the-art underwater image decoupling pipeline to extensive publicly available real-world underwater datasets. This process yields a vast collection of pristine, medium-free underwater scene images, paired with descriptive captions generated via an image captioning pipeline.

During training, we freeze the VAE and original U-Net weights and fine-tune only the LoRA layers inserted into the attention blocks. We then optimize the model with the standard diffusion denoising objective, using the *clean* scene images as training targets. By calculating the gradients solely against the clean scene latents rather than the degraded underwater images, we establish a strong inductive bias. This forces the diffusion model to internalize "water-free" underwater semantics (e.g. coral structures, divers) while implicitly treating any potential residual media artifacts as noise to be removed. This strategy effectively preserves the high-frequency spatial fidelity of the latent space, preventing the latent space degradation typically associated with fine-tuning on optically degraded domain data.

3.4.3. Stage 2: Noise-Injected Medium-Conditioned Decoding

The training of the **Multi-scale Medium-Conditioned Decoder** is isolated from the semantic diffusion backbone. Based on the insight that media simulation is orthogonal to semantic generation, we treat the decoding process as a standalone pixel-space reconstruction task. This isolation gives us the flexibility to utilize large-scale terrestrial datasets—which are free from underwater domain biases—to rigorously train the physical simulation capabilities of the decoder.

We construct a specific training pipeline using high-quality terrestrial images as ground truth scenes J. To ensure physical diversity, we project these scenes through the underwater image formation model using water parameters $\Phi = \{B^\infty, \beta^D, \beta^B\}$ sampled from a database of real-world measurements. This gives us precise training triplets (J, I, Φ) with infinite variability, providing the decoder with exact physical supervision.

To ensure that the decoder is robust to the unavoidable residual media artifacts present in the Stage 1 clean latents, we employ a **Stochastic Media Noise Injection** strategy. During training, we apply random and slight physical degradations to the input images before encoding. This produces "perturbed" latents that mimic the imperfect output of the diffusion model. This mechanism forces the decoder to

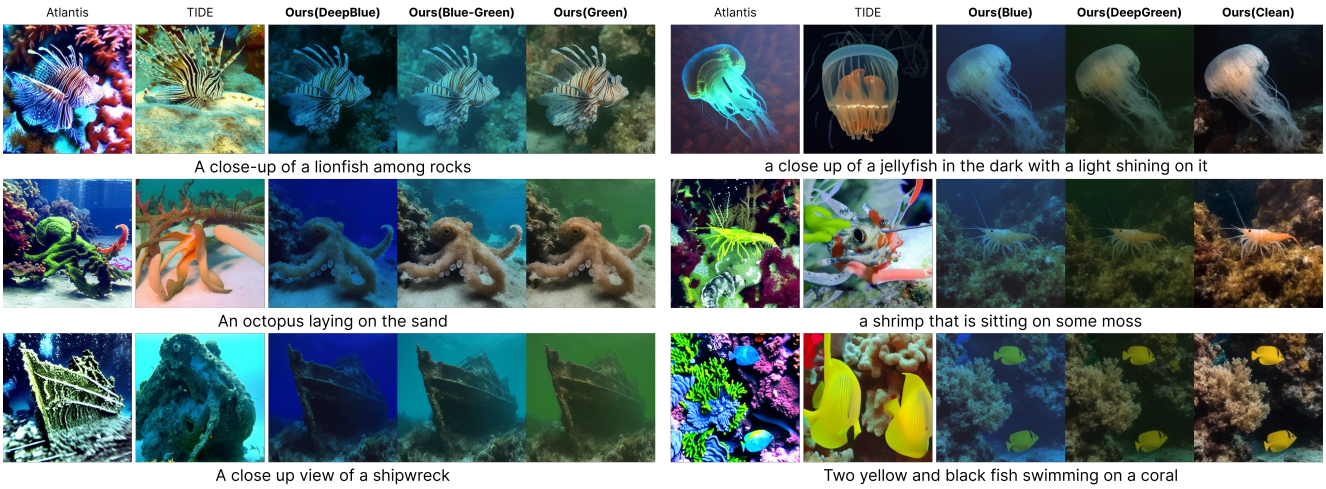

Figure 4. **Qualitative results on synthesis fidelity and diverse water types compared to Atlantis [52] and TIDE [30].** We visualize our pipeline's outputs across six common water types (DeepBlue, Blue, Blue-Green, Green, DeepGreen, and Clean) from the UF7D dataset [40]. Since the baseline methods lack precise control over medium degradation. TIDE is conditioned only on text prompts. For fairness, we provide Atlantis with depth maps obtained from [3] using our generated images.

disregard the implicit noise or "style" in the latent embedding and condition the reconstruction strictly on the injected physical parameters Φ, ensuring that the model learns the underlying physical laws rather than a simple identity mapping.

We apply a dual-forward pass mechanism for each latent z in every training iteration. For the first pass, the medium conditions are set to a non-degraded state ($\Phi = \mathbf{0}$). The decoder must reconstruct the water-free image $\hat{J}$. For the second pass, The medium conditions are set to the sampled degradation parameters Φ_{input}. The decoder must synthesize the corresponding physically degraded image $\hat{I}$. To rigorously enforce the physical validity of the synthesis, we introduce a Bidirectional Physics-Consistency Constraint. We not only minimize the reconstruction error of the forward simulation ($\hat{J} \rightarrow I_{gt}$) but also enforce an inverse consistency constraint ($\hat{I} \rightarrow J_{gt}$).

$$\mathcal{L}_{UIFM} = \left| \left(\hat{J} \cdot T + B \right) - I_{gt} \right|_1 + \left| \frac{\hat{I} - B}{T} - J_{gt} \right|_1 \quad (2)$$

where $T = e^{-\beta_c^D d}$, $B = B_c^\infty \cdot (1 - e^{-\beta_c^B d})$. The total objective is formulated as:

$$\mathcal{L}_{stage2} = \mathcal{L}_{rec}(I) + \mathcal{L}_{rec}(J) + \lambda_{uifm}\mathcal{L}_{UIFM} \quad (3)$$

where the reconstruction term $\mathcal{L}_{rec}$ is defined as a weighted sum: $\mathcal{L}_{rec} = \lambda_1\mathcal{L}_1 + \lambda_{ssim}\mathcal{L}_{SSIM} + \lambda_{lpips}\mathcal{L}_{LPIPS}$. In our implementation, we set $\lambda_1 = 1.0$, $\lambda_{ssim} = 1.0$, $\lambda_{lpips} = 0.5$, and $\lambda_{uifm} = 0.3$.

4. Experiment Results

4.1. Datasets and Experimental Setup

We initialize our latent diffusion model using SDXL (Base) [35], while both the medium encoder and the image encoder-decoder are initialized with the pre-trained SDXL VAE weights. Low-Rank Adaptation (LoRA) [10] is applied to the diffusion U-Net, configuring the adapters with a rank of 32 and an alpha of 16. For Stage 1 training, we introduce WaterGen-Clean, which contains text-captioned, degradation-free underwater images. Raw images sourced from six standard training sets (UIIS10K [24], USIS10K [28], SUIM [11], USOD10K [9], UIEB [22], and UF7D [40]) are restored using the state-of-the-art SLURPP [48] model and captioned via pre-trained BLIP-2 [26] to align with TIDE [30] and Atlantis [52] baselines. In particular, while the SLURPP restorations are not universally devoid of degradation artifacts, the model consistently yields high-quality outputs. For Stage 2 training, we synthesize underwater degradation on high-quality terrestrial images via an Underwater Image Formation Model (Eq. 1). The resulting (J, I, Φ) triplets provide the medium-conditioned decoder with precisely annotated physical medium data. The Stage 1 fine-tuning model and the Stage 2 conditional latent decoder are trained sequentially on a single NVIDIA A6000 GPU. Both stages employ an identical learning rate of 10^{-5}, with each training stage takes approximately 1 day to converge.

4.2. Underwater Image Generation

We quantitatively and qualitatively compare our method with two state-of-the-art underwater image generation mod-

Table 1. Quantitative comparison against Atlantis [52] and TIDE [30] on clear underwater scene generation. Visual quality and text-image alignment are evaluated using UIQM [33]/MUSIQ [19] and CLIP [36] Score, respectively. To isolate scene visual fidelity from underwater degradation, we compare our degradation-free generations ($B = 0, T = 1$) against baseline outputs that have been post-processed by a state-of-the-art underwater image restoration model SLURPP [48]. The numbers presented in the table denote the mean and standard deviation (mean $\pm$ std) calculated across five distinct random seeds.

Method	UIQM ↑	MUSIQ ↑	CLIP Score ↑	Controllability
Atlantis [52]	2.83 ± 0.19	67.54 ± 1.73	0.24 ± 0.02	Text-only
TIDE [30]	2.37 ± 0.38	66.43 ± 2.17	0.23 ± 0.01	Text-only
Ours	$\mathbf{3.02 \pm 0.13}$	$\mathbf{69.26 \pm 0.88}$	$\mathbf{0.26 \pm 0.01}$	**Text + Medium**

els, Atlantis [52] and TIDE [30], utilizing 5,451 underwater scene captions from the SynTIDE [30] dataset. The CLIP Score [36] is used to evaluate the semantic consistency of the generated images. Because synthesized underwater degradation intrinsically lowers standard visual quality scores, we isolate this confounding factor to fairly assess the visual fidelity of the underlying scene generation. Specifically, we compare our degradation-free outputs ($B = 0, T = 1$) with baseline generations post-processed by the state-of-the-art SLURPP restoration model, using two standard reference-free metrics: UIQM [33] and MUSIQ [19]. Specifically, UIQM focuses on evaluating underwater-specific attributes such as colorfulness, sharpness, and contrast, whereas MUSIQ provides a comprehensive, deep-learning-based assessment of overall multi-scale perceptual quality.

As shown in Table 1, compared to the baselines—which suffer from decreased semantic adherence after fine-tuning on heavily degraded underwater images—our WaterGen achieves significantly higher visual fidelity in native clear scene generation while maintaining robust semantic consistency. More importantly, owing to our scene-medium decoupled design, we not only prevent underwater degradation effects from corrupting the latent space during fine-tuning but also enable the injection of pixel-level medium conditions during the decoding of latent embeddings. This facilitates physically accurate, degradation-controlled generation and re-editing.

As illustrated in Fig. 4, in contrast to the baselines, our pipeline can decode a single latent diffusion embedding under varying medium degradation conditions, thereby yielding highly realistic underwater degradations across a diverse array of water types. TIDE [30] can produce multimodal outputs (underwater images, depth maps, and masks), but it often fails to generate coherent scene and object structures. Meanwhile, Atlantis leverages text prompts for semantic control and depth maps for scene structure. Although it preserves the scene structure relatively well, the generated textures exhibit overly vibrant and stylized artifacts. Notably, both methods are trained on degraded underwater images without decoupling water effects, both visual quality and text-image alignment of its outputs are suboptimal, corroborating the quantitative findings in Table 1.

Table 2. Quantitative comparison on UIIS10K [24] dataset using UIQM [33] and MUSIQ [19] reference-free metrics. Higher is better (↑).

Metric	WaterNet [23]	Phaseformer [20]	DeepWaveNet [39]	Histoformer [34]
UIQM Baseline	3.06	2.23	2.68	2.99
UIQM Ours	**3.14**	**2.27**	**2.73**	**3.05**
MUSIQ Baseline	66.86	67.40	66.84	66.17
MUSIQ Ours	**68.32**	**69.16**	**67.93**	**66.92**

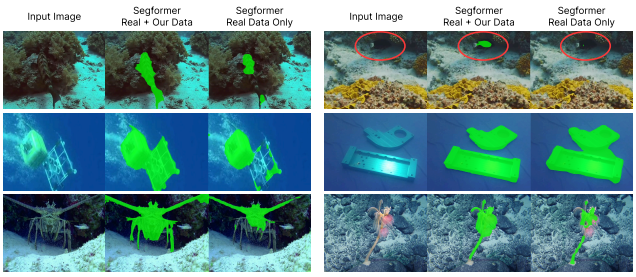

Figure 5. **Qualitative comparison under strong underwater degradation.** From left to right: input image, SegFormer [49] trained on *Real + Our Data*, and SegFormer trained on *Real Data Only*. Adding our data produces cleaner and more complete segmentation masks with sharper boundaries and fewer missed detections, particularly under challenging conditions such as strong turbidity, color shift, backscattering.

4.3. Downstream Applications

4.3.1. Underwater Image Restoration

To validate the practicality of WaterGen beyond data synthesis, we further evaluate its impact on underwater image restoration. Using our scene–medium decoupled framework, we generate over 20,000 paired underwater images by varying physically meaningful water parameters while preserving scene semantics. This synthetic data is then used as additional training augmentation, exposing restoration models to a broader range of attenuation and backscattering conditions than typically available in real datasets. Quantitative evaluation on a real-world benchmark (Table 2) shows that training with our generated data consistently improves restoration performance across multiple representative baselines under UIQM [33] and MUSIQ [19].

4.3.2. Degradation Robust Underwater Image Segmentation

We further evaluate our pipeline for underwater semantic segmentation. We generate $40,000$ paired clean and degraded underwater images using underwater scene prompts

Table 3. Quantitative results of underwater semantic segmentation. We calculate the mean Intersection over Union (mIoU) over six categories (Fish, Reefs, Aquatic Plants, Wreck) following TIDE [30]. Higher is better (↑).

Dataset	Setting	Segformer (MiT-B4)	Mask2Former (Swin-B)	ViT-Adapter (B)
UIIS [27]	Real only	70.2	72.7	73.5
	Real+SynTIDE [30]	75.4(+5.2)	74.3(+1.6)	75.1(+1.6)
	Real+Ours	75.6(+5.4)	74.0(+1.3)	75.3(+1.8)
USIS10K [28]	Real only	74.6	76.1	74.6
	Real+SynTIDE [30]	76.1(+1.5)	77.1(+1.0)	76.7(+2.1)
	Real+Ours	76.7(+2.1)	76.9(+0.9)	76.9(+2.3)

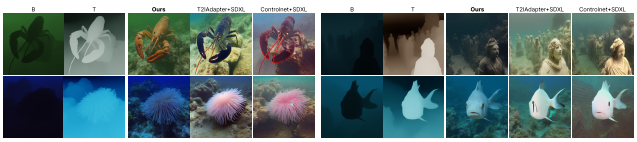

Figure 6. **Qualitative ablation study on different medium injection mechanisms.** While capable of dictating high-frequency structures, ControlNet [53] and T2I-Adapter [32] fail to accurately control low-frequency scattering color and attenuation.

and randomly sampled water medium parameters. We then apply off-the-shelf underwater segmentation models [27, 28] to the **clean images** to extract pseudo-masks of object classes following TIDE [30] (Fish, Reefs, Aquatic Plants, Wrecks, Human Divers, and Robots). These masks serve as reliable ground-truth supervision for training the models on their corresponding degraded counterparts, enabling efficient construction of a synthetic dataset with accurate annotations in diverse and challenging underwater conditions. As shown in Tab. 3, training with our synthetic data consistently improves segmentation performance across architectures and benchmarks. Compared to models trained exclusively on real data, adding our synthetic samples yields clear gains on both UIIS and USIS10K test datasets. Qualitative results, shown in Fig. 5, further show cleaner, more complete masks, sharper boundaries, and fewer missed detections under severe degradations. These results show the effectiveness of our pipeline for generating high-quality supervision and improving segmentation robustness.

4.4. Ablation Study

4.4.1. Medium Injection Mechanism

Achieving accurate medium degradation control depends critically on where and how medium conditions are injected. We therefore compare our method with two widely used conditioning architectures, ControlNet [53] and T2I-Adapter [32], using 5,451 SynTIDE [30] captions and five randomly sampled medium conditions per caption. We evaluate medium control accuracy by extracting background light intensity [42] from the generated image and compare it with the input condition. As shown in Fig. 6 and Tab. 4, both baselines do not accurately control scattering color and attenuation intensity. Unlike our framework, they do not decouple semantic generation from medium degradation but instead entangle both within the diffusion process. As a result, the medium conditions cannot be enforced faithfully, even when the scene structure is preserved. Moreover, modeling the effects of water degradation in the denoising U-Net conflicts with the pre-trained "denoise-to-clear" prior, distorting the latent distribution and weakening the overall generative quality.

Table 4. Quantitative ablation study on medium injection mechanisms and internal components. To systematically evaluate the accuracy of medium control, we compare our full pipeline against alternative condition injection mechanisms, specifically standard adapters, as well as training design choices, validating both our architectural design and component choices. We measure the Root Mean Square Error (RMSE), Mean Angular Error (MAE), and the CIEDE2000 color difference (ΔE_{00}).

Medium Injection Method	RMSE (↓)	MAE (↓)	ΔE_{00} (↓)
ControlNet [53]+SDXL [35]	0.45	21.05°	43.70
T2I-Adapter [32]+SDXL [35]	0.30	9.96°	30.10
Ours (w/o Degradation inject)	0.07	7.01°	7.27
Ours (w/o Bidir UIFM Consistency)	0.07	7.31°	6.90
Ours (Full)	**0.06**	**4.42°**	**5.44**

4.4.2. Bidirectional UIFM Consistency and Stochastic Media Noise Injection

As demonstrated in Table 4, our random medium degradation injection and bidirectional UIFM consistency self-supervised loss jointly compel the model to ignore the influence of any residual medium information within the latent space during decoding. This prevents the conditional decoder from degenerating into a naive weighted superposition of medium maps, which would otherwise lead to error accumulation. Furthermore, the bidirectional UIFM consistency loss ensures that the decoder learns a consistent, physically grounded degradation process rather than merely overfitting the synthetic data distribution.

5. Conclusion

We introduced WaterGen, a physically grounded underwater image generation framework that decouples scene semantics from water-medium effects within a latent diffusion pipeline. Our decoupled two-stage design combines clean underwater scene latent diffusion with medium-conditioned decoding, enabling precise control over attenuation and backscattering without sacrificing semantic fidelity or scene diversity. Beyond realistic and controllable image synthesis, WaterGen serves as a scalable data engine for producing paired underwater data with accurate medium variation. Extensive experiments demonstrate that our method outperforms existing underwater generative baselines in fidelity and controllability, and that the resulting synthetic data provides consistent benefits for downstream restoration and segmentation tasks.

References

[1] Akkaynak, D., Treibitz, T.: A Revised Underwater Image Formation Model. In: CVPR. pp. 6723–6732 (2018) 3

[2] Blasinski, H., Farrell, J.: A three parameter underwater image formation model. Electronic Imaging **2016**(18), 1–8 (2016). https://doi.org/10.2352/ISSN.2470-1173.2016.18.DPMI-252 3

[3] Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. arXiv preprint arXiv:2410.02073 (2024) 5, 6

[4] Cai, H., Huang, T.W., Gehlot, S., Feng, B.Y., Shah, S., Su, G.M., Metzler, C.: Parametric shadow control for portrait generation in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18207–18217 (2025) 3

[5] Desai, C., Benur, S., Patil, U., Mudenagudi, U.: Rsuigm: Realistic synthetic underwater image generation with image formation model. ACM Transactions on Multimedia Computing, Communications and Applications **21**(1), 1–22 (2024) 3

[6] Desai, C., Tabib, R.A., Reddy, S.S., Patil, U., Mudenagudi, U.: Ruig: Realistic underwater image generation towards restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2181–2189 (2021) 3

[7] Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 3

[8] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3

[9] Hong, L., Wang, X., Zhang, G., Zhao, M.: Usod10k: a new benchmark dataset for underwater salient object detection. IEEE transactions on image processing **34**, 1602–1615 (2023) 6

[10] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022) 6

[11] Islam, M.J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., Enan, S.S., Sattar, J.: Semantic segmentation of underwater imagery: Dataset and benchmark. In: 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 1769–1776. IEEE (2020) 6

[12] Islam, M.J., Li, A.Q., Girdhar, Y.A., Rekleitis, I.: Computer vision applications in underwater robotics and oceanography. In: Computer Vision, pp. 173–204. Chapman and Hall/CRC (2024) 3

[13] Islam, M.J., Xia, Y., Sattar, J.: Fast underwater image enhancement for improved visual perception. IEEE robotics and automation letters **5**(2), 3227–3234 (2020) 3

[14] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 3

[15] Jaffe, J.S.: Computer modeling and the design of optimal underwater imaging systems. IEEE journal of oceanic engineering **15**(2), 101–111 (1990) 3

[16] Jia, Y., Hoyer, L., Huang, S., Wang, T., Van Gool, L., Schindler, K., Obukhov, A.: Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In: European Conference on Computer Vision. pp. 91–109. Springer (2024) 3

[17] Kaneko, R., Ueda, T., Higashi, H., Tanaka, Y.: Phiswid: Physics-inspired underwater image dataset synthesized from rgb-d images. arXiv preprint arXiv:2404.03998 (2024) 3

[18] Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A., Schindler, K.: Marigold: Affordable adaptation of diffusion-based image generators for image analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 3, 4

[19] Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: ICCV. pp. 5148–5157 (2021) 7

[20] Khan, M., Negi, A., Kulkarni, A., Phutke, S.S., Vipparthi, S.K., Murala, S.: Phaseformer: Phase-based attention mechanism for underwater image restoration and beyond. arXiv preprint arXiv:2412.01456 (2024) 7, 12, 13

[21] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 3

[22] Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., Tao, D.: An underwater image enhancement benchmark dataset and beyond. IEEE transactions on image processing **29**, 4376–4389 (2019) 6

[23] Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., Tao, D.: An underwater image enhancement benchmark dataset and beyond. IEEE transactions on image processing **29**, 4376–4389 (2019) 7

[24] Li, H., Lian, S., Li, Z., Cong, R., Li, C., Yang, L.T., Zhang, W., Kwong, S.: Advancing marine research: Uwsam framework and uiis10k dataset for precise underwater instance segmentation. arXiv preprint arXiv:2505.15581 (2025) 6, 7

[25] Li, J., Skinner, K.A., Eustice, R.M., Johnson-Roberson, M.: Watergan: Unsupervised generative network to enable real-time color correction of

monocular underwater images. IEEE Robotics and Automation letters **3**(1), 387–394 (2017) 3

[26] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023) 5, 6

[27] Lian, S., Li, H., Cong, R., Li, S., Zhang, W., Kwong, S.: Watermask: Instance segmentation for underwater imagery. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1305–1315 (2023) 8

[28] Lian, S., Zhang, Z., Li, H., Li, W., Yang, L.T., Kwong, S., Cong, R.: Diving into underwater: Segment anything model guided underwater salient instance segmentation and a large-scale dataset. In: Proceedings of the 41st International Conference on Machine Learning. pp. 29545–29559 (2024), https://proceedings.mlr.press/v235/lian24c.html 6, 8

[29] Lin, H., Liang, D., Qi, Z., Bai, X.: A unified image-dense annotation generation model for underwater scenes. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 961–970 (2025) 3

[30] Lin, H., Liang, D., Qi, Z., Bai, X.: A unified image-dense annotation generation model for underwater scenes. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 961–970 (2025) 6, 7, 8

[31] Lv, Q., Dong, J., Li, Y., Chen, S., Yu, H., Zhang, S., Wang, W.: Uwstereo: A large synthetic dataset for underwater stereo matching. IEEE Transactions on Circuits and Systems for Video Technology (2025) 3

[32] Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4296–4304 (2024) 3, 4, 8

[33] Panetta, K., Gao, C., Agaian, S.: Human-visual-system-inspired underwater image quality measures. IEEE Journal of Oceanic Engineering **41**(3), 541–551 (2015) 7

[34] Peng, Y.T., Chen, Y.R., Chen, G.R., Liao, C.J.: Histoformer: Histogram-based transformer for efficient underwater image enhancement. IEEE Journal of Oceanic Engineering (2024) 7

[35] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 5, 6, 8

[36] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 7

[37] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 3

[38] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022) 3

[39] Sharma, P., Bisht, I., Sur, A.: Wavelength-based attributed deep neural network for underwater image restoration. ACM TOMM (2023) 7

[40] Siddique, M.A.B., Ramesh, V., Liu, J., Singh, P., Islam, M.J.: Ustyle: Waterbody style transfer of underwater scenes by depth-guided feature synthesis. IEEE Journal of Oceanic Engineering (JOE) (2025). https://doi.org/10.1109/JOE.2025.3647814 3, 6

[41] Siddique, M.A.B., Wu, J., Rekleitis, I., Islam, M.J.: Aquafuse: Waterbody fusion for physics-guided view synthesis of underwater scenes. IEEE Robotics and Automation Letters (2025). https://doi.org/10.1109/LRA.2025.3550816 3

[42] Song, W., Wang, Y., Huang, D., Tjondronegoro, D.: A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration. In: Pacific rim conference on multimedia. pp. 678–688. Springer (2018) 8, 12

[43] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021) 3

[44] Ueda, T., Yamada, K., Tanaka, Y.: Underwater image synthesis from rgb-d images and its application to deep underwater image restoration. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 2115–2119. IEEE (2019) 3

[45] Wang, N., Zhou, Y., Han, F., Zhu, H., Yao, J.: Uwgan: Underwater gan for real-world underwater color restoration and dehazing. arXiv preprint arXiv:1912.10269 (2019) 3

[46] Wang, T., Kanakis, M., Schindler, K., Van Gool, L., Obukhov, A.: Breathing new life into 3d assets with generative repainting. arXiv preprint arXiv:2309.08523 (2023) 3

[47] Wang, T., Xie, M., Cai, H., Shah, S., Metzler, C.A.: Flash-split: 2d reflection removal with flash cues and latent diffusion separation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5688–5698 (2025) 3

[48] Wu, J., Wang, T., Siddique, M.A.B., Islam, M.J., Fermuller, C., Aloimonos, Y., Metzler, C.A.: Single-step latent diffusion for underwater image restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 5, 6, 7

[49] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021) 7

[50] Xu, D., Zhou, J., Liu, Y., Min, X.: Underwater image enhancement based on hybrid enhanced generative adversarial network. Journal of Marine Science and Engineering **11**(9), 1657 (2023) 3

[51] Yang, D., Zhang, T., Li, B., Li, M., Chen, W., Li, X., Wang, X.: Underwater image translation via multi-scale generative adversarial network. Journal of Marine Science and Engineering **11**(10), 1929 (2023) 3

[52] Zhang, F., You, S., Li, Y., Fu, Y.: Atlantis: Enabling underwater depth estimation with stable diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11852–11861 (2024) 3, 6, 7

[53] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023) 3, 4, 8

[54] Zhao, Q., Xin, Z., Yu, Z., Zheng, B.: Unpaired underwater image synthesis with a disentangled representation for underwater depth map prediction. Sensors **21**(9), 3268 (2021) 3

[55] Zhao, Q., Zheng, Z., Zeng, H., Yu, Z., Zheng, H., Zheng, B.: The synthesis of unpaired underwater images for monocular underwater depth prediction. Frontiers in Marine Science **8**, 690962 (2021) 3

[56] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017) 3