Self-supervised Monocular Underwater Depth Recovery, Image Restoration, and a Real-sea Video Dataset
(Supplementary Material)

Nisha Varghese Ashish Kumar A. N. Rajagopalan
Indian Institute of Technology Madras, India
nishavarghese15@gmail.com askjovial005@gmail.com raju@ee.iitm.ac.in

The figures and tables in this supplementary material are numbered using the prefix S, and are arranged as follows:

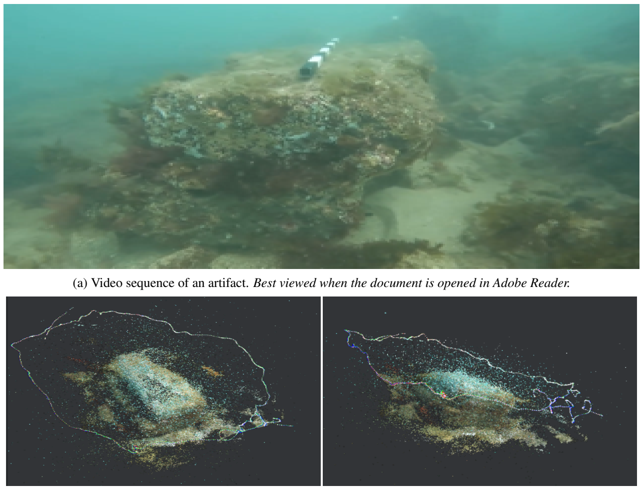
1. A sample video sequence of an artifact from DRUVA and visualization of the 3D point cloud of the artifact along with the camera poses.
2. The detailed network architecture of the proposed USe-ReDI-Net.
3. Test phase.
4. Additional ablations.
5. Additional qualitative and quantitative results.

S1. DRUVA dataset

Our Dataset of Real-world Underwater Videos of Artifacts (DRUVA) contains video sequences of 20 different artifacts with almost full azimuthal coverage of each artifact. A part of one such video sequence is included in Fig. S1(a) (*please use Adobe Reader to view the video*) whose frames were passed through USe-ReDI-Net. The restored frames from our network were used to reconstruct a 3D point cloud (shown in Fig. S1(b,c)) using Meshroom [12], an open-source 3D reconstruction software. The camera trajectory is also indicated as a solid path around the artifact.

S2. Network Architecture

The detailed structure of each block of USe-ReDI-Net is given in Table S1. USe-ReDI-Net disentangles input UW image I_1 into its latent components: scene radiance (J), transmission maps (T^d and T^b), and global background light (A). A is estimated analytically from I_1 using a Gaussian blur-kernel. USe-ReDI-Net has three main network blocks: scene-radiance network (SR-Net), transmission map network (TM-Net), and channel-wise extinction-coefficient ($\beta_c^* : c = \{R, G, B\} : * = \{d, b\}$) estimation network (Beta-Net). PoseNet is used to estimate the relative pose between two adjacent frames. For PoseNet, we use the same network structure which was followed by monocular depth estimation methods for terrestrial images [11, 27]. It is an encoder-decoder architecture where the encoder is a standard ResNet-18 [15], and the decoder outputs a 6-dimensional camera pose output ($\mathcal{T}$) consisting of 3 rotation angles and 3 translations. Following [17, 9], we use a non-degenerative structure for SR-Net where we use only stride-1 convolutions. Works [17, 6, 9] use the same structure for disentangling transmission maps also. But, we use a different structure for TM-Net. In order to increase the receptive field in estimating the transmission map, we use stride-2 convolution in TM-Net and use skip connections to combine the features at two different resolutions. We use the same network structure of TM-Net for both T^d and T^b. Depth is estimated analytically using transmission map and β_c^* where β_c^* is estimated from Beta-Net. For Beta-Net, we use a regression network that predicts a 6-dimensional output (corresponding to both T_d and T_b) from the features extracted from the input image using a convolutional neural network (CNN). In the CNN, we perform progressive downsampling using stride-2 convolutions to arrive at β_c^*.



(a) Video sequence of an artifact. *Best viewed when the document is opened in Adobe Reader.*

(b) Top-view

(c) Side-view

Figure S1: (a): Sample video sequence of an artifact from DRUVA with 360° azimuthal view; (b,c): 3D visualization of the artifact (a).

S3. Test phase

SR-Net, TM-Net, and Beta-Net in USe-ReDI-Net are trained by utilizing cues from haze as well as geometry from neighboring frames. During test time, only a single image is needed which is passed through SR-Net, TM-Net, and Beta-

Layer	Input	Block Structure	Config.	Output
		SR-Net		
Input	I_1	-	c:3	out$_1$
Block 1 to 4	out$_1$	$\begin{bmatrix}\text{Conv}\\ \text{Inst. Norm.}\\ \text{ReLU}\end{bmatrix}$ x4	k: 3x3 s: 1 c: 64	out$_2$
Output	out$_2$	$\begin{bmatrix}\text{Conv}\\ \text{Sigmoid}\end{bmatrix}$	k: 1x1 s: 1 c: 3	J
		TM-Net		
Input	I_1	-	c:3	out$_1$
Block 1 to 2	out$_1$	$\begin{bmatrix}\text{Conv}\\ \text{Inst. Norm.}\\ \text{ReLU}\end{bmatrix}$ x2	k: 3x3 s: 1 c: 64	out$_2$
Block 3	out$_2$	$\begin{bmatrix}\text{Conv}\\ \text{Inst. Norm.}\\ \text{ReLU}\end{bmatrix}$	k: 3x3 s: 2 c: 64	out$_3$
Block 4	out$_3$	$\begin{bmatrix}\text{Conv}\\ \text{Inst. Norm.}\\ \text{ReLU}\end{bmatrix}$	k: 3x3 s: 1 c: 64	out$_4$
Block 5	out$_4$	$\begin{bmatrix}\text{Deconv}\end{bmatrix}$	k: 3x3 s: 2 c: 64	out$_5$
Output	$\begin{bmatrix}\text{out}_5\\ \text{out}_2\end{bmatrix}$	$\begin{bmatrix}\text{Conv}\\ \text{Sigmoid}\end{bmatrix}$	k: 3x3 s: 1 c: 3	T_c
		Beta-Net		
Input	I/p image	-	c:3	out$_1$
Block 1	out$_1$	$\begin{bmatrix}\text{Conv}\\ \text{Inst. Norm.}\\ \text{ReLU}\end{bmatrix}$	k: 3x3 s: 1 c: 64	out$_2$
Block 2 to 4	out$_2$	$\begin{bmatrix}\text{Conv}\\ \text{Inst. Norm.}\\ \text{ReLU}\end{bmatrix}$ x3	k: 3x3 s: 2 c: 64	out$_3$
Block 5	out$_3$	Flatten	-	out$_4$
Block 6	out$_4$	$\begin{bmatrix}\text{Sigmoid}\end{bmatrix}$	n: 250	out$_5$
Output	out$_5$	$\begin{bmatrix}\text{Sigmoid}\end{bmatrix}$	n: 6	β_c^*
		PoseNet		
Input	$\begin{bmatrix}I_1\\ I_2\end{bmatrix}$	$\begin{bmatrix}\textit{ResNet-18}\text{[15]}\\ \text{ReLU}\end{bmatrix}$	c:512	out$_1$
Block 1	out$_1$	$\begin{bmatrix}\text{Conv}\\ \text{ReLU}\end{bmatrix}$	k: 1x1 s: 1 c: 256	out$_2$
Block 2 to 3	out$_2$	$\begin{bmatrix}\text{Conv}\\ \text{ReLU}\end{bmatrix}$ x2	k: 3x3 s: 1 c: 256	out$_3$
Block 4	out$_3$	$\begin{bmatrix}\text{Conv}\\ \text{ReLU}\end{bmatrix}$	k: 3x3 s: 1 c: 6	out$_4$
Output	out$_4$	$\begin{bmatrix}\text{Average}\end{bmatrix}$	n: 6	$\mathcal{T}$

Table S1: Network structure of USe-ReDI-Net. [c: number of output channels in the block, k: kernel size, s: stride, n: number of output nodes, Inst. Norm.: Instance normalization]

Net to get the restored image, transmission maps, and β_c^*. From any single channel of the transmission map T_c, we derive depth map $D = -\log(T_c)/\beta_c$. Thus, USe-ReDI-Net returns both the depth map and the restored image simultaneously from a single UW image in real-time (55 fps as mentioned in the main paper).

S4. Additional ablations

S4.1. Joint vs sequential estimation

In our approach, we jointly estimate the depth map and the clean image. Another possibility is to sequentially compute 1) restored image followed by depth map, and 2) depth map followed by the restored image using other SoTA methods.

The results for the case of estimating clean image followed by depth map have been given in the main paper under Mono2$_\text{d}$ where restoration was done using UIEC^2-Net[25] which was followed by depth estimation using Monodepth2 [11]. We have shown that its performance is inferior to our method. In order to compare with a more recent method, we provide results for HR-Depth [21] (a depth estimation method for terrestrial images) in Fig. S2 and Table S2. HR-Depth [21] is trained using the restored images (restored using [25]) from DRUVA dataset. It can be seen that the depth map obtained using HR-Depth [21] is not good, especially at higher depths. Quantitative metrics also show that our method is superior. Both HR-Depth [21] and Monodepth2 [11] yield comparable quantitative results.

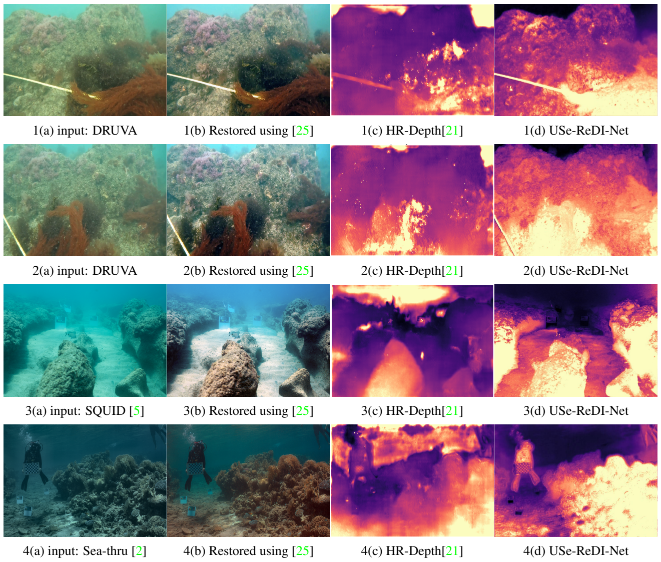

1(a) input: DRUVA 1(b) Restored using [25] 1(c) HR-Depth[21] 1(d) USe-ReDI-Net
2(a) input: DRUVA 2(b) Restored using [25] 2(c) HR-Depth[21] 2(d) USe-ReDI-Net
3(a) input: SQUID [5] 3(b) Restored using [25] 3(c) HR-Depth[21] 3(d) USe-ReDI-Net
4(a) input: Sea-thru [2] 4(b) Restored using [25] 4(c) HR-Depth[21] 4(d) USe-ReDI-Net

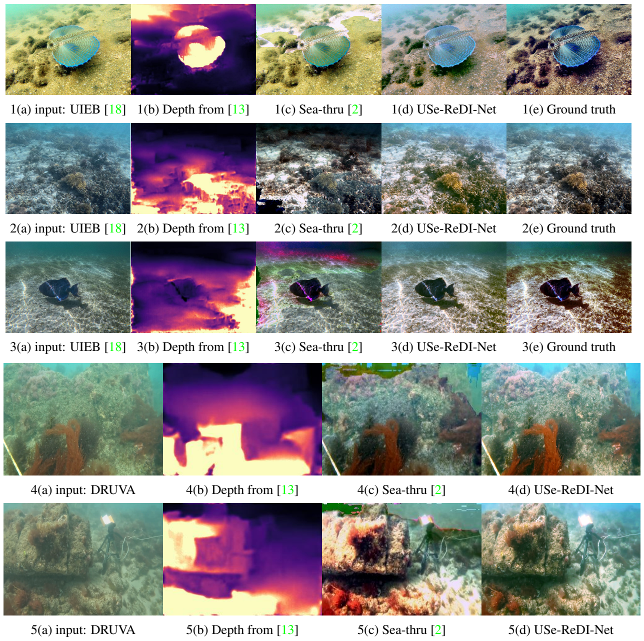
Figure S2: (a) Input UW images from three different datasets; (b) the restored outputs using UIEC^2-Net [25]; depth map from (c) sequential approach (restoration [25] followed by depth estimation [21]), and (d) USe-ReDI-Net.

Method	Sequential	Ours: USe-ReDI-Net
$\rho\uparrow$ /SI-MSE $\downarrow$	0.33/0.20	0.55/0.16

Table S2: Quantitative comparisons between the sequential approach (restoration [25] followed by depth estimation [21]) and USe-ReDI-Net for SQUID [5] dataset.

In order to perform the other sequential process of first estimating depth map followed by the restored image, we find the depth map using UW-Net [13]. The recovered depth map is then used to estimate the restored image using the method described in Sea-thru [2]. We use a third-party implementation [16] for Sea-thru [2] since the original implementation from the authors is not available. The results are given in Fig. S3 and Table S3. The restored output of the sequential method contains color deviations and is far from the ground truth. For DRUVA, the sequential method outputs artifacts in the restored image. The output from our method is close to the ground truth for UIEB [18] dataset. Our PSNR and SSIM values are significantly higher.

As expected, sequential processing yields suboptimal results as compared to joint estimation of depth and image.


1(a) input: UIEB [18] 1(b) Depth from [13] 1(c) Sea-thru [2] 1(d) USe-ReDI-Net 1(e) Ground truth

2(a) input: UIEB [18] 2(b) Depth from [13] 2(c) Sea-thru [2] 2(d) USe-ReDI-Net 2(e) Ground truth

3(a) input: UIEB [18] 3(b) Depth from [13] 3(c) Sea-thru [2] 3(d) USe-ReDI-Net 3(e) Ground truth

4(a) input: DRUVA 4(b) Depth from [13] 4(c) Sea-thru [2] 4(d) USe-ReDI-Net

5(a) input: DRUVA 5(b) Depth from [13] 5(c) Sea-thru [2] 5(d) USe-ReDI-Net

Figure S3: (a) Input UW images; (b) depth map from UW-Net [13]; restored outputs from the (c) sequential approach (depth map estimation [13] followed by restoration [2]), and (d) from USe-ReDI-Net.

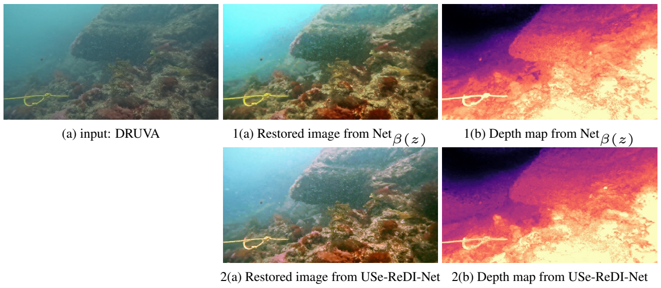
Method	Sequential	Ours: USe-ReDI-Net
PSNR↑/SSIM↑	14.4/0.58	18.9/0.70

Table S3: Quantitative comparisons between the sequential approach and USe-ReDI-Net for UIEB [18] dataset.

S4.2. Depth-dependency of β_c^d

Strictly speaking, β_c^d is depth-dependent as shown by Akkayanak *et al.* [1, 2, 3]. However, we chose to model it as a constant due to the following reasons. 1) In [1, 3], it is shown that variation of β_c^d for a depth range $z > 3$ m (which is a more practical depth range in real experiments) is $\Delta\beta_c^d < 0.1$ for the RED channel, and $\Delta\beta_c^d < 0.05$ for GREEN and BLUE channels. In our experiment, the depth values are mostly at higher depths (> 3 m); 2) They [2] use a parametric model to express the dependency of β_c^d on depth (z). The coefficients of this model can be determined only if the ground truth depth map is known. Ours is a self-supervised method that does not require knowledge of the original depth map.

Nevertheless, we modified the network structure of Beta-Net to return β^d and β^b separately where β^d is a 3-channel output (using a network with the same structure of TM-Net) with the same dimension as the input UW image, and β^b is a vector of dimension 3. Depth map from T^d is calculated by element-wise division with β^d, and the same channel-wise depth consistency loss is used for training. We refer to this network as Net$_{\beta(z)}$. We give comparison results of the estimated depth and restored image from Net$_{\beta(z)}$ and USe-ReDI-Net in Fig. S4 and Table S4. From the qualitative and quantitative results, it can be seen that both Net$_{\beta(z)}$ and USe-ReDI-Net provide almost similar outputs for both the restored image and the depth map. Net$_{\beta(z)}$ performs very similar to USe-ReDI-Net with a slightly lesser SSIM metric value for image restoration on UIEB [18] dataset. It is quite possible that a depth-dependent β_c^d is more useful when it can take full advantage in a supervised setup.


(a) input: DRUVA 1(a) Restored image from Net$_{\beta(z)}$ 1(b) Depth map from Net$_{\beta(z)}$

2(a) Restored image from USe-ReDI-Net 2(b) Depth map from USe-ReDI-Net

Figure S4: (a) Input UW image from DRUVA, (a) restored image and (b) depth map estimated from (1) Net$_{\beta(z)}$, and (2) USe-ReDI-Net.

Method	Net$_{\beta(z)}$	Ours: USe-ReDI-Net
$\rho\uparrow$ /SI-MSE $\downarrow$	0.55/0.16	0.55/0.16
PSNR↑/SSIM↑	18.9/0.68	18.9/0.70

Table S4: Quantitative comparisons b/w Net$_{\beta(z)}$ and USe-ReDI-Net with ρ/SI-MSE for SQUID [5] dataset, and PSNR/SSIM for UIEB [18] dataset.

S5. Additional Results

A sample output video from our USe-ReDI-Net with both the restored image sequence and the estimated depth maps is given in Fig. S5.

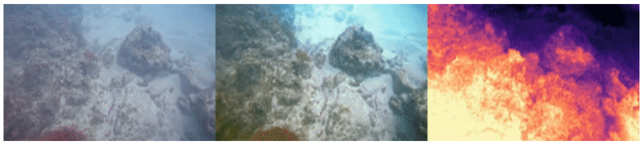
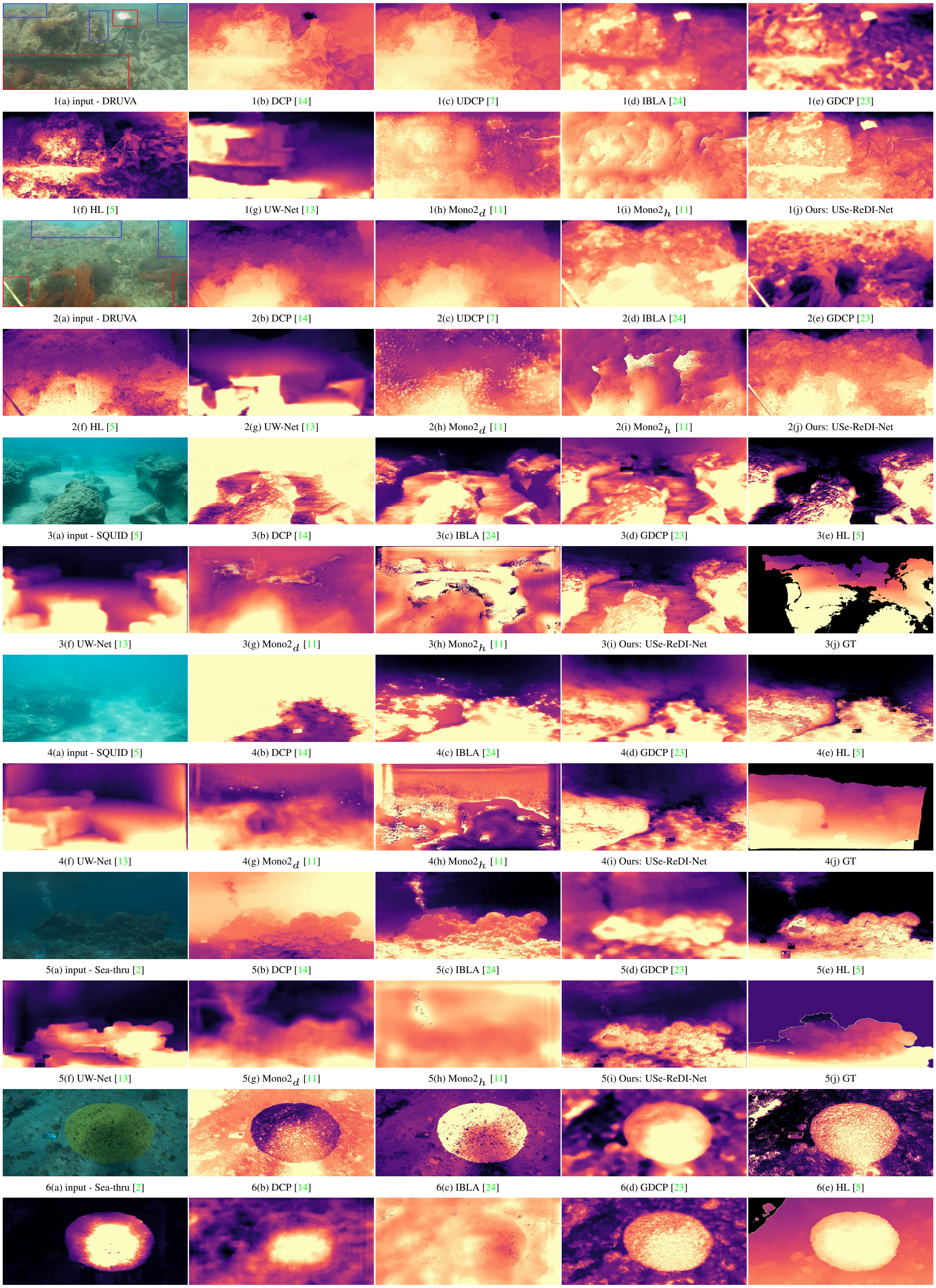


Figure S5: Sample output from USe-ReDI-Net for a short video sequence. UW image, restored output, and estimated depth map are in columns 1, 2, and 3, respectively. *Best viewed when the document is opened in Adobe Reader.*

We provide additional qualitative and quantitative evaluations in Fig. S6, Fig. S7, and Fig. S8.

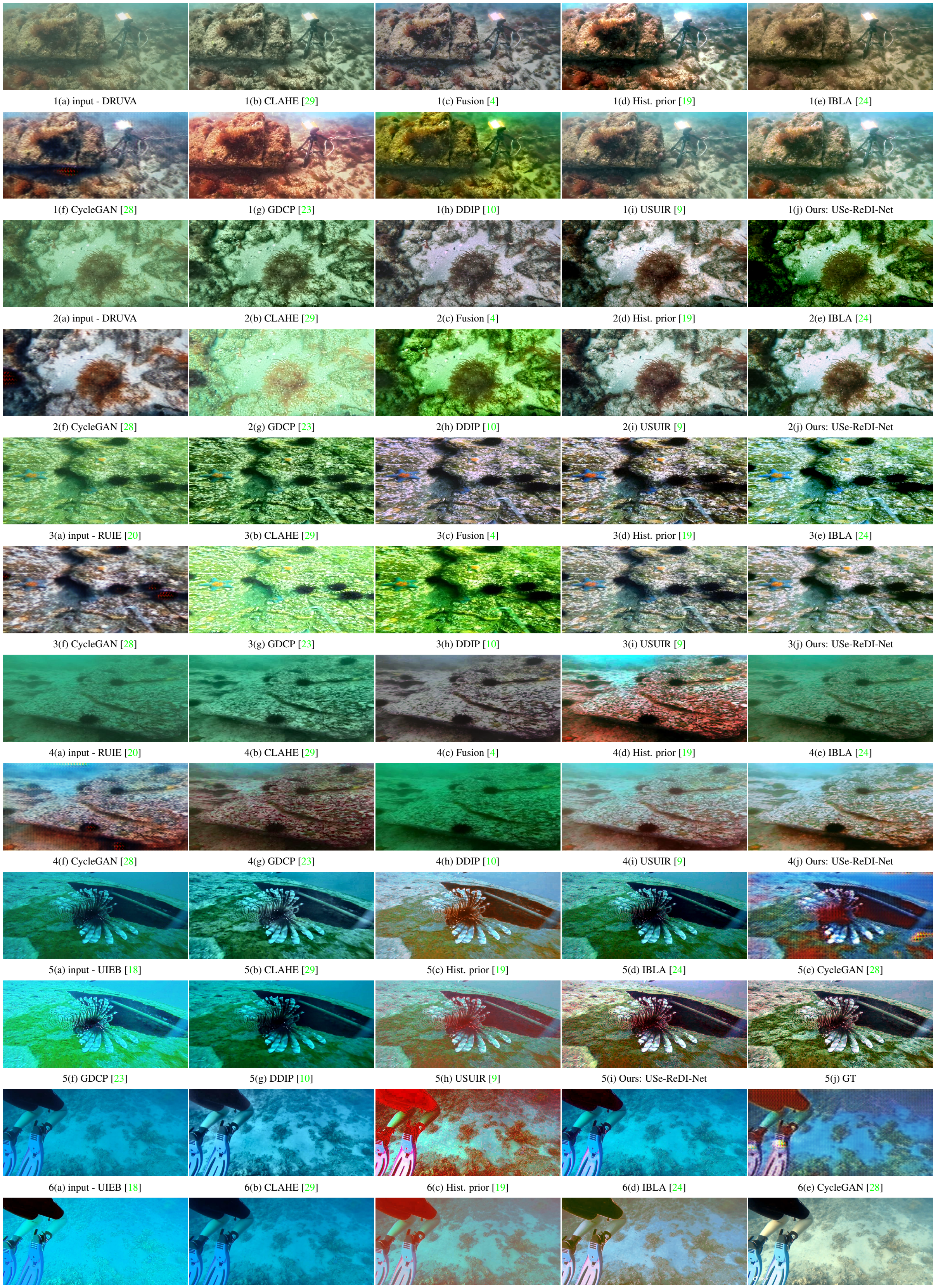
Datasets and metrics used for comparison

Details of each UW dataset used for comparisons are given in Table 1 in the main paper. UIEB dataset [18] con-



1(a) input - DRUVA
1(b) DCP [14]
1(c) UDCP [7]
1(d) IBLA [24]
1(e) GDCP [23]
1(f) HL [5]
1(g) UW-Net [13]
1(h) Mono2$_d$ [11]
1(i) Mono2$_h$ [11]
1(j) Ours: USe-ReDI-Net
2(a) input - DRUVA
2(b) DCP [14]
2(c) UDCP [7]
2(d) IBLA [24]
2(e) GDCP [23]
2(f) HL [5]
2(g) UW-Net [13]
2(h) Mono2$_d$ [11]
2(i) Mono2$_h$ [11]
2(j) Ours: USe-ReDI-Net
3(a) input - SQUID [5]
3(b) DCP [14]
3(c) IBLA [24]
3(d) GDCP [23]
3(e) HL [5]
3(f) UW-Net [13]
3(g) Mono2$_d$ [11]
3(h) Mono2$_h$ [11]
3(i) Ours: USe-ReDI-Net
3(j) GT
4(a) input - SQUID [5]
4(b) DCP [14]
4(c) IBLA [24]
4(d) GDCP [23]
4(e) HL [5]
4(f) UW-Net [13]
4(g) Mono2$_d$ [11]
4(h) Mono2$_h$ [11]
4(i) Ours: USe-ReDI-Net
4(j) GT
5(a) input - Sea-thru [2]
5(b) DCP [14]
5(c) IBLA [24]
5(d) GDCP [23]
5(e) HL [5]
5(f) UW-Net [13]
5(g) Mono2$_d$ [11]
5(h) Mono2$_h$ [11]
5(i) Ours: USe-ReDI-Net
5(j) GT
6(a) input - Sea-thru [2]
6(b) DCP [14]
6(c) IBLA [24]
6(d) GDCP [23]
6(e) HL [5]
6(f) UW-Net [13]
6(g) Mono2$_d$ [11]
6(h) Mono2$_h$ [11]
6(i) Ours: USe-ReDI-Net
6(j) GT

Figure S6: Input UW image (a) from datasets: (1,2) - DRUVA, (3,4) - SQUID [5], (5,6) - from Sea-thru [2] with ground truth (3(j), 4(j), 5(j), and 6(j)) and the depth maps obtained from different methods. Note that USe-ReDI-Net gives visually plausible depth maps for all the three datasets.



1(a) input - DRUVA 1(b) CLAHE [29] 1(c) Fusion [4] 1(d) Hist. prior [19] 1(e) IBLA [24]

1(f) CycleGAN [28] 1(g) GDCP [23] 1(h) DDIP [10] 1(i) USUIR [9] 1(j) Ours: USe-ReDI-Net

2(a) input - DRUVA 2(b) CLAHE [29] 2(c) Fusion [4] 2(d) Hist. prior [19] 2(e) IBLA [24]

2(f) CycleGAN [28] 2(g) GDCP [23] 2(h) DDIP [10] 2(i) USUIR [9] 2(j) Ours: USe-ReDI-Net

3(a) input - RUIE [20] 3(b) CLAHE [29] 3(c) Fusion [4] 3(d) Hist. prior [19] 3(e) IBLA [24]

3(f) CycleGAN [28] 3(g) GDCP [23] 3(h) DDIP [10] 3(i) USUIR [9] 3(j) Ours: USe-ReDI-Net

4(a) input - RUIE [20] 4(b) CLAHE [29] 4(c) Fusion [4] 4(d) Hist. prior [19] 4(e) IBLA [24]

4(f) CycleGAN [28] 4(g) GDCP [23] 4(h) DDIP [10] 4(i) USUIR [9] 4(j) Ours: USe-ReDI-Net

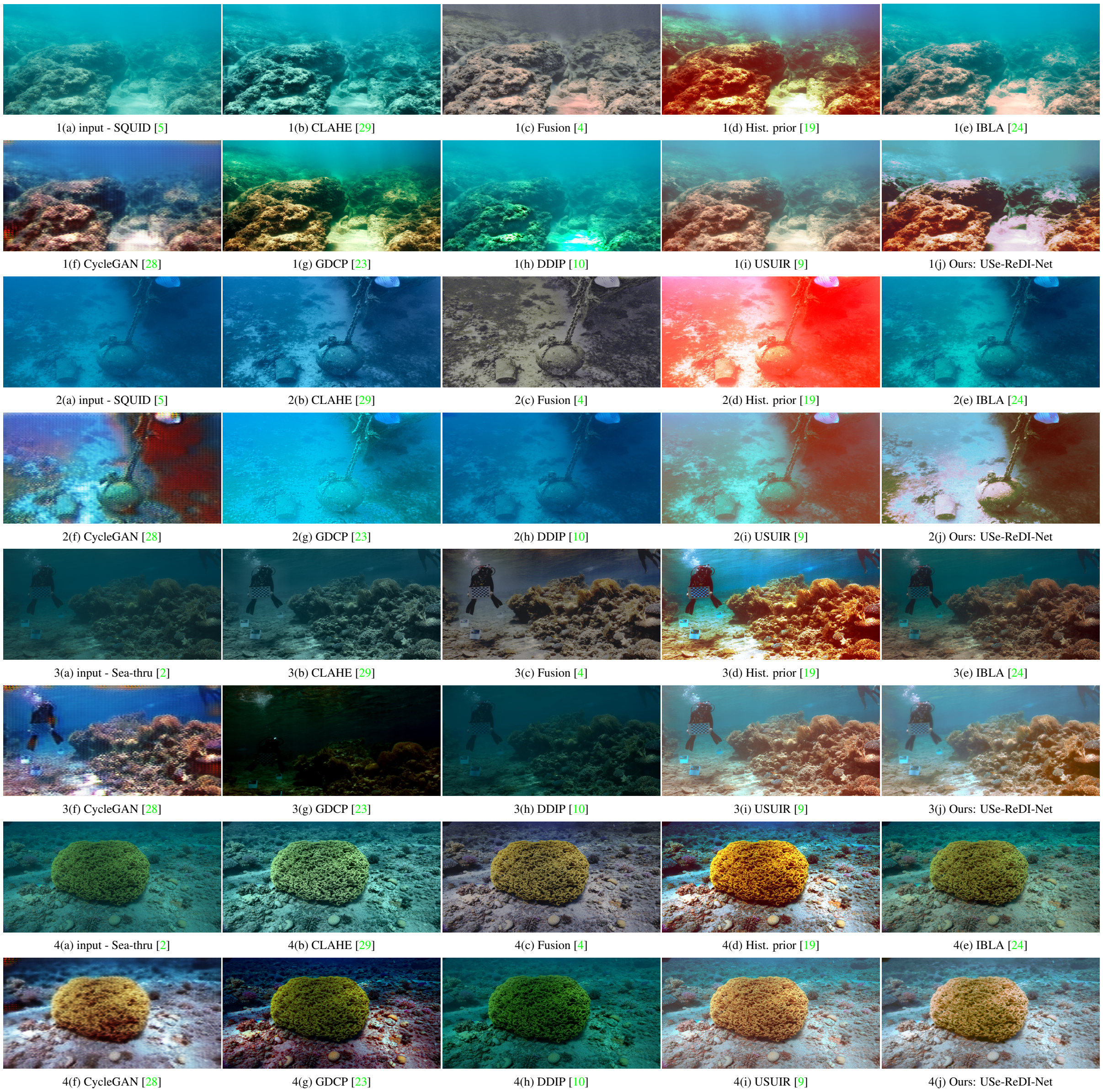
5(a) input - UIEB [18] 5(b) CLAHE [29] 5(c) Hist. prior [19] 5(d) IBLA [24] 5(e) CycleGAN [28]

5(f) GDCP [23] 5(g) DDIP [10] 5(h) USUIR [9] 5(i) Ours: USe-ReDI-Net 5(j) GT

6(a) input - UIEB [18] 6(b) CLAHE [29] 6(c) Hist. prior [19] 6(d) IBLA [24] 6(e) CycleGAN [28]

6(f) GDCP [23] 6(g) DDIP [10] 6(h) USUIR [9] 6(i) Ours: USe-ReDI-Net 6(j) GT

Figure S7: Input UW image (a) from datasets: (1,2) - DRUVA, (3,4) - RUIE [20], (5,6) - UIEB [18] with ground truth (5(j) and 6(j)) for UIEB and the enhanced images obtained from different methods. Note that our enhanced image results are visually good. For UIEB dataset, our output is quite close to the ground truth.



1(a) input - SQUID [5] 1(b) CLAHE [29] 1(c) Fusion [4] 1(d) Hist. prior [19] 1(e) IBLA [24]

1(f) CycleGAN [28] 1(g) GDCP [23] 1(h) DDIP [10] 1(i) USUIR [9] 1(j) Ours: USe-ReDI-Net

2(a) input - SQUID [5] 2(b) CLAHE [29] 2(c) Fusion [4] 2(d) Hist. prior [19] 2(e) IBLA [24]

2(f) CycleGAN [28] 2(g) GDCP [23] 2(h) DDIP [10] 2(i) USUIR [9] 2(j) Ours: USe-ReDI-Net

3(a) input - Sea-thru [2] 3(b) CLAHE [29] 3(c) Fusion [4] 3(d) Hist. prior [19] 3(e) IBLA [24]

3(f) CycleGAN [28] 3(g) GDCP [23] 3(h) DDIP [10] 3(i) USUIR [9] 3(j) Ours: USe-ReDI-Net

4(a) input - Sea-thru [2] 4(b) CLAHE [29] 4(c) Fusion [4] 4(d) Hist. prior [19] 4(e) IBLA [24]

4(f) CycleGAN [28] 4(g) GDCP [23] 4(h) DDIP [10] 4(i) USUIR [9] 4(j) Ours: USe-ReDI-Net

Figure S8: Input UW image (a) from datasets: (1,2) - SQUID [5], (3,4) - Sea-thru [2] and the enhanced images obtained from different methods. Note that our enhanced image results are visually pleasing without any color deviations.

	SQUID [5]					Sea-thru [2]				
Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	Abs Rel↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$
DCP [14]	2.96	2.80	0.95	1.30	0.34	0.54	0.36	0.58	0.69	0.19
UDCP [7]	1.43	1.13	0.54	0.98	0.43	0.51	0.31	0.52	0.86	0.31
GDCP [23]	1.31	1.10	0.69	0.92	0.42	0.49	0.28	0.52	0.85	0.23
IBLA [24]	1.89	1.59	0.64	1.13	0.68	0.48	0.29	0.50	0.75	0.23
HL [5]	2.89	3.66	0.93	1.31	0.54	0.50	0.36	0.50	0.60	0.21
UW-Net [13]	2.38	2.22	0.74	1.17	0.36	0.54	0.33	0.53	0.99	0.21
USUIR [9]	2.47	2.35	0.78	1.19	0.05	0.73	0.55	0.77	1.47	0.14
Mono2_h [11]	2.74	2.81	0.76	1.25	0.04	0.83	0.68	0.85	1.86	0.15
Mono2_d [11]	1.94	1.41	0.56	1.05	0.43	0.51	0.32	0.55	0.55	0.28
USe-ReDI-Net	1.38	1.09	0.54	0.85	0.71	0.48	0.31	0.41	0.45	0.44

Table S5: Quantitative comparisons of depth estimation accuracy on SQUID [5] dataset and the dataset proposed by Sea-thru [2] using depth evaluation metrics.

tains real-UW images with pseudo-ground-truth for image restoration. We have used UIEB [18] dataset for comparing performance on image restoration using PSNR and SSIM. RUIE dataset [20] is a large-scale UW dataset with only raw UW images which are divided into three subsets according to specific tasks of image restoration algorithms: UW image quality set (UIQS), UW color cast set (UCCS), and UW higher-level task-driven set (UHTS). We have used images from UIQS for inference. DRUVA is the only *video* dataset from among all UW datasets. SQUID [5] is a stereo dataset, and all the other datasets contain independent raw UW images. For DRUVA, RUIE [20], and UIEB [18], we have used no-reference image quality assessment metrics UIQM [22], and UCIQE [26]. SQUID dataset [5] comes with camera calibration files, and ground truth depth maps. The dataset proposed by Sea-thru [2] contains raw underwater images with range maps. We use both SQUID [5] and Sea-thru [2] dataset for comparing depth prediction accuracy using scale invariant metrics: Pearson correlation coefficient (ρ) [5] and scale-invariant mean squared error (SI-MSE) [8] as all the methods predict depth map up to a scale. After optimizing for the scale factor, we have also found absolute relative error (Abs Rel), squared relative error (Sq Rel), root mean squared error (RMSE), and root mean squared logarithmic error (RMSE log) between the ground truth depth map and the estimated depth map, and the accuracy ($\delta < threshold$) [8] for different methods. These values are given in Table S5.

Qualitative and quantitative evaluation

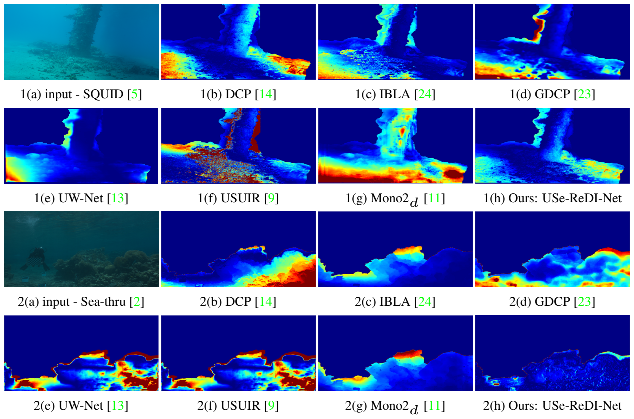
For evaluating depth accuracy, we have included additional results for every method on two images from DRUVA, SQUID [5], and dataset by Sea-thru [2] in Fig. S6. For SQUID [5] and Sea-thru [2] datasets, we have given their ground truth depth maps also where black regions indicate undefined regions in depth map since their ground truth depth map is derived from stereo and SFM, respectively. SQUID [5] contains some spurious depth values also. From Fig. S6, it can be seen that, for DRUVA dataset, USe-ReDI-Net performs better and returns plausible depth maps as compared to other methods (see Fig. S6 1(a) and 2(a) with red rectangle portions indicating a lesser depth and blue rectangle regions indicating transitions in depth). When comparing depth maps for SQUID [5] and Sea-thru [2], the output from USe-ReDI-Net is closer to ground truth.

Additional results for comparing image restoration performance on two images from 5 datasets (DRUVA, RUIE [20], UIEB [18], SQUID [5], and Sea-thru [2]) are included in Fig. S7 and Fig. S8. It can be seen that USe-ReDI-Net provides more realistic restored outputs without color deviations. For UIEB dataset [18], outputs from USe-ReDI-Net are close to ground truth. For the dataset from Sea-thru [2], both USUIR [9] and USe-ReDI-Net perform well. The visual quality of USUIR [9] is poor on other datasets especially on SQUID [5] and UIEB [18]. Our method performs consistently well on all five datasets.

In the main paper, we gave quantitative results for depth using the most used metrics ρ and SI-MSE. Here, we give additional quantitative results on the metrics Abs Rel, Sq Rel, RMSE, and RMSE log for comparing depth prediction accuracy on SQUID [5] and dataset from Sea-thru [2]. These metric values for different methods are given in Table S5. It can be seen that USe-ReDI-Net has the best metric values. Traditional methods, GDCP [23] and IBLA [24] are also near to USe-ReDI-Net, but their qualitative results are not good, especially on DRUVA (see Fig. S6).

Error map of estimated depth

In order to compare the accuracy of the depth map, we have included error maps (using jet colormap) of depths estimated by different methods on two images from SQUID [5] and Sea-thru [2] in Fig. S9. Note that our depth errors are smaller.



1(a) input - SQUID [5] 1(b) DCP [14] 1(c) IBLA [24] 1(d) GDCP [23]

1(e) UW-Net [13] 1(f) USUIR [9] 1(g) Mono2$_d$ [11] 1(h) Ours: USe-ReDI-Net

2(a) input - Sea-thru [2] 2(b) DCP [14] 2(c) IBLA [24] 2(d) GDCP [23]

2(e) UW-Net [13] 2(f) USUIR [9] 2(g) Mono2$_d$ [11] 2(h) Ours: USe-ReDI-Net

Figure S9: Input UW image (a) from datasets: (1) - SQUID [5], (2) - Sea-thru [2] and the error maps of depths obtained from different methods. Note that USe-ReDI-Net has lesser errors.

References

[1] Derya Akkaynak and Tali Treibitz. A revised underwater image formation model. In *CVPR*, pages 6723–6732, 2018. 3

[2] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *CVPR*, pages 1682–1691, 2019. 2, 3, 4, 6, 7

[3] Derya Akkaynak, Tali Treibitz, Tom Shlesinger, Yossi Loya, Raz Tamir, and David Iluz. What is the space of attenuation coefficients in underwater computer vision? In *CVPR*, pages 568–577, 2017. 3

[4] Cosmin Ancuti, Codruta Orniana Ancuti, Tom Haber, and Philippe Bekaert. Enhancing underwater images and videos by fusion. In *CVPR*, pages 81–88, 2012. 5, 6

[5] Dana Berman, Deborah Levy, Shai Avidan, and Tali Treibitz. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE PAMI*, 43(8):2822–2837, 2021. 2, 3, 4, 6, 7

[6] Shu Chai, Zhenqi Fu, Yue Huang, Xiaotong Tu, and Xinghao Ding. Unsupervised and untrained underwater image restoration based on physical image formation model. In *ICASSP*, pages 2774–2778, 2022. 1

[7] Paulo L.J. Drews, Erickson R. Nascimento, Silvia S.C. Botelho, and Mario Fernando Montenegro Campos. Underwater depth estimation and image restoration based on single images. *IEEE CG&A*, 36(2):24–35, 2016. 4, 6

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NIPS*, 27, 2014. 7

[9] Zhenqi Fu, Huangxing Lin, Yan Yang, Shu Chai, Liyan Sun, Yue Huang, and Xinghao Ding. Unsupervised underwater image restoration: From a homology perspective. *AAAI*, 36(1):643–651, Jun. 2022. 1, 5, 6, 7

[10] Yosef Gandelsman, Assaf Shocher, and Michal Irani. "double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *CVPR*, pages 11018–11027, 2019. 5, 6

[11] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3827–3837, 2019. 1, 2, 4, 6, 7

[12] Carsten Griwodz, Simone Gasparini, Lilian Calvet, Pierre Gurdjos, Fabien Castan, Benoit Maujean, Gregoire De Lillo, and Yann Lanthony. Alicevision Meshroom: An open-source 3D reconstruction pipeline. In *ACM MMSys '21*, 2021. 1

[13] Honey Gupta and Kaushik Mitra. Unsupervised single image underwater depth estimation. In *ICIP*, pages 624–628, 2019. 3, 4, 6, 7

[14] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *PAMI*, 33(12):2341–2353, 2011. 4, 6, 7

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 2

[16] Hai Nguyen Hong. Sea-thru: Implementation of sea-thru by derya akkaynak and tali treibitz, https://github.com/hainh/sea-thru. 3

[17] Boyun Li, Yuanbiao Gou, Shuhang Gu, Jerry Zitao Liu, Joey Tianyi Zhou, and Xi Peng. You Only Look Yourself: Unsupervised and Untrained Single Image Dehazing Neural Network. *IJCV*, pages 1–14, 2021. 1

[18] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *TIP*, 29:4376–4389, 2020. 3, 5, 7

[19] Chong-Yi Li, Ji-Chang Guo, Run-Min Cong, Yan-Wei Pang, and Bo Wang. Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior. *TIP*, 25(12):5664–5677, 2016. 5, 6

[20] Risheng Liu, Xin Fan, Ming Zhu, Minjun Hou, and Zhongxuan Luo. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *CSVT*, 30(12):4861–4875, 2020. 5, 7

[21] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. *AAAI*, 35(3), 2021. 2

[22] Karen Panetta, Chen Gao, and Sos Agaian. Human-visual-system-inspired underwater image quality measures. *IEEE J. Ocean. Eng.*, 41(3):541–551, 2016. 7

[23] Yan-Tsung Peng, Keming Cao, and Pamela C. Cosman. Generalization of the dark channel prior for single image restoration. *TIP*, 27(6):2856–2868, 2018. 4, 5, 6, 7

[24] Yan-Tsung Peng and Pamela C. Cosman. Underwater image restoration based on image blurriness and light absorption. *TIP*, 26(4):1579–1594, 2017. 4, 5, 6, 7

[25] Yudong Wang, Jichang Guo, Huan Gao, and Huihui Yue. Uiec$\hat{2}$-net: Cnn-based underwater image enhancement using two color space. *Signal Process. Image Commun.*, 96:116250, 2021. 2

[26] Miao Yang and Arcot Sowmya. An underwater color image quality evaluation metric. *TIP*, 24(12):6062–6071, 2015. 7

[27] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 6612–6619, 2017. 1

[28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2242–2251, 2017. 5, 6

[29] Karel Zuiderveld. Contrast limited adaptive histograph equalization. In *Graphic Gems IV. San Diego: Academic Press Professional*, page 474–485, 1994. 5, 6