sensors

MDPI

Communication

Unpaired Underwater Image Synthesis with a Disentangled Representation for Underwater Depth Map Prediction

Qi Zhao [1], Zhichao Xin [1], Zhibin Yu [1,2,*] and Bing Zheng [1,2]

[1] College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China; zq4984@stu.ouc.edu.cn (Q.Z.); xinzhichao@stu.ouc.edu.cn (Z.X.); bingzh@ouc.edu.cn (B.Z.)
[2] Sanya Oceanographic Institution, Ocean University of China, Sanya 572000, China
* Correspondence: yuzhibin@ouc.edu.cn

Abstract: As one of the key requirements for underwater exploration, underwater depth map estimation is of great importance in underwater vision research. Although significant progress has been achieved in the fields of image-to-image translation and depth map estimation, a gap between normal depth map estimation and underwater depth map estimation still remains. Additionally, it is a great challenge to build a mapping function that converts a single underwater image into an underwater depth map due to the lack of paired data. Moreover, the ever-changing underwater environment further intensifies the difficulty of finding an optimal mapping solution. To eliminate these bottlenecks, we developed a novel image-to-image framework for underwater image synthesis and depth map estimation in underwater conditions. For the problem of the lack of paired data, by translating hazy in-air images (with a depth map) into underwater images, we initially obtained a paired dataset of underwater images and corresponding depth maps. To enrich our synthesized underwater dataset, we further translated hazy in-air images into a series of continuously changing underwater images with a specified style. For the depth map estimation, we included a coarse-to-fine network to provide a precise depth map estimation result. We evaluated the efficiency of our framework for a real underwater RGB-D dataset. The experimental results show that our method can provide a diversity of underwater images and the best depth map estimation precision.

Keywords: underwater image synthesis; underwater depth map estimation; image-to-image translation

Citation: Zhao, Q.; Xin, Z.; Yu, Z.; Zheng, B. Unpaired Underwater Image Synthesis with a Disentangled Representation for Underwater Depth Map Prediction. *Sensors* **2021**, *21*, 3268. https://doi.org/10.3390/s21093268

Academic Editor: Raul Marin Prades

Received: 22 March 2021
Accepted: 6 May 2021
Published: 9 May 2021

1. Introduction

In 3D computer vision, a depth map refers to a frame in which each pixel represents the distances of the surfaces of objects in a scene from a viewpoint. There are a number of uses for depth maps, including machine vision, 3D reconstruction, and shadow mapping [1]. As an important branch of underwater vision, underwater depth map estimation plays an important role in many fields, including underwater landform surveys, vehicle navigation, and underwater hull cleaning. Although considerable progress has been achieved in screening-laser-technology-based underwater 3D reconstruction [2], many approaches have the limitation that the patterns cannot be changed online [3]. In addition, calibration-based methods can be affected by the index of refraction transformation [4]. Some in-air depth map estimation devices, such as the Kinect [5], Lidar [6], or monocular lenses [7], can only obtain a limited effect in an underwater environment [8]. The major challenge comes from the complicated underwater environment. Most underwater images are captured with low contrast due to the scattering and absorption degradation caused by underwater particulates [9]. Inhomogeneous illumination further intensifies the problem of color distortion in underwater images.

While deep-learning-based methods have achieved great success in the field of computer vision [10,11], the progress is still considerably limited in the field of image-based underwater depth map estimation. The lack of data is a major challenge when deploying a

deep learning model with supervised learning for underwater depth map estimation. Collecting underwater images is expensive and time consuming, as is the collection of paired underwater RGB-D data containing underwater images and corresponding depth maps. The success of generative adversarial networks (GANs) in the field of image-to-image translation [12–15] provides a feasible way to translate images between two domains or multiple domains in an unsupervised manner.

At present, many researchers are attempting to synthesize underwater images with in-air RGB-D images to build paired datasets for underwater image color restoration [16–18] or depth map estimation [10,11,19]. For instance, WaterGAN [16] and UWGAN [20] input a paired in-air RGB-D image into a physical-model-based generator such that the final output is a synthesized underwater image produced by the generator [10,11]. However, these methods adopt a two-stage training strategy in which the modules for underwater depth map estimation and synthesis of underwater images are isolated, thus ignoring the latent relationship between visual images and depth information.

In a recent work, a method called UW-Net [11] was constructed in a single-stage network with two generators to simultaneously synthesize an underwater image and estimate an underwater depth map. However, all of these models attempted to build a function for mapping from the synthetic images to the target domain by using one single network, which led to poor performance in terms of both depth map estimation and image synthesis tasks. Moreover, none of the methods mentioned above could generate various underwater images with disentangled representations, which may lead to an inefficient use of training data and a lack of diversity in underwater image synthesis. In order to solve these problems, we propose a novel image-to-image translation framework for underwater image synthesis and depth map estimation. A discussion of our motivations is presented in the following.

In practice, it is relatively easy to obtain unlabeled underwater images from the internet. These images may include rich information on various underwater conditions, which may help our synthetic framework in generating underwater images with a rich diversity. However, labeling these images is a time-consuming task. Inspired by the success of InfoGAN [21] and its extensions [22], we redesigned the loss functions of our framework to include interpretative disentangled representations of various underwater conditions, including the illumination and water color.

Due to the decreased visibility and lack of references, another practical problem of our underwater depth map estimation task is that objects at different distances cannot show uniformly show precise information. Therefore, we adopted a multi-depth estimator mechanism to accomplish coarse-to-fine adjustment. As Figure 1 shows, our two depth generators are responsible for the global–coarse depth map estimation and local–fine depth map estimation, respectively. With the depth map passing though these two generators, depth information is refined and forces the generators to pay attention to nearby objects. Overall, the main contributions of this paper are summarized as follows:

- We propose a novel end-to-end framework that applies image-to-image translation to underwater depth map estimation and further boosts current underwater depth map estimation research.
- To enrich our synthesized underwater dataset, we propose a disentangled representation loss along with style diversification loss to identify interpretable and meaningful representations from the unlabeled underwater dataset and the synthesized underwater images with a rich diversity.
- Following the coarse-to-fine principle, and inspired by the work of Eigen et al. [23] and Skinner et al. [19], our approach adopted global–local generators for the estimation of coarse and fine depth maps, respectively. We evaluated our model on a real underwater RGB-D dataset and achieved better results than those of other state-of-the-art models.

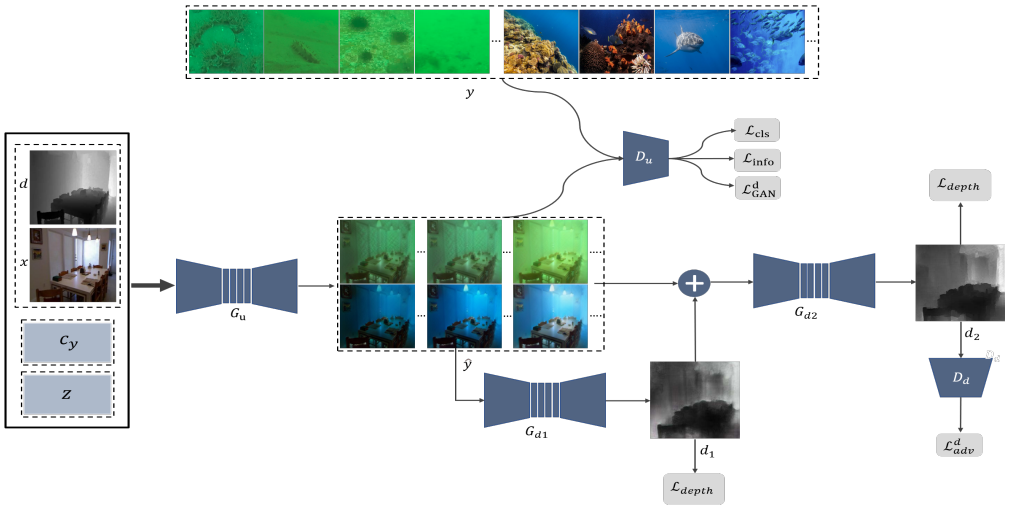


Figure 1. The network framework of our proposed model was designed to synthesize multiple underwater images and estimate underwater depth maps. We used the generator G_u and the discriminator D_u to synthesize various underwater images in the given underwater domain c_y. We designed the generators G_{d1} and G_{d2} and the discriminator D_d to learn to estimate underwater depth maps based on the synthesized underwater RGB-D dataset.

2. Methods

2.1. Overall Framework

Because supervised learning could not be directly performed due to the lack of paired underwater RGB-D images, we designed a two-stage model, as described in Figure 1. Our model includes two cascades: an underwater image synthesis module and an underwater depth map estimation module. The first underwater image synthesis module can translate an original in-air image with its corresponding depth into the underwater domain with disentangled representations to generate various underwater RGB-D pseudo-pairs. The synthetic pseudo-pairs were further used to provide the underwater depth map estimation module with supervised learning through a coarse-to-fine process. Our overall framework consists of three generators, namely, $G_u : (x, d, c_y, z) \rightarrow \tilde{y}$, $G_{d1} : \tilde{y} \rightarrow d_1$, and $G_{d2} : (\tilde{y}, d_1) \rightarrow d_2$, where x represents the original in-air images, d is the corresponding depth map, c_y is the target underwater domain, z is the continuous noise vector, $\tilde{y}$ is the generated underwater image, d_1 represents the global results of the underwater depth map estimation, and d_2 is the final estimated depth map. According to the two tasks, we also designed two discriminators, D_u and D_d. D_u aims to distinguish real and fake underwater images and classify their corresponding domains in the real and fake underwater images. The discriminator D_d only aims to distinguish real and fake underwater depth maps.

Underwater image synthesis with disentangled representation. We referred to StarGAN [15] and InfoGAN [21] to design the underwater image synthesis module. We defined a random noise vector (z) and target domain label vector (c_y) to produce multiple outputs in a specific domain. To ensure that the generated underwater images preserved the original depth information after translation, the inputs of our module included four parts, namely, the in-air image (x), the corresponding depth (d), the target underwater label (c_y), and the noise vector (z), to synthesize an underwater image $\tilde{y} = G_u(C(x, d, c_y, z))$, where C represents depth-wise concatenation. The generator G_u was taken from CycleGAN [12] and StarGAN [15]. To guarantee that the synthetic image $\tilde{y}$ belonged to the target domain c_y, we designed the discriminator D_u by following the PatchGAN [13] with three branches (domain classification, computation of naturalness, and limit of the coupling of noise (Z)). The domain classification loss L_{cls} was designed for the classification task of recognizing the underwater domain attributions (c_y) of the synthesized image $\tilde{y}$ and real underwater images y. Notably, y did not have the corresponding depth annotation due to the lack of an underwater ground truth. Furthermore, to force the noise vector z to represent and

control the disentangled information from the underwater environment, we also defined an auxiliary discriminator Q, which refers to InfoGAN [21].

The coarse-to-fine underwater depth map estimation process. According to the characteristics of underwater depth map estimations, we designed a coarse-to-fine generative adversarial network that includes two identical generators, G_{d1} and G_{d2}. Following the work on UW-Net [11], we also chose DenseNet [24] for the generators. Differently from UW-Net [11], each dense block [24] has five layers with eight filters. In the training stage, we took the synthetic underwater images $\tilde{y}$ from the synthetic module as the input of the coarse network G_{d1}. To obtain a broadly correct result, we adopted the L_1 norm, which makes equal contributions to distant and nearby points in a scene. Then, the output of the coarse generator $G_{d1}(\tilde{y})$ and the generated underwater images $\tilde{y}$ were used as the input of the fine generator G_{d2} to obtain a better depth map $G_{d2}(C(G_{d1}(\tilde{y}), \tilde{y}))$. Unlike the coarse prediction task in G_{d1}, we also introduced the L_{depth} loss to guide the fine generator G_{d2} for more in-depth observations. Specifically, the discriminator D_d was a PatchGAN [13] with only one discrimination output.

2.2. Loss Functions

Adversarial Loss. As an extension of a conditional GAN, the conditional generative adversarial loss [25] was used as a basic component of our loss functions. During the training process, the generator G_u took hazy in-air RGB-D image pairs (x, d), the target domain label c_y, and the continuous noise vector z as inputs, and it learned to generate underwater images G_u(x, d, c_y, z) through adversarial loss [26]. L^u_{GAN} can be expressed as follows:

$$\begin{aligned} \mathcal{L}^u_{GAN} = \min_G \max_D \{ & \mathbb{E}_{x,y\sim P_{data}(x,y)}[(D_u(y)-1)^2] \\ & + \mathbb{E}_{x\sim P_{data}(x)}[(D_u(\tilde{y}))^2]\}, \\ & where \quad \tilde{y} = G_u(\mathcal{C}(x,d,c_y,z)), \end{aligned} \tag{1}$$

where G_u aims to synthesize the multiple underwater images $G_u(\mathcal{C}(x,d,c_y,z))$ belonging to the target domain c_y. The discriminator D_u learns to distinguish the real underwater image y and the synthesized underwater image $\tilde{y}$. For underwater depth map estimation, the adversarial loss $\mathcal{L}^d_{GAN}$ is described as:

$$\begin{aligned} \mathcal{L}^d_{GAN} = \min_G \max_D \{ & \mathbb{E}_{\tilde{y},d\sim P_{data}(\tilde{y},d)}[(D_d(d)-1)^2] \\ & + \mathbb{E}_{\tilde{y}\sim P_{data}(\tilde{y})}[(D_d(d_2))^2]\}, \\ & where \quad d_2 = G_{d2}(\mathcal{C}(G_{d1}(\tilde{y}),\tilde{y})), \end{aligned} \tag{2}$$

where the G_{d1} output is a global depth map d_1 from the synthesized underwater images $\tilde{y}$. Based on the output of G_{d1}, G_{d2} attempts to fine-tune the results. D_d learns to recognize the estimated depth output d_2 from the inputs.

Feature-matching loss. In the process of underwater image synthesis, to preserve the object content of the original in-air images and to pair the contents of the synthesized underwater images and their corresponding in-air depth maps, a feature-level loss function [14,27] was introduced, which is called $\mathcal{L}_{feat}$. The loss is based on a pre-trained VGG19 network [28] that extracts the feature representations from fake and real underwater images. It can effectively preserve the content of the objects between the original images x and the generated underwater images $\tilde{y}$. Moreover, it only changes the domain-related parts of the original images and does not have any negative effects on underwater image synthesis. $\mathcal{L}_{feat}$ is expressed as follows:

$$\mathcal{L}_{feat} = \sum_{i=0}^{N} \frac{1}{M_i}[||\Phi^{(i)}(x) - \Phi^{(i)}(G_u(x,d,c_y,z))||_1]. \tag{3}$$

where $\Phi^{(i)}$ denotes the feature maps at the i-th layer with M_i elements of a pre-trained VGG19 network [28]. The parameters that we set can be found in the work of Kupyn et al. [29].

Domain classification loss. Our model aims to generate multi-style underwater images and continuous outputs in a given underwater style. It involves two domain classification losses: discrete domain classification loss and continuous domain classification loss. Here, the domain classification loss is used to classify discrete domains. Inspired by UMGAN [10] and StarGAN [15], we included an optional domain classification loss to handle a classic domain classification task, which forces the synthetic sample $\tilde{y}$ to be generated in the target domain c_y. The domain classification loss L^r_{cls} is defined as follows:

$$\mathcal{L}^r_{cls} = \mathbb{E}_{y,c'}[-\log D_u(c'|y)]. \tag{4}$$

where the discriminator D_u learns to classify the real underwater images to their original domain c'. For generator G_u, the loss function for the domain classification of the synthetic underwater images is defined as:

$$\mathcal{L}^f_{cls} = \mathbb{E}_{\tilde{y},c_y}[-\log D_u(c_y|\tilde{y})]. \tag{5}$$

where the discriminator D_u attempts to classify the generated underwater images to their target underwater domain c_y.

Disentangled representation loss. To output continuous underwater images in a given underwater style, a continuous domain classification loss—namely disentangled representation loss—was designed. Inspired by InfoGAN [21], we included the disentangled representation loss to make the generator G_u extract various representations from real underwater images with a random noise vector z. The vector z could be set to either a binary or a decimal value according to the different tasks. In the test stage, the generator G_u could generate a controllable synthetic underwater image $\tilde{y}$ by using a specified latent vector z. The disentangled representation loss L_{info} can be expressed as:

$$\mathcal{L}_{info} = [||Q_u(\tilde{y}) - z||_2]. \tag{6}$$

Similarly to the model setting in InfoGAN, here, Q_u is a sub-network of the discriminator D_u.

Style diversification loss. As a supplement to the disentangled representation loss, we referred to StarGANv2 [30] and the style diversification loss L_{dis} to maximize the intra-domain distance in order to stabilize the training process and produce various outputs for a given input image pair (x, d) in a target domain c_y. We maximized the loss term and minimized the info loss force of G_u to generate multiple controllable underwater images in a given domain. The style diversification loss L_{dis} can be written as follows:

$$\mathcal{L}_{dis} = [||G_u(x, d, c_y, z_i) - G_u(x, d, c_y, z_j)||_1]. \tag{7}$$

where z_i and z_j represent the latent vectors of two samples.

Reconstruction loss. For unpaired image-to-image translation, the cycle consistency loss [12] is commonly used to preserve domain-invariant characteristics and stabilize the training process. In our model of underwater image synthesis, the reconstruction loss L_{rec} between the hazy in-air images x and reconstructed image $\tilde{x}$ is defined as follows:

$$\begin{aligned} \mathcal{L}_{rec} &= \mathbb{E}_{x,c_y,c_x}[||x - \hat{x}||_1], \\ \hat{x} &= G_u(\mathcal{C}(G_u(\mathcal{C}(x, d, c_y, z)), d, c', z)), \end{aligned} \tag{8}$$

Depth loss. Our coarse network G_{d1} estimates a global and coarse depth map d_1 from the generated underwater image $\tilde{y}$. Here, we adopted the general L_1 norm between the generated depth map d_1 and its ground truth d. The L_1 norm has an equal contribution between distant and nearby points in a scene. Separately, the fine network should pay more

attention to nearby points [31]. Therefore, we explored a loss to guide our coarse-to-fine network. So, the loss L_{depth} can be expressed as follows:

$$\mathcal{L}_{depth} = [||G_{d1}(\tilde{y}) - d||_1 + \frac{1}{n}\sum_{n}^{i=1} ln(||d_2 - d||_1 + 1)], \quad d_2 = G_{d2}(\mathcal{C}(G_{d1}(\tilde{y}), \tilde{y})), \tag{9}$$

where G_{d1} tries to globally estimate the depth map from the generated underwater images $\tilde{y}$. G_{d2} tries to locally fine-tune the depth map d_1. The final results are d_2 after fine-tuning.

Full objective. Our full objective functions can be written as follows:

$$\mathcal{L}_{D_u} = \mathcal{L}^u_{GAN} + \alpha\mathcal{L}^r_{cls} \tag{10}$$

$$\mathcal{L}_{G_u} = \mathcal{L}^u_{GAN} + \alpha\mathcal{L}^f_{cls} + \eta\mathcal{L}_{feat} + \gamma\mathcal{L}_{info} - \theta\mathcal{L}_{dis} + \beta\mathcal{L}_{rec} \tag{11}$$

$$\mathcal{L}_{D_d} = \mathcal{L}^d_{GAN} \tag{12}$$

$$\mathcal{L}_{G_d} = \mathcal{L}^d_{GAN} + \lambda\mathcal{L}_{depth} \tag{13}$$

where α, η, γ, θ, β, and λ are the hyperparameters for each term. We optimized these parameters with a greedy search and set $\alpha = 1, \eta = 1, \gamma = 0.1, \theta = 0.1, \beta = 1$, and $\lambda = 50$ in all of our experiments. The optimization of our model was successful.

3. Results

3.1. Datasets and Implementation Details

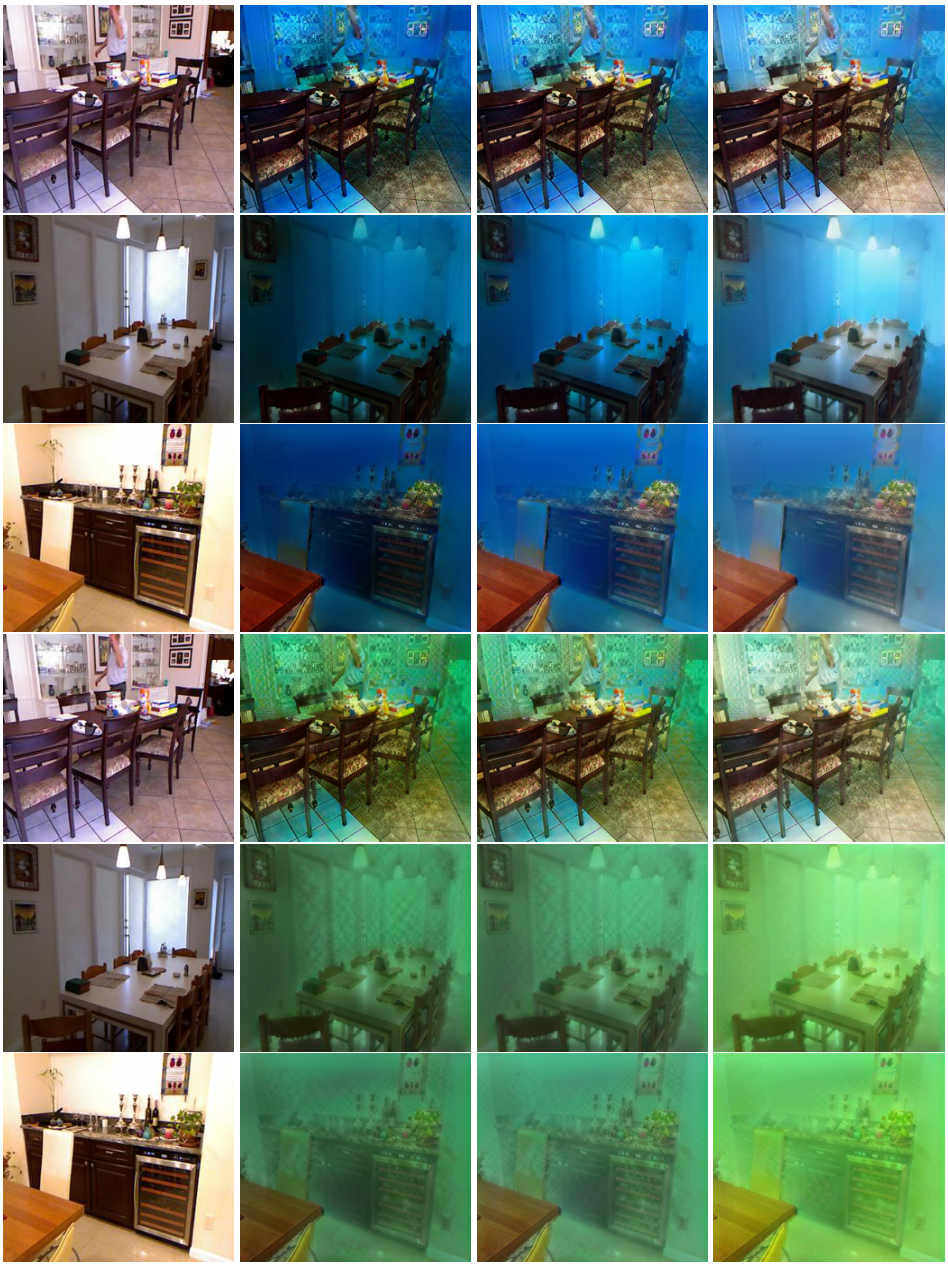
Our experiments mainly involved two tasks: underwater image synthesis and underwater depth map estimation. For the first task, we synthesized underwater images from hazy in-air RGB-D images and evaluated the image qualities with multiple image generation models, including WaterGAN [16], CycleGAN [12], StarGAN [15], UW-Net [11], and NICE-GAN [32]. For the second task, we evaluated our depth map estimation results with a real underwater RGB-D dataset. We compared the depth map estimation results obtained using the methods of dark channel prior (DCP) [33], underwater dark channel prior (UDCP) [34], Berman et al. [35], and Gupta et al. [11], as well as our method of underwater depth map estimation. Following the experimental setting of UW-Net [11], we also chose the D-Hazy dataset [36] as the in-air RGB-D images for the inputs. Note that both UW-Net and our model can be fine-tuned on the dataset of Berman et al.. The real underwater datasets for training contained 1031 blue and 1004 green underwater images from the SUN [37], URPC (http://www.cnurpc.org/ (accessed on 5 August 2019)), and Fish datasets (http://www.fishdb.co.uk/ (accessed on 7 October 2018)). We randomly chose 1400 images for the training dataset from the D-Hazy dataset [36], which includes 1449 paired in-air RGB-D images. The remaining pairs were used for evaluation. We took 128×128 patches for training and 256×256 complete images for testing. The training took about 40 h on one Nvidia GeForce GTX 1070 (8GB) using the Pytorch framework. To avoid mode collapse, we also introduced spectral normalization [38]. Following the work of BigGAN [39] and SAGAN [40], the learning rates were set to 0.0002 in the discriminators and 0.00005 in the generators. We set the batch size to 10, and the model was trained for 80,000 iterations in our experiments.

3.1.1. Qualitative Evaluation

To evaluate the effectiveness of the synthetic underwater images, we compared our method with other approaches on the NYU v2 [41] and D-Hazy datasets [36]. To show how close our synthetic images were to the real underwater images, we present some synthetic images in Figure 2. WaterGAN [16] refers to the underwater imaging process and takes in-air RGB-D images as input to synthesize underwater images. As shown in Figure 2b, the results of WaterGAN [16] are close to the in-air images and lack underwater

characteristics. In Figure 2c, the underwater images generated by CycleGAN [12] seem better than those of WaterGAN [16]. However, the results of CycleGAN [12] include serious structural distortions, such as the vase in the fifth row of Figure 2c. StarGAN [15] can simultaneously synthesize multi-style underwater images (Figure 2d), but the results still do not meet expectations due to the lack of depth information and clear structural information. In addition, the results retain many artifacts, such as the desk in the last row of Figure 2d. To retain the depth information for better underwater depth map estimation, UW-Net [11] takes the hazy in-air RGB-D images as input and uses DenseNet [24] for the generators, as shown in Figure 2e; this method shows a fuzzy structure. The results of NICE-GAN [32] can be seen in Figure 2f, and there are many artifacts in the results. Furthermore, most of the methods, including WaterGAN [16], CycleGAN [12], UW-Net [11], and NICE-GAN [32], are in two domains, and only StarGAN [15] can synthesize multi-style images. None of the above-mentioned methods consider the diversities in a given style. The synthetic underwater images from our method are shown in Figure 2g; the structure and depth information is well preserved. Our methods can simultaneously synthesize multi-style underwater images and use the noise z to produce multiple outputs with a target style, as shown in Figure 3. Here, we set z = 1, 0, −1. Overall, for underwater image synthesis, our method performed better and generated more diverse outputs than the other methods.

Figure 2. Comparison of the visual quality of the synthetic underwater images using the following methods: WaterGAN [16], CycleGAN [12], StarGAN [15], UW-Net [11], NICE-GAN [32], and our method.



(**a**) In-air images (**b**) z = 1 (**c**) z = 0 (**d**) z = −1

Figure 3. Sample results of our method for underwater image synthesis. The continuous noise z was used to generate multiple underwater images with a specific domain. (**a**) In-air images, (**b**–**d**) multiple underwater images generated in two specific domains (blue and green).

Following the work of UW-Net [11], we used the dataset from Berman et al. [35] to compare our method with other methods. Some results are shown in Figure 4. The former three methods are based on traditional physical processes that rely on pre-estimated parameters. Comparing them with the deep-learning-based UW-Net [11] and our method, we note that the latter two were able to obtain depth maps with smoother predictions. The predicted depth map of our method seems to be more accurate than that of UW-Net [11]. More qualitative results can be seen in Figure 5.

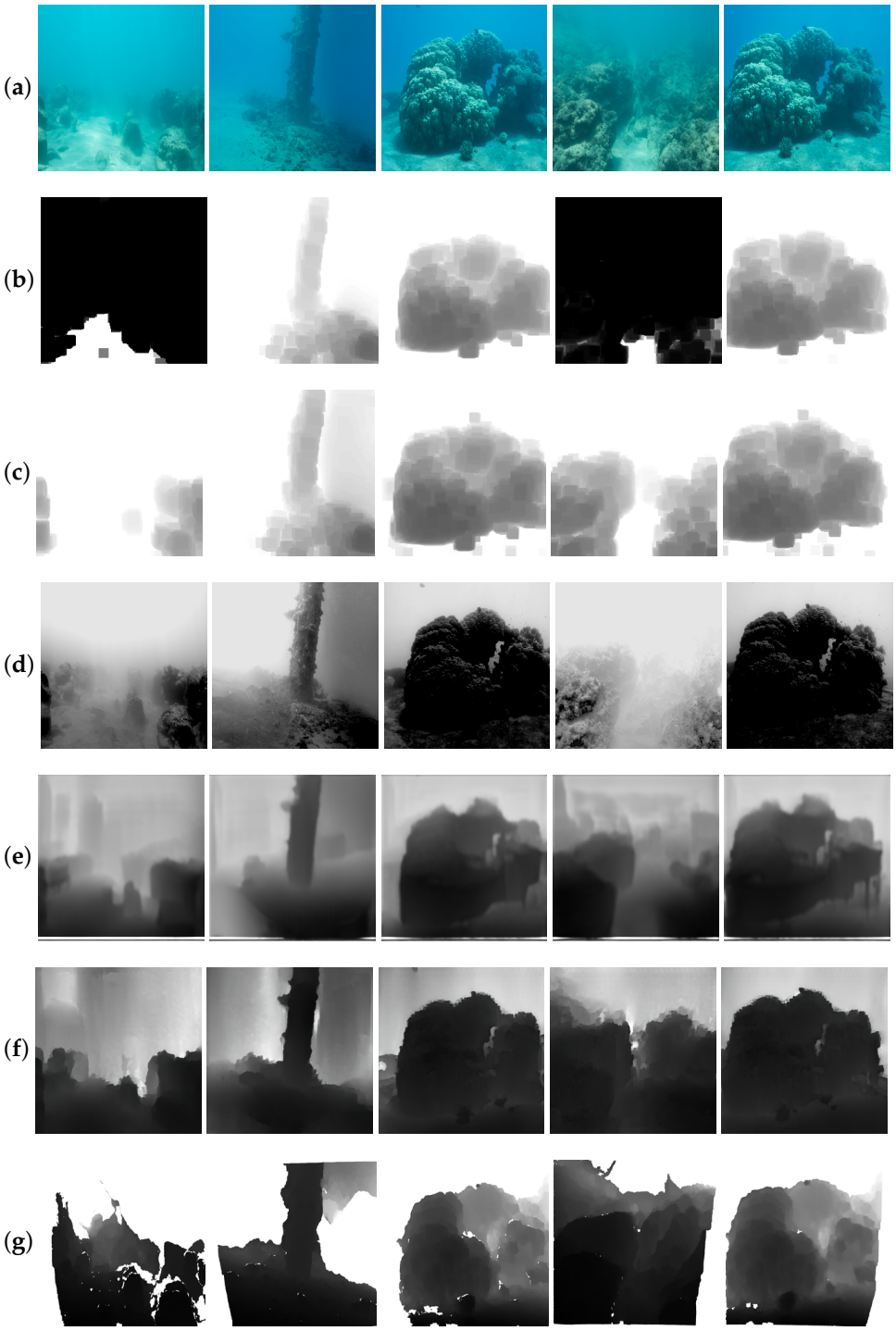


Figure 4. Comparison of our method with other methods for underwater depth map estimation. (**a**) Teal underwater images. (**b**–**g**) Results of DCP [33], UDCP [34], Berman et al. [35], UW-Net [11], and our method, as well as the ground truth.

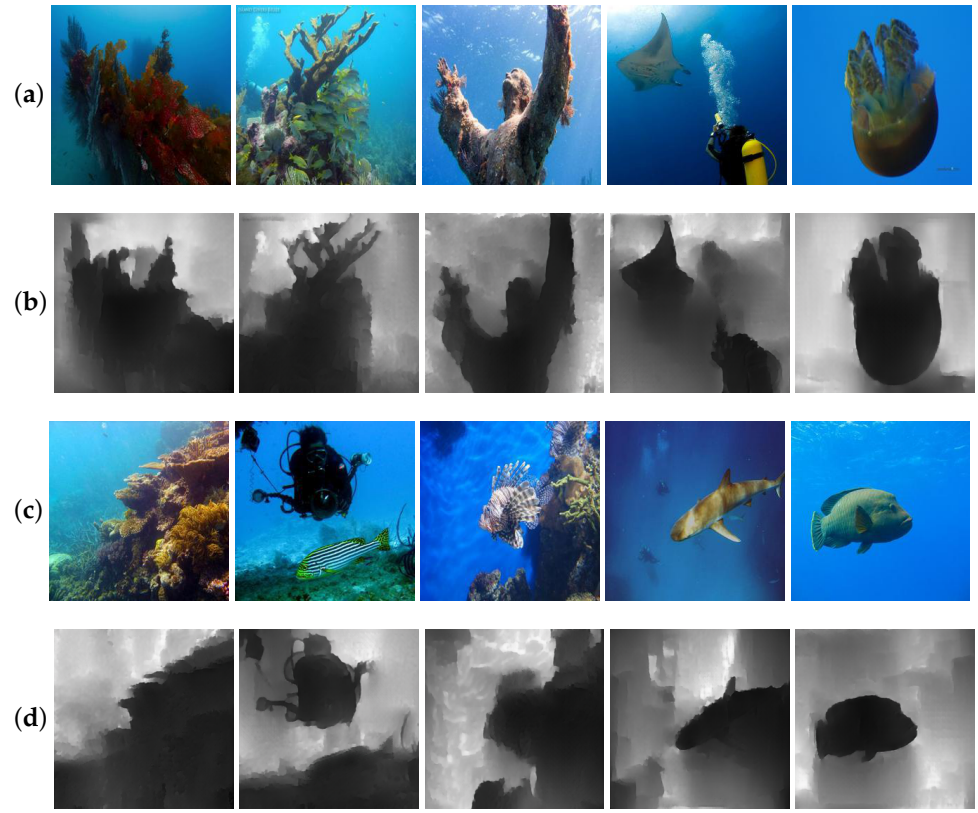


Figure 5. Estimation of multiple underwater depth maps. (**a**,**c**) are real underwater images. (**b**,**d**) are their predicted depth maps. Note that there is no ground truth.

3.1.2. Quantitative Evaluation

To quantitatively evaluate our model, we adopted two metrics for comparison: log scale-invariant mean squared error (SI-MSE [1]) and the Pearson correlation coefficient (ρ) with the dataset from Berman et al. [35]. Higher ρ values and lower SI-MSE [1] values represent better results. Due to the limitations of the Berman dataset, the ground truth was not fully provided in each depth map. We only evaluated the pixels with a distance value that was defined in the ground truth (GT). Comparing our method with other approaches, namely, DCP [33], UDCP [34], Berman et al. [35], and UW-Net [11], we observed that our method obtained the lowest scale-invariant error (SI-MSE [1]) and the highest Pearson correlation coefficient (ρ) (Table 1).

Table 1. Quantitative comparison of our method and other methods on the dataset of Berman et al. [35]. FT represents a fine-tuned (FT) underwater model. Lower SI-MSE [1] values and higher ρ values are better for underwater depth map estimation.

	DCP	UDCP	Berman et al.	UW-Net	UW-Net (FT)	Ours	Ours (FT)
SI-MSE	1.3618	0.6966	0.6755	0.4404	0.3708	0.3199	**0.2486**
ρ	0.2968	0.4894	0.6448	0.6202	0.6451	0.7018	**0.7600**

3.2. Ablation Study

The lack of diversity is the main obstacle in obtaining a precise underwater depth map with a data-driven model. We believe that the disentangled representation and the coarse-to-fine strategy play key roles in increasing the diversity of synthetic underwater images and enhancing the depth map prediction results. We evaluated the effectiveness of each proposed component, as shown in Table 2. Our framework included the underwater image synthesis module and the underwater depth map estimation module. Theoretically,

underwater image synthesis with disentangled representation can be used to generate realistic underwater images that are rich in diversity. A coarse-to-fine pipeline can further help our model to obtain better estimation results. From Table 2, we can observe that synthesizing multiple underwater images with disentangled representation and adopting a coarse-to-fine pipeline can practically help our model to obtain the best scores for SI-MSE and ρ in the final depth map estimation task.

Table 2. Ablation study of our method.

	Proposed	**w/o Disentangled Representation**	**w/o Coarse-to-Fine Pipeline**
SI-MSE	**0.2486**	0.2797	0.2707
ρ	**0.7600**	0.7136	0.7117

4. Discussion

4.1. Cross-Domain Underwater Image Synthesis with Disentangled Representation

In this section, we further explore the potential of our model for underwater image generation. With the help of the disentangled representation loss, our model can generate the intermediate information between two domains with semi-supervised learning. In this experiment, we removed the discrete conditional vector c_y. Instead, we assigned a three-dimensional vector (z_1, z_2, z_3) with decimal values for our task, where z_1 and z_2 were used for semi-supervised learning to control the underwater color, and z_3 was a free latent variable. To control the synthesized water color in a continuous manner, we manually labeled 20% of the underwater images from each underwater domain (blue and green). The deep blue images are labeled (1, 0), and the deep green images are labeled (0, 1). Both the labeled (20%) and unlabeled (80%) underwater images were used for training. The unlabeled underwater images were labeled by the classification branch from the discriminator D_u, which was introduced in Section 2.1. The results are shown in Figure 6. We can observe that our model can perform a gradual transition from the blue style to the green style according to the values of z_1 and z_2. Without any ground truth for the illumination, we found that our model could also perform a gradual transition from dark to bright according to the value of the free latent variable z_3.

We also evaluated the effectiveness of the synthesized underwater images for underwater depth map estimation, as shown in Figure 7. The quantitative results can be found in Table 3. The experiments show that the cross-domain synthetic strategy can also practically improve the performance in underwater depth map estimation. Our model with the cross-domain synthesis (Ours-C) setting obtained a lower SI-MSE score and an improved ρ score compared to Berman et al.'s dataset, which indicates that the cross-domain synthesis task can practically increase the diversity of the synthetic images and the generalization ability of our model. Although both models (Ours-C(fine-tuned (FT)), Ours(FT)) had a similar performance when they were fine-tuned on the unlabeled test dataset, note that one might not always have the opportunity to obtain a test dataset before deploying the model.

Table 3. Quantitative comparison of our method and other methods using the dataset of Berman et al. [35]. FT represents a fine-tuned (FT) underwater model. Ours-C is the method proposed in this section.

	DCP	**UDCP**	**Berman et al.**	**UW-Net**	**UW-Net(FT)**	**Ours-C**	**Ours-C(FT)**
SI-MSE	1.3618	0.6966	0.6755	0.4404	0.3708	0.3526	**0.2447**
ρ	0.2968	0.4894	0.6448	0.6202	0.6451	0.6823	**0.7423**

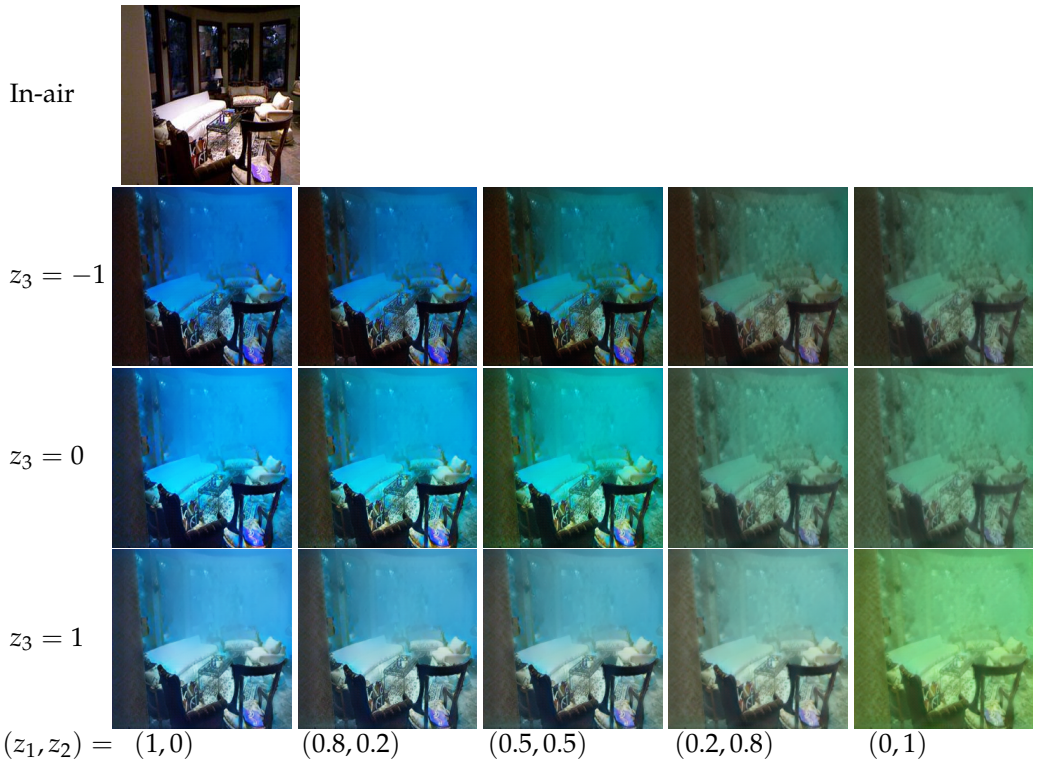


Figure 6. The continuous process of synthesizing underwater images. We used pseudo-labels to represent domain attributes. The two dimensions represent the green style (deep green when $z_1 = 0, z_2 = 1$) and blue style (deep blue when $z_1 = 1, z_2 = 0$), respectively. The first row on the top is the source in-air image, and the remaining images show the gradual transition from the blue style to the green style with different latent variables z_3.

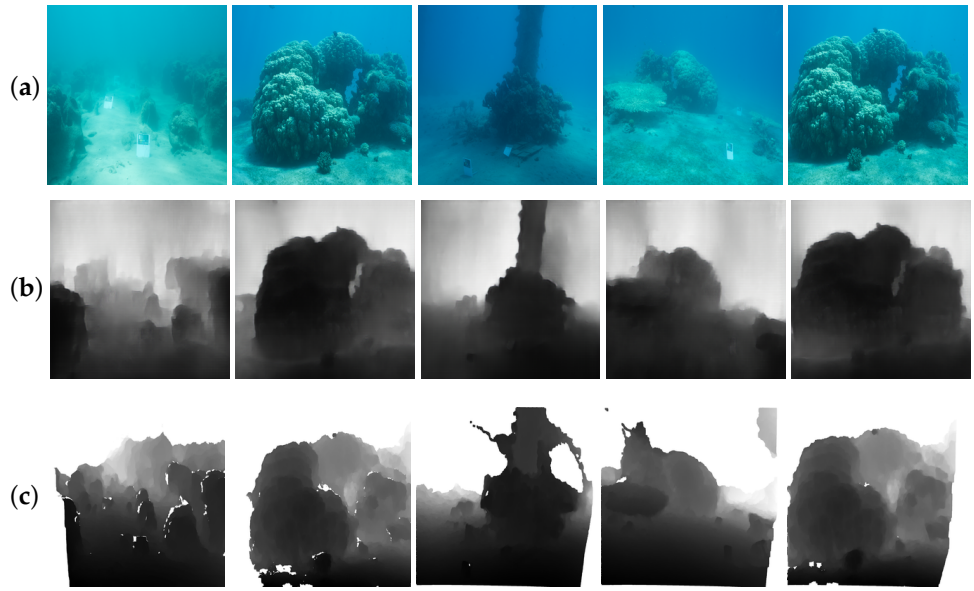
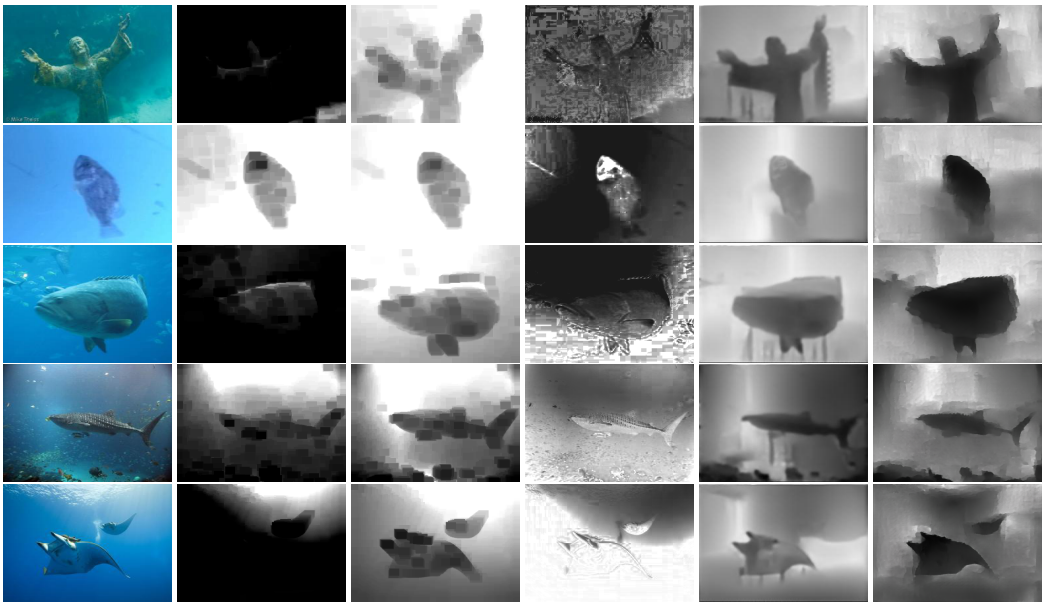


Figure 7. Sample results of our method using pseudo-labels for underwater depth map estimation. (**a**) Real underwater images from the dataset provided by Berman et al. [35]. (**b**,**c**) are the results of our model for depth map estimation and the ground truth.

4.2. Challenges of Underwater Scenes with Inhomogeneous Illumination

Due to the reflections and the angular changing of illuminants, many real underwater images show bad visibility with inhomogeneous illumination, as seen in Figure 8a. These factors usually bring negative effects for detection, segmentation, and depth map estimation

in real underwater images. The inhomogeneous illumination can easily cause a domain shift and mislead the feature extraction process. As seen in Figure 8, we show some results of DCP [33], UDCP [34], Berman et al. [35], UW-Net [11], and our method. As seen in the first two rows of Figure 8, the objects are difficult to accurately recognize from the real underwater images, which have a low contrast. Compared to the other methods, our model has a lower error ratio. However, our model still achieves inaccurate background depth map prediction results. Domain adaptation [42] might be a solution for improving our model in order to overcome this obstacle. We will consider this in our future research.



(**a**) Input (**b**) DCP (**c**) UDCP (**d**) Berman et al. (**e**) UW-Net (**f**) Ours

Figure 8. Comparison of the results of underwater depth map estimation in various underwater images with bad visibility by using different methods. We compared the results of dark channel prior (DCP) [33], underwater dark channel prior (UDCP) [34], Berman et al. [35], and Gupta et al. [11] with those of our method.

5. Conclusions

In this paper, we proposed an end-to-end system for underwater image synthesis and underwater depth map estimation. Our model can synthesize underwater images in a continuous manner to construct RGB-D pairs with disentangled representations. The coarse-to-fine pipeline can practically increase the performance for the task of underwater depth map estimation. We adopted a series of experiments for comparisons with the existing state-of-the-art methods. Both qualitative and quantitative results proved the efficiency of our method in both tasks.

Author Contributions: Q.Z. performed the experiments and manuscript. Z.X. draw the figures and performed a literature search. Z.Y. reviewed and edited the manuscript. B.Z. provided advices and GPU devices for parallel computing. All authors read and approved the manuscript.

Funding: This work was supported by the finance science technology project of 630 Hainan province of China under Grant Number ZDKJ202017, National Natural Science Fundation of China of Grant Number 61701463.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eigen, D.; Puhrsch, C.; Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2366–2374.
2. Chi, S.; Xie, Z.; Chen, W. A laser line auto-scanning system for underwater 3D reconstruction. *Sensors* **2016**, *16*, 1534. [CrossRef] [PubMed]
3. Palomer, A.; Ridao, P.; Ribas, D.; Forest, J. Underwater 3D laser scanners: The deformation of the plane. In *Sensing and Control for Autonomous Vehicles*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 73–88.
4. Xi, Q.; Rauschenbach, T.; Daoliang, L. Review of underwater machine vision technology and its applications. *Mar. Technol. Soc. J.* **2017**, *51*, 75–97. [CrossRef]
5. Dancu, A.; Fourgeaud, M.; Franjcic, Z.; Avetisyan, R. Underwater reconstruction using depth sensors. In *SIGGRAPH ASIA Technical Briefs*; ACM: New York, NY, USA, 2014; pp. 1–4.
6. Churnside, J.H.; Marchbanks, R.D.; Lembke, C.; Beckler, J. Optical backscattering measured by airborne lidar and underwater glider. *Remote Sens.* **2017**, *9*, 379. [CrossRef]
7. Deris, A.; Trigonis, I.; Aravanis, A.; Stathopoulou, E. Depth cameras on UAVs: A first approach. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *42*, 231–236. [CrossRef]
8. Ahamed, J.R.; Abas, P.E.; De Silva, L.C. Review of underwater image restoration algorithms. *IET Image Process.* **2019**, *13*, 1587–1596.
9. Massot-Campos, M.; Oliver-Codina, G. Optical sensors and methods for underwater 3D reconstruction. *Sensors* **2015**, *15*, 31525–31557. [CrossRef] [PubMed]
10. Li, N.; Zheng, Z.; Zhang, S.; Yu, Z.; Zheng, H.; Zheng, B. The synthesis of unpaired underwater images using a multistyle generative adversarial network. *IEEE Access* **2018**, *6*, 54241–54257. [CrossRef]
11. Gupta, H.; Mitra, K. Unsupervised single image underwater depth estimation. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 624–628.
12. 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, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
13. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
14. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8798–8807.
15. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8789–8797.
16. 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 Robot. Autom. Lett.* **2017**, *3*, 387–394. [CrossRef]
17. Cao, K.; Peng, Y.T.; Cosman, P.C. Underwater image restoration using deep networks to estimate background light and scene depth. In Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation, Las Vegas, NV, USA, 8–10 April 2018; pp. 1–4.
18. Li, C.; Anwar, S.; Porikli, F. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognit.* **2020**, *98*, 107038. [CrossRef]
19. Skinner, K.A.; Zhang, J.; Olson, E.A.; Johnson-Roberson, M. Uwstereonet: Unsupervised learning for depth estimation and color correction of underwater stereo imagery. In Proceedings of the International Conference on Robotics and Automation, Montreal, QC, Canada, 20–24 May 2019; pp. 7947–7954.
20. Wang, N.; Zhou, Y.; Han, F.; Zhu, H.; Zheng, Y. UWGAN: Underwater GAN for real-world underwater color restoration and dehazing. *arXiv* **2019**, arXiv:1912.10269.
21. Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv* **2016**, arXiv:1606.03657.
22. Spurr, A.; Aksan, E.; Hilliges, O. Guiding InfoGAN with semi-supervision. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 119–134.
23. Eigen, D.; Fergus, R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015; pp. 2650–2658.
24. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
25. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
26. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.

27. Wang, C.; Xu, C.; Wang, C.; Tao, D. Perceptual adversarial networks for image-to-image transformation. *IEEE Trans. Image Process.* **2018**, *27*, 4066–4079. [CrossRef] [PubMed]
28. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–13.
29. Kupyn, O.; Martyniuk, T.; Wu, J.; Wang, Z. DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–29 October 2019; pp. 8878–8887.
30. Choi, Y.; Uh, Y.; Yoo, J.; Ha, J.W. StarGAN v2: Diverse image synthesis for multiple domains. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8188–8197.
31. Hu, J.; Ozay, M.; Zhang, Y.; Okatani, T. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1043–1051.
32. Abady, L.; Barni, M.; Garzelli, A.; Tondi, B. GAN generation of synthetic multispectral satellite images. In *Image and Signal Processing for Remote Sensing XXVI*; International Society for Optics and Photonics: London, UK, 2020; Volume 11533, pp. 122–133.
33. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 2341–2353. [PubMed]
34. Drews, P.L.; Nascimento, E.R.; Botelho, S.S.; Campos, M.F.M. Underwater depth estimation and image restoration based on single images. *IEEE Comput. Graph. Appl.* **2016**, *36*, 24–35. [CrossRef] [PubMed]
35. Berman, D.; Treibitz, T.; Avidan, S. Diving into haze-lines: Color restoration of underwater images. In Proceedings of the British Machine Vision Conference (BMVC), London, UK, 9–12 September 2017; Volume 1, pp. 1–12.
36. Ancuti, C.; Ancuti, C.O.; De Vleeschouwer, C. D-hazy: A dataset to evaluate quantitatively dehazing algorithms. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 2226–2230.
37. Xiao, J.; Hays, J.; Ehinger, K.A.; Oliva, A.; Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3485–3492.
38. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral normalization for generative adversarial networks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April 30–3 May 2018; pp. 1–26.
39. Brock, A.; Donahue, J.; Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April 30–3 May 2018; pp. 1–35.
40. Zhang, H.; Goodfellow, I.; Metaxas, D.; Odena, A. Self-attention generative adversarial networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 7354–7363.
41. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 746–760.
42. Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; Van Gool, L. Domain adaptive faster R-CNN for object detection in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3339–3348.