Sea-ing in Low-light

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

Abstract

Underwater (UW) robotics applications require depth and restored images simultaneously in real-time, irrespective of whether the UW images are captured in good lighting conditions or not. Most of the UW image restoration and depth estimation methods have been devised for images under normal lighting. Consequently, they struggle to perform on poorly lit images. Even though artificial illumination can be used when there is insufficient ambient light, it can introduce non-uniform lighting artifacts in the restored images. Hence, the recovery of depth and restored images directly from Low-Light UW (LLUW) images is a critical requirement in marine applications. While a few works have attempted LLUW image restoration, there are no reported works on joint recovery of depth and clean image from LLUW images. We propose a Self-supervised Low-light Underwater Image and Depth recovery network (SelfLUID-Net) for joint estimation of depth and restored image in real-time from a single LLUW image. We have collected an Underwater Low-light Stereo Video (ULVStereo) dataset which is the first-ever UW dataset with stereo pairs of low-light and normally-lit UW images. For the dual tasks of image and depth recovery from a LLUW image, we effectively utilize the stereo data from ULVStereo that provides cues for both depth and illumination-independent clean image. We harness a combination of the UW image formation process, the Retinex model, and constraints enforced by the scene geometry for our self-supervised training. To handle occlusions, we additionally utilize monocular frames from our video dataset and propose a masking scheme to prevent dynamic transients, that do not respect the underlying scene geometry, from misguiding the learning process. Evaluations on five LLUW datasets demonstrate the superiority and generalization ability of our proposed SelfLUID-Net over existing state-of-the-art methods. The dataset ULVStereo is available at https://github.com/nishavarghese15/ULVStereo.

1. Introduction

Underwater (UW) image and depth recovery pose difficulties due to color cast and haziness caused by absorption and scattering of light in water. The problem is compounded when these images are captured in dim ambient light influenced by factors such as cloudiness, time of the season, higher depths, etc. UW exploration [53, 58, 90] requires restoration of images and depth maps from the captured UW observations, regardless of whether they are well-lit or not. Existing methods for UW depth [25, 62, 72] and restoration [5, 15, 47, 71] that work in adequately illuminated UW images are not well-suited for dimly-lit scenarios. Although low-light enhancement methods [23, 36, 45, 55] for terrestrial images have shown significant advancements, they cannot be directly applied to low-light UW (LLUW) images due to the presence of haze and color-cast. Our work enables 'sea-ing' in low-light by real-time recovery of images and depth maps from dimly lit underwater observations.

There exist a few works for LLUW image enhancement [29, 34, 56, 84]. However, to the best of our knowledge, there has been no attempt to jointly recover both depth and clean image from a LLUW image. Another fundamental challenge in LLUW research is the need for a proper dataset. The works [84, 95] propose paired UW low-light datasets where low-illumination is added synthetically to existing UW dataset [43]. But this does not characterize a real UW low-illumination scenario. [57] proposes a real low-light UW dataset. But it is quite small with only 183 images and the scenes are not diverse. [29] proposes a dataset NUID by collecting real low-light UW images from several existing UW datasets [31, 43, 57] and from the internet. But these datasets [29, 57] contain only LLUW images, which are better-suited for inference.

Down-stream tasks such as UW object classification, target recognition, and tracking require restored images and depth maps in real-time. Although traditional methods [5, 62, 63] based on the UW image formation model perform both UW image restoration and depth map estimation, their processing time is very high. Deep-learning (DL) methods are better-suited to meet the real-time requirements of these tasks since they need only a single pass during inference. Due to the unavailability of *real* paired datasets with Ground Truth (GT) in UW scenario, supervised networks [43, 44, 79, 84], trained on synthetic or paired dataset with pseudo-GT, are not reliable for UW image restoration and depth estimation. Unpaired learning-based [25] and Generative Adversarial Network (GAN)-based [26] methods, that use either terrestrial RGBD datasets [68] or synthetically generated datasets for training, also pose domain gap issues for UW images. To circumvent the need for real paired or unpaired datasets, several works have emerged with self-supervision [8, 15, 72, 88] where only real UW images are used for training. But, they are devised for normally-lit UW images.

In this paper, we propose a self-supervised learning strategy for joint estimation of depth and latent image from a single LLUW image in real-time. We refer to it as Self-supervised Low-light Underwater Image and Depth recovery network (SelfLUID-Net). Our method is primarily applicable for near-shore situations and when i) natural light is insufficient while operating at low depths or ii) under normal ambient lighting but at relatively higher depths. To the best of our knowledge, [72] is the only other self-supervised approach that estimates both depth and restored image simultaneously from a single UW image. Akin to [72], we also make use of cues from haze and geometry for estimating depth. But there are major differences. [72] was devised only for well-lit UW images. Our method handles LLUW images by additionally utilizing Retinex theory. For better depth estimation, we model scene geometry using the illumination-independent reflectance component whereas [72] uses geometry between scene radiance without accounting for illumination effects. Our self-supervision framework utilizes both stereo and monocular cues, unlike [72] that uses only monocular frames for training. We additionally propose a moving pixel masking scheme to account for moving transients (such as plants, fishes, etc.) in monocular frames which do not respect the static scene geometry. [72] does not address this issue.

We have captured a UW stereo dataset, Underwater Low-light Stereo Video (ULVStereo), using two cameras, with different exposure settings. The cameras are positioned with a known baseline on a stereo-rig which is designed ruggedly to prevent any camera shake during data capture under water. ULVStereo is the first-ever low-light UW dataset containing low-light and normally-lit stereo pairs of real UW scenes. To facilitate LLUW image restoration and depth estimation, we judiciously leverage the inherent characteristics of ULVStereo. We effectively utilize the stereo pair to extract cues for both depth and clean image. We combine the physics of UW image formation, Retinex theory, and constraints induced by the scene geometry for self-supervised training. The input LLUW image is disentangled into its latent components to get the scene illumination, reflectance of the scene (which is independent of the scene illumination) which we take to be the clean image (following [55][50][93]), transmission maps, and global background light. According to Retinex model, a low-light and a normally-lit image of a scene share the same reflectance. This view-consistency of reflectances disentangled from the stereo pair of our dataset is enforced during self-supervision. Depth is first estimated directly from the transmission map based on their relationship, which is refined using the geometric relationship between reflectance of image pairs. Since the view-consistency constraint inherently involves depth as well as reflectance, their coupling is mutually beneficial during self-supervision. To accommodate any occluded pixels in stereo view, we additionally utilize the geometry between monocular frames from the LLUW video of ULVStereo. However, monocular frames may contain small moving objects that do not respect the geometry of the underlying static scene. To address this issue, we propose a moving pixel masking scheme.

Our main contributions are given below.

1. We propose a self-supervised network (SelfLUID-Net) that effectively utilizes constraints from the physics of image formation model (UW image formation and Retinex theory) and scene geometry (relation connecting two images via depth map) for LLUW image and depth recovery.
2. Unlike existing UW works that ignore the effect of illumination in modeling scene geometry, we utilize the illumination-independent image components in the geometry constraint to accurately estimate the depth map.
3. Ours is the first work to *jointly* recover image and depth from a single LLUW image in real-time. It is the first to propose depth estimation *directly* from a LLUW image.
4. We propose a masking scheme to prevent moving transients from hampering the training process.
5. We propose the first-ever Underwater Low-light Stereo Video (ULVStereo) dataset with low-light and normally-lit UW image pairs that can be used by researchers for diverse underwater applications.
6. SelfLUID-Net is computationally very efficient (62 fps) and outperforms the state-of-the-art methods for image restoration as well as depth estimation. Its generalization ability is verified on real LLUW datasets captured under actual lowlight conditions (in shallow water with insufficient light and at higher depths upto 10m).

2. Related works

2.1. UW image restoration and depth estimation

For UW image restoration, traditional methods either use UW image formation model [12] or they simply enhance the visual image quality using Rayleigh-stretching [19], contrast correction [30, 92, 99], Retinex [14, 91, 98], etc. Works that utilize UW image formation model [5, 10, 11, 46, 63, 78, 83] estimate both depth and restored image using suitably-chosen priors in their optimization process. [2] utilizes the depth map for restoration. Traditional methods are time-consuming and return inaccurate results due to mismatch between the adopted prior and actual UW conditions.

Supervised DL-based methods for UW image restoration and depth are unreliable due to scarcity of real UW datasets with GT. Supervised UW image enhancement works such as [43], [44], [79], [37] and [94] used the paired UW dataset UIEB [43] which has subjectively selected ground truth for supervised training. [27] proposed a paired dataset HICRD with image GT generated from the measured attenuation coefficients and assumed depth. [74] and [86] used synthetically generated UW datasets for supervision. The depth estimation methods [89] and [73] used a paired UW dataset USOD10K [28] that contains GT depth maps of UW images returned from a transformer model for terrestrial depth estimation [65]. However, GT in the paired datasets used by these aforementioned methods is not real. To circumvent the requirement of paired data, unsupervised or GAN-based methods for image restoration ([24, 42, 47]) and depth ([25, 26]) utilizing unpaired datasets have evolved. However, these unpaired datasets do not fully depict real UW situations. Recently, several self-supervised methods [8, 15, 71, 72] have been proposed that utilize only the input UW images for training. Image restoration methods [8, 15, 71] utilize UW image formation model for supervision. [72] returns both the depth map and restored image simultaneously by integrating the UW image formation model with the geometry constraint between neighboring frames in a UW video. Recent works [3, 72, 75, 88] utilizes a self-supervised approach proposed in [20] for UW depth estimation. But all these networks are designed only for normally-lit images.

Several attempts have been made to address dynamic scenes in self-supervised depth prediction from terrestrial images. [20] proposes a masking scheme to remove stationary pixels that move with the same velocity as the camera. [35] utilizes information from several source images. [39] uses semantic guidance, [17] utilizes radar information, and [21] makes use of optical flow.

2.2. LL image enhancement and depth estimation

Traditional low-light image enhancement methods [59, 67, 70] that use modifications of histogram equalization suffer from saturation effects. Retinex decomposition [41, 77] improves contrast by decomposing the input image into illumination and reflectance. [23] first refines the initial illumination map and then combines it with the denoised reflectance to obtain the output. Supervised DL-based methods that utilize Retinex decomposition [7, 52, 81, 82] heavily depend on paired data. The unsupervised method of [55] proposes a self-calibrated illumination learning framework for fast and robust image restoration. [50] proposes an architecture search strategy. A zero-shot work [9] uses different no-reference training losses. The self-supervised work of [96] incorporates bilateral learning into the Retinex model. Without any paired or unpaired data, the works [22, 45] design a network to learn an image-specific curve, that can effectively map the low-light image to an enhanced image.

Self-supervised approaches have been proposed for depth estimation from dark or night-time images. [69] addresses it by learning a cross-domain dense feature representation. By exploiting radar as a supervision signal, [17] shows better performance in night scenes. [76] uses an image enhancement module and a prior-based regularization. A partially shared network for day and night images is used by [49]. [18] proposes a training strategy to enhance the robustness of depth estimation models in diverse conditions.

2.3. Low-light UW image enhancement

Recently, DL-based works [34, 84, 85, 95, 97] and traditional methods [29, 48, 56, 57, 60] have been proposed for LLUW image enhancement. Traditional methods [56] and [57] utilize the UW image formation model and the local contrast information in the input image patches. [29] incorporates the Retinex model within an illumination-channel sparsity prior (ICSP) guided optimization framework. [60] addresses non-uniform illumination in UW by equalizing the illumination component, using non-linear guided filtering. An inverted UW image is used to model the low-light characteristics of UW images in [48]. [84] and [95] are DL-based supervised networks for LLUW image restoration where they synthetically add low-illumination to real UW images for the training data. Supervised network [34] synthesized LLUW-paired data by introducing haziness to terrestrial low-light datasets. However, the low illumination or haziness in these datasets is not real. The unsupervised approach in [97] proposes a GAN-based network that learns the mapping from LLUW images to normal terrestrial images. But this mapping to land images is not realistic for UW images. [85] includes a low-light enhancement branch similar to [45] with no-reference loss functions for training. However, it has a pre-defined exposure level at the output that is not generalizable to different images.

To the best of our knowledge, there are no reported works for depth estimation directly from a LLUW image. A recent work [51] achieves depth estimation from a LLUW image but it does so in two steps. It first estimates the clean image from a LLUW image using the pre-trained weights from a UW image restoration method [16] (strictly speaking this is not correct as [16] is not devised for LLUW images), and then estimates depth map from the clean image utilizing monocular depth estimation methods for terrestrial images [66] [20]. Hence, the performance suffers as the quality of the estimated depth map is directly linked to the output of [16] which is poor for low-light input images.

3. ULVStereo dataset

Underwater Low-light Stereo Video (ULVStereo) dataset captured by us is the first-ever dataset of normally lit and low-light UW stereo-videos. It contains pairs of videos of 10 different submerged UW 3D structures found at a depth of 4-7 m from the sea surface. All videos were captured using two GoPro Hero10 cameras at 30 fps. The two cameras are separated by a known baseline and fixed to a rugged stereo rig. With a minimum ISO, exposure time of both cameras is adjusted (the least exposure time 1/480 sec for Camera 1 and Auto mode in Camera 2) such that one captures LLUW videos while the other captures the corresponding normally lit counterpart. Exposure control is a common practice in paired data generation for LL terrestrial image restoration [4, 40, 81]. When ambient light is dim, one can capture a normally-lit image by increasing the exposure time of the camera. But such an image will contain motion blur due to the transients in water. It may be noted that the normally-lit image in the pair is not the exact GT of the clean image corresponding to LLUW image, but is a UW hazy image captured in normal light. Sample frames and the statistics of the 10 pairs of videos from ULVStereo dataset are given in the supplementary. A stereo image pair is shown in Fig. 1.

Strictly speaking, in ULVStereo, LLUW images captured by altering the exposure settings of the camera do not depict exact LLUW conditions. However, acquiring real low light and the corresponding normally lit image pairs from the same viewpoint is not practically feasible, more so in UW scenarios. Moreover, safety concerns preclude capturing UW images in the deep sea or at night. The captured videos from both cameras are time-synchronized with DaVinci Resolve software [6] utilizing sound waves. We manually selected pairs (normally lit and low-light) of video segments from the synchronized videos where each segment contains a single UW structure. The two cameras were calibrated with a 10x7 checkerboard pattern. It was found that the stereo rig has a small but non-negligible translation in directions other than the intended baseline. Along with the dataset, we also provide camera intrinsics as well as extrinsic parameters (obtained from MATLAB stereo camera calibrator app). For evaluation, we have used real LLUW images captured at depth of upto 10m from four other datasets along with ULVStereo.

4. Proposed method

For a scene radiance (clean image) J and global background light A, with transmission maps T_D and T_B corresponding to direct signal and backscatter, respectively, the observed UW image I at a pixel x is given by [1]

$$I(x) = J(x)T_D(x) + (1 - T_B(x))A \tag{1}$$

If $\mathcal{D}(x)$, β_D^c and β_B^c are the depth of the image pixel x, and channel-wise extinction coefficients for direct signal and backscatter, respectively, then $T_D(x) = e^{-\beta_D^c \mathcal{D}(x)}$ and $T_B(x) = e^{-\beta_B^c \mathcal{D}(x)}$.

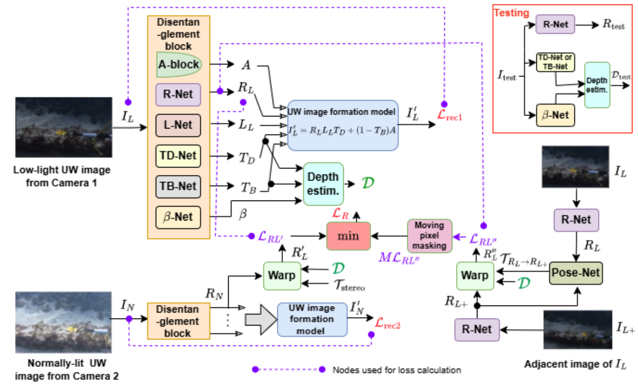


Figure 1. Schematic of our SelfLUID-Net. Input LLUW image I_L from Camera 1 is disentangled into its latent components (A, R_L, L_L, T_D, and T_B) using the disentanglement block where R_L is the restored image. Depth $\mathcal{D}$ of I_L is estimated directly from the transmission maps and β. The reflectance R_N from the corresponding normally-lit UW image I_N from Camera 2 is warped to the viewpoint of Camera 1 using $\mathcal{D}$ and the known relative pose $\mathcal{T}_{\text{stereo}}$ between the two cameras. The consistency between R_L and the warped reflectance from R_N is used for self-supervision. The consecutive frames from the low-light video are used to solve for occluded pixels and a masking scheme is proposed to address dynamic transients in the video. The network structure of each block is given in the supplementary.

According to classical Retinex theory [23, 41], the observed scene radiance can be represented as a product of reflectance R and illumination L where the reflectance is the desired clean image (which is independent of the scene illumination). Inspired by this theory, LLUW image enhancement works [29, 84] represent the original scene radiance of a LLUW image J as the product of its reflectance R_L and illumination L_L. Following this representation, a low-light UW image I_L can be written as

$$I_L(x) = R_L(x)L_L(x)T_D(x) + (1 - T_B(x))A \tag{2}$$

The block diagram of our SelfLUID-Net is given in Fig. 1. We aim to determine reflectance R_L, which we treat as the clean image and depth $\mathcal{D}$ from a single LLUW image I_L. Consider a time-synchronized image pair (I_L and I_N) from our ULVStereo dataset. The corresponding reflectances disentangled from I_L and I_N are R_L and R_N. We impose reflectance consistency constraint on R_L and R_N, i.e., the low-light/normally-lit pair must share the same reflectance. Since the images in a stereo pair are related by a known geometry (with relative pose $\mathcal{T}_{\text{stereo}}$), R_L can be predicted from the viewpoint of R_N utilizing depth $\mathcal{D}$. Hence, $\mathcal{D}$ also gets refined by the reflectance consistency constraint. Training with only the stereo image pairs suffers from issues due to occlusions and out-of-view pixels. We address this by utilizing adjacent frames from the low-light UW monocular video and propose a pixel masking scheme to prevent moving transients from adversely influencing the learning process. An outline of our training strategy is given below:

1. Input LLUW image I_L is disentangled into its latent components. R-Net, L-Net, TD-Net, and TB-Net estimate R_L, L_L, T_D, and T_B, respectively. The A-block estimates the global background light A by using Gaussian filtering with a high standard deviation [15, 72]. We

impose illumination smoothness constraint [22, 81] to derive L_L. To accomodate wavelength dependent attenuation of light intensity that falls on UW objects, we assume a three channel illumination map L_L. R_L, T_D, T_B, and A also contain three separate channels. The disentangled components are combined using the LLUW image formation model (Eq. 2) to get I_L back.

2. Depth of I_L is estimated from transmission maps T_D and T_B using the relation $\mathcal{D} = \{-\log(T_*)/\beta_* : * = D \text{ or } B\}$ where β, the extinction coefficient is estimated using β-Net. β-Net returns a 6-valued vector as each 3-value vector is used with T_D or T_B. We constrain depth maps, derived from all the pairs of β and transmission map channels, to be equal [72].
3. **Reflectance consistency constraint:** We utilize the illumination-independent reflectance to model the scene geometry. Reflectance R_N corresponding to the normally-lit UW image I_N is also disentangled using the same R-Net. R_N, which we refer to as source reflectance, if viewed from the pose of the low-light camera, must resemble R_L (derived from I_L), which we refer to as target reflectance. Using the estimated depth $\mathcal{D}$, the known intrinsic matrices of two cameras K_1 and K_2, and the relative pose between the two cameras $\mathcal{T}_{stereo}$, R_N is warped to R_L using the reprojection formula [20],

$$R'_L = R_N\langle \text{proj}(\mathcal{D}, \mathcal{T}_{\text{stereo}}, K_1, K_2)\rangle \tag{3}$$

where proj() maps the target image coordinate x_{R_L} to the source image coordinate x_{R_N} using the relation,

$$x_{R_N} = K_2 \mathcal{T}_{\text{stereo}} \mathcal{D}(x_{R_L}) K_1^{-1} x_{R_L} \tag{4}$$

We use locally sub-differentiable bilinear sampling from a spatial transformer network [32] to sample source images. We impose reflectance consistency constraint between R_L and warped reflectance R'_L which forces a) R-Net to return an improved reflectance, and b) TD-Net, TB-Net, and β-Net to estimate a better depth map $\mathcal{D}$.
4. **Addressing occluded or out-of-view pixels:** In the stereo data, there can be pixels that are visible in the low-light target image but are not visible (are either occluded or out-of-view) in the corresponding normally-lit source image. During training, even if our network estimates the correct depth in the target image, the absence of such pixels in the source image can adversely affect the reprojection error (difference between the target pixel intensity and warped pixel intensity from source image). [20] introduces a per-pixel minimum reprojection error where they calculate reprojection error for several source images, and for each pixel, they take the minimum photometric reprojection error. In a similar vein, for any target reflectance R_L, we also take two source reflectances. One is R_N which is from the normally-lit image captured and the other is R_{L+} which is the reflectance from the adjacent monocular frame of the LLUW image I_L. For warping R_{L+}, we use equations similar to Eqns. (3) and (4), but with $K_1 = K_2$ and the relative camera pose between consecutive frames ($\mathcal{T}_{R_L \to R_{L+}}$) estimated from Pose-Net. By utilizing the geometry cue from both stereo and monocular video, we take full advantage of our stereo video dataset.
5. **Moving pixel masking scheme:** Although our training data from the ULVStereo dataset contains images with a majority of static pixels, there are frames with small and dynamic transients such as fish, swaying plants, etc. Because we incorporate monocular video frames for training (to account for occluded pixels as explained in the previous point), we need to mitigate the effect of moving pixels on the reprojection error. Here, we introduce a masking scheme that removes the pixels with a high reprojection error from the loss calculation. We introduce a threshold $\tau = \mu + k\sigma$ where k is a hyperparameter, while μ and σ are the mean and standard deviation of the reprojection errors for all the pixels in the target image. Pixels that have per-pixel reprojection error greater than τ are masked for the final loss calculation. Hence, our final per-pixel photometric reprojection loss is taken to be the minimum of per-pixel reprojection loss from stereo $\mathcal{L}_{RL'}$ and the masked per-pixel reprojection loss $M\mathcal{L}_{RL''}$ from monocular video frames (see Fig. 1).

4.1. Loss functions

Reconstruction Loss, $\mathcal{L}_{\text{rec}}$: We combine the disentangled components of the input low-light UW image I_L using Eq. 2 to get I'_L. Similarly, we obtain I'_N from the disentangled components of I_N. Hence, the reconstruction loss $\mathcal{L}_{\text{rec}}$ is given by

$$\mathcal{L}_{\text{rec}} = ||I'_L - I_L||_2^2 + ||I'_N - I_N||_2^2 \tag{5}$$

Photometric reprojection loss, $\mathcal{L}_R$: The reflectance consistency constraint between R_L and R'_L can be written as the per pixel reprojection loss from stereo $\mathcal{L}_{RL'} = |R_{Lp} - R'_{Lp}|$ for any pixel p. Similarly, R''_L is estimated from R_{L+} using a similar relation as in Eq. 3, and the corresponding per pixel reprojection loss from monocular video is given by $\mathcal{L}_{RL''} = |R_{Lp} - R''_{Lp}|$. Our proposed mask for removing moving pixels, M can be written as $M = [\![\mathcal{L}_{RL''} < \tau]\!]$ where $[\![\ ]\!]$ is the Iversion bracket. Our final per pixel reprojection loss is

$$\mathcal{L}_{Rp} = \min\left(\mathcal{L}_{RL'}, M\mathcal{L}_{RL''}\right) \tag{6}$$

$\mathcal{L}_{Rp}$ is averaged over all pixels to get final photometric reprojection loss $\mathcal{L}_R$.

Illumination smoothness loss, $\mathcal{L}_{\text{is}}$: We enforce smoothness of illumination L using the loss $\mathcal{L}_{\text{is}}$ which is given by,

$$\mathcal{L}_{\text{is}} = \sum_{i=\text{low light, normally lit}} ||\nabla L_i||_1 \tag{7}$$

Edge-aware depth smoothness loss, $\mathcal{L}_{\text{ds}}$: Edge-aware depth smoothness loss $\mathcal{L}_{\text{ds}}$ [20, 72] imposes depth smoothness except at image gradients.

$$\mathcal{L}_{\text{ds}} = |\partial_x D^*| e^{-|\partial_x R_L|} + |\partial_y D^*| e^{-|\partial_y R_L|} \tag{8}$$

where D^*, the mean-normalized depth is used to avoid shrinkage of depth [20].

Channel-wise depth consistency loss, $\mathcal{L}_{\text{dc}}$: To arrive at a single depth $\mathcal{D}$ from the 3 channel transmission maps (T_D and T_B) and 6-valued vector β, as in [72], we use channel-wise depth consistency loss $\mathcal{L}_{\text{dc}}$ as

$$\mathcal{L}_{\text{dc}} = \sum_{x=\{\mathcal{R},\mathcal{G},\mathcal{B}\};y=\{D,B\}} ||\mathcal{D}_D^R - \mathcal{D}_y^x||_1 \tag{9}$$

where $\mathcal{D}_y^x = -\log(T_y^x)/\beta_x$.

Color loss, $\mathcal{L}_{\text{clr}}$: To correct for potential color deviations in R, a color loss $\mathcal{L}_{\text{clr}}$ [15] is added based on the gray-world assumption of natural image statistics. If μ denotes mean and $\Omega = \{\mathcal{R}, \mathcal{G}, \mathcal{B}\}$ is the color channels in R,

$$\mathcal{L}_{\text{clr}} = \sum_{c\in\Omega} ||\mu(R^c) - 0.5||_2^2 \tag{10}$$

Spatial consistency loss, $\mathcal{L}_{\text{spa}}$: To mitigate the effect of noise in the reflectance R, we impose spatial consistency loss that preserves the difference of neighboring regions between the scene radiance J ($= R{\cdot}L$) and its restored version R and this is given by,

$$\mathcal{L}_{\text{spa}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in\mathcal{S}(i)}\left(|(J_i^{\text{avg}} - J_j^{\text{avg}}| - |(R_i^{\text{avg}} - R_j^{\text{avg}})|\right)^2 \tag{11}$$

where i is a local region (N number of local regions) and $j \in \mathcal{S}(i)$ are its four neighbouring regions. J^{avg} and R^{avg} are the average intensity values of the local region in J and R, respectively. We choose the size of local region as 4×4.

Total Loss: The total loss of our network is given by

$$\mathcal{L} = \alpha\mathcal{L}_{\text{rec}} + \gamma\mathcal{L}_R + \zeta\mathcal{L}_{\text{is}} + \eta\mathcal{L}_{\text{ds}} + \lambda\mathcal{L}_{\text{dc}} + \delta\mathcal{L}_{\text{clr}} + \omega\mathcal{L}_{\text{spa}} \tag{12}$$

where α, γ, ζ, η, λ, δ, and ω are the weights corresponding to different losses. We set $\alpha = 1.5$, $\gamma = 0.05$, $\zeta = 3$, $\eta = 50$, $\lambda = 0.02$, $\delta = 0.2$, and $\omega = 0.001$ using grid-search.

During test time, a single low-light UW image is passed to R-Net to get the restored image R and the depth $\mathcal{D}$ is estimated from the transmission map (T_D or T_B returned from TD-Net or TB-Net) and β returned by β-Net.

5. Experiments

5.1. Training and evaluation settings

Our model is trained using cropped patches of 800 × 800 pixels with a learning rate of 3×10^{-6} for 20 epochs using Adam optimizer with a batch size of 1. We took $k = 1$ in calculating $\tau = \mu + k\sigma$ as it was found to be good empirically. Experiments are conducted on a PC with an Intel Xeon CPU, 24 GB RAM, and an NVIDIA GeForce RTX3090 GPU.

Datasets: For training, we have used 6000 frames from both low-light and normally-lit videos of ULVStereo dataset. For testing, we have used: ULVStereo (65 images), Sea-thru [2] (50 images), NUID [29] (115 images), FLSea [64] (50 images), and UIEB$_{\text{dark}}$ where UIEB$_{\text{dark}}$ is formed from 50 dark images of UIEB [43] dataset. Sea-thru and FLSea has GT for depth and UIEB$_{\text{dark}}$ has pseudo-GT (selected subjectively) for restored images. It is to be noted that the datasets that we have used for testing (NUID, UIEB$_{\text{dark}}$, Sea-thru, and FLSea) are real UW datasets with poorly illuminated images which are captured under normal camera settings in actual low light conditions whereas our training dataset ULVStereo contains LLUW images captured by adjusting the exposure settings of the camera.

Table 1. Quantitative comparisons of enhanced image quality on datasets ULVStereo, Sea-thru [2], NUID [29], and UIEB$_{\text{dark}}$ [43]. PSNR is in dB. The best and the second-best entries are highlighted in red and blue, respectively. Trad.: Traditional, SS.: Self-supervised, UnS.: Unsupervised. URN: USe-ReDI-Net, ZD: ZeroDCE.

Dataset		Ours: ULVStereo		Sea-thru [2]		NUID [29]		UIEB$_{\text{dark}}$ [43]	
Category	Method	UCIQE↑	UIQM↑	UCIQE↑	UIQM↑	UCIQE↑	UIQM↑	PSNR↑	SSIM↑
Trad. UW	HL [5]	0.63	4.65	0.68	4.56	0.61	3.65	15.38	0.55
	HLRP [98]	0.65	2.47	0.64	4.43	0.63	2.04	12.34	0.27
	UNTV [83]	0.56	2.84	0.63	4.65	0.51	2.49	16.01	0.50
	MMLE [92]	0.61	3.16	0.66	4.00	0.58	2.56	17.01	0.55
	CBLA [33]	0.62	3.08	0.58	4.10	0.55	1.98	16.70	0.53
SS. UW	USUIR [15]	0.61	2.83	0.62	4.81	0.60	2.63	16.92	0.54
	URN [72]	0.52	2.13	0.52	4.44	0.60	2.01	16.13	0.50
Trad. LLUW	ICSP [29]	0.61	4.17	0.58	2.67	0.58	2.11	14.72	0.52
	L^{2}UWE [56]	0.63	4.84	0.59	2.86	0.60	2.63	14.82	0.54
Trad. LL	LIME [23]	0.52	2.12	0.46	1.67	0.54	1.28	13.24	0.41
UnS. LL	ZD [22]	0.39	2.01	0.44	5.02	0.48	1.32	12.11	0.49
	ZD++ [45]	0.44	1.96	0.45	2.29	0.51	2.73	12.45	0.40
	RUAS [50]	0.52	1.61	0.47	1.58	0.56	1.37	8.28	0.33
	SCI [55]	0.54	2.46	0.50	2.82	0.55	1.58	13.96	0.52
Ours	SelfLUID-Net	0.66	2.80	0.74	4.82	0.63	3.18	17.21	0.58

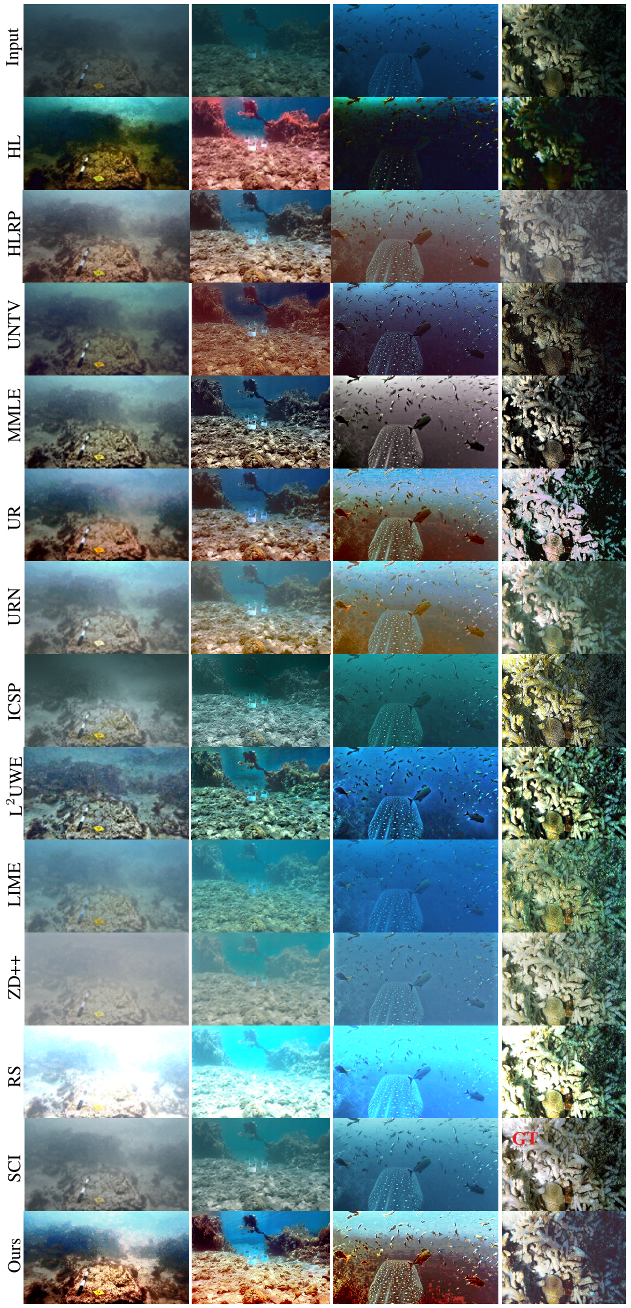
Methods for comparison are provided in Tables 1 and 2. All DL methods are re-trained using images from ULVStereo dataset. We have not taken any supervised methods since those methods cannot be trained using ULVStereo due to the unavailability of GT. Since there is no other real dataset with LL and normally-lit UW image pairs, SelfLUID-Net cannot be trained on any other dataset. Results of all baselines are obtained from the source codes provided by the respective authors.

Evaluation metrics: For image quality assessment, we have used 1) no-reference metrics: UIQM [61] and UCIQE [87] for datasets without GT; 2) PSNR/SSIM for UIEB$_{\text{dark}}$. Since all methods produce depth maps up to a scale, for evaluating depth prediction accuracy, two scale invariant metrics are used. 1) SI-MSE: scale-invariant mean squared error [13]; 2) Pearson correlation coefficient (ρ) [5]: $\rho_{D_1,D_2} = \frac{\text{Cov}(D_1,D_2)}{\sigma_{D_1}\sigma_{D_2}}$; $\text{Cov}(D_1, D_2)$ is the covariance between depth maps D_1 and D_2, and σ is standard deviation.

5.2. Qualitative and quantitative evaluations

In the figures and tables, we have used these abbreviations for different methods. UR: USUIR[15], URN: USe-ReDI-Net[72], ZD: ZeroDCE[22], RS: RUAS[50], Mn2: Mono2[20], HD: HR-Depth[54], MD: Manydepth[80], UWN: UW-Net[25].

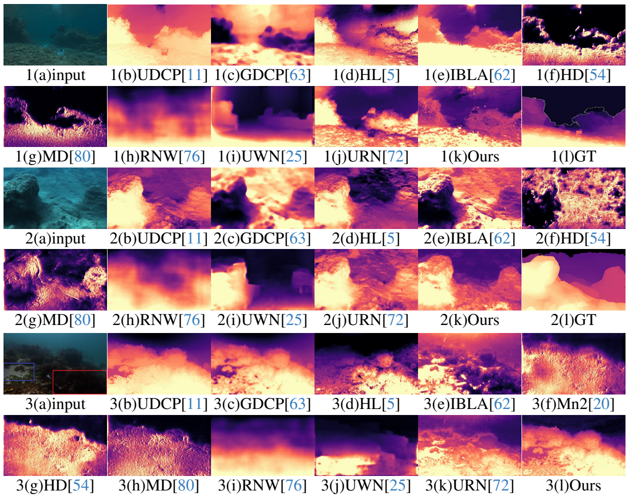
Image restoration: In Fig. 2, we provide restoration results from different methods for one LLUW image each from ULVStereo, Sea-thru [2], NUID [29], and UIEB$_{\text{dark}}$ [43] datasets. Traditional UW image restoration methods HL [5], HLRP [98], UNTV [83], and MMLE [92] struggle to handle LLUW images. Lowlight image restoration meth-



ULVStereo Sea-thru [2] NUID [29] UIEB$_{\text{dark}}$ [43]

Figure 2. Comparison of enhanced underwater images from different methods. Pseudo-GT for UIEB$_{\text{dark}}$ is given in the second-to-last row instead of SCI output.

ods LIME [23], ZeroDCE++ [45], RUAS [50], and SCI [55] remove lowlight effects, but are unable to remove UW haze from images. The output of the LLUW image restoration method ICSP [29] is not good whereas L^{2}UWE [56] brightens the darker areas, but is unable to remove haze. Self-supervised methods USe-ReDI-Net [72] and USUIR [15] struggle in darker regions. Only our method enhances darker regions along with color restoration for all four images. For the first two images, our output has the most visibility compared to others. For the third image, our method reveals some orange-colored fishes and background regions. For the fourth image (UIEB$_{\text{dark}}$ dataset), our output is close to GT (provided in the second-to-last row).



1(a)input 1(b)UDCP[11] 1(c)GDCP[63] 1(d)HL[5] 1(e)IBLA[62] 1(f)HD[54]
1(g)MD[80] 1(h)RNW[76] 1(i)UWN[25] 1(j)URN[72] 1(k)Ours 1(l)GT
2(a)input 2(b)UDCP[11] 2(c)GDCP[63] 2(d)HL[5] 2(e)IBLA[62] 2(f)HD[54]
2(g)MD[80] 2(h)RNW[76] 2(i)UWN[25] 2(j)URN[72] 2(k)Ours 2(l)GT
3(a)input 3(b)UDCP[11] 3(c)GDCP[63] 3(d)HL[5] 3(e)IBLA[62] 3(f)Mn2[20]
3(g)HD[54] 3(h)MD[80] 3(i)RNW[76] 3(j)UWN[25] 3(k)URN[72] 3(l)Ours

Figure 3. Input UW image (a) from datasets: (1) - Seathru [2], (2) - FLSea [64], and (3) - ULVStereo. Ground truth depth map (1(l) and 2(l)) is given for Seathru (GT is obtained using stereo images) and FLSea datasets. The depth map obtained from different methods are shown. Note that SelfLUID-Net returns plausible depth maps [(1(k) and 2(k)) are closer to GT (1(l) and 2(l))]. It is to be noted that GT image for Seathru dataset contains black regions where depth is undefined due to the inability of stereo to predict GT depth for the homogeneous sky region. Such undefined regions must be discarded while comparing the results.

For quantitative comparisons, average metric values are given in Table 1. Our SelfLUID-Net has the best PSNR and SSIM scores for UIEB$_{\text{dark}}$. Our UCIQE [87] values are the best. We have the second best UIQM values for Sea-thru [2] and NUID [29] datasets. It is well-known that, for an accurate assessment of image quality, UIQM should be used alongside other metrics and subjective evaluations as well. For ULVStereo, even though L^{2}UWE [56] and HL [5] have the highest UIQM scores, their outputs are not visually good. Our method consistently gives excellent results (both visually and quantitatively) on all four datasets.

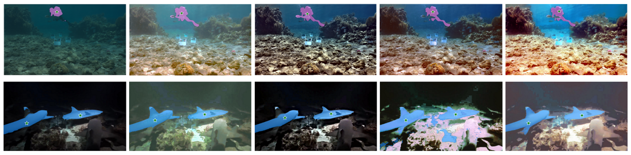
Depth estimation: The depth map estimated from different methods are given in Fig. 3 and the average metric values are given in Table 2. Depth prediction of IBLA is wrong for FLSea [64] (Fig. 3:2(e)) and ULVStereo (Fig. 3:3(e)) datasets whereas GDCP returns wrong depth maps for Sea-thru [2] (Fig. 3:1(c)) and FLSea [64] (Fig. 3:2(c)) datasets. UDCP and HL struggle in predicting good depth map for Sea-thru dataset (Fig. 3:1(b,d)). HR-Depth [54] and Many-depth [80] predict wrong depth values in dark regions, especially for Sea-thru and FLSea datasets (Fig. 3:(1,2)(f,g)). RNW [76] is devised for depth estimation from LL terrestrial images and it struggles to predict good depth map from LLUW images (see Fig. 3:(1,2)(h) and (3)(i)). The depth maps returned by UW-Net [25] are also not good. USe-ReDI-Net [72] struggles to output the transitions in depth. Our method returns plausible depth maps for ULVStereo dataset (see Fig. 3:(3)(a) with a red rectangle at a lower depth and a blue rectangle where the sea-bed is at a higher

Table 2. Quantitative comparisons of depth estimation accuracy on Sea-thru [2] and FLSea [64] datasets for the methods UDCP [11], GDCP [63], HL [5], IBLA [62], Mn2 [20], HD [54], MD [80], RNW [76], UWN [25], URN [72], and Ours. Trad.: Traditional, SST: Self-supervised method for Terrestrial images, UnS.: Unsupervised.

Category		Trad. UW				SST			SST-LL	UnS. UW		
Method		[11]	[63]	[5]	[62]	[20]	[54]	[80]	[76]	[25]	[72]	Ours
Sea-thru	$\rho\uparrow$	0.39	0.33	0.44	0.49	-0.11	0.04	0.08	0.25	0.42	0.43	0.52
	SI-MSE↓	0.46	0.56	0.58	0.43	0.92	0.74	0.81	0.62	0.82	0.42	0.40
FLSea	$\rho\uparrow$	0.49	-0.01	0.61	0.29	-0.06	-0.02	-0.04	0.33	0.49	0.59	0.71
	SI-MSE↓	0.25	0.55	0.23	0.24	0.52	0.78	0.75	0.49	0.28	0.22	0.18

Table 3. Execution time in milliseconds for a 512x512 image.

UR	URN	ZD	ZD+	RS	SCI	Mn2	HD	MD	RNW	UWN	Ours
14	18	8	5	28	30	25	40	54	20	25	16



(a) LLUW image (b) URN [72] (c) MMLE [92] (d) UR [15] (e) Ours

Figure 4. Segmentation output (pink in row 1 and blue in row 2) from SAM using point-prompts (green stars indicated for diver in the first row and fishes in the second row) on (a) LLUW images and (b)-(e) restored outputs from different methods. Please note that segmentation mask generated is better on our restored images.

depth.) Our outputs are closer to GT for Seathru [2] and FLSea [64]. In Table 2, our method has the best metrics scores for both datasets.

Execution time for all DL methods is given in Table 3. Our method provides both depth and image in just 16ms. [72] also returns both outputs in 18ms, but our results are significantly better.

To demonstrate the effectiveness of SelfLUID-Net on a downstream task, we performed prompt based segmentation using the foundation model SAM [38] on two input LLUW images, and the corresponding restored images from the top 4 methods: USeReDINet, MMLE, USUIR, and ours. The results are given in Fig. 4. Segmentation on the first LLUW image only segments half of the diver. In the outputs from USeReDINet and MMLE, SAM merges fishes also in the segmented output. Compared to USUIR, segmentation output from our restored output is better as it could capture the hand region of the diver clearly. For the second example, segmentation on only the restored image from SelfLUID-Net has the full structure of fish, including its fins. Segmented output on the restored image from USUIR erroneously contains other portions also along with the fishes.

5.3. Ablation studies

We conducted ablation studies specifically on the photometric reprojection loss, the effect of stereo and monocular cues, proposed mask, and spatial consistency loss. Reconstruction loss, illumination smoothness loss, and channel-wise depth consistency loss cannot be excluded from loss calculations since they enforce constraints on the physics of image formation model. Edge-aware depth smoothness loss and color loss are commonly used in literature [15, 20, 72]. The subfigures (2) to (6) in Fig. 5 are the outputs obtained after removing the specific losses that are given in the figure caption. The average PSNR (in dB)/SSIM on UIEB$_{\text{dark}}$ [43] and ρ/SI-MSE on FLSea [64] dataset for each configuration is given in Table 4. If we remove $\mathcal{L}_R$ (Fig. 5:(2)), the network does not utilize the geometry cue for depth estimation and struggles to predict depth only from haze. Without the reflectance and stereo geometry cue from the normally-lit UW image (Fig. 5:(3)), the network is unable to produce a good reflectance and depth from the LLUW image. Fig.5:(4) shows that, if we only utilize stereo pairs for training, the performance is marginally less than that with both stereo and monocular frames. If we do not use the proposed masking scheme (Fig. 5:(5)), the network performance is lower. Spatial consistency loss $\mathcal{L}_{\text{spa}}$ improves the result (Fig. 5:(6)) by removing noise in the reflectance returned from R-Net. In Fig. 5 (right-side), the mask predicted for a target image is also shown. It can be seen that the mask removes the adverse effect of moving pixels in the areas of rope and moving plants inside the sea.

Table 4. Ablation study. Average PSNR (in dB)/SSIM on UIEB$_{\text{dark}}$ [43] is given in the first row and ρ/SI-MSE on FLSea [64] dataset is given in the second row.

SelfLUID-Net	Remove $\mathcal{L}_R$	Remove $\mathcal{L}'_{R1}$	Remove $M\mathcal{L}''_{R1}$	Remove mask M	Remove $\mathcal{L}_{spa}$
17.21/0.58	16.33/0.52	16.20/0.50	16.82/0.53	16.93/0.55	16.50/0.53
0.71/0.18	0.32/0.30	0.50/0.27	0.58/0.23	0.64/0.20	0.68/0.19

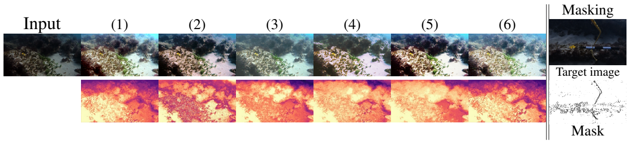


Figure 5. Left (ablation study): Input UW image from ULVStereo and outputs (restored images in the first row and depth maps in the second row) from different combinations of our proposed losses (1-6). (1): SelfLUID-Net, (2): Remove $\mathcal{L}_R$, (3): Remove $\mathcal{L}'_{R1}$, (4): Remove $M\mathcal{L}''_{R1}$, (5): Remove mask M, (6): Remove $\mathcal{L}_{spa}$. Right: For a target image, the predicted mask to avoid moving pixels is shown.

In supplementary, a discussion on lowlight underwater scenarios, results on more LLUW images, network complexity, additional ablations, and limitations are included. We also show that sequential processing of LLUW images (UW haze removal followed by LL restoration, and vice versa, and depth estimation from restored images) yields suboptimal results compared to our method.

6. Conclusions

In this paper, we addressed the twin tasks of image restoration and depth estimation from a single LLUW image using a self-supervised network that was trained with our proposed ULVStereo dataset that contains time-synchronized UW stereo image pairs captured under low-light and normal illumination. We exploit the physics of UW image formation model in conjunction with the Retinex model to disentangle the input LLUW image into its latent components. The constraint that reflectance is invariant to change in illumination settings is enforced for self-supervision, to refine the reflectance (clean image) as well as the depth map. Extensive experiments and evaluations, on LLUW datasets captured under actual lowlight environments, reveal the efficacy of our proposed method. Our proposed dataset can be harnessed 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. 4

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

[3] Shlomi Amitai, Itzik Klein, and Tali Treibitz. Self-supervised monocular depth underwater. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1098–1104, 2023. 3

[4] Nantheera Anantrasirichai, Ruirui Lin, Alexandra Malyugina, and David Bull. Bvi-lowlight: Fully registered benchmark dataset for low-light video enhancement, 2024. 4

[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, 6, 7, 8

[6] Blackmagicdesign. Davinci resolve. 4

[7] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12504–12513, 2023. 3

[8] 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

[9] Yongqiang Chen, Chenglin Wen, Weifeng Liu, and Wei He. A depth iterative illumination estimation network for low-light image enhancement based on retinex theory. *Scientific Reports*, 13:19709, 2023. 3

[10] John Y. Chiang and Ying-Ching Chen. Underwater image enhancement by wavelength compensation and dehazing. *TIP*, 21(4):1756–1769, 2012. 2

[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. 2, 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. 2

[13] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NIPS*, 27, 2014. 6

[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. 2

[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, 2022. 1, 2, 3, 4, 6, 7, 8

[16] Zhenqi Fu, Wu Wang, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Uncertainty inspired underwater image enhancement. In *Computer Vision – ECCV 2022*, pages 465–482, Cham, 2022. Springer Nature Switzerland. 3

[17] Stefano Gasperini, Patrick Koch, Vinzenz Dallabetta, Nassir Navab, Benjamin Busam, and Federico Tombari. R4dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes. In *2021 International Conference on 3D Vision (3DV)*, pages 751–760, 2021. 3

[18] Stefano Gasperini, Nils Morbitzer, HyunJun Jung, Nassir Navab, and Federico Tombari. Robust monocular depth estimation under challenging conditions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8143–8152, 2023. 3

[19] 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. 2

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

[21] Vitor Campanholo Guizilini, Kuan-Hui Lee, Rares Ambrus, and Adrien Gaidon. Learning optical flow, depth, and scene flow without real-world labels. *IEEE Robotics and Automation Letters*, PP:1–1, 2022. 3

[22] Chunle Guo Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1780–1789, 2020. 3, 5, 6

[23] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2017. 1, 3, 4, 6, 7

[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, 3, 6, 7, 8

[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. 2, 3

[27] Junlin Han, Mehrdad Shoeiby, Tim Malthus, Elizabeth Botha, Janet Anstee, Saeed Anwar, Ran Wei, Mohammad Ali Armin, Hongdong Li, and Lars Petersson. Underwater image restoration via contrastive learning and a real-world dataset. *Remote Sensing*, 14(17), 2022. 3

[28] Lin Hong, Xin Wang, Gan Zhang, and Ming Zhao. Usod10k: A new benchmark dataset for underwater salient object detection. *IEEE Transactions on Image Processing*, pages 1–1, 2023. 3

[29] Guojia Hou, Nan Li, Peixian Zhuang, Kunqian Li, Haihan Sun, and Chongyi Li. Non-uniform illumination underwater image restoration via illumination channel sparsity prior. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. 1, 3, 4, 6, 7

[30] 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. 2

[31] Md Jahidul Islam, Youya Xia, and Junaed Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5(2):3227–3234, 2020. 1

[32] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, page 2017–2025, Cambridge, MA, USA, 2015. 5

[33] Manvi Jha and Ashish Kumar Bhandari. Cbla: Color-balanced locally adjustable underwater image enhancement. *IEEE Transactions on Instrumentation and Measurement*, 73:1–11, 2024. 6

[34] Kai Ji, Weimin Lei, and Wei Zhang. A deep retinex network for underwater low-light image enhancement. *Mach. Vision Appl.*, 34(6), 2023. 1, 3

[35] Hualie Jiang, Laiyan Ding, Zhenglong Sun, and Rui Huang. Unsupervised monocular depth perception: Focusing on moving objects. *IEEE Sensors Journal*, 21(24):27225–27237, 2021. 3

[36] Praveen Kandula, Maitreya Suin, and A. N. Rajagopalan. Illumination-adaptive unpaired low-light enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):3726–3736, 2023. 1

[37] Md Raqib Khan, Priyanka Mishra, Nancy Mehta, Shruti S. Phutke, Santosh Kumar Vipparthi, Sukumar Nandi, and Subrahmanyam Murala. Spectroformer: Multi-domain query cascaded transformer network for underwater image enhancement. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1443–1452, 2024. 3

[38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. 8

[39] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-Supervised Monocular Depth Estimation: Solving the Dynamic Object Problem by Semantic Guidance. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[40] Mohit Lamba, Kranthi Kumar Rachavarapu, and Kaushik Mitra. Harnessing multi-view perspective of light fields for low-light imaging. *IEEE Transactions on Image Processing*, 30:1501–1513, 2021. 4

[41] Edwin H Land. The retinex theory of color vision. *Scientific American*, 237(6):108–128, 1977. 3, 4

[42] 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. 3

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

[44] 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. 1, 3

[45] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4225–4238, 2022. 1, 3, 6, 7

[46] 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. 2

[47] 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. 1, 3

[48] Yujie Li, Jianru Li, Yun Li, Hyoungseop Kim, and Seiichi Serikawa. Low-light underwater image enhancement for deep-sea tripod. *IEEE Access*, 7:44080–44086, 2019. 3

[49] Lina Liu, Xibin Song, Mengmeng Wang, Yong Liu, and Liangjun Zhang. Self-supervised monocular depth estimation for all day images using domain separation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12717–12726, 2021. 3

[50] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10556–10565, 2021. 2, 3, 6, 7

[51] Tong Liu, Sainan Zhang, and Zhibin Yu. Redefining accuracy: Underwater depth estimation for irregular illumination scenes. *Sensors*, 24(13), 2024. 3

[52] Xinyi Liu, Qi Xie, Qian Zhao, Hong Wang, and Deyu Meng. Low-light image enhancement by retinex-based algorithm unrolling and adjustment. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2023. 3

[53] M. Ludvigsen, G. Johnsen B. Sortland, and H. Singh. Applications of geo-referenced underwater photo mosaics in marine biology and archaeology. In *Oceanography*, page 140–149, 2007. 1

[54] 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. 6, 7, 8

[55] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image

enhancement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5627–5636, 2022. 1, 2, 3, 6, 7

[56] Tunai Porto Marques and Alexandra Branzan Albu. L2uwe: A framework for the efficient enhancement of low-light underwater images using local contrast and multi-scale fusion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2286–2295, 2020. 1, 3, 6, 7

[57] Tunai Porto Marques, Alexandra Branzan Albu, and Maia Hoeberechts. A contrast-guided approach for the enhancement of low-lighting underwater images. *Journal of Imaging*, 5, 2019. 1, 3

[58] 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

[59] Mila Nikolova and Gabriele Steidl. Fast ordering algorithm for exact histogram specification. *IEEE Transactions on Image Processing*, 23(12):5274–5283, 2014. 3

[60] Yu Ning, Yong-Ping Jin, You-Duo Peng, and Jian Yan. Low-illumination underwater image enhancement based on non-uniform illumination correction and adaptive artifact elimination. *Frontiers in Marine Science*, 10, 2023. 3

[61] Karen Panetta, Chen Gao, and Sos Agaian. Human-visual-system-inspired underwater image quality measures. *IEEE J. Ocean. Eng.*, 41(3):541–551, 2016. 6

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

[63] 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, 7, 8

[64] Yelena Randall and Tali Treibitz. Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets, 2023. 6, 7, 8

[65] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12159–12168, 2021. 3

[66] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(03):1623–1637, 2022. 3

[67] Debdoot Sheet, Hrushikesh Garud, Amit Suveer, Manjunatha Mahadevappa, and Jyotirmoy Chatterjee. Brightness preserving dynamic fuzzy histogram equalization. *IEEE Transactions on Consumer Electronics*, 56(4):2475–2480, 2010. 3

[68] 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

[69] Jaime Spencer, Richard Bowden, and Simon Hadfield. Defeat-net: General monocular depth via simultaneous unsupervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14402–14413, Virtual, 2020. IEEE/CVF. 3

[70] Gabriel Thomas, Daniel Flores-Tapia, and Stephen Pistorius. Histogram specification: A fast and flexible method to process digital images. *IEEE Transactions on Instrumentation and Measurement*, 60(5):1565–1578, 2011. 3

[71] Nisha Varghese and Rajagopalan N Ambasamudram. Re-degradation and contrastive learning for zero-shot underwater image restoration. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023. 1, 3

[72] Nisha Varghese, Ashish Kumar, and A. N. Rajagopalan. Self-supervised monocular underwater depth recovery, image restoration, and a real-sea video dataset. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12214–12224, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[73] Chen Wang, Haiyong Xu, Gangyi Jiang, Mei Yu, Ting Luo, and Yeyao Chen. Underwater monocular depth estimation based on physical-guided transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 3

[74] Hao Wang, Weibo Zhang, Lu Bai, and Peng Ren. Metalantis: A comprehensive underwater image enhancement framework. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–19, 2024. 3

[75] Junting Wang, Xiufen Ye, Yusong Liu, Xinkui Mei, and Jun Hou. Underwater self-supervised monocular depth estimation and its application in image enhancement. *Engineering Applications of Artificial Intelligence*, 120:105846, 2023. 3

[76] Kun Wang, Zhenyu Zhang, Zhiqiang Yan, Xiang Li, Baobei Xu, Jun Li, and Jian Yang. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16035–16044, 2021. 3, 7, 8

[77] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, 22(9):3538–3548, 2013. 3

[78] 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. 2

[79] 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. 1, 3

[80] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1164–1174, 2021. 6, 7, 8

[81] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*. British Machine Vision Association, 2018. 3, 4, 5

[82] Xinxu Wei, Xi Lin, and Yongjie Li. Da-drn: A degradation-aware deep retinex network for low-light image enhancement. *Digital Signal Processing*, 144:104256, 2024. 3

[83] Jun Xie, Guojia Hou, Guodong Wang, and Zhenkuan Pan. A variational framework for underwater image dehazing and deblurring. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3514–3526, 2022. 2, 6

[84] Yaofeng Xie, Zhibin Yu, Xiao Yu, and Bing Zheng. Lighting the darkness in the sea: A deep learning model for underwater image enhancement. *Frontiers in Marine Science*, 9, 2022. 1, 3, 4

[85] Zhichao Xin, Zhe Wang, Zhibin Yu, and Bing Zheng. Ull-slam: underwater low-light enhancement for the front-end of visual slam. *Frontiers in Marine Science*, 10, 2023. 3

[86] Geonmo Yang, Gilhwan Kang, Juhui Lee, and Younggun Cho. Joint-id: Transformer-based joint image enhancement and depth estimation for underwater environments. *IEEE Sensors Journal*, 24(3):3113–3122, 2024. 3

[87] Miao Yang and Arcot Sowmya. An underwater color image quality evaluation metric. *TIP*, 24(12):6062–6071, 2015. 6, 7

[88] Xuewen Yang, Xing Zhang, Nan Wang, Guoling Xin, and Wenjie Hu. Underwater self-supervised depth estimation. *Neurocomputing*, 2022. 2, 3

[89] Boxiao Yu, Jiayi Wu, and Md Jahidul Islam. Udepth: Fast monocular depth estimation for visually-guided underwater robots. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3116–3123, 2023. 3

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

[91] Shu Zhang, Ting Wang, Junyu Dong, and Hui Yu. Underwater image enhancement via extended multi-scale retinex. *Neurocomputing*, 245:1–9, 2017. 2

[92] 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. 2, 6, 8

[93] Yu Zhang, Xiaoguang Di, Bin Zhang, Ruihang Ji, and Chunhui Wang. Better than reference in low-light image enhancement: Conditional re-enhancement network. *IEEE Transactions on Image Processing*, 31:759–772, 2022. 2

[94] Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8281–8291, 2024. 3

[95] Wenfeng Zhao, Shenghui Rong, Jiankang Ma, and Bo He. Nonuniform illumination correction for underwater images through a pseudo-siamese network. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1329–1335, 2022. 1, 3

[96] Zunjin Zhao, Hexiu Lin, Daming Shi, and Guoqing Zhou. A non-regularization self-supervised retinex approach to low-light image enhancement with parameterized illumination estimation. *Pattern Recognition*, 146:110025, 2024. 3

[97] Wang L. Zheng Z, Huang X. Underwater low-light enhancement network based on bright channel prior and attention mechanism. *PLoS One*, 18(2), 2023. 3

[98] Peixian Zhuang, Jiamin Wu, Fatih Porikli, and Chongyi Li. Underwater image enhancement with hyper-laplacian reflectance priors. *IEEE Transactions on Image Processing*, 31:5442–5455, 2022. 2, 6

[99] Karel Zuiderveld. Contrast limited adaptive histograph equalization. In *Graphic Gems IV. San Diego: Academic Press Professional*, page 474–485, 1994. 2