Atlantis: Enabling Underwater Depth Estimation with Stable Diffusion

Fan Zhang[1] Shaodi You[2] Yu Li[3] Ying Fu[1†]
[1]Beijing Institute of Technology [2]University of Amsterdam
[3]International Digital Economy Academy

Figure 1. **Left & Right:** vivid underwater scenes generated in *Atlantis*, retaining the scene layout of terrestrial depth with diverse variations. **Middle:** depth models trained on *Atlantis* can well handle unseen real underwater scenes and predict reliable depth maps.

Abstract

Monocular depth estimation has experienced significant progress on terrestrial images in recent years thanks to deep learning advancements. But it remains inadequate for underwater scenes primarily due to data scarcity. Given the inherent challenges of light attenuation and backscatter in water, acquiring clear underwater images or precise depth is notably difficult and costly. To mitigate this issue, learning-based approaches often rely on synthetic data or turn to self- or unsupervised manners. Nonetheless, their performance is often hindered by domain gap and looser constraints. In this paper, we propose a novel pipeline for generating photorealistic underwater images using accurate terrestrial depth. This approach facilitates the supervised training of models for underwater depth estimation, effectively reducing the performance disparity between terrestrial and underwater environments. Contrary to previous synthetic datasets that merely apply style transfer to terrestrial images without scene content change, our approach uniquely creates vivid non-existent underwater scenes by leveraging terrestrial depth data through the innovative Stable Diffusion model. Specifically, we introduce a specialized Depth2Underwater ControlNet, trained on prepared {Underwater, Depth, Text} data triplets, for this generation task. Our newly developed dataset, Atlantis, enables terrestrial depth estimation models to achieve considerable improvements on unseen underwater scenes, surpassing their terrestrial pretrained counterparts both quantitatively and qualitatively. Moreover, we further show its practical utility by applying the improved depth in underwater image enhancement, and its smaller domain gap from the LLVM perspective. Code and dataset are publicly available at https://github.com/zkawfanx/Atlantis.

1. Introduction

Precise underwater depth acquisition is essential for human exploration of the sea. This holds particularly true for fields such as autonomous underwater vehicles (AUV) [8, 39], underwater robotics [56], marine biology, ecology [20] and archaeology [4, 11]. Unlike costly and operationally complex active ranging equipment, such as underwater LiDARs [17, 63], monocular depth estimation [25, 34] offers a more cost-effective and convenient deployment solution. In

†Corresponding author: fuying@bit.edu.cn

spite of remarkable progress in terrestrial depth estimation [16, 18, 21, 22, 42], underwater depth estimation remains challenging due to factors like light attenuation, backscatter, and water turbidity [2, 5, 29], which lead to poor image quality and imprecise depth data. The scarcity of data hampers the training of powerful learning-based models, which is a prevalent challenge among various tasks in deep learning era, *e.g.*, image restoration [9, 52, 57, 61, 62] and adverse weather removal [26, 32, 33, 58, 59].

While some datasets like Sea-thru [2] and SQUID [5] offer real underwater data, they are costly to acquire thus limited in scene diversity and scale. Their depth data, derived from stereo pairs or video sequences, is often sparse and not always reliable. GAN-based methods [24, 30] have emerged as an alternative to the data scarcity issue, by transferring terrestrial scenes to underwater styles using image formation models [2, 15, 38]. Despite the advantanges of easier acquisition and relatively larger scale and diversity, their domain gap and lack of realism limit their efficacy.

To address these challenges, our paper introduces a novel pipeline to generate underwater depth dataset, comprising diverse and realistic underwater images and accurate depth. Compared to real datasets, it is inexpensive and easy to obtain, featuring large diversity and theoretically unlimited scale. Meanwhile, it is more realistic and possesses smaller domain gap than GAN-based datasets. Using Stable Diffusion (SD) [44] and ControlNet [60], this approach can generate underwater imagery following the scene structure and layout of terrestrial depth. Despite the widespread applications in AI-generated content, they have rarely been used for generating training data. We present *Atlantis*, a dataset that combines the accuracy of terrestrial depth with the lifelike depiction of underwater scenes, offering a robust resource for training depth models for underwater scenes.

Specifically, we first construct a dataset comprising underwater images, estimated depths, and captions describing the image content. Then we train a *Depth2Underwater* ControlNet targeting realistic underwater image generation using depth map. With the pretrained SD and our trained ControlNet, we construct the *Atlantis* dataset with realistic underwater images and accurate depth, enabling the training of terrestrial depth models for underwater depth estimation. Their performance are largely improved both quantitatively and qualitatively over the terrestrial counterparts of KITTI [19] and NYU Depthv2 [46]. Moreover, we show the utility of *Atlantis* by applying the trained depth model in underwater image enhancement and reveal its smaller domain gap than GAN-based datasets using large language vision model. It is worth noting that our goal is not necessarily to surpass the results of robust depth models trained with million-scale data and various training tricks, *e.g.*, MiDaS [42], on underwater scenes, but to enable existing depth models on underwater scenes with our data and simple

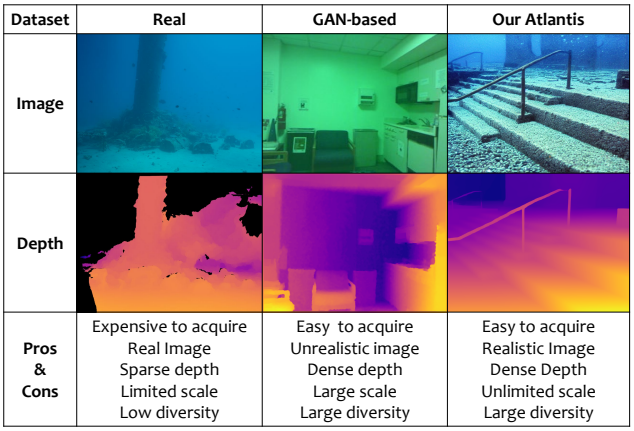


Figure 2. Comparisions of real dataset [5], GAN-based synthetic dataset [24] and ours proposed underwater depth dataset *Atlantis*.

training. To summarize, our contributions are three-fold:

- We are the first, to the best of our knowledge, proposing to construct paired dataset for underwater depth estimation training, utilizing newly emerged SD and ControlNet.
- The proposed dataset, *Atlantis*, which comprises realistic underwater images and reliable depth, is easy to collect and extend, and features large diversity, theoretically unlimited scale and smaller domain gap.
- We propose to improve the performance of existing depth models on unseen underwater scenes using *Atlantis* for supervised training. The improved depth can further be applied for underwater image enhancment, which highlights the effectiveness and utility of our dataset.

2. Related Work

In this section, we briefly review the key developments in terrestrial monocular depth estimation, current underwater depth estimation techniques, and methods integrating underwater depth estimation with image enhancement.

2.1. Terrestrial Depth Estimation

Eigen *et al*. [16] pioneered the coarse-to-fine network approach for end-to-end monocular depth estimation, a significant breakthrough. Their Scale-Invariant log loss was widely adopted in subsequent methods. Monodepth [21] and Monodepth2 [22] achieved impressive self-supervised performance and robustness. DORN [18] and Adabins [6] represented methods that treat depth estimation as ordinal regression and classification by discretizing depth. Recently, MiDaS [42] set a new benchmark for robust zero-shot depth estimation by training on million-scale multi-source data with various optimization techniques. DPT [43] and ZoeDepth [7] further enhanced the performance in relative and absolute depth metrics. NeWCRFs [55] and iDisc [40] introduced fully-connected CRFs and an Internal Discretization module into depth estimation, respectively.

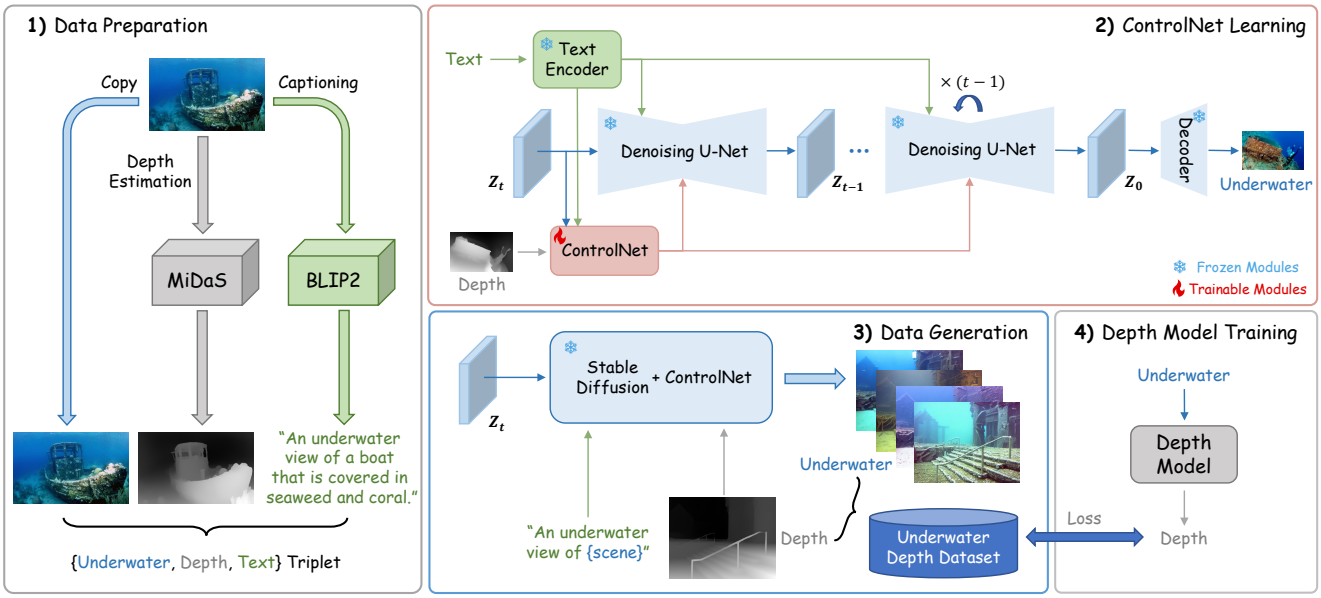


Figure 3. Overview of our method for generating the underwater depth dataset. The process begins by creating an intermediate dataset containing underwater images, depth maps, and text descriptions. This dataset is used to train the *Depth2Underwater* ControlNet for generating underwater images from depth maps. The resulting dataset, namely *Atlantis*, facilitates the training and performance improvement of terrestrial depth models for unseen underwater scenes.

IEBins [45] introduced iterative elastic bins along the line of classification-regression-based MDE and VA-DepthNet [35] imposed first-order variational constraints in the scene space. However, these models' performance in underwater scenes is limited due to domain gap and data scarcity.

2.2. Underwater Depth Estimation

Underwater, light attenuation and backscatter depend on the distance light travels through water. Image formation models [2, 15, 28, 38] that elucidate these relationships aid in estimating parameters such as attenuation coefficients and transmission. Intriguingly, depth information often emerges as a secondary product of this process. Traditional techniques of DCP family [14, 27], therefore, can estimate depth. Gupta and Mitra [23] proposed UW-Net that utilizes GAN for unsupervised training. Li *et al*. [30] and Hambarde *et al*. [24] proposed to synthesize different types of underwater images using the image formation model [10] and NYU Depthv2 [46], focusing on image enhancement and depth estimation, respectively. Recent work has also explored lightweight models [54] and self-supervised learning [3, 53]. Despite their effectiveness, these methods still lag behind terrestrial models in performance, underscoring the need for novel datasets that enable the training of powerful terrestrial depth estimation models.

2.3. Underwater Image Enhancement

Unlike underwater depth estimation, underwater image enhancement has been an actively investigated field since the era of traditional techniques, focusing on color correction, contrast enhancement, and backscatter removal. Early methods predominantly relied on physical models and handcrafted priors [2, 14, 27], often integrating depth-related aspects. Recent learning-based methods [13, 47, 48] have shown a preference for jointly estimating underwater depth and image recovery. A notable advancement is Akkaynak and Treibitz's revised image formation model [1] and their Sea-thru algorithm [2], which achieves effective dewatering results using range maps. To highlight the effectiveness and utility of *Atlantis*, We further apply the trained depth models in underwater image enhancement.

3. Method

In this section, we first detail the motivation, then introduce our pipeline for data generation as depicted in Figure 3.

3.1. Motivation

In the pursuit of accurate underwater depth estimation, one of the primary challenges is the labor-intensive and complex task of collecting real underwater data, including both imagery and precise depth information. Existing datasets like Sea-thru [2] and SQUID [5], although valuable, are limited in the diversity of scenes and scale due to the acquisition difficulty. The depth data obtained from stereo pairs in these datasets is often sparse and compromised in reliability due to the inherently low quality of underwater images.

As an alternative, GAN-based methods [24, 30] have been utilized to synthesize underwater images by transferring the style of terrestrial images, combining terrestrial

depth and image formation models, aiming to alleviate the scarcity of real underwater data. However, this approach, while being less costly and in larger diversity and scale, typically results in unrealistic synthetic images with significant domain gap, as the transformation is more akin to style transfer than to the creation of authentic underwater scenes.

This is where *Atlantis* comes into play. We offer a solution that generates vivid, non-existent underwater scenes using only depth maps and textual descriptions. This approach not only provides an infinite range of sampling possibilities but also ensures the ease of depth map acquisition. The resulting images exhibit a smaller domain gap compared to traditional methods (Section 4.4). Our dataset, therefore, stands out for its advantages in terms of easy acquisition, diversity and scale, realism, and practicality, marking a significant improvement over existing datasets and underwater imagery synthesis methods.

3.2. Underwater Depth Dataset: Atlantis

In the creation of our underwater depth dataset, as illustrated in Figure 3, we initiate by constructing an intermediate dataset that is instrumental in training a specialized ControlNet [60]. This tailored ControlNet is then utilized to guide the pretrained Stable Diffusion v1.5 [44] in generating underwater images informed by outdoor depth maps.

Data Preparation. Our process begins with the utilization of the robust MiDaS [42] model to estimate inverse relative depth for images from the UIEB dataset [29], following ControlNet [60] procedure. For each underwater image U, a corresponding depth map D is obtained as follows:

$$D = \mathcal{F}_{MiDaS}(U), \tag{1}$$

where $\mathcal{F}_{MiDaS}$ denotes the pretrained MiDaS model. Additionally, each image U undergoes captioning using the pretrained BLIP2 model [31] to generate descriptive text T:

$$T = \mathcal{F}_{BLIP2}(U). \tag{2}$$

This leads to the formation of our intermediate dataset, comprising {Underwater, Depth, Text} triplets. Here, the depth map D serves as the conditioning input, with U as the target image and T providing the textual narrative for SD's content generation. During the training stage, only ControlNet is set as trainable and other parts of SD are freezed in the whole process.

Data Generation. Post training our *Depth2Underwater* ControlNet, we can now generate underwater images based on provided depth maps. For instance, with a text prompt *"an underwater view of Atlantis"* and a corresponding outdoor depth map D, a vivid non-existent underwater scene is created. The process is as follows:

$$c = \mathcal{F}_{CtrlNet}(z_t, D, T), \tag{3}$$

where $\mathcal{F}_{CtrlNet}$ represents our trained ControlNet and c is conditioning feature extracted from the depth map. t denotes the t-th step of the backward diffusion process. This feature c is then utilized in the SD generation process:

$$\bar{U} = \mathcal{F}_{SD}(z_t, T|c), \tag{4}$$

yielding the generated underwater image $\bar{U}$. $\mathcal{F}_{SD}(\cdot|\cdot)$ denotes the generation process of pretrained SD conditioned by a ControlNet. This methodology allows for the creation of a diverse array of underwater images, all adhering to the predetermined scene structure but with varied appearances.

Underwater Depth Dataset. The final dataset is produced by conditioning the generation process of the pretrained SD model with our *Depth2Underwater* ControlNet. Utilizing 400 terrestrial images from the DIODE-outdoor dataset [49] for depth estimation, we employ text prompts such as *"an underwater view of Atlantis"* and *"a corner of lost Atlantis"* to guide the generation of unique underwater scenes. Sampling four times for each prompt and depth map results in a dataset comprising 3,200 data pairs. This dataset is pivotal in training and enhancing the performance of state-of-the-art terrestrial depth estimation models, particularly for unseen underwater scenes. The final output is an estimated depth map D' for any given unseen underwater image U':

$$D' = \mathcal{F}_{Depth}(U'), \tag{5}$$

where $\mathcal{F}_{Depth}$ denotes the depth estimation model trained on our dataset.

3.3. Implementation Details

This subsection outlines the key implementation aspects of our data generation pipeline, ensuring a comprehensive understanding of the process and techniques involved.

Data Preparation. We leverage the training set of UIEB dataset [29], which consists of 700 underwater images, for initial depth estimation and captioning. The robust MiDaS model [42] is employed for depth estimation, while the BLIP2 model [31] facilitates image captioning. These steps result in 700 data triplets comprising underwater images, depth maps, and textual descriptions, which forms the foundation of our training data for ControlNet.

ControlNet Training and Deployment. We utilize the `diffusers` library [50] for the modification and efficient deployment of both SD and ControlNet. We train the ControlNet using standard training settings. For inference, we set the guidance scale to 5 to avoid unrealistic lighting styles, and sample for 20 steps for each image generation.

Depth Estimation Model Training. For the training of depth estimation models, we employ recent iDisc [40], NeWCRFs [55], IEBins [45] and VA-DepthNet [35]. These models are trained on our generated underwater depth

Table 1. Quantitative comparisons on real underwater images from D3 and D5 subsets of Sea-thru dataset [2].

Models	Training Data	$RMSE\downarrow$	$RMSE_{log}\downarrow$	$A.Rel\downarrow$	$S.Rel\downarrow$	$log_{10}\downarrow$	$SI_{log}\downarrow$	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$
iDisc [40]	KITTI	5.891	1.192	4.702	44.288	0.489	35.846	0.093	0.241	0.359
	NYU Depthv2	3.144	0.845	**0.819**	**2.471**	0.338	37.296	0.215	0.403	0.504
	Atlantis	**1.371**	**0.354**	1.630	14.279	**0.109**	**34.654**	**0.553**	**0.850**	**0.955**
NeWCRFs [55]	KITTI	3.251	0.934	2.874	15.768	0.365	42.341	0.213	0.375	0.465
	NYU Depthv2	3.390	0.955	**0.770**	**2.350**	0.372	47.667	0.179	0.365	0.479
	Atlantis	**1.435**	**0.378**	1.683	14.764	**0.120**	**37.101**	**0.476**	**0.837**	**0.952**
IEBins [45]	KITTI	4.217	1.072	3.648	25.007	0.427	44.031	0.159	0.311	0.417
	NYU Depthv2	3.287	0.901	**0.814**	**2.373**	0.357	44.753	0.151	0.350	0.489
	Atlantis	**1.597**	**0.425**	1.687	13.766	**0.139**	**41.090**	**0.425**	**0.762**	**0.919**
VA-DepthNet [35]	KITTI	7.842	1.326	5.999	76.830	0.555	31.574	0.025	0.129	0.257
	NYU Depthv2	2.969	0.777	**0.969**	**2.626**	0.315	38.286	0.143	0.279	0.494
	Atlantis	**1.204**	**0.292**	1.781	19.937	**0.086**	**28.739**	**0.648**	**0.939**	**0.985**

Table 2. Quantitative comparisons on real underwater images from SQUID dataset [5].

Models	Training Data	$RMSE\downarrow$	$RMSE_{log}\downarrow$	$A.Rel\downarrow$	$S.Rel\downarrow$	$log_{10}\downarrow$	$SI_{log}\downarrow$	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$
iDisc [40]	KITTI	7.265	0.736	1.039	8.040	0.289	35.827	0.156	0.349	0.555
	NYU Depthv2	8.752	1.638	0.737	6.454	0.683	41.097	0.016	0.046	0.093
	Atlantis	**2.663**	**0.277**	**0.249**	**0.920**	**0.094**	**27.221**	**0.637**	**0.900**	**0.960**
NeWCRFs [55]	KITTI	6.692	0.779	0.579	3.930	0.294	52.091	0.197	0.381	0.541
	NYU Depthv2	8.957	1.764	0.766	6.740	0.734	46.791	0.013	0.029	0.064
	Atlantis	**2.563**	**0.256**	**0.229**	**0.830**	**0.088**	**25.189**	**0.675**	**0.902**	**0.964**
IEBins [45]	KITTI	7.353	0.780	1.059	9.476	0.289	52.793	0.207	0.412	0.581
	NYU Depthv2	8.839	1.674	0.740	6.532	0.692	47.271	0.013	0.041	0.094
	Atlantis	**2.896**	**0.296**	**0.263**	**0.992**	**0.100**	**29.209**	**0.615**	**0.870**	**0.951**
VA-DepthNet [35]	KITTI	8.753	0.827	1.299	12.381	0.328	38.362	0.148	0.308	0.461
	NYU Depthv2	8.274	1.349	0.657	5.747	0.558	35.518	0.042	0.112	0.205
	Atlantis	**2.666**	**0.239**	**0.204**	**0.703**	**0.082**	**23.337**	**0.705**	**0.915**	**0.970**

dataset. Given that MiDaS outputs inverse relative depth, we cap the depth values at a maximum of 20 meters. This aligns with the understanding that scene radiance in underwater environments is predominantly affected by backscatter beyond this range [2].

Hardware and Accessibility. All experiments and model trainings are conducted on an NVIDIA RTX 3090 GPU. Both the intermediate triplet data and Atlantis, as well as the *Depth2Underwater* ControlNet will be released, contributing to the broader research community in this field.

4. Experiments

In this section, we demonstrate the effectiveness of *Atlantis* in training supervised depth estimation models. We compare models trained from scratch on our dataset with their officially pretrained counterparts on terrestrial datasets, which is evaluated on unseen underwater datasets. Additionally, we apply the Sea-thru [2] algorithm[1] for underwater image enhancement with estimated depth, to showcase the practical application of depth models trained on *Atlantis*. Finally, we investigate into the issue of domain gap from the perspective of Large Language Vision Model (LLVM), which evidently shows the smaller domain gap of *Atlantis* compared to previous synthetic datasets.

[1]https://github.com/hainh/sea-thru

Due to the space limitation, we provide more samples of *Atlantis*, qualitative comparisons as well as underwater image enhancement results in the supplementary material.

Experimental Setup. We focus on four models: iDisc [40], NeWCRFs [55], IEBins [45] and VA-DepthNet [35]. All models are trained from scratch on *Atlantis* and evaluated against their official pretrained counterparts on KITTI [19] and NYU Depthv2 [46]. They all utilize the Swin-L model [37] pretrained on ImageNet-22k [12] for encoder initialization. Quantitatively, we conducted evaluations using the D3 and D5 subsets of Sea-thru [2] and the SQUID dataset [5], which consist of underwater images and depth obtained via Structure-from-Motion (SfM) algorithm. For qualitative comparison, we addtionally include the test set of UIEB dataset [29] to complement the diversity of tested scenes. The metrics used for quantitative evaluation encompass root mean square error ($RMSE$) and its log variant ($RMSE_{log}$), absolute error in log-scale (Log_{10}), absolute ($A.Rel$) and squared ($S.Rel$) mean relative error, the percentage of inlier pixels (δ_i) with threshold 1.25^i, and scale-invariant error in log-scale (SI_{log}): $100\sqrt{Var(\epsilon_{log})}$.

4.1. Quantitative Results

The results, as detailed in Tables 1 and 2, demonstrate a significant domain gap for models pretrained on terres-

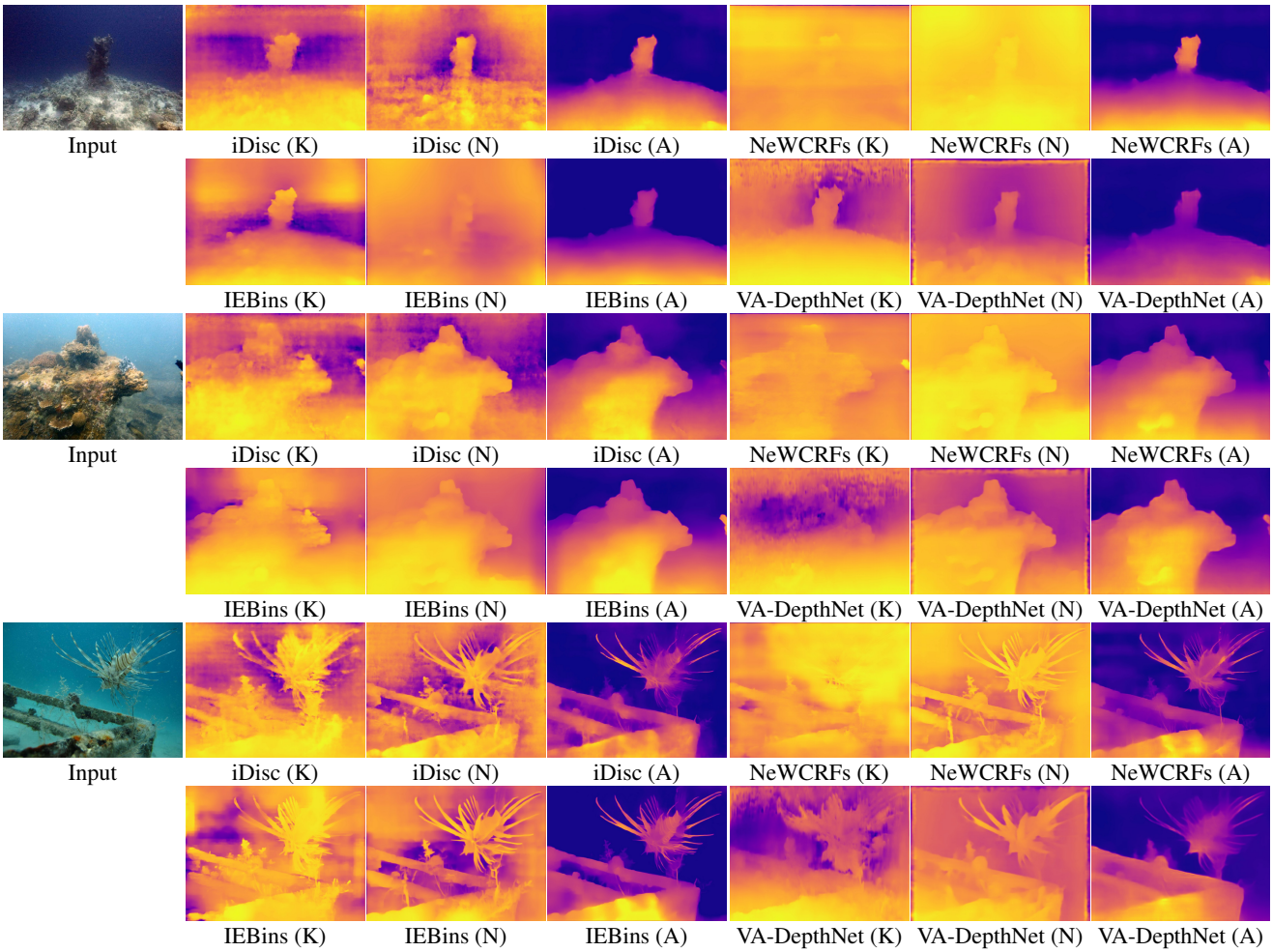


Figure 4. Qualitative results on the test set of UIEB dataset [29]. K and N denote models pretrained on KITTI [19] and NYU Depthv2 [46] while A represents models trained on *Atlantis*. Depth results are evidently improved after training on our dataset.

trial datasets of KITTI [19] and NYU Depthv2 [46] when they are applied to underwater images. This domain gap, which adversely affects the performance across most metrics, is evident for all four models, underscoring the inherent challenges in directly applying supervised monocular depth models to underwater scenes. Conversely, when these models are trained from scratch on *Atlantis*, they all exhibit substantial improvements across the majority of quantitative metrics. The improvements are consistent across evaluations on both Sea-thru [2] and SQUID [5] datasets, affirming the efficacy of *Atlantis* in supervisely training depth models and enhancing the performance of monocular depth estimation for unseen underwater scenes. This outcome suggests that training on *Atlantis* effectively reduces the domain gap. It is noteworthy that *Atlantis*, despite being smaller in size compared to the terrestrial datasets, has shown significant potential in this context. This suggests that further expanding its scale and diversity or employing hybrid training approaches might yield even more pronounced improvements.

4.2. Qualitative Results

Figures 4 and 5 showcase visual comparisons that highlight the stark contrast in depth estimation performance. All pretrained models on terrestrial datasets, including iDisc [40], NeWCRFs [55], IEBins [45] and VA-DepthNet [35], produce significantly erroneous results on underwater images. These inaccuracies manifest as heavy haze artifacts in water body areas and incorrect relative scene layout distances. In sharp contrast, after training on *Atlantis*, all four models exhibit a remarkable improvement in interpreting underwater scenes. Notably, they accurately identify and appropriately assign distance to water body areas, demonstrating enhanced discriminative capabilities. The transitions in scene content are marked by clear borders, and the models adeptly handle transparent water with varying color casts. Overall, the layout of underwater scenes is more accurately rendered, and depth ambiguities, particularly in water bodies, are substantially reduced. This improvement underscores the effectiveness of *Atlantis* in enabling depth estimation models to better differentiate water bodies and adapt

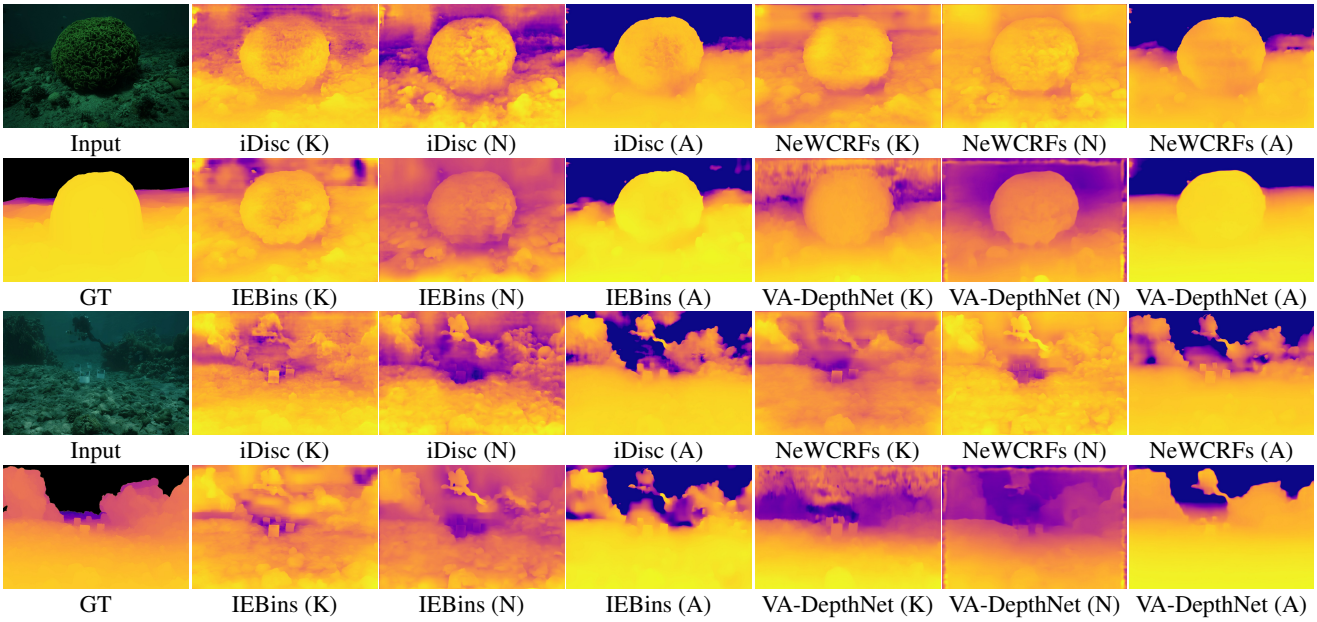


Figure 5. Qualitative results on Sea-thru dataset [2]. K and N denote models pretrained on KITTI [19] and NYU Depthv2 [46] datasets while A represents the models trained on *Atlantis*. Depth results are evidently improved after training on our dataset.

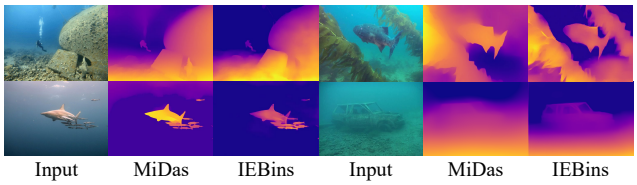


Figure 6. Some MiDaS results for reference.

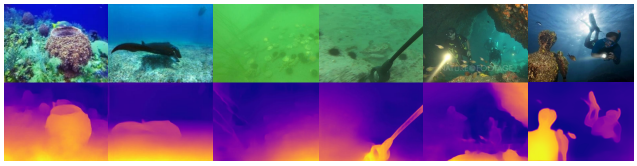


Figure 7. Results on various water types and artificial lights.

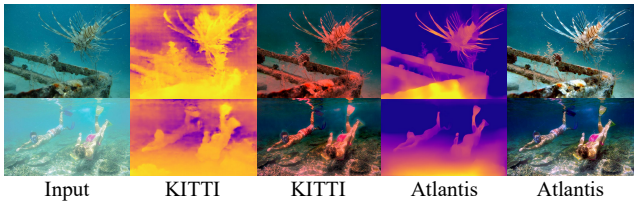


Figure 8. Effects of different depth on UIE. iDisc is used here.

to diverse underwater conditions, including color casts and lighting variations. It's worth noting that the underwater images used in these comparisons were unseen during training since *Atlantis* consists of all non-existent scenes generated from scene layout of terrestrial depth. This further emphasizes the generalizability of *Atlantis* in training depth models that effectively adapt to real underwater scenes.

We provide results of MiDaS for additional reference in Figure 6. Here IEBins [45] is selected for illustration. Interestingly, we can find that IEBins is better at recognizing water bodies, producing cleaner and sharper depth without hazy artifacts on different water types. This also highlights the improvement brought by *Atlantis* on underwater scenes. Moreover, we visualize the results on various water types as well as the existence of artificial lights in Figure 7 to show the robustness of model trained on *Atlantis*. Model (e.g., iDisc) trained on *Atlantis* shows good generalizability on different water bodies with different haze heaviness and in the presence of artificial lights. It validates that our generation pipeline can implicitly learn the depth-dependent effects in a data-driven manner.

4.3. Improved Depth for UIE

Sea-thru [2], known for its ability to remove water effects with precise range maps derived from stereo pairs or video sequences, can be extended to single underwater images using depth by models trained on *Atlantis*. It fails the enhancement with reddish color and artifacts when using inaccurate depth (Figure 8), while oppositely, it produces impressive enhancements (Figure 9) when equipped with depth by models trained on *Atlantis*. It evidently shows the improvement of depth accuracy brought by *Atlantis* and further suggests its practical utility in real-world applications.

4.4. Domain Gap from LLVM Perspective

The advent of Large Language Vision Models (LLVM) [36, 41] has revolutionized the alignment between textual and visual features, opening new avenues for synthetic data analysis. We utilize recent LLaVA v1.5 model [36] to reveal the misalignment problem of previous style transfer

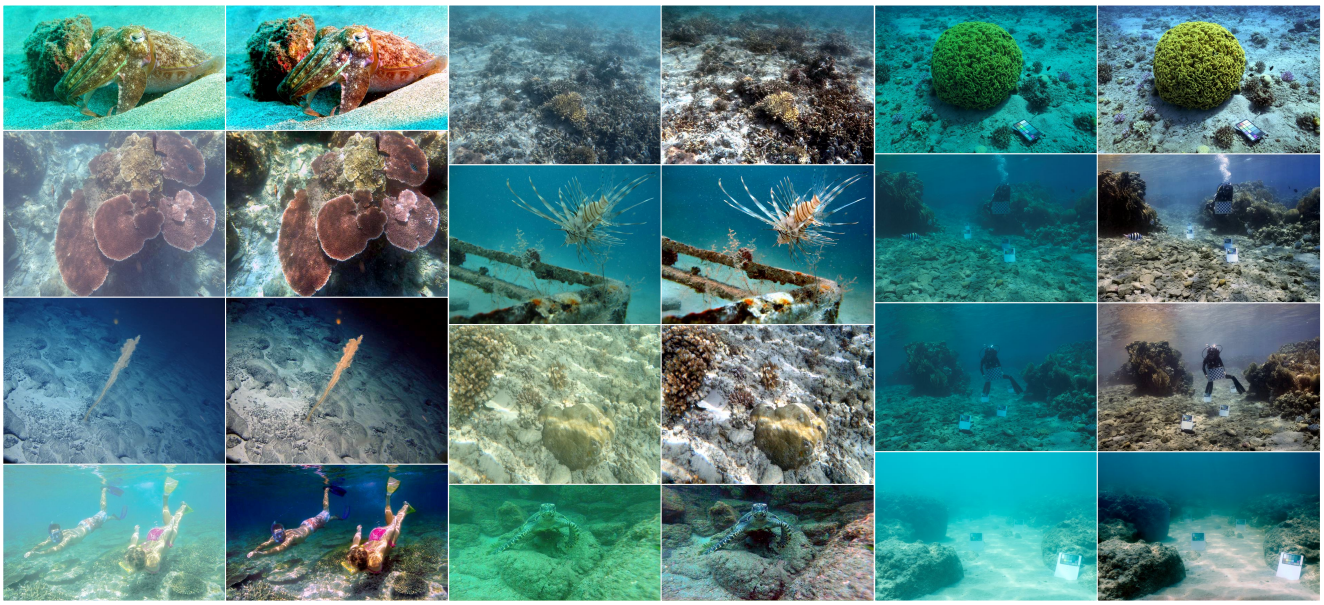
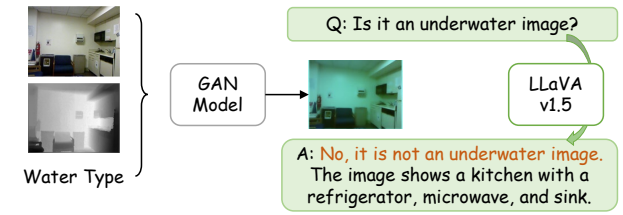
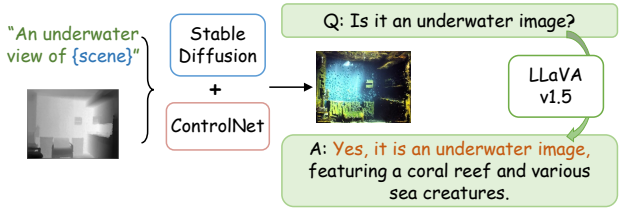


Figure 9. Qualitative results of the improved depth result applied to downstream underwater image enhancement. **Left & Middle:** UIEB dataset [29]. **Right:** Sea-thru dataset [2] (the above three) and SQUID dataset [5] (the bottom one). Enhancement outputs well show the effectiveness of the proposed dataset on training depth models for reliable underwater depth estimation.



(a) UW-GAN



(b) Ours

Figure 10. (a) GAN style transfers images with the same scene content, possessing large domain gap. (b) Our method generates non-existent underwater scenes following the scene structure, correctly identified as underwater scenes by LLVM.

synthesis (Figure 10). Specifically, we evaluate 14490 images from UW-GAN [30, 51] and our method on how much an LLVM will agree with the generation to be *underwater*. Surprisingly, 100% of our images pass the test while only 18% of images in UW-GAN own the agreement of underwater scene, which evidently confirms the smaller domain gap of *Atlantis*. It also suggests the great advantage of SD [44] and ControlNet [60] over previous synthesis methods in serving as the novel engine for training data synthesis.

5. Conclusion

In this paper, we introduce a novel pipeline utilizing Stable Diffusion and a specialized ControlNet for generating realistic underwater images with accurate depth. We propose a dataset, *Atlantis*, to enable the training of terrestrial depth models for underwater depth estimation, which significantly enhances their performance on underwater scenes. The proposed dataset comprises realistic underwater images and accurate depth, featuring easy acquisition, large diversity, theoretically unlimited scale and smaller domain gap. Our experiments, encompassing both quantitative and qualitative analyses, demonstrate the superiority of models trained on our dataset compared to those pretrained on terrestrial datasets. Notably, the application of trained models in underwater image enhancement showcase their practical utility and highlight the value of our dataset. Our study reveals the potential of SD to be a new source of high-quality training data. As future work, expanding the dataset and utilizing hybrid training could unlock greater improvements in model performance and generalization.

Acknowledgments This work was supported by the National Natural Science Foundation of China (62331006, 62171038, 62088101), the R&D Program of Beijing Municipal Education Commission (KZ202211417048), and the Fundamental Research Funds for the Central Universities.

References

[1] Derya Akkaynak and Tali Treibitz. A revised underwater image formation model. In *CVPR*, 2018. 3

[2] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *CVPR*, 2019. 2, 3, 5, 6, 7, 8

[3] Shlomi Amitai, Itzik Klein, and Tali Treibitz. Self-supervised monocular depth underwater. In *ICRA*, 2023. 3

[4] Geoffrey N Bailey and Nicholas C Flemming. Archaeology of the continental shelf: marine resources, submerged landscapes and underwater archaeology. *Quaternary Science Reviews*, 27(23-24):2153–2165, 2008. 1

[5] Dana Berman, Deborah Levy, Shai Avidan, and Tali Treibitz. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE TPAMI*, 43(8):2822–2837, 2020. 2, 3, 5, 6, 8

[6] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021. 2

[7] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2

[8] D Richard Blidberg. The development of autonomous underwater vehicles (auv); a brief summary. In *ICRA*, 2001. 1

[9] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *IJCV*, 131(8): 2198–2218, 2023. 2

[10] John Y Chiang and Ying-Ching Chen. Underwater image enhancement by wavelength compensation and dehazing. *IEEE TIP*, 21(4):1756–1769, 2011. 3

[11] Dwight F Coleman, James B Newman, and Robert D Ballard. Design and implementation of advanced underwater imaging systems for deep sea marine archaeological surveys. In *OCEANS MTS/IEEE Conference and Exhibition*, 2000. 1

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[13] Paulo LJ Drews, Erickson R Nascimento, Silvia SC Botelho, and Mario Fernando Montenegro Campos. Underwater depth estimation and image restoration based on single images. *IEEE computer graphics and applications*, 36(2):24–35, 2016. 3

[14] Paulo LJ Drews, Erickson R Nascimento, Silvia SC Botelho, and Mario Fernando Montenegro Campos. Underwater depth estimation and image restoration based on single images. *IEEE computer graphics and applications*, 36(2):24–35, 2016. 3

[15] Seibert Q Duntley. Light in the sea. *JOSA*, 53(2):214–233, 1963. 2, 3

[16] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 2014. 2

[17] Andrew Filisetti, Andreas Marouchos, Andrew Martini, Tara Martin, and Simon Collings. Developments and applications of underwater lidar systems in support of marine science. In *OCEANS MTS/IEEE Charleston*, 2018. 1

[18] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 2

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2, 5, 6, 7

[20] R Gibson, R Atkinson, and J Gordon. A review of underwater stereo-image measurement for marine biology and ecology applications. *Oceanography and marine biology: an annual review*, 47:257–292, 2016. 1

[21] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 2

[22] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 2

[23] Honey Gupta and Kaushik Mitra. Unsupervised single image underwater depth estimation. In *ICIP*, 2019. 3

[24] Praful Hambarde, Subrahmanyam Murala, and Abhinav Dhall. Uw-gan: Single-image depth estimation and image enhancement for underwater images. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021. 2, 3

[25] Zhixiang Hao, Yu Li, Shaodi You, and Feng Lu. Detail preserving depth estimation from a single image using attention guided networks. In *3DV*, 2018. 1

[26] Zhixiang Hao, Shaodi You, Yu Li, Kunming Li, and Feng Lu. Learning from synthetic photorealistic raindrop for single image raindrop removal. In *ICCVW*, 2019. 2

[27] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE TPAMI*, 33(12): 2341–2353, 2010. 3

[28] Jules S Jaffe. Computer modeling and the design of optimal underwater imaging systems. *IEEE Journal of Oceanic Engineering*, 15(2):101–111, 1990. 3

[29] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE TIP*, 29: 4376–4389, 2019. 2, 4, 5, 6, 8

[30] Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *PR*, 98:107038, 2020. 2, 3, 8

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 4

[32] Kunming Li, Yu Li, Shaodi You, and Nick Barnes. Photorealistic simulation of road scene for data-driven methods in bad weather. In *ICCVW*, 2017. 2

[33] Ruoteng Li, Xiaoyi Zhang, Shaodi You, and Yu Li. Learning to dehaze from realistic scene with a fast physics-based dehazing network. *arXiv preprint arXiv:2004.08554*, 2020. 2

[34] Siyuan Li, Yue Luo, Ye Zhu, Xun Zhao, Yu Li, and Ying Shan. Enforcing temporal consistency in video depth estimation. In *ICCV*, 2021. 1

[35] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. Va-depthnet: A variational approach to single image depth prediction. In *ICLR*, 2022. 3, 4, 5, 6

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 7

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5

[38] BL McGlamery. A computer model for underwater camera systems. In *Ocean Optics VI*, pages 221–231. SPIE, 1980. 2, 3

[39] Liam Paull, Sajad Saeedi, Mae Seto, and Howard Li. Auv navigation and localization: A review. *IEEE Journal of oceanic engineering*, 39(1):131–149, 2013. 1

[40] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. idisc: Internal discretization for monocular depth estimation. In *CVPR*, 2023. 2, 4, 5, 6

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 7

[42] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3):1623–1637, 2020. 2, 4

[43] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 2

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 4, 8

[45] Shuwei Shao, Zhongcai Pei, Xingming Wu, Zhong Liu, Weihai Chen, and Zhengguo Li. Iebins: Iterative elastic bins for monocular depth estimation. *arXiv preprint arXiv:2309.14137*, 2023. 3, 4, 5, 6, 7

[46] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 3, 5, 6, 7

[47] Wei Song, Yan Wang, Dongmei Huang, and Dian Tjondronegoro. A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration. In *Pacific Rim Conference on Multimedia*, 2018. 3

[48] Nisha Varghese, Ashish Kumar, and AN Rajagopalan. Self-supervised monocular underwater depth recovery, image restoration, and a real-sea video dataset. In *ICCV*, 2023. 3

[49] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 4

[50] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 4

[51] Nan Wang, Yabin Zhou, Fenglei Han, Haitao Zhu, and Jingzheng Yao. Uwgan: underwater gan for real-world underwater color restoration and dehazing. *arXiv preprint arXiv:1912.10269*, 2019. 8

[52] Kaixuan Wei, Angelica Aviles-Rivero, Jingwei Liang, Ying Fu, Hua Huang, and Carola-Bibiane Schönlieb. Tfpnp: Tuning-free plug-and-play proximal algorithms with applications to inverse imaging problems. *Journal of Machine Learning Research*, 23(16):1–48, 2022. 2

[53] Xuewen Yang, Xing Zhang, Nan Wang, Guoling Xin, and Wenjie Hu. Underwater self-supervised depth estimation. *Neurocomputing*, 514:362–373, 2022. 3

[54] Boxiao Yu, Jiayi Wu, and Md Jahidul Islam. Udepth: Fast monocular depth estimation for visually-guided underwater robots. In *ICRA*, 2023. 3

[55] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *CVPR*, 2022. 2, 4, 5, 6

[56] Junku Yuh and Michael West. Underwater robotics. *Advanced Robotics*, 15(5):609–639, 2001. 1

[57] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *CVPR*, 2021. 2

[58] Fan Zhang, Shaodi You, Yu Li, and Ying Fu. Gtav-nightrain: Photometric realistic large-scale dataset for night-time rain streak removal. *arXiv preprint arXiv:2210.04708*, 2022. 2

[59] Fan Zhang, Shaodi You, Yu Li, and Ying Fu. Learning rain location prior for nighttime deraining. In *ICCV*, 2023. 2

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 4, 8

[61] Tao Zhang, Ying Fu, and Cheng Li. Hyperspectral image denoising with realistic data. In *ICCV*, 2021. 2

[62] Tao Zhang, Ying Fu, and Jun Zhang. Guided hyperspectral image denoising with realistic data. *IJCV*, 130(11):2885–2901, 2022. 2

[63] Guoqing Zhou, Chenyang Li, Dianjun Zhang, Dequan Liu, Xiang Zhou, and Jie Zhan. Overview of underwater transmission characteristics of oceanic lidar. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8144–8159, 2021. 1