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

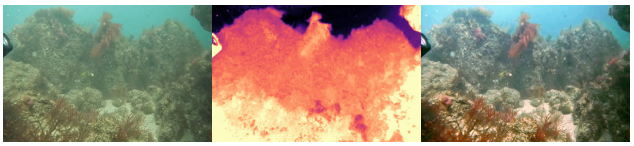
Nisha Varghese Ashish Kumar A. N. Rajagopalan
Indian Institute of Technology Madras, India
nishavarghese15@gmail.com askjovial005@gmail.com raju@ee.iitm.ac.in

Abstract

Underwater (UW) depth estimation and image restoration is a challenging task due to its fundamental ill-posedness and the unavailability of real large-scale UW-paired datasets. UW depth estimation has been attempted before by utilizing either the haze information present or the geometry cue from stereo images or the adjacent frames in a video. To obtain improved estimates of depth from a single UW image, we propose a deep learning (DL) method that utilizes both haze and geometry during training. By harnessing the physical model for UW image formation in conjunction with the view-synthesis constraint on neighboring frames in monocular videos, we perform disentanglement of the input image to also get an estimate of the scene radiance. The proposed method is completely self-supervised and simultaneously outputs the depth map and the restored image in real-time (55 fps). We call this first-ever Underwater Self-supervised deep learning network for simultaneous Recovery of Depth and Image as USe-ReDI-Net. To facilitate monocular self-supervision, we collected a Dataset of Real-world Underwater Videos of Artifacts (DRUVA) in shallow sea waters. DRUVA is the first UW video dataset that contains video sequences of 20 different submerged artifacts with almost full azimuthal coverage of each artifact. Extensive experiments on our DRUVA dataset and other UW datasets establish the superiority of our proposed USe-ReDI-Net over prior art for both UW depth and image recovery. The dataset DRUVA is available at https://github.com/nishavarghese15/DRUVA.

1. Introduction

Underwater (UW) depth recovery and image restoration is very important in ocean exploration applications such as marine biology [40], marine archaeology [38], UW robotics [54], etc. 3D reconstruction of UW structures warrants depth as a fundamental requirement. Current UW depth estimation methods can be divided into active and passive [54]. Active methods, which include different kinds of sonar [42, 23, 8], UW laser line-scanning [45, 16], range-gated imaging systems [39], and LiDAR [46] are usually bulky. Also, their performance is limited by the scattering of light in water [51]. On the other hand, passive methods use images captured from a camera. These include traditional [5, 44, 11, 4] as well as deep-learning (DL) based approaches [25, 53, 26]. Traditional UW depth estimation methods find depth from haze by utilizing the UW image formation model. These methods return erroneous estimates when there is a mismatch between the adopted prior and actual scene conditions, and are time-consuming too. Supervised deep networks for UW depth estimation are cumbersome to design due to the unavailability of paired large-scale UW depth datasets. Unpaired-learning based [25] and Generative Adversarial Network (GAN)-based [26] methods exist for depth. But they use either terrestrial RGBD datasets or synthetically generated datasets for training. As a result, there is a domain gap with real UW images.



(a) Input from *DRUVA* (b) Depth map (c) Enhanced image

Figure 1: (a) An example image from our *DRUVA* dataset. (b) Depth map, and (c) enhanced image obtained using our *USe-ReDI-Net*.

Self-supervised depth estimation networks for terrestrial images [21, 19, 22, 57, 59] utilize geometry cues between adjacent frames of a monocular video or stereo pairs where a target view is synthesized from a source view using the relative pose and the estimated depth. These methods cannot be directly applied to UW images due to the presence of haze. A recent method [53] estimates depth from UW images using the geometry cue. Because [53] performs self-supervision based on the depth and pose derived directly from hazy UW images, the depth map lacks details.

Since light is absorbed/scattered in water, UW images suffer from color distortions and poor contrast. Restoration of UW images has attracted a lot of attention. [44, 43, 1, 35, 56] estimate the parameters of the UW imaging model using image priors. Due to the inconsistency

between the prior and actual image conditions, the results are not always satisfactory [37]. DL-based methods include supervised [34, 33, 31] and unsupervised approaches [6, 15] for UW image enhancement. Supervised methods rely largely on either synthetic UW datasets generated from RGBD datasets like NYU Depth v2 [47] or real underwater datasets like UIEB [33] with (subjective) pseudo-ground-truth. The unavailability of real UW datasets with actual ground truth continues to pose challenges for supervised UW image restoration. Inspired by an unsupervised method for image dehazing [30], the works [15, 6] leverage UW image formation model for self-supervision by disentangling the input image.

In this paper, we propose a unified learning framework for joint monocular UW depth estimation and image restoration based on self-supervision that runs in real-time. We refer to it as Underwater Self-supervised network for simultaneous Recovery of Depth and Image (USe-ReDI-Net). This is the first UW work to use both haze and geometry as cues for depth. We harness the UW image formation model [1] to disentangle the input hazy image to get an estimate of the clean image and the transmission map. We find depth analytically from the transmission map and invoke an additional view-consistency constraint from a neighboring frame to facilitate proper disentanglement. When we perform self-supervision by view-synthesis, the warped location is governed by the estimated depth while the photometric value at the warped location comes from the restored input. Joint estimation of depth and image is mutually beneficial in the sense that a refined depth map aids the image restoration process, and an improved image estimate in turn helps to recover a better estimate of depth. The strength of our method lies in encapsulating this strong coupling between the two tasks. In our method, we cannot estimate one without estimating the other. To the best of our knowledge, this is the first DL-based work to estimate depth as well as restored image jointly in an end-to-end manner from an underwater image. We perform extensive experiments on several real-world UW image datasets to establish the effectiveness of our USe-ReDI-Net for both UW depth estimation and image restoration.

In order to advance the state-of-the-art, UW datasets are essential but these are quite difficult to capture. Considering the critical need for real UW video sequences with camera intrinsics, we collected our own Dataset of Real-world Underwater Videos of Artifacts (DRUVA) using a GoPro Hero 10 camera. This dataset contains video sequences of 20 different artifacts in shallow waters where the diver goes around the artifacts to acquire an almost 360° azimuthal view. Details of the dataset are given in Sec. 4. An example image from DRUVA is given in Fig. 1 along with the depth map and the enhanced output obtained using USe-ReDI-Net.

Our main contributions are as follows.

1. We propose a self-supervised deep network (USe-ReDI-Net) for monocular underwater depth estimation and image restoration that runs in real-time (55 fps).
2. USe-ReDI-Net is the first **end-to-end** DL method to simultaneously recover depth and latent image from an underwater observation. By jointly solving for image and depth, we judiciously invoke relevant losses governing the scene radiance as well as the depth map to better constrain the problem.
3. Our work is the first attempt to utilize cues from both haze and geometry to recover depth in UW images.
4. USe-ReDI-Net is a fully self-supervised approach that outperforms state-of-the-art methods both in terms of output quality as well as computational speed.
5. We have collected a unique dataset (DRUVA) which, to the best of our knowledge, is the first-ever UW video dataset containing real underwater image sequences of submerged artifacts. We shall release this dataset for the research community to harness it in a multitude of ways.

2. Related Works

2.1. UW depth estimation

Traditional UW depth estimation methods use image formation model for image restoration and depth estimation. Typically, depth is estimated as a by-product of the image restoration task. [27] was the first to introduce dark channel prior (DCP) for single-image depth estimation. Various researchers [11, 17, 4, 43] have used variations of DCP in UW depth estimation. Peng *et al*. [44] use image blurriness while Berman *et al*. [5] recover UW scenes by considering different spectral profiles of various water types. Traditional methods find depth using the relation between the transmission map and depth. In cases of severe attenuation, these methods do not yield satisfactory results. Also, they are quite time-consuming. Prior-based methods return erroneous values when there is a mismatch between prior and the image conditions.

Depth estimation from terrestrial images using supervised DL-based methods has achieved good success. These methods use extensive amounts of training data containing depth ground truth. But this is not possible in the case of UW as there is not a single large-scale UW depth dataset that can support supervised training. An unsupervised network proposed in [48] estimates depth from stereo UW images using the geometry cue. Gupta and Mitra [25] propose an unsupervised single-image UW depth estimation network by learning a mapping between unpaired RGBD hazy terrestrial images and arbitrary UW images. This method relies on the depth of terrestrial hazy images, which may not actually correspond to the characteristics of UW images. Hambarde *et al*. [26] propose a GAN-based UW depth esti-

mation network where they use synthetic UW images generated from NYU dataset [47]. But such images do not fully depict real UW situations.

Along the lines of monocular self-supervised depth estimation from clean terrestrial images [22], a recent work [53] proposes a self-supervised UW depth estimation network by leveraging real UW video sequences taken from the internet. They extract depth from the input UW image using DepthNet and use photometric reprojection loss by warping the restored images obtained from contrast limited adaptive histogram equalization (CLAHE). Since depth and pose are directly derived from hazy UW images, details of the depth map are lacking.

2.2. Underwater image enhancement

Traditional approaches can be broadly classified as model-free color correction methods and model-based enhancement methods. Model-free methods modify each pixel irrespective of the underlying image formation model such as contrast correction and color adjustment [28], Rayleigh-stretching [20], retinex [14, 55], etc. Physical model-based enhancement methods utilize the UW image formation model for estimating the parameters using prior information. Different modifications of DCP [27] have been applied for UW image restoration [7, 11, 43]. Chau *et al*. [50] propose an adaptive attenuation-curve prior and Li *et al*. [35] propose a histogram distribution prior. Berman *et al*. [5] consider different types of water and distinct spectral profiles for each type to refine the restored images. Akkayanak *et al*. [2] rely on the depth map for restoration.

While considering supervised DL-based methods, UW image restoration is challenging mainly because of the scarcity of real datasets with ground truth. To partially overcome this challenge, GAN-based methods [36, 34, 24] have emerged. UWCNN [32] trains 10 image enhancement models corresponding to each water type. [33] created a paired real underwater dataset (UIEB) where pseudo-ground-truth is subjectively selected based on human perception of the outputs of different enhancement techniques. They proposed a gated-fusion network for image enhancement. Methods such as [31] and [49] have used UIEB dataset for supervision. [9] proposes an UW restoration network that uses depth as a cue where depth is estimated from a pre-trained SoTA model [21]. Considering the attenuation coefficient as a cue, [10] proposes a generative model to restore UW images. Along with depth estimation, the unsupervised methods of [26] and [48] find enhanced images using the UW image formation model by utilizing the depth returned from their network. These methods [26, 48] are not end-to-end and they perform image restoration as a post-processing task. Unsupervised method Chai *et al*. [6] and Fu *et al*. [15] perform physics-based disentanglement of underwater images. The work [15] uses a homology constraint on the enhanced image for self-supervision.

2.3. Differences with existing works

Along the lines of UW image restoration works [6] and [15], we disentangle the UW image to perform self-supervision; but there are distinct differences. [15] performs disentanglement based on the simple underwater image formation model [12]. These works [6, 15] do not impose any explicit constraint on the transmission map. During training, they rely on the information from a single frame to output the restored image. In contrast, our method invokes relevant constraints on both the scene radiance and transmission map using adjacent frames to facilitate proper disentanglement. We employ a separate module to obtain the extinction coefficients which are needed to compute depth.

Self-supervised monocular depth estimation works, developed for terrestrial images [22, 57], also use view-synthesis strategy. But their performance on UW images is poor (see Sec. 5). They estimate depth from a fixed clean input image using DepthNet which has a general U-Net architecture. Instead, we derive depth analytically from the transmission map which is estimated by disentangling the UW image. Unlike [22, 57], the input to our PoseNet is the disentangled scene radiance.

Unlike [53] which also utilizes view-synthesis for depth estimation, our loss is calculated on the disentangled scene radiance which is more faithful than using CLAHE. [53] does not use haze as a cue for depth. We use view-synthesis loss on the hazy images also to preserve the relative geometry after restoration.

3. Proposed Method

The basic image formation model for UW images which is based on Koschmieder's light scattering model is as given in [12]. Akkayanak *et al*. [1] observe that direct signal and backscatter are governed by two distinct attenuation coefficients whereas [12] treats them to be the same. The revised UW image formation model as proposed in [1] is given by

$$I(x) = J(x)T^d(x) + (1 - T^b(x))A \tag{1}$$

where x is pixel location, I is the UW image, J is the scene radiance which is the underlying clean image, A is global background light, while T^d and T^b are the transmission maps corresponding to direct signal and backscatter, respectively. Transmission map and distance from the source $D(x)$ are related by $T^d(x) = e^{-\beta_c^d D(x)}$ and $T^b(x) = e^{-\beta_c^b D(x)}$ where β_c^d and β_c^b are the channel-wise extinction coefficients for direct signal and backscatter, respectively.

From an UW image I_1, we aim to estimate both depth D and original image J_1 simultaneously in an end-to-end manner. To train our USe-ReDI-Net, we utilize two adjacent frames I_1 and I_2 from an UW video sequence. Self-supervision is carried out by disentangling the UW image

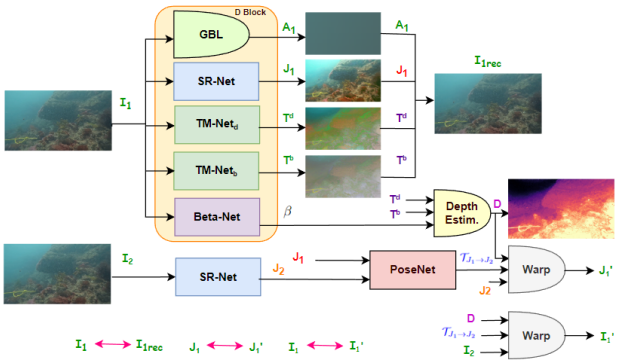


Figure 2: Schematic of our USe-ReDI-Net. Input UW image is disentangled into its latent components: scene radiance, transmission maps, and global background light using scene radiance network (SR-Net), transmission map networks (TM-Net$_d$ and TM-Net$_b$), and global background light module (GBL), respectively. The input is reconstructed using these components. Depth is estimated from the transmission map analytically. Using adjacent frames, the network enforces consistency of view-synthesis to refine depth which in turn improves the disentanglement.

into its latent components based on Eq. (1), and predicting the appearance of a restored target image J_1 from the viewpoint of a restored reference image J_2. During self-supervision, we ensure that the relative geometry between input UW images I_1 and I_2 is preserved in the restored image pair J_1 and J_2. A detailed block diagram of our approach is given in Fig. 2.

3.1. Self-supervision by disentangling UW image, and estimation of depth from haze

We use three networks, namely, scene-radiance network (SR-Net) to estimate J, and transmission map networks (TM-Net$_d$ and TM-Net$_b$) to find T^d and T^b, respectively. SR-Net does not have any downsampling operations to prevent loss of details [30]. TM-Net$_d$ and TM-Net$_b$ find the transmission maps corresponding to direct signal and backscatter respectively, from which we derive the depth map D which can be considered as the depth derived from the haze cue. Unlike [6, 15], we use stride-2 convolution and skip connection in TM-Net to extract features at two different resolutions which help to increase the receptive field in estimating transmission map. For estimating global background light A, we use a global background light estimation module (GBL) which blurs the input image with a Gaussian filter of high standard deviation [15]. GBL renders A to be smooth so that image details are mostly contained in scene radiance and transmission map. To facilitate self-supervision, the disentangled quantities J, T^d, T^b, and A are combined using Eq. (1) to get a reconstructed image I_{rec} (which should ideally be close to I).

Transmission map T is different for all three channels since the attenuation of light in water depends on the wavelength. In order to find depth D from T^d and T^b, we need channel-wise extinction coefficients β_c^d and $\beta_c^b : c = \{R, G, B\}$ which has a dimension 6, and can be estimated from the input image using Beta-Net. Depth D can be written as $D = -\log(T_c^*)/\beta_c^* : * = d$ or b, where T_c is the transmission map for channel c. Details of the network structure of each block are provided in the supplementary.

3.2. Self-supervision by view synthesis, and refining depth from haze using geometry

We leverage self-supervision on view-synthesis to achieve proper disentanglement. We use adjacent frame I_2 which is passed through SR-Net to get the restored image J_2. Thus, we have two restored images; J_2 as the source image and J_1 as the target image from SR-Net. Given target image J_1 and source image J_2, the system is trained to predict the relative camera pose $\mathcal{T}_{J_1 \to J_2}$ between J_1 and J_2, and to refine depth D which is obtained from the transmission maps T^d and T^b. This is how the geometry cue refines the depth obtained from haze. The model predicts the target image J_1 using $\mathcal{T}_{J_1 \to J_2}$ and D from source image J_2 using the reprojection formula,

$$J_1' = J_2 \langle \text{proj}(D, \mathcal{T}_{J_1 \to J_2}, K) \rangle \tag{2}$$

where K is the intrinsic camera matrix, and proj() is the transformation function which maps the target image coordinate x_{J_1} to the source image coordinate x_{J_2} using the relation

$$x_{J_2} = K \mathcal{T}_{J_1 \to J_2} D(x_{J_1}) K^{-1} x_{J_1} \tag{3}$$

We use bilinear sampling from a spatial transformer network (STN) [29] to sample the source images, which is locally sub-differentiable. The predicted target image from Eq. (2) can be used to impose a constraint on the photometric reprojection error of J_1. This constraint forces TM-Net to return an improved estimate of the underlying transmission map which in turn yields a better depth map and proper disentanglement.

In order to preserve the geometry between input images I_1 and I_2 after enhancement, we constrain the pose and depth returned from the enhanced images J_1 and J_2 to be respected by the input UW images I_1 and I_2 also. Hence, we provide another self-supervision by view synthesis on the input images.

We introduce relevant loss functions for self-supervision, as explained next.

3.3. Loss functions

Since we use self-supervision on disentanglement and view-synthesis, we mainly employ two losses for this purpose. We use an edge-aware loss to encourage depth to be locally smooth. In order to enforce depth consistency across R, G, B channels obtained from both T^d and T^b, we use an additional channel-wise depth consistency loss.

Reconstruction Loss, $\mathcal{L}_{rec}$

Using GBL, SR-Net, TM-Net$_d$ and TM-Net$_b$, we disentangle the input image I_1 into its latent components A_1, J_1, T^d, and T^b. In order to perform self-supervision on disentanglement, we combine these components using Eq. (1) to

estimate the input image I_{1rec}. Hence, our reconstruction loss $\mathcal{L}_{rec}$ can be written as

$$\mathcal{L}_{rec} = ||I_{1rec} - I_1||_2^2 \tag{4}$$

View synthesis loss, $\mathcal{L}_{VS}$

For self-supervision on view-synthesis, we use restored images J_1 and J_2. Target image J_1 is predicted as J_1' using camera pose $\mathcal{T}_{J_1 \to J_2}$ between J_1 and J_2, and depth D as given in Sec. 3.2 to derive a loss $\mathcal{L}_{VSd}$. Along with that, we constrain D and $\mathcal{T}_{J_1 \to J_2}$ to be followed by input images I_1 and I_2 as well to derive a loss $\mathcal{L}_{VSh}$. Thus, our total view synthesis loss $\mathcal{L}_{VS}$ can be written as

$$\mathcal{L}_{VS} = \mathcal{L}_{VSd} + \mathcal{L}_{VSh} = ||J_1 - J_1'||_1 + ||I_1 - I_1'||_1 \tag{5}$$

Edge-aware depth smoothness loss, $\mathcal{L}_{ds}$

Depth is mostly smooth except at image gradients. This factor can be used to smoothen depth map D by providing image gradient weightage to the depth gradients. Edge-aware depth smoothness loss $\mathcal{L}_{ds}$ can be written as

$$\mathcal{L}_{ds} = |\partial_x D^*| e^{|-\partial_x J_1|} + |\partial_y D^*| e^{|-\partial_y J_1|} \tag{6}$$

where D^* is the mean-normalized depth which is used to avoid shrinkage of estimated depth [22].

Channel-wise depth consistency loss, $\mathcal{L}_{dc}$

We find depth D analytically from the transmission maps T^d and T^b. There are three different channels for each transmission map since the transmission values depend on the wavelength of light. All three channel-wise extinction coefficients $\beta_c : \{c = (R, G, B)\}$ for both β_c^d and β_c^b are estimated using Beta-Net. The 3-channel transmission maps T_c^d and $T_c^b : \{c = (R, G, B)\}$ are estimated using TM-Net$_\text{d}$ and TM-Net$_\text{b}$, respectively. We then find, $D_R = -\log(T_R)/\beta_R$, $D_G = -\log(T_G)/\beta_G$, and $D_B = -\log(T_B)/\beta_B$ corresponding to both T^d and T^b. In order to arrive at a single depth map, we force them all to be equal. For this, we use channel-wise depth consistency loss $\mathcal{L}_{dc}$ as

$$\mathcal{L}_{dc} = \sum_{x=\{R,G,B\}; y=\{d,b\}} ||D_R^d - D_x^y||_1 \tag{7}$$

where $D_x^y = -\log(T_x^y)/\beta_x$

Total Loss

The total loss of our network is given by

$$\mathcal{L} = \alpha \mathcal{L}_{rec} + \gamma \mathcal{L}_{VS} + \eta \mathcal{L}_{ds} + \lambda \mathcal{L}_{dc} \tag{8}$$

where α, γ, η, and λ are the weights corresponding to different losses. We empirically set $\alpha = 1$, $\gamma = 0.1$, $\eta = 0.005$, and $\lambda = 0.02$.

During test time, only a single image is needed. The test image is passed through SR-Net, TM-Net, and Beta-Net to get the restored image, and the depth map.

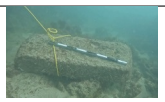
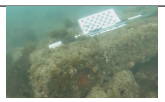
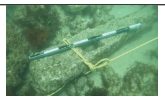
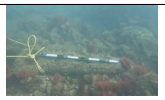
A. No.	Dur. (sec)	Sample frame	A. No.	Dur. (sec)	Sample frame
A1	68		A11	53	
A2	37		A12	50	
A3	69		A13	50	
A4	62		A14	77	
A5	53		A15	28	
A6	73		A16	45	
A7	50		A17	57	
A8	77		A18	73	
A9	54		A19	62	
A10	74		A20	70	

Table 1: Dataset summary with sample frames from each artifact. A. No.: Artifact number, Dur.: Duration.

4. Dataset DRUVA

DRUVA contains videos of 20 different artifacts in shallow waters. The duration and a sample frame from each of these artifacts are given in Table 1. We provide camera intrinsics also with the dataset which is estimated from the MATLAB Camera Calibrator App using a total of 121, 10x7 checkerboard images. All videos were captured under natural illumination at a depth of 3-6 m from the sea surface using a GoPro Hero 10 Black camera of 30 fps, 1920×1080 resolution, with a minimum - maximum ISO: 100-1600, and auto-exposure settings. The artifacts are mainly rocks with shapes ranging from circular to oblong shapes and with dimensions of 0.5-1.5 m. The level of water turbidity fluctuates between being clear and slightly cloudy. Divers went around each artifact when capturing the videos to get full azimuthal coverage. Existing UW datasets [33, 37, 5] contain independent frames with little or no temporal informa-

Dataset	GT available	Type	Task	Phase	# imgs.
DRUVA	×	Video	Depth & Restoration	Training Testing	6000 110
SQUID [5]	✓ (Depth)	Stereo	Depth	Testing	72
Sea-thru [2]	✓ (Depth)	Single	Depth	Testing	50
RUIE [37]	×	Single	Restoration	Testing	100
UIEB [33]	✓ (Pseudo)	Single	Restoration	Testing	190

Table 2: Details of datasets used for comparison. Note that all are real UW datasets. The training dataset from DRUVA contains a total of 6000 frames where a set of 500 frames corresponds to one continuous video sequence of one artifact, and the test dataset is from video sequences that were not used for training. Following [53], we used 72 images from SQUID [5]. From the dataset of Sea-thru [2], we use 50 images with significant depth variations.

Depth estimation methods					
Traditional	DCP [27], UDCP [11], IBLA [44], GDCP [43], HL [5]				
DL-based Training	UW-Net [25] DRUVA$_{raw}$+ NYU [47]	USUIR [15] DRUVA$_{raw}$	Mono2$_h$ [22] DRUVA$_{raw}$	Mono2$_d$ [22] DRUVA$_{res}$	Ours DRUVA$_{raw}$
Testing	DRUVA$_{res}$, SQUID$_{res}$ [5], and Sea-thru$_{res}$ [2] are used for Mono2$_d$. DRUVA$_{raw}$, SQUID [5], and Sea-thru [2] for all the other methods.				
Restoration methods					
Traditional	CLAHE [60], Fusion [3], Hist.prior [35], GDCP [43], IBLA [44]				
DL-based Training	CycleGAN [58] DRUVA$_{raw}$+ UIEB [33] GT	DDIP [18] ×(zero-shot)	USUIR [15] DRUVA$_{raw}$	Ours DRUVA$_{raw}$	
Testing	DRUVA$_{raw}$, RUIE [37], and UIEB [33] for all the methods.				

Table 3: Details of methods used for comparison, and their retraining strategies. The dataset used for training DL-based methods is given at the bottom of each method. DRUVA$_{raw}$ is UW images from DRUVA. DRUVA$_{res}$, SQUID$_{res}$, and Sea-thru$_{res}$ are the restored images from datasets DRUVA, SQUID [5], and Sea-thru [2] using a state-of-the-art UW image restoration method [49]. Note that UW-Net[25], and CycleGAN[58] use two datasets for training. Mono2$_d$, and Mono2$_h$ are two versions (trained differently) of Monodepth2 [22] which is a self-supervised monocular depth estimation method for terrestrial clean images.

tion. DRUVA holds tremendous potential to be harnessed by the research community and can be used for diverse research applications such as UW 3D reconstruction, UW novel view-synthesis using neural radiance fields (NeRFs), UW video interpolation, and extrapolation, to name a few. A video of an artifact from DRUVA along with its 3D reconstruction is included in the supplementary.

5. Experiments

In this section, we first discuss implementational aspects along with the details of the datasets used for comparison. Then we include qualitative and quantitative evaluations of our results. We consider state-of-the-art methods for comparison, both for UW depth estimation and UW image restoration. Ablation studies are included to verify the effectiveness of the different modules of USe-ReDI-Net.

5.1. Datasets and implementation details

Details of the datasets used for comparisons are given in Table 2. For training, we use only DRUVA since USe-ReDI-Net needs neighboring frames. All other datasets are used only for testing. Training is done for 50 epochs using Adam optimizer with a learning rate of 0.0001 on cropped patches of 800 × 800 pixels and batch size of 1. For our framework, intrinsic camera matrix K is available from camera calibration. We conduct our experiments on a PC with Intel Xeon CPU, 24 GB RAM, and an NVIDIA GeForce RTX3090 GPU.

5.2. Performance comparison

Details of the depth estimation and restoration methods used for comparison are given in Table 3. We do not compare supervised methods since they cannot be trained using DRUVA due to the unavailability of ground truth. Also, USe-ReDI-Net cannot be trained on any other real UW dataset since none of them is a video dataset. SQUID [5] is a stereo dataset, but it contains only 57 stereo pairs, which is insufficient for training purpose. The results of all baselines are obtained from the source codes provided by the respective authors.

5.2.1 Qualitative and quantitative evaluation of depth

In Fig. 3, we give comparison results of the estimated depth map from different methods for an image from DRUVA, two images from SQUID [5], and one image from the dataset of Sea-thru [2]. For better visual comparison, some portions of the input images from DRUVA dataset (1(a)) are marked with red and blue rectangles where red rectangle portions are visually at a lesser depth and blue rectangles show regions with transitions in depth. On DRUVA dataset, DCP [27] and UDCP [11] do not emphasize the nearer portions, and the depth at the transition regions is also not proper. IBLA [44] fails to detect segments at low depths. Performance of GDCP [43] and HL [5] is poor at artifact boundaries. UW-Net [25] does not detect nearer regions and its depth map is visually unrealistic and somewhat resembles the depth map of indoor objects in NYU Depth v2 [47] which is the secondary dataset used for its training along with the primary dataset DRUVA. The depth map obtained from the transmission map returned from USUIR [15] is erroneous. Mono2$_h$ [22] does not perform well as it is trained directly on UW images. We have included its visual results in the supplementary. As given in Table 3, Mono2$_d$ [22] was trained on the restored frames (using [49]) of DRUVA. Depthmap returned from Mono2$_d$ [22] is also not good. This can be considered as a sequential process, where restoration is done first followed by depth estimation. Mono2$_d$ [22] does not use haze as a cue as it only deals with restored images. Only the proposed USe-ReDI-Net performs consistently well over the entire range of depth values. For SQUID [5] dataset and the dataset of Sea-thru [2], only the depth map returned by our method is close to ground truth. UW-Net [25] returns depth maps that have an unrealistic appearance.

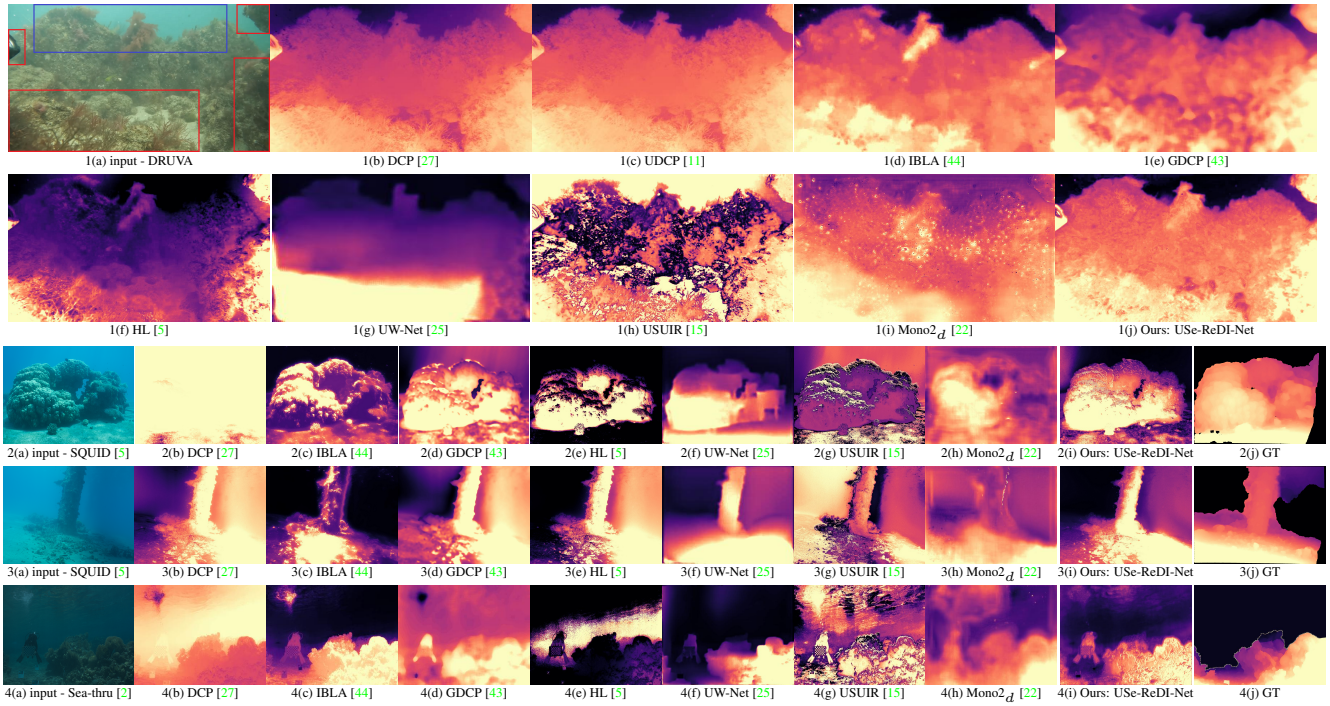


Figure 3: Input UW image (a) from datasets: (1) - DRUVA, (2,3) - SQUID [5], (4) - from Sea-thru [2] with ground truth (2(j), 3(j), and 4(j)) and the depth map obtained from different methods. Note that our USe-ReDI-Net gives visually plausible depth maps for all three datasets.

In order to quantitatively evaluate the depth map accuracy, we used 72 images from SQUID dataset [5] and 50 images from the dataset proposed by Sea-thru [2] which have ground truth depth maps. Since all the methods produce depth maps up to a scale, two scale invariant metrics are used for evaluating depth prediction accuracy. 1) SI-MSE: scale-invariant mean squared error (in log space) [13], and 2) Pearson correlation coefficient (ρ) [5] defined as $\rho_{D_1,D_2} = \frac{\text{Cov}(D_1,D_2)}{\sigma_{D_1}\sigma_{D_2}}$ where $\text{Cov}(D_1, D_2)$ is the covariance between two depth maps D_1 and D_2, and σ is the standard deviation. ρ and SI-MSE are calculated using the evaluation code provided by [5], and the average metric values obtained for different methods are given in Table 4. It can be seen that USe-ReDI-Net has the highest ρ and the least SI-MSE which is desirable. UW-Net [25] has satisfactory quantitative scores for SQUID [5], but it has a large SI-MSE value on the dataset of Sea-thru [2]. As noted earlier, their depth map is visually unrealistic. The performance of Mono2$_d$ [22] is not good even though it was trained and tested with restored images. Mono2$_h$ [22] trained on our raw UW images does poorly as it gives very low ρ and high SI-MSE values. The performance of USUIR [15] is not acceptable. Among traditional methods, IBLA [44] has reasonable metric values but our method scores the best.

5.2.2 Qualitative and quantitative evaluation of image restoration

For evaluating the performance on image restoration, we provide restored images from different methods in Fig. 4. These include one image from DRUVA, two from RUIE [37], and an image from UIEB [33]. The restoration quality of CLAHE [60], Fusion [3], IBLA [44], GDCP [43], and DDIP [18] is visually poor. Results of [35] are visually pleasing and of high contrast, but there are some color artifacts due to over-restoration. CycleGAN [58] also produces artifacts in the restored images. The results of USUIR [15], which is an unsupervised method, have a foggy appearance. For all four images, USe-ReDI-Net delivers very good visual quality. For the image from UIEB [33] dataset, our output 4(i) is close to ground truth 4(j) while all the other methods perform poorly.

Performance on image restoration is evaluated using two no-reference UW image quality assessment metrics UIQM [41] and UCIQE [52] on test images from DRUVA and RUIE dataset [37] since both do not have ground truth images. On images from UIEB dataset [33] with pseudo-ground-truth, we use full-reference image quality assessment metrics PSNR and SSIM. Average metric values calculated for different methods are given in Table 5. It can be observed that histogram prior [35] has the highest no-reference metric scores on test images from DRUVA as well as RUIE. However, the no-reference metrics have to be judged along with the qualitative results. [35] has significant color deviations at the output as shown in Fig. 4. Even though CycleGAN [58] and DDIP [18] have higher UIQM and UCIQE values on our dataset, their outputs have color deviations as well as artifacts. GDCP [43] has less PSNR and SSIM for UIEB. USe-ReDI-Net, and USUIR [15] have

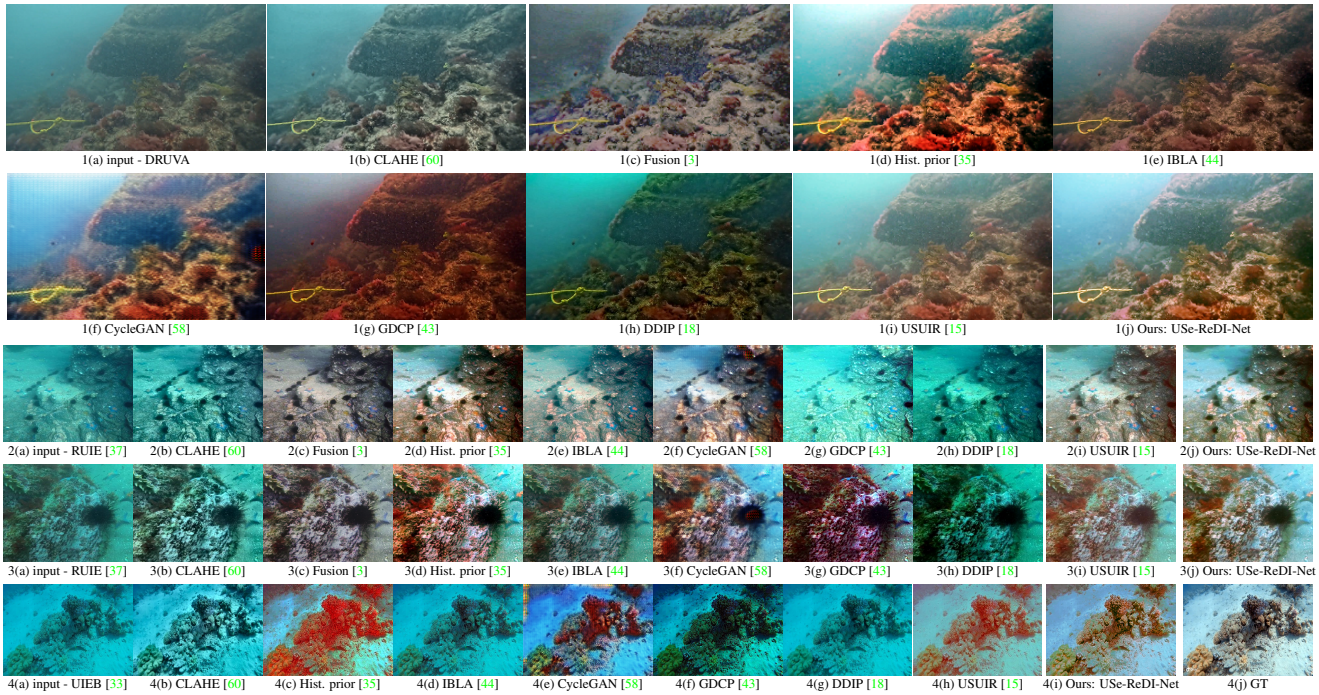


Figure 4: Input UW image (a) from datasets: (1) - DRUVA, (2,3) - RUIE [37], (4) - UIEB [33] with ground truth (4(j)) for UIEB and the enhanced images obtained from different methods. Note that our output results are visually good, and for UIEB dataset, our output (4(i)) is closer to ground truth (4(j)).

Dataset	Method	DCP [27]	UDCP [11]	GDCP [43]	IBLA [44]	HL [5]	UW-Net [25]	USUIR [15]	Mono2$_d$ [22]	Mono2$_h$ [22]	USe-ReDI-Net
SQUID[5]	$\rho\uparrow$	-0.07	0.03	0.28	0.33	0.27	0.49	0.14	0.31	-0.1	0.55
	SI-MSE$\downarrow$	0.72	0.54	0.25	0.20	0.29	0.18	0.48	0.19	0.73	0.16
Sea-thru[2]	$\rho\uparrow$	-0.48	0.39	0.33	0.44	0.45	0.50	0.09	0.37	0.03	0.51
	SI-MSE$\downarrow$	0.69	0.47	0.56	0.47	0.58	1.13	0.69	0.43	0.53	0.40

Table 4: Quantitative comparisons of depth estimation accuracy on SQUID [5] dataset and the dataset proposed by Sea-thru [2] using scale invariant metrics.

Dataset	Ours: DRUVA		RUIE [37]		UIEB [33]	
Method	UIQM$\uparrow$	UCIQE$\uparrow$	UIQM$\uparrow$	UCIQE$\uparrow$	PSNR$\uparrow$	SSIM$\uparrow$
CLAHE [60]	2.67	0.52	1.3	0.45	16.7	0.60
Fusion [3]	3.46	0.52	3.49	0.48	17.5	0.61
Hist. prior [35]	3.41	0.65	4.15	0.67	18.5	0.59
GDCP [43]	2.84	0.54	2.62	0.53	13.3	0.55
IBLA [44]	2.56	0.56	1.73	0.51	14.3	0.57
CycleGAN [58]	3.12	0.62	2.75	0.59	17.0	0.52
DDIP [18]	3.31	0.57	1.85	0.47	12.4	0.38
USUIR [15]	2.86	0.57	3.04	0.57	18.9	0.69
USe-ReDI-Net	2.84	0.59	3.15	0.65	18.9	0.70

Table 5: Quantitative comparisons of enhanced image quality on datasets DRUVA, RUIE [37], and UIEB [33] using image quality assessment metrics. PSNR is in dB.

DDIP[18]	CycleGAN[58]	USUIR[15]	Mono2[22]	UW-Net [25]	USe-ReDI-Net
245 sec	28ms	14ms	25ms	3.5 sec	18ms

Table 6: Execution time of DL-based depth estimation and image restoration methods

the best PSNR and SSIM scores. As observed before, the results of USUIR [15] have some residual fog even at closer depths. Our method performs consistently well (both visually and quantitatively) on all three datasets.

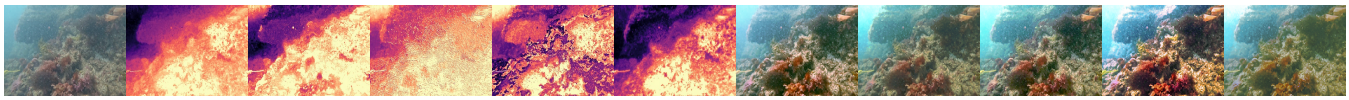
It is important to note that USe-ReDI-Net emerges as the best for depth recovery as well as image restoration as compared to methods devised individually for these twin tasks. More comparison results are given in the supplementary.

5.2.3 Time complexity

The processing time for different DL-based methods on an image of size 512×512 executed on NVIDIA GeForce RTX3090 GPU is given in Table 6. We do not consider traditional methods for comparison as those can be run on a CPU and typically take at least 1 sec for execution. Table 6 shows that the proposed USe-ReDI-Net is computationally most efficient as it returns both the enhanced image as well as depth map in just 18ms. USUIR [15] takes 14ms but it outputs only the restored image. The closest depth estimation method UW-Net [25] takes 3.5 sec to output the depth map from a single frame which is very high.

5.3. Ablation studies

To study the effectiveness of our proposed modules and losses, we conduct ablation studies on UIEB dataset for UW image enhancement, and on SQUID dataset for UW depth estimation. We formed 5 networks, Net1 to Net5 as shown in Table 7. Net2 does not use view-synthesis loss on input UW images; Net3 does not use view-synthesis loss on the enhanced images; Net4 is with only $\mathcal{L}_{rec}$ without any view-synthesis loss; and in Net5, we remove Beta-Net and find depth directly by averaging depth maps derived from the

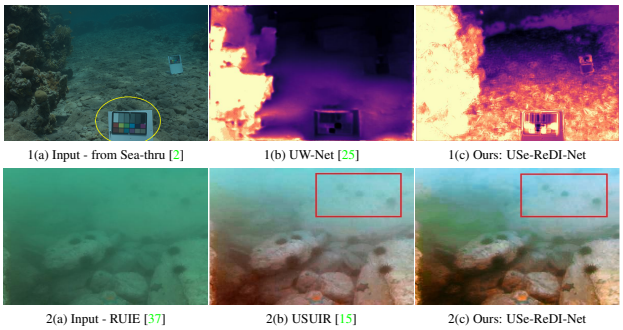


(a) Input (b) Net1 (c) Net2 (d) Net3 (e) Net4 (f) Net5 (g) Net1 (h) Net2 (i) Net3 (j) Net4 (k) Net5

Figure 5: Ablation studies on DRUVA dataset for different configurations and with different combinations of our proposed losses.

N/w	$\mathcal{L}_{rec}$	$\mathcal{L}_{VSd}$	$\mathcal{L}_{VSh}$	Beta-Net	$\rho\uparrow$ /SI-MSE$\downarrow$	PSNR/SSIM$\uparrow$
Net1	✓	✓	✓	✓	0.55/0.16	18.9/0.70
Net2	✓	✓	×	✓	0.38/0.28	18.7/0.68
Net3	✓	×	✓	✓	0.31/0.20	18.5/0.67
Net4	✓	×	×	✓	0.18/0.31	17.2/0.64
Net5	✓	✓	✓	×	0.31/0.17	17.4/0.58

Table 7: Ablation studies on SQUID [5] for depth, and UIEB [33] for enhancement.



1(a) Input - from Sea-thru [2] 1(b) UW-Net [25] 1(c) Ours: USe-ReDI-Net

2(a) Input - RUIE [37] 2(b) USUIR [15] 2(c) Ours: USe-ReDI-Net

Figure 6: Failure case for (1) depth estimation on the dataset of Sea-thru [2], and (2) image restoration on RUIE [37] dataset.

6 transmission maps. Hence, we do not use channel-wise depth consistency loss, $\mathcal{L}_{dc}$, in Net5. An input UW image from our DRUVA dataset, and the corresponding depth map output and enhanced image output for Net1 to Net5 are given in Fig. 5. If we do not constrain depth and pose to be respected by the input UW images (Net2), the depth map obtained is not good (see Fig. 5 and Table 7). By comparing Net2 and Net3, we note that $\mathcal{L}_{VSd}$ is more important than $\mathcal{L}_{VSh}$. Qualitative and quantitative results of Net5 reveal the importance of Beta-Net and $\mathcal{L}_{dc}$ loss. Without Beta-Net, both the depth prediction accuracy and enhancement quality are reduced. Net4 can be considered as the case which utilizes only the haze cue for depth estimation without using the geometry cue. From the results, it is evident that geometry cue (view-synthesis loss) plays an important role in depth prediction which in turn aids image enhancement.

Figure 6 shows a failure case. Results are given for USe-ReDI-Net and closest SoTA method (UWNet [25] for depth and USUIR [15] for restoration). USe-ReDI-Net returns inconsistent depths for objects which have reflective surfaces (see encircled region). UW-Net [25] also struggles. In fact, its depth prediction is poor in most places. At image portions with a highly foggy appearance, all the methods struggle, as expected. Notably, USe-ReDI-Net does not introduce any color deviations.

6. Conclusions

In this work, we dealt with the problem of self-supervised UW depth estimation and image restoration from a single UW image utilizing cues from haze and geometry of UW images. For self-supervision, our USe-ReDI-Net uses the physical model of UW image formation and view-synthesis based on the depth map and camera pose. We constrain both the enhanced image and transmission map using neighboring frames in the input UW video. Experiments demonstrate that USe-ReDI-Net surpasses SoTA methods in terms of depth accuracy, visual quality, and execution speed as well. The proposed UW video dataset DRUVA, with its unique features, can be greatly leveraged by the research community.

Acknowledgement

Support provided by the Department of Science and Technology, India through project No. EE1920271DSTX005001 is gratefully acknowledged. We thank Dr. Sundaresh and his team from National Institute of Oceanography, Goa for helping us in collecting the underwater data.

References

[1] Derya Akkaynak and Tali Treibitz. A revised underwater image formation model. In *CVPR*, pages 6723–6732, 2018. 1, 2, 3

[2] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *CVPR*, pages 1682–1691, 2019. 3, 6, 7, 8, 9

[3] Cosmin Ancuti, Codruta Orniana Ancuti, Tom Haber, and Philippe Bekaert. Enhancing underwater images and videos by fusion. In *CVPR*, pages 81–88, 2012. 6, 7, 8

[4] Codruta O. Ancuti, Cosmin Ancuti, Christophe De Vleeschouwer, Laszlo Neumann, and Rafael Garcia. Color transfer for underwater dehazing and depth estimation. In *ICIP*, pages 695–699, 2017. 1, 2

[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. 1, 2, 3, 5, 6, 7, 8, 9

[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. 2, 3, 4

[7] John Y. Chiang and Ying-Ching Chen. Underwater image enhancement by wavelength compensation and dehazing. *TIP*, 21(4):1756–1769, 2012. 3

[8] Enrique Coiras, Yvan Petillot, and David M. Lane. Multiresolution 3-d reconstruction from side-scan sonar images. *TIP*, 16(2):382–390, 2007. 1

[9] Chaitra Desai, Sujay Benur, Ramesh Ashok Tabib, Ujwala Patil, and Uma Mudenagudi. Depthcue: Restoration of underwater images using monocular depth as a clue. In *WACVW*, pages 196–205, 2023. 3

[10] Chaitra Desai, Badduri Sai Sudheer Reddy, Ramesh Ashok Tabib, Ujwala Patil, and Uma Mudenagudi. Aquagan: Restoration of underwater images. In *CVPRW*, pages 295–303, 2022. 3

[11] 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. 1, 2, 3, 6, 7, 8

[12] P. Drews Jr, E. do Nascimento, F. Moraes, S. Botelho, and M. Campos. Transmission estimation in underwater single images. In *ICCVW*, pages 825–830, 2013. 3

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

[14] Xueyang Fu, Peixian Zhuang, Yue Huang, Yinghao Liao, Xiao-Ping Zhang, and Xinghao Ding. A retinex-based enhancing approach for single underwater image. In *ICIP*, pages 4572–4576, 2014. 3

[15] 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. 2, 3, 4, 6, 7, 8, 9

[16] Y. Lim G. Kim, W. Kim J. Park, and H. Cho D. Lee. Single-energy material decomposition in radiography using a three-dimensional laser scanner. In *Journal of the Korean Physical Society*, volume 75, pages 153–159, 2019. 1

[17] Adrian Galdran, David Pardo, Artzai Picón, and Aitor Alvarez-Gila. Automatic red-channel underwater image restoration. *Vis. Commun. Image Represent.*, 26:132–145, 2015. 2

[18] Yosef Gandelsman, Assaf Shocher, and Michal Irani. "double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *CVPR*, pages 11018–11027, 2019. 6, 7, 8

[19] Ravi Garg, Vijay Kumar B.G., Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *ECCV*, pages 740–756, Cham, 2016. 1

[20] Ahmad Shahrizan Abdul Ghani and Nor Ashidi Mat Isa. Underwater image quality enhancement through rayleigh-stretching and averaging image planes. *Int. J. Nav. Archit. Ocean Eng.*, 6(4):840–866, 2014. 3

[21] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 6602–6611, 2017. 1, 3

[22] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3827–3837, 2019. 1, 3, 5, 6, 7, 8

[23] Yinbin Guo. 3d underwater topography rebuilding based on single beam sonar. In *ICSPCC*, pages 1–5, 2013. 1

[24] Yecai Guo, Hanyu Li, and Peixian Zhuang. Underwater image enhancement using a multiscale dense generative adversarial network. *IEEE J. Ocean. Eng.*, 45(3):862–870, 2020. 3

[25] Honey Gupta and Kaushik Mitra. Unsupervised single image underwater depth estimation. In *ICIP*, pages 624–628, 2019. 1, 2, 6, 7, 8, 9

[26] Praful Hambarde, Subrahmanyam Murala, and Abhinav Dhall. Uw-gan: Single-image depth estimation and image enhancement for underwater images. *IEEE Trans. Instrum. Meas.*, 70:1–12, 2021. 1, 2, 3

[27] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *PAMI*, 33(12):2341–2353, 2011. 2, 3, 6, 7, 8

[28] Kashif Iqbal, Michael Odetayo, Anne James, Rosalina Abdul Salam, and Abdullah Zawawi Hj Talib. Enhancing the low quality images using unsupervised colour correction method. In *2010 IEEE International Conference on Systems, Man and Cybernetics*, pages 1703–1709, 2010. 3

[29] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, NIPS'15, page 2017–2025, Cambridge, MA, USA, 2015. 4

[30] 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. 2, 4

[31] Chongyi Li, Saeed Anwar, Junhui Hou, Runmin Cong, Chunle Guo, and Wenqi Ren. Underwater image enhancement via medium transmission-guided multi-color space embedding. *TIP*, 30:4985–5000, 2021. 2, 3

[32] Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognition*, 98:107038, 2020. 3

[33] 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. 2, 3, 5, 6, 7, 8, 9

[34] Chongyi Li, Jichang Guo, and Chunle Guo. Emerging from water: Underwater image color correction based on weakly supervised color transfer. *SPL*, 25(3):323–327, 2018. 2, 3

[35] 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. 1, 3, 6, 7, 8

[36] Jie Li, Katherine A. Skinner, Ryan M. Eustice, and Matthew Johnson-Roberson. Watergan: Unsupervised generative network to enable real-time color correction of monocular underwater images. *IEEE Robotics and Automation Letters*, 3(1):387–394, 2018. 3

[37] 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. 2, 5, 6, 7, 8, 9

[38] M. Ludvigsen, G. Johnsen B. Sortland, and H. Singh. Applications of geo-referenced underwater photo mosaics in ma-

rine biology and archaeology. In *Oceanography*, volume 20, page 140–149, 2007. 1

[39] Patrizio Mariani, Iñaki Quincoces, Karl H. Haugholt, Yves Chardard, Andre W. Visser, Chris Yates, Giuliano Piccinno, Giancarlo Reali, Petter Risholm, and Jens T. Thielemann. Range-gated imaging system for underwater monitoring in ocean environment. *Sustainability*, 11(1), 2019. 1

[40] Charles H. Mazel. In situ measurement of reflectance and fluorescence spectra to support hyperspectral remote sensing and marine biology research. In *OCEANS 2006*, pages 1–4, 2006. 1

[41] 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

[42] Kaustubh Pathak, Andreas Birk, and Narunas Vaskevicius. Plane-based registration of sonar data for underwater 3d mapping. In *IROS*, pages 4880–4885, 2010. 1

[43] 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. 1, 2, 3, 6, 7, 8

[44] Yan-Tsung Peng and Pamela C. Cosman. Underwater image restoration based on image blurriness and light absorption. *TIP*, 26(4):1579–1594, 2017. 1, 2, 6, 7, 8

[45] Luke K. Rumbaugh, Erik M. Bollt, William D. Jemison, and Yifei Li. A 532 nm chaotic lidar transmitter for high resolution underwater ranging and imaging. In *2013 OCEANS*, pages 1–6, 2013. 1

[46] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, pages 770–779, 2019. 1

[47] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, Berlin, Heidelberg, 2012. 2, 3, 6

[48] Katherine A. Skinner, Junming Zhang, Elizabeth A. Olson, and Matthew Johnson-Roberson. Uwstereonet: Unsupervised learning for depth estimation and color correction of underwater stereo imagery. In *ICRA*, pages 7947–7954, 2019. 2, 3

[49] 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. 3, 6

[50] Yi Wang, Hui Liu, and Lap-Pui Chau. Single underwater image restoration using adaptive attenuation-curve prior. *IEEE Trans. Circuits Syst. I Regul. Pap.*, 65(3):992–1002, 2018. 3

[51] Yi Wu, Yaqin Zhou, Shangjing Chen, Yunpeng Ma, and Qingwu Li. Defect inspection for underwater structures based on line-structured light and binocular vision. *Appl. Opt.*, 60(25):7754–7764, Sep 2021. 1

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

[53] Xuewen Yang, Xing Zhang, Nan Wang, Guoling Xin, and Wenjie Hu. Underwater self-supervised depth estimation. *Neurocomputing*, 2022. 1, 3, 6

[54] J. Yuh and M. West. Underwater robotics. *Advanced Robotics*, 15(5):609–639, 2001. 1

[55] Shu Zhang, Ting Wang, Junyu Dong, and Hui Yu. Underwater image enhancement via extended multi-scale retinex. *Neurocomputing*, 245:1–9, 2017. 3

[56] Weidong Zhang, Peixian Zhuang, Hai-Han Sun, Guohou Li, Sam Kwong, and Chongyi Li. Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement. *TIP*, 31:3997–4010, 2022. 1

[57] 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, 3

[58] 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. 6, 7, 8

[59] Yuliang Zou, Pan Ji, Quoc-Huy Tran, Jia-Bin Huang, and Manmohan Chandraker. Learning monocular visual odometry via self-supervised long-term modeling. In *ECCV*, pages 710–727, Cham, 2020. 1

[60] Karel Zuiderveld. Contrast limited adaptive histograph equalization. In *Graphic Gems IV. San Diego: Academic Press Professional*, page 474–485, 1994. 6, 7, 8