

Atlantis: Enabling Underwater Depth Estimation with Stable Diffusion

Fan Zhang¹ Shaodi You² Yu Li³ Ying Fu¹

¹Beijing Institute of Technology ²University of Amsterdam

³International Digital Economy Academy

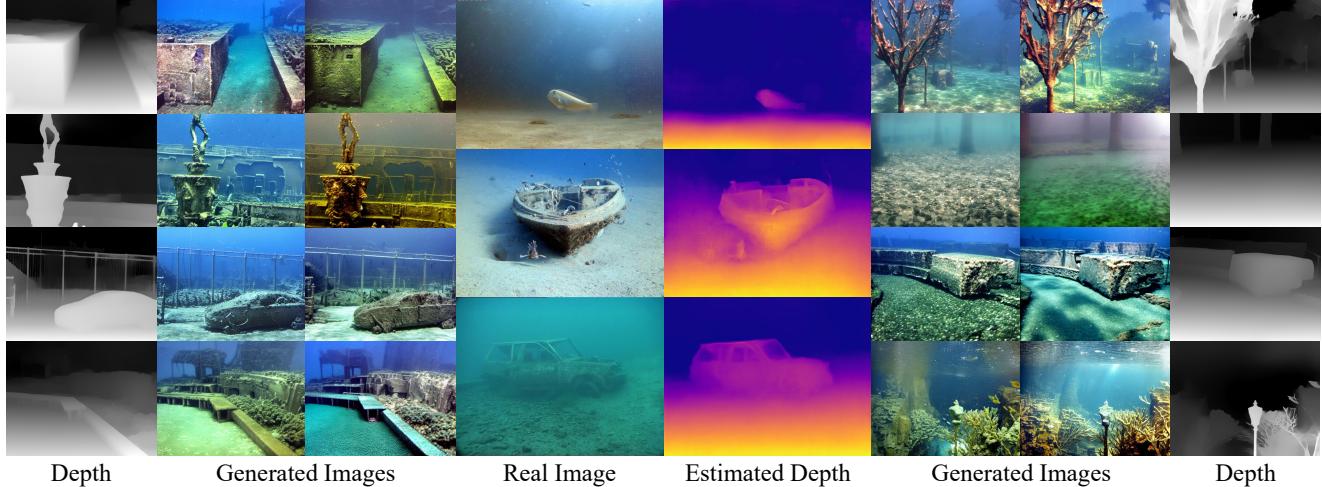


Figure 1. The proposed method can generate various vivid non-existent underwater scenes by sampling, following the scene layout of conditioning depth map (on the left and right) for training depth estimation models. The depth model trained on the proposed dataset can well handle unseen real underwater scenes and get reliable depth maps (in the middle).

Abstract

Monocular depth estimation has experienced significant progress on terrestrial images in recent years, largely due to deep learning advancements. However, it remains inadequate for underwater scenes, primarily because of data scarcity. Given the inherent challenges of light attenuation and backscattering in water, acquiring clear underwater images or precise depth information is notably difficult and costly. Consequently, learning-based approaches often rely on synthetic data or turn to unsupervised or self-supervised methods to mitigate this lack of data. Nonetheless, the performance of these methods is often constrained by the domain gap and looser constraints. In this paper, we propose a novel pipeline for generating photorealistic underwater images using accurate terrestrial depth data. This approach facilitates the training of supervised models for underwater depth estimation, effectively reducing the performance disparity between terrestrial and underwater environments. Contrary to prior synthetic datasets that merely apply style transfer to terrestrial images without altering the scene content, our approach uniquely creates vibrant, non-existent underwater scenes by leveraging terrestrial depth data

through the innovative Stable Diffusion model. Specifically, we introduce a unique Depth2Underwater Control-Net, trained on specially prepared {Underwater, Depth, Text} data triplets, for this generation task. Our newly developed dataset enables terrestrial depth estimation models to achieve considerable improvements, both quantitatively and qualitatively, on unseen underwater images, surpassing their terrestrial pre-trained counterparts. Moreover, the enhanced depth accuracy for underwater scenes also aids underwater image restoration techniques that rely on depth maps, further demonstrating our dataset's utility. The dataset will be publicly available at <https://github.com/zkawfanx/Atlantis>.

1. Introduction

Occupying over two-thirds of Earth's surface, the sea is crucial for human exploration, where precise underwater depth acquisition is essential. This holds particularly true for fields such as autonomous underwater vehicles (AUV) [8, 32], underwater robotics [47], marine biology, ecology [19] and archaeology [4, 10]. Unlike costly and operationally complex active ranging equipment, such as under-

water LiDARs [16, 49], monocular depth estimation offers a more cost-effective and convenient deployment solution. Despite significant advancements in monocular depth estimation for terrestrial applications [7, 15, 17, 20, 21, 36], underwater depth estimation remains challenging due to factors like light attenuation, backscatter, and water turbidity [2, 5, 26], which lead to poor image quality and imprecise depth data. The scarcity of data hampers the training of powerful learning based models.

While some datasets like Sea-thru [2] and SQUID [5] offer real underwater data, they are costly to acquire thus are limited in scene diversity and scale. Their depth data, derived from stereo pairs or video sequences, is often sparse and not entirely reliable. GAN-based methods have emerged as an alternative, synthesizing underwater images by transferring styles from terrestrial scenes using image formation models [23, 27]. Despite they provide a remedy for the data scarcity issue because 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 paired with accurate depth data. Compared to aforementioned real datasets, it is inexpensive and easy to obtain, featuring large diversity and theoretically unlimited scale. Utilizing advancements in Stable Diffusion (SD) [38] and ControlNet [48], this approach allows for the generation of underwater imagery following the scene structure and layout of terrestrial depth. Despite their widespread applications in AI-generated content, these technologies have rarely been used for generating training data. We present a dataset that combines the accuracy of terrestrial depth with the lifelike depiction of underwater scenes, offering a robust resource for training reliable depth estimation models for unseen underwater scenes. This dataset not only serves as a bridge between terrestrial and underwater domains but also demonstrates its utility in image restoration techniques like Sea-Thru [2].

Specifically, we first construct a dataset comprising underwater images, estimated depths, and captions that describe the image content. Then we train a *Underwater2Depth* ControlNet targeting realistic underwater image generation using depth map. Using the pretrained SD and our trained ControlNet, we generate an underwater depth dataset comprising realistic underwater images and accurate depth, enabling the training of terrestrial depth estimation models for underwater depth estimation. Compared the terrestrial counterparts of KITTI [18] and NYU Depthv2 [39], their performance are largely improved both quantitatively and qualitatively. We also show the value of our dataset in applying the trained depth model for underwater image enhancement using Sea-thru algorithm [2]. It is important to note that our goal is not necessarily to surpass

the results of robust terrestrial depth models trained with abundant mixed sourced data and training tricks *e.g.*, MiDaS [36] and ZoeDepth [7] on underwater scenes, but to enable existing depth models on underwater scenes with our data and simple 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, is easy to collect and extend, comprising realistic underwater images and reliable depth, and featuring large diversity and theoretically unlimited scale.
- We propose to improve the performance of existing depth models on unseen underwater scenes using our proposed dataset for training. The improved depth can further be applied for underwater image enhancement, which highlighting the effectiveness and utility of our dataset.

2. Related Work

In the evolving field of monocular depth estimation, significant strides have been made through diverse methodologies. This section reviews key developments in terrestrial monocular depth estimation, explores current underwater depth estimation techniques, and introduces methods integrating underwater depth estimation with image restoration.

2.1. Terrestrial Depth Estimation

Eigen *et al.* [15] pioneered the coarse-to-fine network approach for end-to-end monocular depth estimation, a significant breakthrough. Their Scale-Invariant log loss is widely adopted in subsequent methods. Monodepth [20] and Monodepth2 [21] achieved impressive self-supervised performance and robustness. DORN [17] and Adabins [6] represent methods that treat depth estimation as ordinal regression or classification, discretizing depth. Recently, MiDaS [36] set a new benchmark for robust zero-shot depth estimation using multi-source mixed data training and various optimization techniques. DPT [37] and ZoeDepth [7] further enhanced performance in relative and absolute depth metrics, respectively. NeWCRFs [46] and iDisc [33] introduced fully-connected CRFs and an Internal Discretization module, respectively, for depth estimation. However, these models' performance in underwater scenes is limited due to domain gaps and data scarcity.

2.2. Underwater Depth Estimation

Underwater, light attenuation and backscatter depends on the distance light travels through water. Image formation models [2, 14, 25, 31] 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

| Dataset | Real | GAN-based | Our Atlantis |
|-------------|--|---|---|
| Image | | | |
| Depth | | | |
| Pros & Cons | Expensive to acquire Real Image Sparse depth Limited scale Low diversity | Easy to acquire Unrealistic image Dense depth Large scale Large diversity | Easy to acquire Realistic Image Dense Depth Unlimited scale Large diversity |

Figure 2. Comparisons of real dataset [5], GAN-based synthetic dataset [23] and ours proposed underwater depth dataset Atlantis.

techniques of DCP family [13, 24], therefore, can estimate depth. Gupta and Mitra [22] proposed UW-Net that utilizes the GAN for unsupervised training. Li *et al.* [27] and Hambarde *et al.* [23] proposed to synthesize different types of underwater images using the image formation model [9] and NYU Depthv2 [39], focusing on image enhancement and depth estimation, respectively. Recent work has also explored lightweight models [45] and self-supervised learning [3, 44]. 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 modeling techniques.

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, 13, 24], often integrating depth-related aspects. Recent learning-based methods [12, 40, 41] 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 de-watering results using range maps. Our dataset’s potential to enhance depth estimation performance is validated through its application in the Sea-thru algorithm.

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 in scale. 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 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 our proposed dataset 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 synthesized underwater imagery methods.

3.2. Underwater Depth Dataset: Atlantis

In the creation of our underwater depth dataset, illustrated in Figure 3, we initiate by constructing an intermediary dataset that is instrumental in training a specialized ControlNet [48]. This tailored ControlNet is then utilized to guide the pretrained Stable Diffusion v1.5 [38] in generating underwater images informed by outdoor depth maps.

Data Preparation. Our process begins with the utilization of the robust MiDaS [36] model to estimate inverse relative depth for images from the UIEB dataset [26], following ControlNet [48] procedure. For each underwater image U , a corresponding depth map D is obtained as follows:

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

where \mathcal{F}_{MiDaS} denotes the pretrained MiDaS model. Additionally, each image U undergoes captioning using the pre-trained BLIP2 model [28] to generate descriptive text T :

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

This leads to the formation of our intermediate dataset, comprising $\{\text{Underwater}, \text{Depth}, \text{Text}\}$ 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

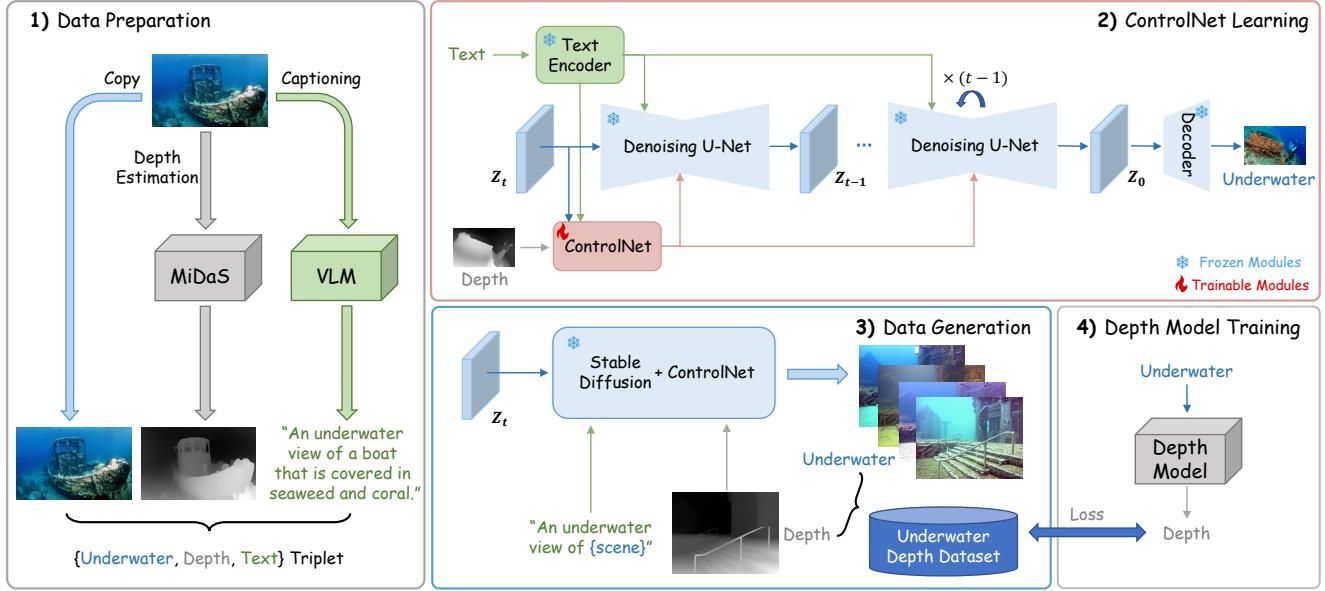


Figure 3. Overview of our method for generating the underwater depth dataset. The process begins by creating an intermediate dataset containing **underwater** images, **depth maps**, and **text** descriptions. This dataset is used to train the *Depth2Underwater* ControlNet for generating underwater images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

content generation. During the training stage, only ControlNet is set as trainable and other parts of SD is 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:

$$\mathbf{c} = \mathcal{F}_{CtrlNet}(\mathbf{z}_t, D, T), \quad (3)$$

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

$$\bar{U} = \mathcal{F}_{SD}(\mathbf{z}_t, T | \mathbf{c}), \quad (4)$$

yielding the generated underwater image \bar{U} . $\mathcal{F}_{SD}(\cdot)$ denotes the generation process of pretrained SD conditioned on 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 pre-trained SD model with our *Depth2Underwater* ControlNet. Utilizing 400 terrestrial images from the DIODE-outdoor dataset [42] 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' :

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

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

3.3. Depth Uncertainty

Despite the excellent performance and robustness of MiDaS [36] model in general depth estimation tasks, handling underwater scenes invariably introduces a domain gap. This gap stems from the distinct and challenging visual characteristics of underwater environments, which are not typically represented in the model’s training data. It can lead to inaccuracies in depth maps estimated for out-of-distribution underwater images. To address this challenge, we propose the Depth Uncertainty (DU) metric as a means to quantify and mitigate the inaccuracies arising from this domain gap.

Depth Uncertainty (DU). The DU metric measures the variance in depth estimations produced by the MiDaS model for both the original underwater images and their horizontally flipped counterparts. This variance reflects the model’s consistency under varied input conditions, as depth models often exhibit inconsistent results for flipped images,

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

| Models | $RMSE \downarrow$ | $RMSE_{log} \downarrow$ | $A.Rel \downarrow$ | $S.Rel \downarrow$ | $log_{10} \downarrow$ | $SI_{log} \downarrow$ | $\delta_1 \uparrow$ | $\delta_2 \uparrow$ | $\delta_3 \uparrow$ |
|--------------------|-------------------|-------------------------|--------------------|--------------------|-----------------------|-----------------------|---------------------|---------------------|---------------------|
| IDisc-KITTI | 5.891 | 1.192 | 4.702 | 44.288 | 0.489 | 35.846 | 0.093 | 0.241 | 0.359 |
| IDisc-NYUDepthv2 | 3.144 | 0.845 | 0.819 | 2.471 | 0.338 | 37.296 | 0.215 | 0.403 | 0.504 |
| IDisc-Atlantis | 1.371 | 0.354 | 1.630 | 14.279 | 0.109 | 34.654 | 0.553 | 0.850 | 0.955 |
| NewCRFs-KITTI | 3.251 | 0.934 | 2.874 | 15.768 | 0.365 | 42.341 | 0.213 | 0.375 | 0.465 |
| NewCRFs-NYUDepthv2 | 3.390 | 0.955 | 0.770 | 2.350 | 0.372 | 47.667 | 0.179 | 0.365 | 0.479 |
| NewCRFs-Atlantis | 1.435 | 0.378 | 1.683 | 14.764 | 0.120 | 37.101 | 0.476 | 0.837 | 0.952 |

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

| Models | $RMSE \downarrow$ | $RMSE_{log} \downarrow$ | $A.Rel \downarrow$ | $S.Rel \downarrow$ | $log_{10} \downarrow$ | $SI_{log} \downarrow$ | $\delta_1 \uparrow$ | $\delta_2 \uparrow$ | $\delta_3 \uparrow$ |
|--------------------|-------------------|-------------------------|--------------------|--------------------|-----------------------|-----------------------|---------------------|---------------------|---------------------|
| IDisc-KITTI | 7.265 | 0.736 | 1.039 | 8.040 | 0.289 | 35.827 | 0.156 | 0.349 | 0.555 |
| IDisc-NYUDepthv2 | 8.752 | 1.638 | 0.737 | 6.454 | 0.683 | 41.097 | 0.016 | 0.046 | 0.093 |
| IDisc-Atlantis | 2.663 | 0.277 | 0.249 | 0.920 | 0.094 | 27.221 | 0.637 | 0.900 | 0.960 |
| NewCRFs-KITTI | 6.692 | 0.779 | 0.579 | 3.930 | 0.294 | 52.091 | 0.197 | 0.381 | 0.541 |
| NewCRFs-NYUDepthv2 | 8.957 | 1.764 | 0.766 | 6.740 | 0.734 | 46.791 | 0.013 | 0.029 | 0.064 |
| NewCRFs-Atlantis | 2.563 | 0.256 | 0.229 | 0.830 | 0.088 | 25.189 | 0.675 | 0.902 | 0.964 |

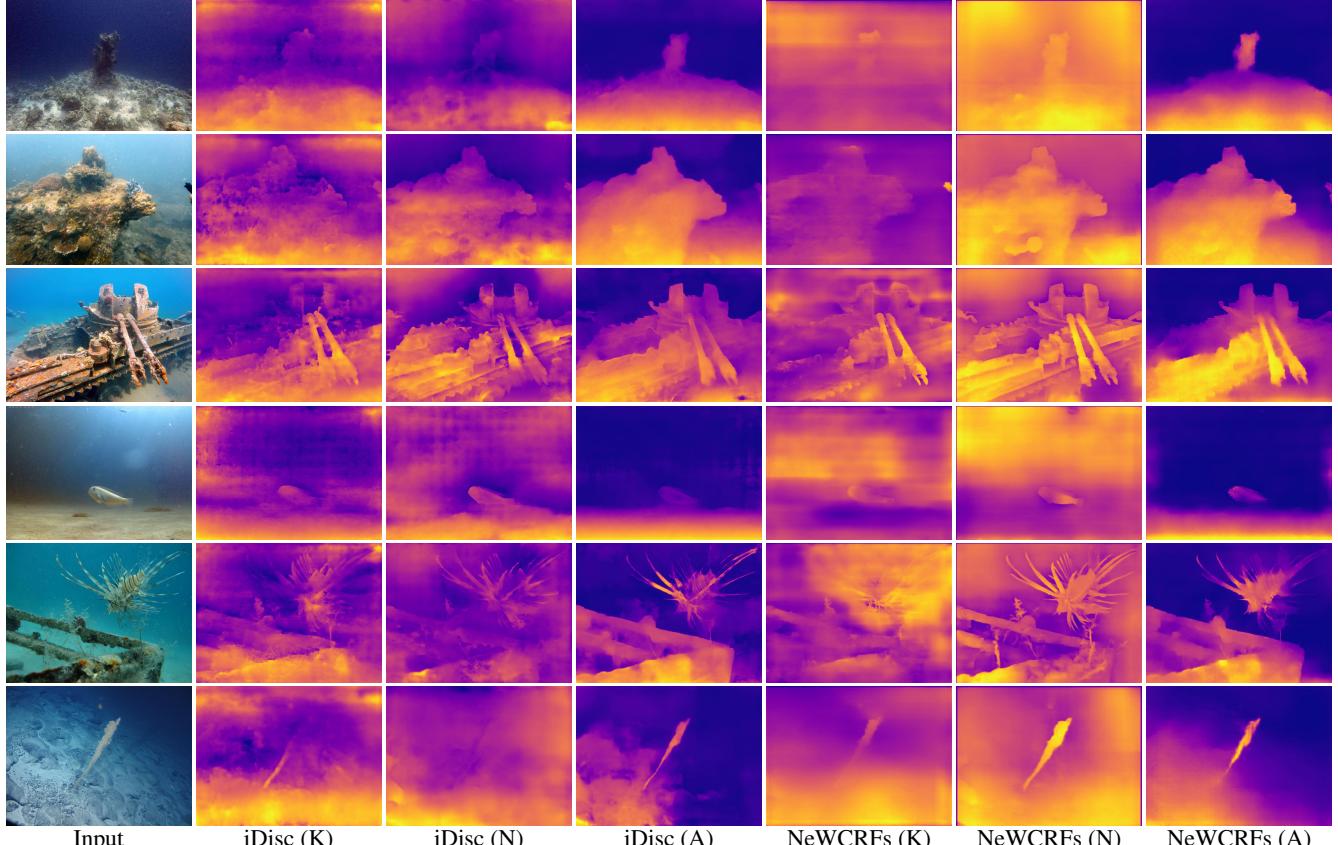


Figure 4. Qualitative results on test set of UIEB dataset [26]. K and N denote models pretrained on KITTI [18] and NYU Depthv2 [39] datasets. A represents the models trained on our dataset Atlantis. Our method gets the best visual results. Please zoom in for details.

a phenomenon leveraged in self-supervised training methods like Monodepth [20]. This is a crucial factor considering depth models like MiDaS are typically less exposed to underwater imaging conditions. For each original underwater image \bar{U} and its flipped version \bar{U}^{lr} , the DU for each pixel location p is calculated as follows:

$$DU_p = Var(D_p, D_p^{lr}), \quad (6)$$

where DU_p represents the per-pixel variance between the depth maps \bar{D} and \bar{D}^{lr} . This variance provides a quantitative measure of the depth estimation reliability [34] in the face of the domain gap.

Validity Mask. To ensure the reliability of our depth data, we introduce a Validity Mask, filtering out depth values at pixel locations where the DU exceeds a certain threshold.

This threshold is empirically set at 0.15, allowing us to identify and discard depth values with high uncertainty. Consequently, only depth values with $DU < 0.15$ are considered reliable, enhancing the overall quality and dependability of the depth information used in our dataset.

3.4. 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 leveraged the training set of UIEB dataset [26], which consists of 700 underwater images, for initial depth estimation and captioning. The robust MiDaS model [36] was employed for depth estimation, while the BLIP2 model [28] facilitated image captioning. These steps resulted in 700 data triplets comprising underwater images, depth maps, and textual descriptions, forming the foundation of our training data for ControlNet.

ControlNet Training and Deployment. We utilized the diffusers library [43] 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, avoiding unrealistic lighting styles, and sample for 20 steps for each image generation.

Depth Estimation Model Training. For the training of depth estimation models, we employed recent iDisc [33] and NeWCRFs [46]. These models were trained on our generated underwater depth dataset. Given that MiDaS outputs inverse relative depth, we capped the depth values at a maximum of 20 meters. This aligns with the understanding that scene radiance in underwater environments is predominantly affected by backscattering beyond this depth [2].

Hardware and Accessibility. All experiments and model trainings were conducted on an NVIDIA RTX 3090 GPU. We plan to make both the intermediate triplet data and underwater depth dataset, as well as the customized *Depth2Underwater* ControlNet publicly available, contributing to the broader research community in this field.

4. Experiments

In this section, we demonstrate the effectiveness of our novel underwater depth dataset in training supervised depth estimation models, both quantitatively and qualitatively. We compare models trained from scratch on our dataset, specifically iDisc [33] and NeWCRFs [46], with their officially pretrained counterparts on terrestrial datasets, namely KITTI [18] and NYU Depthv2 [39]. This comparison is conducted on unseen underwater datasets to highlight performance differences. Additionally, to showcase the practical application of depth models trained on our dataset, we utilize the Sea-thru algorithm [2], originally designed for underwater image enhancement using depth maps obtained

from stereo images or video sequences, and apply it to single images with estimated depth. Finally, we conduct a comparison to show smaller domain gap of our dataset compared to previous synthetic dataset. Due to limited space, we provide more visual results in supplementary material.

Experimental Setup. We focus on two models: iDisc [33] and NeWCRFs [46]. Both models were trained from scratch on our dataset and evaluated against their official versions pretrained on the KITTI [18] and NYU Depthv2 [39] datasets. Each utilizes the SwinL model [30] pretrained on ImageNet 22k [11] for encoder initialization. Quantitatively, we conducted evaluations using the Sea-thru’s D3 and D5 subsets [2] and the SQUID dataset [5], which includes underwater images with depth maps obtained via Structure-from-Motion (SfM) algorithm. For qualitative assessments, we additionally selected the UIEB dataset’s test set [26] to complement the diversity of tested scenes for visual comparison. 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 terrestrial datasets of KITTI [18] and NYU Depthv2 [39] when applied to underwater images. This domain gap, which adversely affects performance across most metrics, was evident in both iDisc [33] and NeWCRFs [46] models, underscoring the inherent challenges in applying supervised monocular depth models to underwater scenes. Conversely, when these models were trained from scratch on our underwater depth dataset, both iDisc and NeWCRFs exhibited substantial improvements across the majority of quantitative metrics. This improvement was consistent across evaluations on the Sea-thru [2] and SQUID [5] datasets, affirming the efficacy of our dataset in enhancing monocular depth estimation for unseen underwater scenes. This outcome suggests that training on our dataset effectively reduces the domain gap. It is noteworthy that our dataset, despite being smaller in size compared to the terrestrial datasets, has shown significant potential in this context. This indicates that expanding the dataset further or employing hybrid training approaches could 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. For un-

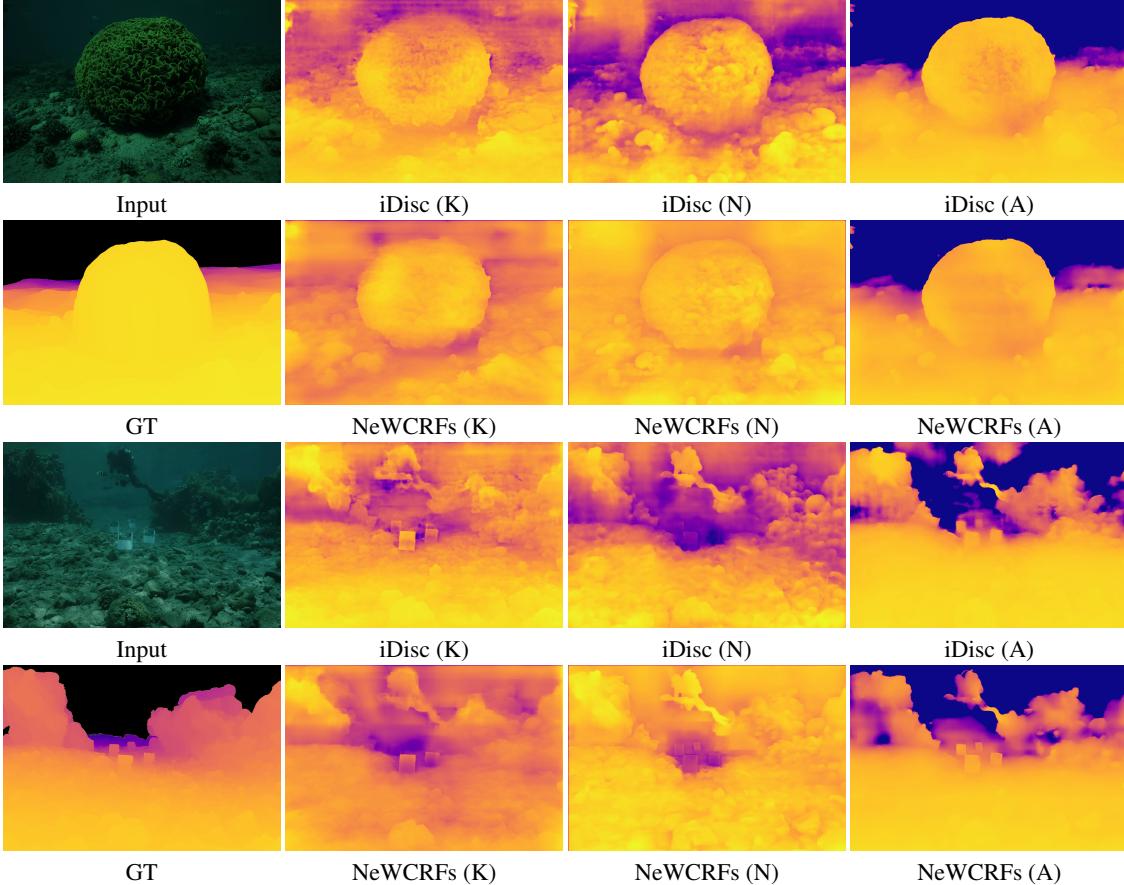


Figure 5. Qualitative results on test set of Sea-thru dataset [2]. K and N denote models pretrained on KITTI [18] and NYU Depthv2 [39] datasets. A represents the models trained on our dataset Atlantis. Our method gets the best visual results. Please zoom in for details.

derwater images, pretrained models on terrestrial datasets, including both iDisc [33] and NeWCRFs [46], produce significantly erroneous results. These inaccuracies manifest as heavy haze artifacts in water body areas and incorrect relative scene layout distances. In sharp contrast, when trained on our specifically designed underwater depth dataset, both models exhibit a remarkable improvement in interpreting underwater scenes. Notably, they accurately identify and appropriately distance water body areas, demonstrating enhanced discrimination 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 our dataset in enabling depth estimation models to better differentiate water bodies and adapt to diverse underwater conditions, including color casts and lighting variations. It's important to note that the underwater images used in these comparisons were not part of the training dataset. This further emphasizes the generalization capability of our dataset in training robust depth estimation

models that effectively adapt to real underwater scenes.

4.3. Sea-thru Enhancement with Depth Models

We demonstrate the application of our dataset in training reliable underwater depth models for effective underwater image enhancement using the Sea-thru algorithm [2]. Known for its ability to remove water effects with precise range maps derived from stereo pairs or video sequences, Sea-thru's capabilities are extended to single underwater images using depth maps estimated by models trained on our dataset. As depicted in Figure 6, the Sea-thru algorithm, when equipped with depth estimates from our models, produces impressive underwater image enhancements. These results not only showcase the models' accuracy in depth estimation but also reaffirm the practical utility and effectiveness of our dataset in real-world applications.

4.4. Domain Gap from LLVM Perspective

The advent of Large Language Vision Models (LLVM) [29, 35] has revolutionized the alignment between textual and visual features, opening new avenues for synthetic data analysis. We utilize recent LLAVA v1.5 model [29] to re-

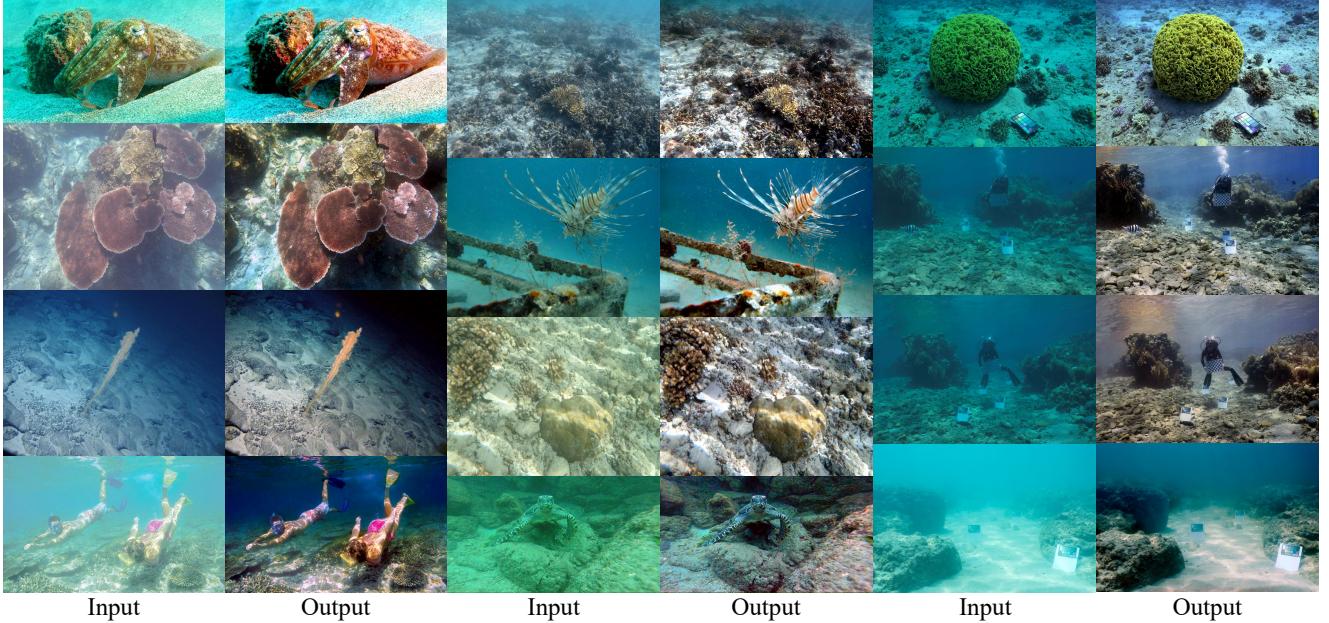


Figure 6. Qualitative results of the improved depth result applied to downstream underwater image enhancement. The left and middle parts are from UIEB dataset [26] and the right part contains images from Sea-thru dataset [2] (the above three) and SQUID dataset [5] (the bottom one). The enhancement method adopts the unofficial Sea-thru algorithm implementation¹. Enhancement outputs well show the effectiveness of the proposed dataset on training depth models for reliable underwater depth estimation.

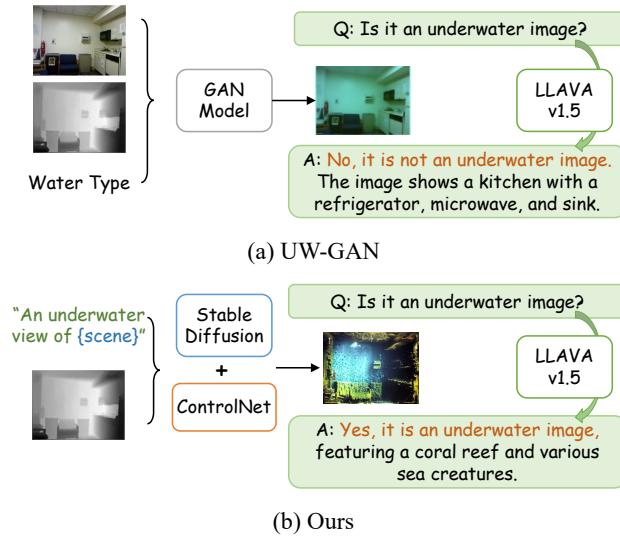


Figure 7. (a) GAN-based methods style transfer images without changing the scene content, resulting in images still recognized as room scenes by LLVM. (b) In contrast, our method generates novel underwater images that maintain the original scene structure, correctly identified as underwater scenes by LLVM.

veal a disconnect in how LLVM perceives synthetic underwater images generated using conventional depth and image formation models. As depicted in Figure 7(a), these images are often not recognized as underwater scenes, signaling a gap in the current synthesis approach. However, SD

[38] and ControlNet [48] stand out in their ability to generate highly realistic images guided by textual prompts, a testament to the advancements in aligning natural language with visual content. Figure 7(b) illustrates how LLAVA effectively recognizes images generated through this method as authentic underwater scenes, confirming smaller domain gap of our proposed dataset.

5. Conclusion

In this paper, we introduced a novel pipeline utilizing Stable Diffusion and a customized ControlNet for generating realistic underwater images with accurate depth. We proposed 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 features easy acquisition, realistic underwater images and accurate depth, large diversity and theoretically unlimited scale. Our experiments, encompassing both quantitative and qualitative analyses, demonstrated the superiority of models trained on our dataset compared to those pretrained on terrestrial datasets. Notably, the application of these models in the Sea-thru algorithm for single underwater image enhancement showcased their practical utility and highlighted 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 exploring hybrid training approaches could unlock greater improvements in model performance and generalization.

¹<https://github.com/hainh/sea-thru>

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 Birk, 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 *Ieee Icra*, 2001. 1
- [9] John Y Chiang and Ying-Ching Chen. Underwater image enhancement by wavelength compensation and dehazing. *IEEE TIP*, 21(4):1756–1769, 2011. 3
- [10] 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
- [11] 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. 6
- [12] 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
- [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] Seibert Q Duntley. Light in the sea. *JOSA*, 53(2):214–233, 1963. 2
- [15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 2014. 2
- [16] 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. 2
- [17] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 2
- [18] 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
- [19] 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
- [20] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017. 2, 5
- [21] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019. 2
- [22] Honey Gupta and Kaushik Mitra. Unsupervised single image underwater depth estimation. In *ICIP*, 2019. 3
- [23] 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
- [24] Kaiming He, Jian Sun, and Xiaou Tang. Single image haze removal using dark channel prior. *IEEE TPAMI*, 33(12):2341–2353, 2010. 3
- [25] Jules S Jaffe. Computer modeling and the design of optimal underwater imaging systems. *IEEE Journal of Oceanic Engineering*, 15(2):101–111, 1990. 2
- [26] 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, 3, 5, 6, 8
- [27] Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *PR*, 98:107038, 2020. 2, 3
- [28] 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. 3, 6
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 7
- [30] 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. 6
- [31] BL McGlamery. A computer model for underwater camera systems. In *Ocean Optics VI*, pages 221–231. SPIE, 1980. 2
- [32] 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
- [33] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. idisc: Internal discretization for monocular depth estimation. In *CVPR*, 2023. 2, 6, 7
- [34] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Maticocchia. On the uncertainty of self-supervised monocular depth estimation. In *CVPR*, 2020. 5

- [35] 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. 2021. 7
- [36] 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, 3, 4, 6
- [37] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 2
- [38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 8
- [39] 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
- [40] 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 *Advances in Multimedia Information Processing – PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part I* 19, 2018. 3
- [41] 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
- [42] 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
- [43] 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. 6
- [44] Xuewen Yang, Xing Zhang, Nan Wang, Guoling Xin, and Wenjie Hu. Underwater self-supervised depth estimation. *Neurocomputing*, 514:362–373, 2022. 3
- [45] Boxiao Yu, Jiayi Wu, and Md Jahidul Islam. Udepth: Fast monocular depth estimation for visually-guided underwater robots. In *ICRA*, 2023. 3
- [46] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *CVPR*, 2022. 2, 6, 7
- [47] Junku Yuh and Michael West. Underwater robotics. *Advanced Robotics*, 15(5):609–639, 2001. 1
- [48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3, 8
- [49] 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. 2