Self-Supervised Monocular Depth Underwater

Shlomi Amitai, Itzik Klein *Senior Member, IEEE*, and Tali Treibitz , The Hatter Dept. of Marine Technologies
Charney School of Marine Sciences, University of Haifa, Haifa, Israel
shlomi.amitai@gmail.com, {kitzik,ttreibitz}@univ.haifa.ac.il

***Abstract*— Depth estimation is critical for any robotic system. In the past years, the estimation of depth from monocular images has shown great improvement. However, in the underwater environment results are still lagging behind due to appearance changes caused by the medium. So far little effort has been invested in overcoming this. Moreover, underwater, there are more limitations to using high-resolution depth sensors, which is a serious obstacle to generating ground truth. So far unsupervised methods that tried to solve this have achieved limited success as they relied on domain transfer from a dataset in the air. We suggest network training using subsequent frames, self-supervised by a reprojection loss, as was demonstrated successfully above water. We propose several additions to the self-supervised framework to cope with the underwater environment and achieve state-of-the-art results on a challenging forward-looking underwater dataset.**

I. Introduction

There is a wide range of target applications for depth estimation, from obstacle detection to object measurement and from 3D reconstruction to image enhancement. Underwater depth estimation (note that here depth refers to the object range, and not to the depth underwater) is important for Autonomous Underwater Vehicles (AUVs) [15] (Fig. 1), localization and mapping, motion planning, and image dehazing [6]. As such inferring depth from vision systems has been widely investigated in the last years. There is a range of sensors and imaging setups that can provide depth, such as stereo, multiple-view, and time-of-flight (ToF) [11], [12], [23]. Monocular depth estimation is different from other vision systems in the sense that it uses a single RGB image with no special setup or hardware, and as such has many advantages. Because of mechanical design considerations, in many AUVs, it is difficult to place a stereo setup with a baseline that is wide enough, so monocular depth there is particularly attractive and can be combined with other sensors (e.g., Sonars) to set the scale.

Monocular depth methods can be trained either supervised or self-supervised. Naturally, supervised methods achieve higher accuracy, however, rely on having a substantial dataset with pairs of images and their ground-truth depth. This is very difficult to achieve underwater as traditional multiple-view methods struggle with appearance changes and are less stable. Additionally, optical properties of water [2] change temporally and spatially, significantly changing scene appearance. Thus, for training supervised methods, a ground-truth dataset is needed for every environment, which is very laborious. Therefore, we chose to develop a self-supervised method, that requires only a set of consecutive frames for training.

Fig. 1: The ALICE autonomous underwater vehicle [15] facing obstacles. Monocular depth maps can aid obstacle avoidance and decision-making.

When testing state-of-the-art monocular depth estimation underwater, new problems arise. Handling underwater scenes requires adding more constraints and using priors. Understanding the physical characteristics of underwater images can assist us in revealing new cues and using them for extracting depth cues from the images.

We improve self-supervised underwater depth estimation with the following contributions: **1)** Examining how the reprojection loss changes underwater, **2)** Handling background areas, **3)** Adding a photometric prior, **4)** Data augmentation specific for underwater. We use the FLSea dataset [27] for training and validation.

II. Related Work

A. Supervised Monocular Depth Estimation

In the supervised monocular depth task a deep network is trained to infer depth from an RGB image using a dataset of paired images with their ground-truth (GT) depth [7], [22]. Reference ground truth can be achieved from a depth sensor or can be generated by classic computer vision methods such as structure from motion (SFM) and stereo. Li et. al [20] suggest collecting the training data by applying SFM on multi-view internet photo collections. Their network architecture is based on an hourglass network structure with suitable loss functions for fine details reconstruction in the depth map. A newer method [3], [28] uses transformers to improve performance.

The research was funded by Israel Science Foundation grant #680/18, the Israeli Ministry of Science and Technology grant #3 − 15621, the Israel Data Science Initiative (IDSI) of the Council for Higher Education in Israel, the Data Science Research Center at the University of Haifa, and the European Union's Horizon 2020 research and innovation program under grant agreement No. GA 101016958.

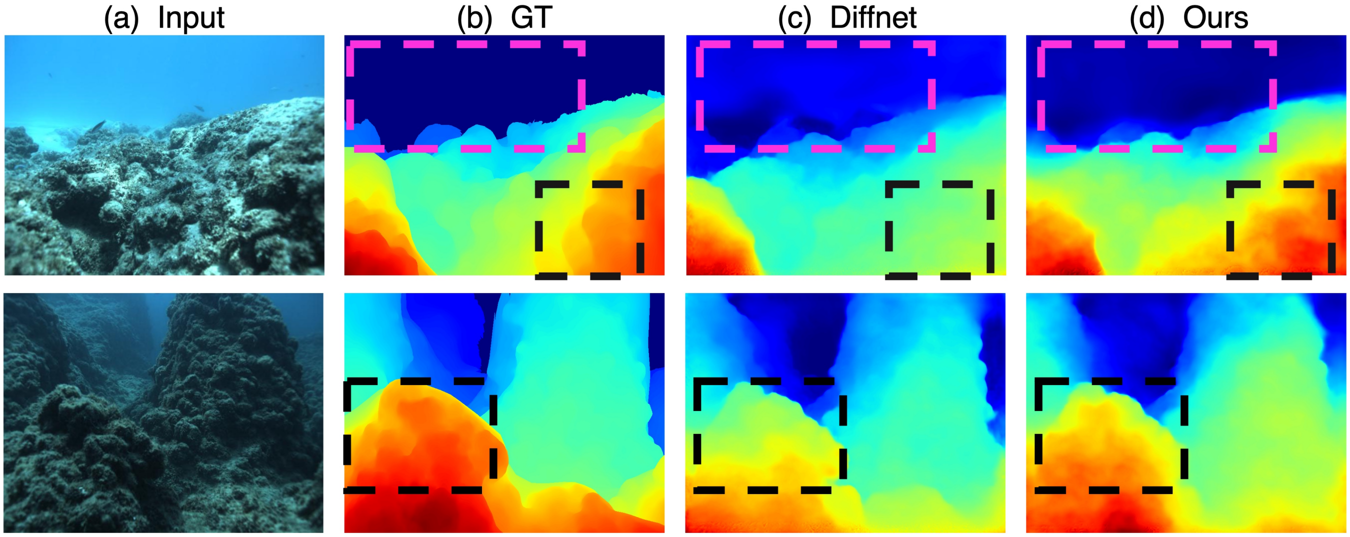


Fig. 2: Example results on two underwater scenes from the FLSea dataset [27]. a) Input scene, b) Ground truth, c) Result of Diffnet [33] and d) our estimated depth map. The magenta rectangle marks the background area where our method significantly improves the results, and black rectangles mark foreground objects where our method improved estimation.

B. *Self-Supervised Monocular Depth Estimation*

To overcome the hurdle of ground-truth data collection, it was suggested [12], [34] to use sequential frames for self-supervised training leveraging the fact that they image the same scene from different poses. The network estimates both the depth and the motion between frames. The estimated camera motion between sequential frames constrains the depth network to predict up-to-scale depth, and the estimated depth constrains the odometry network to predict relative camera pose. The loss is the photometric reprojection error between two subsequent frames using the estimated depth and motion.

Monodepth2 [12] proposed to overcome occlusion artifacts by taking the minimum error between the preceding and following frames. DiffNet [33] is based on monodepth2 [12] with two major differences. They replace the ResNet [18] encoder with high-resolution representations using HRNet [31] which was argued to perform better and added attention modules to the decoder. DiffNet [33] is the current SOTA method on KITTI 2015 stereo dataset [10], the top benchmark for self-supervised monocular depth and also performed the best on our underwater images. Therefore, we base our work on it.

C. *Underwater Depth Estimation*

Underwater, photometric cues have been used for inferring depth from single images, as in scattering media the appearance of objects depends on their distance from the camera. Based on that several priors have been suggested for simultaneously estimating depth and restoring scene appearance.

One line of work is based on the dark channel prior (DCP) [17] and several underwater variants UDCP [5], [8], and the red channel prior [9]. Some methods use the per-patch difference between the red channel and the maximum between the blue and the green as a proxy for distance, termed the maximum intensity prior (MIP) by Carlevaris-Bianco et al. [4]. Song et al. [29] suggested the underwater light attenuation prior (ULAP) that assumes the object distance is linearly related to the difference between the red channel and the maximum blue-green. The blurriness prior [25] leverages the fact that images become blurrier with distance. Peng and Cosman [24] combined this prior with MIP and suggested the image blurring and light absorption (IBLA) prior. Bekerman et al. [2] showed that improving the estimation of the scene's optical properties improves depth estimation.

There have been also attempts at unsupervised learning-based underwater depth estimation. UW-Net [14] uses generative adversarial training by learning the mapping functions between unpaired RGB-D terrestrial images and arbitrary underwater images. UW-GAN [16] also used a GAN to generate depth, using supervision from a synthetic underwater dataset (no code released for comparison). These are supervised methods and none uses geometric cues between subsequent frames for self-supervision as we do. As we show, self-supervision significantly improves the results.

III. Scientific Background

A. *Reprojection Loss*

The reprojection loss is the key self-supervision loss. It uses two sequential frames $[I_{t-1}, I_t]$, where t is the time index, together with the estimated extrinsic rotation, translation, and $\widehat{D}_t$, the estimated depth of frame I_t. These are used to compute the coordinates $\widehat{p}_{t-1}$ in I_{t-1} that are the projection of the coordinates p_t in I_{t-1} [34]:

$$\widehat{p}_{t-1} \sim K\widehat{T}_{t\to t-1}\widehat{D}_t(p_t)K^{-1}p_t \quad . \tag{1}$$

Here $\widehat{T}_{t\to t-1}$ is the inverse transform calculated from the extrinsic parameters and K is the intrinsic camera matrix, known from calibration. Then each pixel in the reprojected image $\widehat{I}_t(p_t)$ is populated with values of $I_{t-1}(\widehat{p}_{t-1})$.

Based on color constancy the reprojection $\widehat{I}_t$ is supposed to be similar to the original frame I_t. Following [12], often the re-projection loss is a combination of two similarity measures, pixel level loss L_1 and single scale structural similarity (SSIM) [32] which compares the image's local structural information by using a sliding window to compare small regions of the two images:

$$L_{\text{reproj}} = \alpha L_1(I_t, \widehat{I}_t) + (1-\alpha) SSIM(I_t, \widehat{I}_t) \quad , \tag{2}$$

weighted by α, commonly set to $= 0.15$ [12].

B. *Underwater Photometry*

As described in [2], the image formation model of a scene pixel x in a participating medium such as underwater is composed of two additive components:

$$I(x) = J(x)t(x) + A(1 - t(x)) \quad , \quad t = e^{-\chi d} \quad . \tag{3}$$

The scene radiance J is attenuated by the medium. The medium transmission t is exponential in the scene depth d and χ, the medium's attenuation coefficient. Backscatter $A(1-t)$ is an additive component that stems from scattering along the line of sight, where A is the global light in the scene.

It is important to note that χ is wavelength dependant, i.e., each color channel attenuates differently with distance from the camera. In most water types the attenuation of red and near-infrared portions in water is much higher than the shorter visible wavelengths [26]. Hence, in underwater scenes, the red channel decreases faster with the distance. Based on this observation the ULAP prior was suggested [29]. It is calculated as the difference between the maximum value of B and G, the blue and green color channels, and the value of R, the red color channel

$$u = \max(B, G) - R \quad . \tag{4}$$

According to [29] the ULAP depth prior u is supposed to be linearly related to the scene depth.

IV. Underwater self-supervised monocular depth estimation framework

A. *Reprojection Loss Underwater*

Following (3) the medium affects the acquired underwater images as a function of object depth. Hence, camera movement underwater might lead to a significant difference between images captured subsequently, questioning the validity of the reprojection loss (2) in this case. One solution to this is to insert the photometric model (3) into the loss function (2). This would require the estimation of additional parameters χ and A and would add complexity. Before doing that, we conducted an experiment to examine the influence of the medium on the reprojection loss, as a function of inter-frame camera motion to check whether in nearby frames the influence of the medium on the loss can be neglected.

Fig. 3 summarizes this analysis in comparison to a similar analysis on the KITTI dataset. The reprojection loss between subsequent frames in our test set is calculated using the predicted depth and camera poses. We repeat the same calculation for an increasing gap between the frames. We see that in nearby subsequent frames, the underwater loss is only slightly larger than the outdoor error in KITTI. As expected, the error increases as the distance between subsequent frames increases. This points to the importance of high frame-rate imaging when acquiring training sets underwater, and confirms our assumption that in our dataset the original loss can be used.

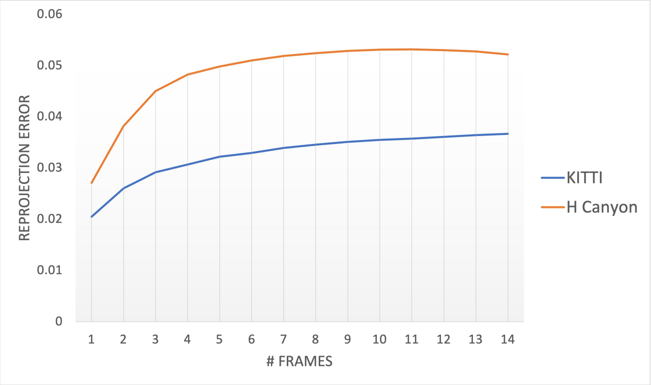


Fig. 3: Reprojection loss error as a function of frame gap in KITTI and an FLSea subset (Horse Canyon). When the gap between frames increases, the error increases as well. This happens in both datasets but is more prominent underwater due to the effect of the medium. In nearby frames, the underwater loss is only slightly larger than the outdoor error in KITTI.

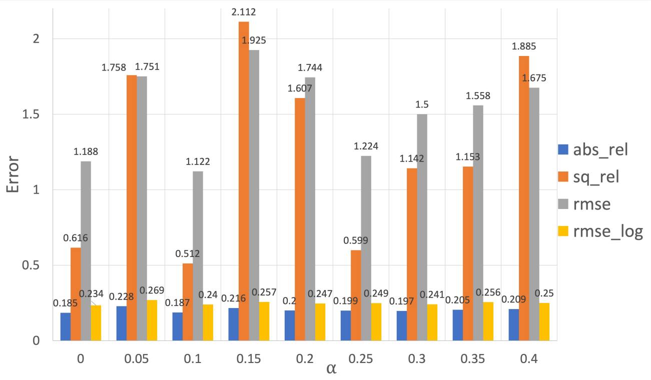


Fig. 4: Reprojection loss error as a function of α.

The loss (2) that is commonly used combines L_1 which is a pixel-wise comparison, with SSIM, which is a more general image quality measure with a weight of $\alpha = 0.15$, i.e., SSIM receives a much larger weight. SSIM should be more robust to illumination changes underwater and therefore we hypothesize that the ideal α value underwater should be lower. To test that, we conducted an experiment in which we trained the baseline method with a range of α values. The results are summarized in Fig. 4. We see that both $\alpha = 0$ and $\alpha = 0.1$ result in lower errors, with a small preference for $\alpha = 0.1$, which we choose to use in our experiments.

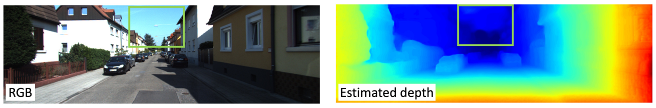


Fig. 5: DiffNet depth prediction results on KITTI. There is no ground truth for sky regions so the error there is not measured. We see that the depth of the sky in the image (green rectangle) is mistakenly predicted to be closer than the trees below it.

B. Inferring Range in Areas Without Objects

The re-projection loss (2) minimizes the misalignment of details in the image. This creates an issue when estimating image areas that have no objects (e.g., sky, water background), since in textureless areas any depth results in a low reprojection loss. In KITTI, there is no ground truth available for the sky as the measurements are LIDAR measurements that only reflect from nearby objects. However, when observing the results qualitatively, it is noticeable that some areas in the sky receive erroneous nearby ranges (see Fig. 5). When using depth inference to guide driving vehicles this is probably not an issue as the vehicles drive on the ground level and values in areas vertically above the car height are less relevant.

However, underwater, vehicles regularly move vertically in a 3D space and require accurate range estimation also in areas that are vertically above them. If an object-less background area is mistakenly assigned a nearby value, it might affect the vehicle motion planning and the vehicle will attempt to bypass it without any reason. Moreover, this issue becomes more severe in underwater scenes, as ambient illumination is non-uniform and the background appearance can change between frames, increasing the reprojection error (e.g., the background noise in Fig. 6b). Thus, this issue becomes critical underwater and we attempt to overcome it.

We want the loss to focus on the visible objects, such that it does not try to explain illumination changes in the object-less areas. For that, we propose the Local Variation Weight (LVW) mask σ_k. We calculate a local variation map over the image (5), which extracts interest areas in the image

$$\sigma_k = \mathbb{E}(x_k^2) - \mathbb{E}(x_k)^2 \quad , \tag{5}$$

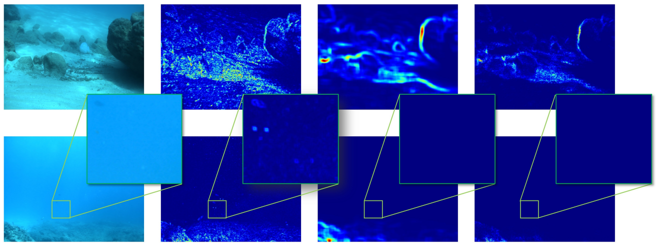
where $\mathbb{E}$ is the expectation operator and x_k is an image region of size $k = 25$. This map is normalized between 0 and 1:

$$\hat{\sigma}_k = \frac{\sigma_k - \min(\sigma_k)}{\max(\sigma_k) - \min(\sigma_k)} \quad . \tag{6}$$

and is used as weights on the original re-projection loss (2) to yield the final re-projection loss $\widehat{L}_{\text{reproj}}$.

$$\widehat{L}_{\text{reproj}} = L_{\text{reproj}} \cdot \hat{\sigma}_k \quad . \tag{7}$$

A similar mask was used in [30] for image segmentation in noisy and textured environments. Fig. 6 demonstrates the effect of LVW on two scenes. The LVW mask reduces some of the effects of flickering, backscatter, and changing the appearance of the rocks due to the combination of different camera orientations and nonuniform illumination.



(a) RGB (b) Reprojection Loss (c) LVW (d) Final Loss

Fig. 6: The effect of local variation map on the reprojection loss. a) An underwater scene. b) The reprojection loss (2) calculated between consecutive frames. (c) The normalized LVW map (6). d) The final loss (7) after multiplication with the normalized LVW. In the original loss (b) reflections and non-uniform illumination introduce errors. The normalized LVW filters out this noise and leaves the real errors of the projection miss-alignment.

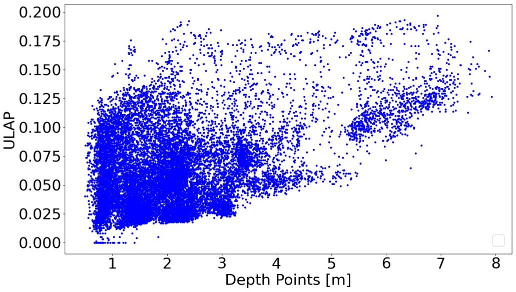


Fig. 7: Correlation of the depth prior ULAP values versus ground truth depth. Pearson Coefficient between ground truth to ULAP equals 0.46.

C. Underwater Light Attenuation Prior (ULAP)

As discussed in Sec. II-C, underwater, photometric cues can aid depth estimation. Here we add the ULAP prior as guidance for the estimation. First, we examine the validity of the prior. In Fig. 7 we show the correlation between both the ground truth depth with the ULAP (4) calculated on our test set images. The correlation is 0.46, which shows some relation but means ULAP by itself cannot be used for depth estimation. Using this insight, we encourage the correlation between the ULAP prior u and our depth estimation d by penalizing scores that are smaller than 1:

$$L_{\text{corr}} = 1 - \frac{\sum(d - \bar{d}) \cdot (u - \bar{u})}{\sqrt{\sum(d - \bar{d})^2 \cdot \sum(u - \bar{u})^2}} \quad , \tag{8}$$

where $\bar{d}$ is the mean depth over the image, and $\bar{u}$ is the mean of u. The weight for this loss was empirically set to 1e^{-5}.

D. Underwater Data Augmentation

Compared to above-water haze-free images, in which the sky is usually uniformly illuminated, the underwater medium introduces light scattering which is changed by distance from the camera, camera orientation, and the direction of the sun. This could greatly affect unsupervised depth estimation since we do not expect to see projection errors in fully aligned regions. To generalize the network to perform well under

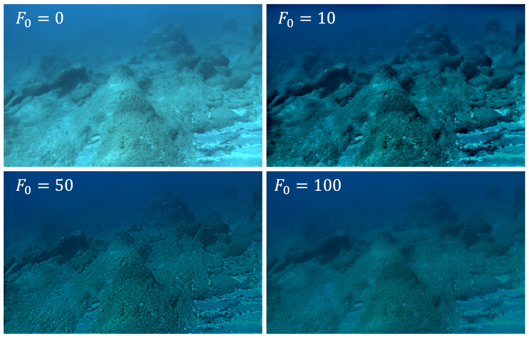


Fig. 8: An example underwater image after homomorphic filtering with different cutoff frequency D_0 values (9). D_0=0 is the original image, higher cutoff frequencies eliminate low frequencies from the image and emphasize high frequencies.

different illuminations, we use dedicated data augmentation in training, using homomorphic filtering.

Homomorphic filtering [1] is an image processing filter that is used for image enhancement, denoising [13] and non-uniform image illumination correction [19]. The homomorphic filter serves as a high-pass filter, reducing low-frequency variations that stem from illumination changes, with a controllable cutoff frequency. We use it to augment the input training images with a randomly parameterized homomorphic filter. Each input image goes through a homomorphic filter with a random uniformly distributed cutoff frequency F_0 with values that range between 0 to 250. Setting $F_0 = 0$ yields the original image. This results in images with more homogeneous illumination (see Fig. 8) and aids training.

The homomorphic filter H is a Butterworth high pass filter (9), initialized with a cutoff frequency F_0

$$H(z,w) = \left\{ 1 + \left[\frac{F_0}{F(z,w)} \right]^{2n} \right\}^{-1} \quad , \tag{9}$$

where $F(z,w)$ is the 2D euclidean distance from the point (z,w) to the center of the frequency space frame. We set n, the order of the filter, to be 2, which generates a moderate transition around the cutoff frequency. To apply the filter the RGB image is converted to YUV (Y– luminance; UV–chrominance) color space $(y,u,v) = RGB2YUV(I)$ and the high pass filter is applied in the Fourier space on the log of the y color channel of the image:

$$\widehat{y} = \exp \mathcal{F}^{-1}(HY) \quad , \quad Y = \mathcal{F}(\log y) \quad . \tag{10}$$

The filtered RGB image $\widehat{I}$ is reconstructed from $(\widehat{y}, u, v)$

$$\widehat{I} = YUV2RGB(\widehat{y}, u, v) \quad . \tag{11}$$

V. Experiment Details

A. Training and Testing

We use the FLSea dataset [27]. It contains 4 scenes: U Canyon, Horse Canyon, Tiny Canyon, and Flatiron, consisting of 2901, 2444, 1082, and 2801 frames respectively. All scenes were acquired in the same region in the Mediterranean Sea. Ground truth depth and camera intrinsics were generated using SFM (with the Agisoft software), and are known to contain some errors. We split each one of the scenes into train (2751, 2651, 932, and 2444 respectively), evaluation (50), and test (last 150 frames of each of the scenes) sets. Horse Canyon was used for training but excluded from the test set due to the apparent low quality of the ground truth. We trained the network using pre-trained weights on KITTI as a starting point, as this yielded better results than training from scratch.

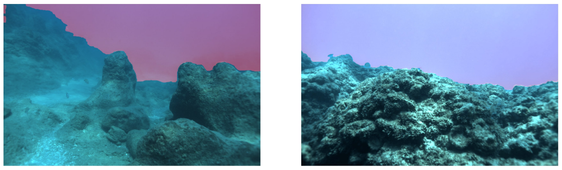


Fig. 9: Examples of background masking of two underwater images taken from Tiny Canyon and flatiron.

B. Background Error Estimation

In most datasets, including ours, there is no ground truth for background areas, as depth is measured only on objects. Thus, performance is not evaluated on background regions. Due to its importance in our case (Sec. IV-B), we specifically added a measure for the background error. The disparity in background areas is expected to be 0, hence, we suggest an error measurement that penalizes pixels in the background that are greater than zero. Our motivation in this error calculation is to give the lowest error to pixels with the lowest disparity S estimation or alternatively farthest depth estimation

$$BGerror = \frac{1}{m} \sum_{i=1}^{m} S_{x \in \mathrm{bg}} \quad , \tag{12}$$

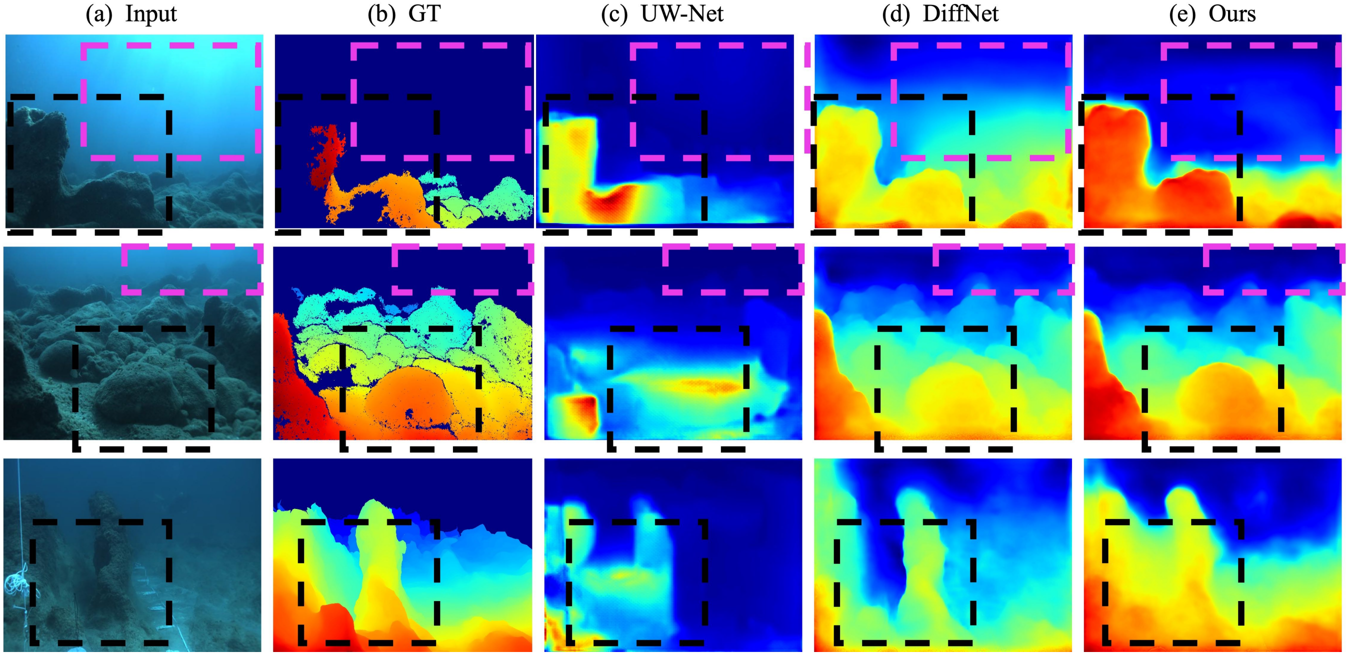
where m is the number of test images and S is the predicted disparity map from the test set. For extracting the open water background bg, we use the method described in [21], originally targeted for sky detection (see examples in Fig. 9).

VI. Results

Table I summarizes the results and the ablation study. The results are reported using the evaluation metrics described in [7]. Since we could achieve good background masking only on Tiny Canyon, we calculate the background error only on this scene. Our method significantly improves the baseline DiffNet in all measures except for $\delta < 1.25$ and $\delta < 1.25^2$. These measures indicate the number of pixels with low errors. This means that our method is less accurate in fine-depth estimation but more accurate in the global depth context, which is manifested in more accurate borders between objects and correct depth decisions of objects with regard to other objects in the scene. Our method is also significantly better in the background error estimation in more than 30%. We see that even the baseline results of the above water method are much better than the dedicated UWNET [14]. Note that [14] is trained on an in-air synthetic

TABLE I: An ablation analysis on the FLSea dataset. All methods perform better than the baseline. Augmentation contributes mostly to background error reduction. Measurement reveals significant improvement in background depth estimation.

L_{corr}	*LVW*	**Augmentation**	α	***AbsRel***	***SqRel***	***RMSE***	***RMSElog***	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	***BGerror***
UW-NET			-	0.527	1.765	1.725	1.961	0.337	0.565	0.699	3.9247
BASELINE			0.15	0.203	1.955	1.546	0.245	0.768	0.923	0.966	1.381
			0.1	0.186	1.828	1.295	0.222	0.793	0.935	0.97	1.396
✓			0.1	0.162	0.245	0.661	0.209	0.78	0.934	0.974	1.213
	✓		0.1	0.158	0.18	0.644	0.218	0.768	0.929	0.963	1.795
		✓	0.1	0.18	0.366	0.775	0.221	0.751	0.924	0.97	1.372
	✓	✓	0.1	0.165	0.212	0.7	0.227	0.774	0.92	0.958	1.714
✓		✓	0.1	0.176	0.399	0.859	0.213	0.771	0.928	0.974	1.098
✓	✓		0.1	0.156	0.146	0.589	0.21	0.775	0.927	0.972	0.79
✓	✓	✓	0.1	0.158	0.149	0.581	0.208	0.778	0.924	0.969	0.783


(a) Input

(b) GT

(c) UW-Net

(d) DiffNet

(e) Ours

Fig. 10: Example results on three underwater scenes from the FLSea dataset [27]. a) Input scene, b) Ground truth, c) UW-Net result, d) Result of Diffnet and e) our estimated depth map. The magenta rectangles mark the background area where our method significantly improves the results, and black rectangles mark foreground objects where our method improved estimation.

dataset together with random internet underwater images, therefore can not be re-trained with our dataset. The ablation study shows that using $\alpha = 0.1$, L_{corr} and LVS always improve results. Augmentation with the homomorphic filter improves some of the measures and especially the background error.

VII. Discussion

So far methods for monocular depth estimation underwater concentrated on leveraging photometric cues in single images, which is challenging to do in a self-supervised manner. We are the first to use self-supervision using subsequent frames, as successfully done above water. We show that using the standard above-water SOTA methods underwater results in decent results but has room for improvement as it is not designed to cope specifically with appearance changes caused by the medium. We analyze the performance of the standard reprojection loss and show that it can be used also underwater given the training set was acquired at a high frame rate. We point to a problem that exists also above water in errors in estimating background areas that do not have ground truth. This was so far ignored above water, but in the three-dimensional underwater realm it cannot be ignored and we suggest a weighed loss to mitigate this issue.

Since photometric priors on the single underwater images contain important information we combine one of them in the loss. In the future, we plan to investigate how to further combine the single image information with the self-supervision obtained from subsequent frames. Lastly, we plan to incorporate this framework into a complete image restoration pipeline. Overall, our method significantly improves the SOTA in underwater monocular depth estimation and can substantially aid vision-based navigation and decision-making in underwater autonomous vehicles.

References

[1] Holger G Adelmann. Butterworth equations for homomorphic filtering of images. *Computers in Biology and Medicine*, 28(2):169–181, 1998.

[2] Yael Bekerman, Shai Avidan, and Tali Treibitz. Unveiling optical properties in underwater images. In *IEEE Int. Conf. on Computational Photography (ICCP)*, 2020.

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, pages 4009–4018, 2021.

[4] Nicholas Carlevaris-Bianco, Anush Mohan, and Ryan M Eustice. Initial results in underwater single image dehazing. In *Oceans Mts/IEEE Seattle*, 2010.

[5] Paul Drews, Erickson Nascimento, Filipe Moraes, Silvia Botelho, and Mario Campos. Transmission estimation in underwater single images. In *ICCV workshops*, pages 825–830, 2013.

[6] Paulo LJ Drews, Erickson R Nascimento, Silvia SC Botelho, and Mario Fernando Montenegro Campos. Underwater depth estimation and image restoration based on single images. *IEEE computer graphics and applications*, 36(2):24–35, 2016.

[7] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[8] Simon Emberton, Lars Chittka, and Andrea Cavallaro. Underwater image and video dehazing with pure haze region segmentation. *Computer Vision and Image Understanding*, 168:145–156, 2018.

[9] Adrian Galdran, David Pardo, Artzai Picón, and Aitor Alvarez-Gila. Automatic red-channel underwater image restoration. *J. of Visual Communication and Image Representation*, 26:132–145, 2015.

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012.

[11] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017.

[12] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019.

[13] Pelin Gorgel, Ahmet Sertbas, and Osman N Ucan. A wavelet-based mammographic image denoising and enhancement with homomorphic filtering. *J. of medical systems*, 34(6):993–1002, 2010.

[14] Honey Gupta and Kaushik Mitra. Unsupervised single image underwater depth estimation. In *IEEE International Conference on Image Processing (ICIP)*, pages 624–628, 2019.

[15] Yevgeni Gutnik, Aviad Avni, Tali Treibitz, and Morel Groper. On the adaptation of an auv into a dedicated platform for close range imaging survey missions. *J. of Marine Science and Engineering*, 10(7):974, 2022.

[16] Praful Hambarde, Subrahmanyam Murala, and Abhinav Dhall. Uw-gan: Single-image depth estimation and image enhancement for underwater images. *IEEE Trans. on Instrumentation and Measurement*, 70:1–12, 2021.

[17] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *PAMI*, 33(12):2341–2353, 2010.

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

[19] Jeffrey W Kaeli, Hanumant Singh, Chris Murphy, and Clay Kunz. Improving color correction for underwater image surveys. In *IEEE MTS/OCEANS*, 2011.

[20] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018.

[21] Orly Liba, Longqi Cai, Yun-Ta Tsai, Elad Eban, Yair Movshovitz-Attias, Yael Pritch, Huizhong Chen, and Jonathan T Barron. Sky optimization: Semantically aware image processing of skies in low-light photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 526–527, 2020.

[22] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, pages 5162–5170, 2015.

[23] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *IEEE international conference on robotics and automation (ICRA)*, pages 4796–4803, 2018.

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

[25] Yan-Tsung Peng, Xiangyun Zhao, and Pamela C Cosman. Single underwater image enhancement using depth estimation based on blurriness. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 4952–4956, 2015.

[26] Robin M Pope and Edward S Fry. Absorption spectrum (380–700 nm) of pure water. ii. integrating cavity measurements. *Applied optics*, 36(33):8710–8723, 1997.

[27] Yelena Randall and Tali Treibitz. Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets. *arXiv preprint arXiv:2302.12772*, 2023.

[28] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021.

[29] Wei Song, Yan Wang, Dongmei Huang, and Dian Tjondronegoro. A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration. In *Pacific Rim Conference on Multimedia*, pages 678–688. Springer, 2018.

[30] Jie Wang, Lili Ju, and Xiaoqiang Wang. Image segmentation using local variation and edge-weighted centroidal voronoi tessellations. *IEEE Trans. on Image Processing*, 20(11):3242–3256, 2011.

[31] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *PAMI*, 43(10):3349–3364, 2020.

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004.

[33] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. *arXiv preprint arXiv:2110.09482*, 2021.

[34] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 1851–1858, 2017.