Physics-informed Knowledge Transfer for Underwater Monocular Depth Estimation

Anonymous CVPR submission

Paper ID 11261

Abstract

*Compared to the in-air case, underwater depth estimation has its own challenges. For instance, acquiring high-quality training datasets with groundtruth poses difficulties due to sensor limitations in aquatic environments. Additionally, the physics characteristics of underwater imaging diverge significantly from the in-air case, the methods developed for in-air depth estimation underperform when applied underwater, due to the domain gap. To address these challenges, our paper introduces a novel transfer-learning-based method - Physics-informed Underwater Depth Estimation (**PUDE**). The key idea is to transfer the knowledge of a pre-trained in-air depth estimation model to underwater settings utilizing a limited dataset without groundtruth measurement, guided by an underwater physics image formation model. We propose novel bound losses based on the underwater physics image formation model to rectify the depth estimations to align with actual underwater physical properties. Finally, in the zero-shot evaluations across multiple datasets, we compare our Physics-informed Underwater Depth Estimation (**PUDE**) model with other existing in-air and underwater approaches. The results reveal that the **PUDE** model excels in both quantitative and qualitative comparisons.*

1. Introduction

Depth estimation is a pivotal task in computer vision, with substantial implications for robotics applications. Traditional techniques typically employ stereo cameras to estimate depth through disparity calculations. However, these methods are hampered by complex camera system settings and limited by the challenging feature matching in environments with low texture or visibility [24]. In contrast, monocular depth estimation provides an alternative solution, necessitating only a single camera and a less complicated setup.

Although monocular depth estimation faces inherent limitations due to the absence of stereoscopic information, many well-performing in-air learning-based depth estimation models have been developed in recent years [5, 6, 13, 22, 28–30]. Benefiting from the massive and various training datasets, some of in-air monocular depth estimation models, such as MiDaS [29] and DPT [30], demonstrate powerful generalization performance. However, the pre-trained in-air models are not fully applicable for underwater cases due to the domain gap and the differences in the image formation [36]. For example, we applied the in-air trained models, e.g., DPT in [30], to underwater images and observed that DPT has the issues of the overestimation for blurry regions, unclear background region detection and loss of depth details.

Additionally, it is challenging to train new models from scratch for underwater settings. Underwater scenarios impose difficulties when capturing a large amount of high-quality images and the groundtruth depth measurements, due to sensor limitations, such as the incapability of LiDAR and the performance drop of stereo matching in water. In contrast, sonar scanning is not affected by water quality and blurriness but suffers from a slow scanning rate and low resolution, lacking detail in scene representation [33].

In response to the distinct challenges of underwater depth estimation mentioned above, this paper proposes a novel transfer-learning-based approach involving an underwater image formation model. Our method transfers knowledge from a pre-trained in-air monocular depth estimation model to underwater settings, guided by the underwater imaging formation model introduced in [10]. Using the newly proposed bound losses, our training method enforces the model to produce estimates that align with the extracted physical properties. Notably, this adaptation requires a low amount of underwater training data without the need for groundtruth depth measurements. Our method benefits from the in-air model's structural knowledge and integrates underwater physics knowledge, thereby enhancing the depth estimation performance underwater. Here are our contributions:

- We first present some experimental observations and limitations when applying an in-air trained model, i.e., DPT in [30], to underwater images. Inspired by these observations, we propose a novel training approach to adapt in-air pre-trained depth estimation models for underwater envi-

ronments. There are two key ingredients in our method: 1) the knowledge transfer using an in-air trained model with good generalisation properties, i.e., DPT in [30], to tackle the challenge of data shortage; 2) the integration of the underwater image formation model proposed in [10] to enforce the depth estimates follow the underwater physics characteristics.

- We propose innovative bound losses to steer the model towards producing depth estimations that adhere to the underwater imaging formation model.
- We introduce the Physics-informed Underwater Depth Estimation (**PUDE**) model, an end-to-end model for underwater monocular depth estimation. In zero-shot evaluations across various datasets, our method demonstrates enhanced performance, both quantitatively and qualitatively, compared to other in-air and underwater methods.

2. Related work

2.1. In-air monocular depth estimation

Compared to LiDAR, infrared, and stereo vision, monocular depth estimation stands out for its cost-effectiveness and simplicity of configuration, making it appealing for depth perception in autonomous systems. Recent methods for in-air monocular depth estimation have exhibited commendable performance, particularly with supervised approaches trained on datasets containing measured groundtruth. They usually require sensors like LiDAR, RGBD cameras, and stereo matching [12, 13, 19, 23]. However, these methods are limited by the lack of variety in the datasets, which restricts their generalization across different real-world scenarios.

In this context, self-supervised learning using continuous image sequences [15] or calibrated stereo image pairs [14] presents an attractive avenue for exploration but performs less compared to supervised learning [24], and its application to dynamic scenes is challenging. Recognizing this limitation, some studies expand the breadth of training data by harnessing web resources to generate depth groundtruth, thereby enlarging the training dataset and encompassing diverse scenes [9, 22, 34]. Building further on that, MiDaS [29] extract depth data from 3D movies to enhance dataset diversity and improve generalization. DPT [30] builds upon the MiDaS training dataset, but with supplementary, and implements an innovative encoder-decoder architecture that uses a vision transformer as its backbone, enabling fine-grained and globally coherent estimations and showing strong zero-shot cross-dataset generalization capabilities.

2.2. Underwater monocular depth estimation

Depth measurement in water introduces new challenges due to the unique characteristics of the underwater environment. The limited availability of sensors, such as LiDAR and infrared, makes the direct acquisition of the depth measurement highly challenging. Meanwhile, the image blurriness highly influences the correspondence matching between images [35], thereby the stereo camera can only provide good depth measurement for the close regions. The existing underwater depth datasets that rely on stereo matching [1, 4], manifest this issue.

Therefore, monocular depth estimation is a potential solution to tackle these issues. Physics-based methods play a pivotal role in addressing the underwater depth estimation challenge. The physics-based underwater image formation model proposed in [10] has been demonstrated to be effective for solving problems in computer vision for underwater scenarios, for example, feature extraction [35], neural radiance fields (NeRFs) [20] and depth estimation [11]. The traditional methods utilize the physics-based image formation model to estimate the depth directly from the underwater images [3, 4, 8, 25–27]. As the foundation of these methods, Dark Channel Prior (DCP) [18], a statistical method initially developed for in-air haze removal, uses the pixel intensity to estimate the medium transmission. A variant for the underwater environment - UDCP [11], claims that the DCP assumption remains applicable to the green and blue channels of underwater images. However, underwater depth estimation, which relies solely on pixel values and the imaging formation model, is particularly vulnerable to variable lighting conditions, such as shadows and sunlight flickers.

Despite the challenges in depth groundtruth acquisition underwater, the learning-based methods show promise. Some works have attempted unsupervised learning, such as UW-Net [16] and UW-GAN [17]. They use the GAN method through color restoration to estimate depth. The recent works [2, 31] adopt self-supervised learning, utilizing the correspondences between subsequent frames. Notably, they incorporate the physics model into the training of the underwater depth estimation models and have shown to be advantageous.

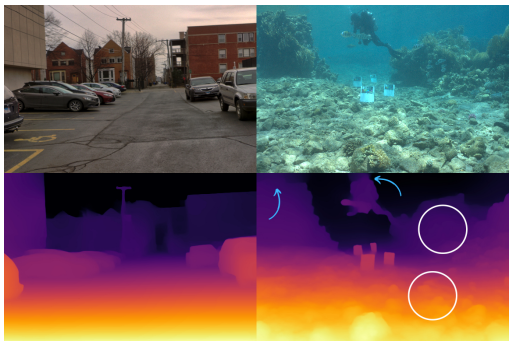
2.3. Underwater imaging formation model

Light transmission in water is subject to absorption, and the presence of minuscule particles within the water precipitates the backscattering, thereby engendering a degree of blur within underwater imagery. A widely used model of underwater image formation is proposed in [10] to describe this physical phenomenon. Consider an underwater image represented by $I_c \in \mathbb{R}^{HW}$, where the subscript c denotes the index of the three color channels {R, G, B}, and H and W represent the image's dimensions in height and width, respectively. Given an image I_c, for each pixel i, $i = 1 \cdots HW$, it holds that $I_{c,i} \in [0, 1]$, and the underwater imaging formation model is:

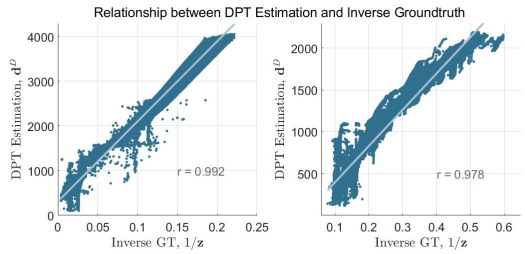
$$I_{c,i} = J_{c,i} \cdot t_{c,i} + B_c^{\infty}(1 - t_{c,i}), \tag{1}$$

and

$$t_{c,i} = e^{-\beta_c \mathbf{z}_i}. \tag{2}$$



(a) DPT depth estimation



(b) Linearity fitting of in-air image (left) and underwater image (right)

Figure 1. **Depth estimation from DPT:** DPT estimation of the underwater image has a curve tendency, with overestimation for distant regions. Meanwhile, the estimated depth has the issue of unclear background detection (blue) and loss of details (white).

In this model, $J_{c,i}$ is the original in-air color value at pixel i. The original in-air image is denoted by $J_c = [J_{c,1}, ..., J_{c,HW}] \in \mathbb{R}^{HW}$. The medium transmission rate denoted by $t_{c,i}$ is associated with the beam attenuation coefficient β_c and the depth $\mathbf{z}_i$. Meanwhile, B_c^∞ is the coefficient of ambient light causing the backscattering from the particles in water. The captured underwater image $I_{c,i}$ is the aggregate of the attenuation of the original image and the additional backscattering from the water environment. The values of β_c and B_c^∞ vary depending on different water environments and can be estimated using the depth information [1].

3. DPT on underwater images: observations

Existing learning-based monocular depth estimation methods, such as DPT [30] and MiDaS [29], generate the inverse relative depth estimation of a single image scene. We denote the inverse depth estimation as $\mathbf{d} \in \mathbb{R}^{HW}$, with $\mathbf{d}_i$ to be the estimate for pixel i, $i = 1, \cdots, HW$. The metric depth of the input image is denoted by $\mathbf{z} \in \mathbb{R}^{HW}$, whose reciprocal ideally has a linear relationship with the inverse relative depth. The relationship between inverse relative depth $\mathbf{d}_i$ and the metric depth $\mathbf{z}_i$ at pixel i is:

$$\mathbf{z}_i = \frac{1}{s\mathbf{d}_i + t}, \quad \forall i \in \{1, \cdots, HW\}. \tag{3}$$

Here, s and t are two unknown scale and shift coefficients. In practice, they can be off-line estimated for specific scenes using sequence images. Although the mapping relationship between the inverse relative depth $\mathbf{d}_i$ and the inverse of metric depth $\mathbf{z}_i$ is linear, when we apply DPT, an in-air model with good generalization, directly to underwater images in the dataset [1, 4], it is observed that in underwater imagery, the estimated inverse depth, denoted by $\mathbf{d}_i^D$, and $\frac{1}{z_i}$ deviates from a linear relationship. Fig. 1 is an example for demonstrating these experimental observations. It compares the depth estimation of DPT on an outdoor in-air image from DIODE [32] and an underwater image from Sea-thru [1]. We can observe that DPT estimation on the in-air image maintains a good linear relationship across all regions (the correlation coefficient $r = 0.992$). In contrast, the estimated $\mathbf{d}_i^D$ on the underwater image shows a curve tendency with a weaker correlation ($r = 0.978$) as the inverse of the metric depth decreases, especially when $\mathbf{d}_i^D$ is small. This observation suggests the DPT tends to overestimate the depth of the distant regions. Furthermore, the background edges on the underwater depth image of DPT, such as the blue arrow pointed regions around corals and the diver, are less clear, exhibiting fogging compared with the estimation on the in-air image. Meanwhile, the estimated depth details in underwater imagery are significantly lacking compared to those derived from in-air imagery, circled by white. Based on these experimental observations, we summarise the primary issues of the depth estimation from DPT in water:

- O1 DPT exhibits a propensity for overestimating depth in regions that are relatively far away from the camera. This overestimation is particularly evident in regions with diminished visual clarity and increased blurriness. DPT loses the estimation detail in visually darker regions at a distance, such as recesses within coral formations or gullies between rocks and corals.
- O2 Secondly, DPT encounters challenges in accurately discerning background regions in underwater images. This issue is especially pronounced for the region far from the camera. Due to the attenuation of the scenes, it is challenging for DPT to define the background area clearly.

To solve these issues, we propose three bound losses (see in Sec. 4.3) and involve them in our training framework.

4. Method

4.1. Training framework overview

Inspired by the observations in Sec. 3, we propose a physics-informed knowledge transfer method for the underwater monocular depth estimation, named the Physics-informed Underwater Depth Estimation (**PUDE**). The main idea of this physics-informed method is to leverage the physical information in underwater imagery to guide knowledge transfer within the neural network, thereby enabling its correct transi-

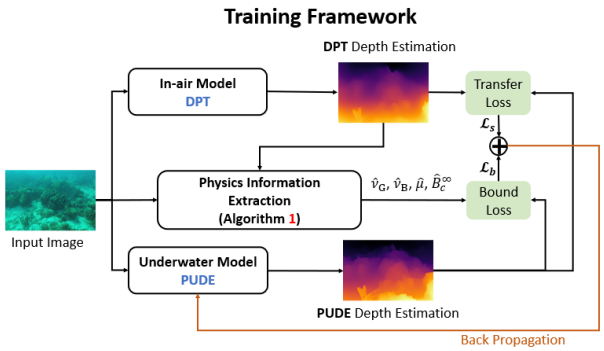


Figure 2. **Overall training framework for PUE.** The overall training framework comprises two components: the direct knowledge transfer from the DPT model and the physic-informed knowledge transfer based on the DPT estimation. Physical information is extracted from DPT's output to establish the bound losses.

tion from in-air to underwater environments. Specifically, we adopt the pre-trained DPT-Hybrid model [30], a state-of-the-art (SOTA) model known for its good generalization, as the transferred model. Benefiting from the massive in-air training dataset, the DPT model has powerful spatial structural knowledge. Despite previously noted issues when applied to underwater images, it can still offer a relatively reasonable depth estimation, especially for the regions that are close to the camera. This estimation serves as our prior depth knowledge for extracting the physical information from the underwater images.

Our training framework consists of two branches shown in Fig. 2. The first branch is the direct knowledge transfer from DPT to PUDE, and we employ a similarity loss to guarantee the similarity between DPT and PUDE in the transfer. We will introduce this branch in Sec. 4.2. The second branch is the physics-informed knowledge transfer. Inspired by the parameter estimation method in Sea-thru [1], we estimate the parameters in Eq. (1) and Eq. (3) for each input underwater image, utilizing DPT depth estimation. We restrict our usage to the green and blue channels, in line with the conclusions from UDCP [11]. Based on that, we propose three novel bound losses, utilizing the physics information contained in the underwater image formation model in Eq. (1) to enforce the model to produce plausible depth estimations that conform to the physics characteristics of underwater imaging. In Sec. 4.3, we will introduce the second branch and corresponding designed bound losses.

4.2. Direct knowledge transfer

In the direct knowledge transfer loop, the goal is to guarantee the similarity between the original DPT and PUDE during the transition to ensure the structural knowledge can be kept in PUDE. We use a similarity loss in the training process to achieve this goal.

In the transfer, PUDE is initialized with DPT. To set the tolerance for the difference between PUDE and the original model DPT during the training, we trim the pixels with the largest estimation differences. We denote the pixel set after trimming as $M_{trim} \subseteq \{1, ..., HW\}$, and T representing the cardinality of M_{trim}. For each pixel $i \in M_{trim}$, we denote the DPT and PUDE estimation as $\mathbf{d}_i^D$ and $\mathbf{d}_i^P$, respectively. The similarity loss L_s is defined as:

$$L_s = \frac{1}{T} \sum_{i \in M_{trim}} |\frac{\mathbf{d}_i^P - \mathbf{d}_i^D}{\mathbf{d}_i^D}|. \tag{4}$$

The trimming sets the tolerance for the changing between PUDE and the original model DPT during the training. In our implementation, we set the trimming rate to be $0.3HW$.

4.3. Physics-informed knowledge transfer

Parameters estiamtion

In the physics-informed knowledge transfer loop, we first extract the parameters in Eq. (1) and Eq. (3) using the preliminary depth estimations from DPT. As shown in Fig. 1, DPT's depth estimations on the underwater images tend to generate a farther estimation in distant and blurry regions but offer more trustable estimation for the closer regions, demonstrating a good linear relationship described in Eq. (3). Therefore, we extract the physics information harnessed from DPT's depth estimations for close regions to correct the biased estimation of the model for the more distant regions. The overall extraction steps are shown in Algorithm 1.

Algorithm 1 Parameters Extraction

Require: $\mathbf{d}^D \in \mathbb{R}^{HW}, I_c \in \mathbb{R}^{HW}, N \in \mathbb{Z}_{++}, c \in \{G, B\}$
Ensure: Pure background exsits in the training images.
1: Find farthest 15% points using $\mathbf{d}^D$, denoted as P^f.
2: Select the darkest 20% points using I_c from P^f, denote as P_c^d. Compute $\hat{B}_c^\infty$=median$(I_{c,i})$ for $i \in P_c^d$.
3: Select N darkest points from the closest $10\% - 40\%$ pixels on each color channel to form the set M_c.
4: Solve least-squares problem per Eq. (6).
5: **return** $\hat{\nu}_G, \hat{\nu}_B, \hat{\mu}$ and $\hat{B}_c^\infty$.

In our implementation, the number of N is set to be 400, and the values of $\hat{\nu}_G$, $\hat{\nu}_B$ and $\hat{\mu}$ is estimated based on the DPT estimation and the input image. In the extraction, we assume that the training underwater images contain certain background areas. We estimate the background light B_c^∞, by utilizing the median values of the darkest 20% pixel values from the furthest 15% points. The estimated background light is noted as $\hat{B}_c^\infty$. The rationale for extracting the darkest points in UDCP is analogous to identifying points where $J_{c,i}$ is assumed to be close to zero and where the color appearance on these pixels is solely due to the backscattering term.

Therefore, we evenly select the N darkest points, on which $J_{c,i} \approx 0$, from the closest 10% to 40% regions and employ these points to estimate the water parameters by solving the least-square problem in Eq. (5). We denote the pixel value on the pixel location i of the color channel c on the input underwater image as $I_{c,i}$, $i = 1 \dots HW$. According to the conclusion in UDCP [11], we only use the color channels blue and green in this process. We denote the sets of selected N points of each color channel as M_c, $c \in \{G, B\}$. Substituting (3) into (1), we define the error ϵ_c for the color channels as:

$$\epsilon_c := \sum_{i \in M_c} ||I_{c,i} - \hat{B}_c^{\infty}(1 - e^{-\frac{\beta_c / s}{\mathbf{d}_i^D + t/s}})||_2^2. \quad (5)$$

To simplify the least-square problem, we set two temporary parameters $\nu_c = \beta_c / s$ and $\mu = t/s$. Then, we solve the following least-square problem to obtain an estimate of the parameters:

$$(\hat{\nu}_G, \hat{\nu}_B, \hat{\mu}) = \underset{\nu_G, \nu_B, \mu}{\operatorname{argmin}} \sum_{c \in \{G,B\}} \epsilon_c, \quad (6)$$

An implementation example of Algorithm 1 is shown in Fig. 3.

Bound losses on medium transmission rate

After estimating the parameters of the underwater imaging formation model in Eq. (1), we are ready to present three bound losses to constrain the PUDE model from predicting depths that contravene the underwater characteristics.

Lower bound 1: The first bound is a lower bound on the estimated medium transmission rate $t_{c,i}^P$, $c \in \{G, B\}$. This bound aids in refining depth estimates in distant regions, preventing from overestimating the depth for the regions that are far from the camera, as described in O1 in Sec. 3. By substituting the obtained parameter $\hat{\nu}_G$, $\hat{\nu}_B$ and $\hat{\mu}$ in Algorithm 1, and using the PUDE depth estimation, we get the estimated $\hat{t}_{c,i}^P$:

$$\hat{t}_{c,i}^P = e^{-\frac{\hat{\nu}_c}{\mathbf{d}_i^D + \hat{\mu}}}. \quad (7)$$

As the original color value of each point $J_{c,i}$ is equal to or greater than zero. Based on Eq. (1), we have:

$$I_{c,i} \geq (1 - t_{c,i}) B_c^{\infty}, \quad (8)$$

substituting the estimated $\hat{B}_c^{\infty}$, the first lower bound is:

$$\hat{t}_{c,i}^P \geq 1 - \frac{I_{c,i}}{\hat{B}_c^{\infty}}. \quad (9)$$

This lower bound on $\hat{t}_{c,i}^P$ is to pull back the depth estimations for the distant regions. When the depth of pixels in these regions is overestimated, it results in an underestimation of the medium transmission rate in Eq. (7). Consequently, the captured pixel value $I_{c,i}$ should not be lower than the estimated backscattering, as shown in Eq. (8).

Upper bound: The second bound is an upper bound on the estimated medium transmission rate $t_{c,i}^P$, $c \in \{G, B\}$ to address the challenge of detecting background edges in blurry underwater images, described in O2.

Considering that the maximum pixel intensity should not surpass that of pure white, namely $J_{c,i} \leq 1$, and referring to Eq. (1), we obtain:

$$I_{c,i} \leq t_{c,i} + (1 - t_{c,i}) B_c^{\infty}. \quad (10)$$

In the regions corresponding to the pure background area, the distance can be seen as infinite. The depth estimation of DPT for these regions is always smaller than infinity. Therefore, the estimated medium transmission for these background regions should be greater than the true value. We denote the estimated medium transmission rate on each pixel location by DPT as $\hat{t}_{c,i}^D$, which is defined as:

$$\hat{t}_{c,i}^D = e^{-\frac{\hat{\nu}_c}{\mathbf{d}_i^D + \hat{\mu}}}. \quad (11)$$

In the distant regions, where the original pixel values $J_{c,i}$ are attenuated, the captured pixel values $I_{c,i}$ are predominantly dictated by backscattering. To select pixel locations likely to be in the background regions, we define the set M_b:

$$M_b = \{i \mid \frac{\hat{B}_c^{\infty}(1 - \hat{t}_{c,i}^D)}{I_{c,i}} \geq \gamma\}. \quad (12)$$

The parameter γ is utilized to identify regions where pixel values are predominantly determined by backscattering, which are likely to be in the background area. In our experiments, γ is set to be 0.6. As mentioned, for the background regions, $\hat{t}_{c,i}^D$ is greater than the true value. Then based on the Eq. (10), we have,

$$I_{c,i} < \hat{t}_{c,i}^D + (1 - t_{c,i}) B_c^{\infty}, \quad i \in M_b, \quad (13)$$

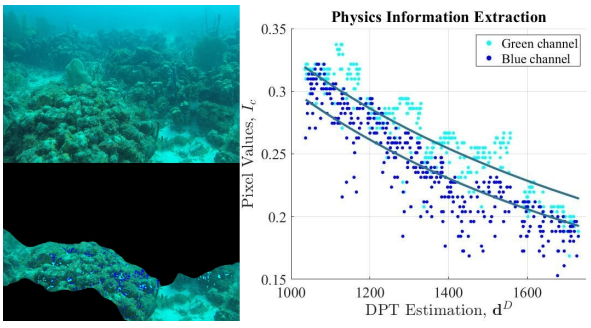


Figure 3. **Physics information extraction.** The left image shows the area and points selected. The right figure shows the fitting of the estimation.

	$\delta < 1.05 \uparrow$	$\delta < 1.05^2 \uparrow$	$\delta < 1.05^3 \uparrow$	AbsRel $\downarrow$	Sq. Rel$\downarrow$	RMSE $\downarrow$	RMSE log$\downarrow$	SILog$\downarrow$
MegaDepth [22]	0.284	0.533	0.721	0.115	0.024	0.205	0.143	0.021
MiDaS [29]	0.341	0.596	0.758	0.131	0.052	0.244	0.166	0.036
DPT [30]	0.509	0.813	0.919	0.065	0.009	0.136	0.089	0.009
UWNET [17]	0.159	0.308	0.453	0.187	0.057	0.299	0.228	0.055
UDepth [37]	0.333	0.598	0.785	0.098	0.017	0.172	0.124	0.016
Ours (PUDE)	**0.570**	**0.841**	**0.934**	**0.059**	**0.008**	**0.123**	**0.081**	**0.007**

Table 1. **Quantitative Evaluation: Sea-thru D3.** D3 contains 68 images, observing an underwater coral at a relatively close distance.

substituting the estimated parameters, we have the bound:

$$\hat{t}^P_{c,i} < \frac{-I_{c,i} + \hat{t}^D_{c,i} + \hat{B}^\infty_c}{\hat{B}^\infty_c}, \quad i \in M_b. \tag{14}$$

This upper bound on $t^P_{c,i}$ in Eq. (14) is to adjust the depth estimation on the background regions. For the pixels on the background regions, the original DPT produce lower depth estimation. Even if the original pixel $J_{c,i}$ is equal to 1 (perfectly white), the captured pixel value $I_{c,i}$ from the background regions should not be greater than the sum of the estimated medium transmission rate from DPT and the backscattering term, as shown in Eq. (13).

Lower bound 2: The third bound loss is again based on Eq. (10). It is to set a lower bound on $t^P_{c,i}$ for the points that are pushed too far due to loose selections in M_b in the upper bound in Eq. (14). In our experiments, we observed that the upper bound sometimes results in an overestimation of depth in the regions that are close to the background regions. This bound ensures that estimations for the points are not excessively reduced by the unilateral constraint of the previously introduced upper bound in Eq. (14). The second lower bound on $t^P_{c,i}$ is as follows:

$$\hat{t}^P_{c,i} \geq \frac{I_{c,i} - \hat{B}^\infty_c}{1 - \hat{B}^\infty_c}. \tag{15}$$

We set the bound losses to enforce the estimated transmission rate $\hat{t}^P_{c,i}$ to satisfy the listed three bounds and optimize the estimated depth of PUDE, d^P_i, in training. To simplify the mathematical expressions in the subsequent losses, we denote the ReLU function as $f : \mathbb{R} \to \mathbb{R}^+$, where $f(x) = \max(0, x)$. According to the bounds in Eq. (9), Eq. (14) and Eq. (15), the three bound losses are:

$$\mathcal{L}_{bl1} = \sum_{c\in\{G,B\}} \sum_{i=1}^{HW} f\Big(- I_{c,i} + (1 - \hat{t}^P_{c,i})\hat{B}^\infty_c \Big), \tag{16}$$

$$\mathcal{L}_{bu} = \sum_{c\in\{G,B\}} \sum_{i\in M_b} f\Big(-(1-\hat{t}^P_{c,i})\hat{B}^\infty_c - \hat{t}^D_{c,i} + I_{c,i} \Big), \tag{17}$$

$$\mathcal{L}_{bl2} = \sum_{c\in\{G,B\}} \sum_{i=1}^{HW} f\Big(-(1-\hat{t}^P_{c,i})\hat{B}^\infty_c - \hat{t}^P_{c,i} + I_{c,i} \Big). \tag{18}$$

The lower bound loss functions L_{bl1} and L_{bl2} are applied to all the pixels in the images, while the upper bound loss L_{bu} is only applied to the pixels in the set M_b. The number of elements in M_b denoted as N_b. We use the parameters α_0, α_1 and α_2 to adjust the weights between losses, which are set to 1, 1, and 2 in our experiment. The final bound loss is then given as,

$$\mathcal{L}_b = \alpha_0 \frac{\mathcal{L}_{bl1}}{HW} + \alpha_1 \frac{\mathcal{L}_{bl2}}{HW} + \alpha_2 \frac{\mathcal{L}_{bu}}{N_b}. \tag{19}$$

4.4. Final loss in the PUDE training framework

We use the parameters β_0 and β_1 to adjust the weights between the similarity loss and the bound loss. β_1 should be in a larger weight to make the physics-informed knowledge transfer dominate the training process. The total loss $\mathcal{L}_{total}$ in the training is:

$$\mathcal{L}_{total} = \beta_0 \mathcal{L}_s + \beta_1 \mathcal{L}_b. \tag{20}$$

In our implementation, β_0 and β_1 are set to 5 and 100, respectively.

	$\delta < 1.05 \uparrow$	$\delta < 1.05^2 \uparrow$	$\delta < 1.05^3 \uparrow$	AbsRel $\downarrow$	Sq. Rel$\downarrow$	RMSE $\downarrow$	RMSE log$\downarrow$	SILog$\downarrow$
MegaDepth [22]	0.369	0.613	0.761	0.102	0.021	0.865	0.141	0.021
MiDaS [29]	0.352	0.593	0.749	0.113	0.030	0.832	0.147	0.027
DPT [30]	0.429	0.684	0.818	0.096	0.027	1.024	0.148	0.024
UWNET [17]	0.205	0.417	0.627	1.458	>5	0.904	0.197	0.045
UDepth [37]	0.410	0.677	0.824	0.091	0.019	0.828	0.131	0.019
Ours (PUDE)	**0.502**	**0.761**	**0.881**	**0.075**	**0.014**	**0.737**	**0.117**	**0.015**

Table 2. **Quantitative Evaluation: Sea-thru D5.** The D5 dataset contains 43 image samples with a wide depth range.

	$\delta < 1.05 \uparrow$	$\delta < 1.05^2 \uparrow$	$\delta < 1.05^3 \uparrow$	AbsRel $\downarrow$	Sq. Rel$\downarrow$	RMSE $\downarrow$	RMSE log$\downarrow$	SILog$\downarrow$
MegaDepth [22]	0.185	0.352	0.489	0.253	0.174	2.419	0.286	0.096
MiDaS [29]	0.196	0.379	0.507	0.296	0.347	2.444	0.309	0.139
DPT [30]	0.340	0.590	0.733	0.142	0.072	3.259	0.211	0.057
UWNET [17]	0.097	0.194	0.296	3.730	> 5	2.750	0.436	0.253
UDepth [37]	0.273	0.463	0.605	0.214	0.163	1.992	0.239	0.084
Ours (PUDE)	**0.348**	**0.602**	**0.750**	**0.120**	**0.038**	**1.595**	**0.163**	**0.030**

Table 3. **Quantitative Evaluation: SQUID.** SQUID contains 57 samples collected in the open ocean environments. Images are more blurry compared with the images in the Sea-thru dataset.

5. Experiments

5.1. Implementation details

The training dataset for the PUDE method is from one high-resolution underwater video from [7], collected in the open ocean environment for object tracking. We choose this dataset since the images contain pure underwater backgrounds, facilitating ambient light estimation. Additionally, a continuous depth variation within the images is also beneficial for accurately estimating the parameters in the underwater image formation model Eq. (1). We selected 100 image samples from the videos, dividing them into 70 for training and 30 for testing. During training, the original images are downsampled to the dimension of 576×384 and fed into the neural network. PUDE is initialized with the DPT-Hybrid [30]. We train the PUDE model for 3 epochs using Adam with the learning rate equal to 10^{-5}. For quantitative evaluation, we choose to use the datasets in [4] and [1], which include the ground truth measurements obtained through stereo-matching. In the qualitative evaluation, the images are sampled from the real underwater dataset [21].

5.2. Zero-shot evaluations

5.2.1 Quantitative evaluation

In the quantitative evaluation, we compare the zero-shot performance of PUDE model with that of the underwater monocular depth estimation methods, UDepth [37], UW-Net [17], and in-air methods with good generalization properties MiDaS [29], MegaDepth [22] and DPT [30].

We employ the alignment strategy described in [29] to compare the estimated inverse relative depth with the gorundtruth depth. This alignment operates in the space of inverse ground truth measurements via solving an optimization problem for getting the shift and scale to align the estimated inverse relative depths with the inverse groundtruth. This alignment is subsequently inverted back to the normal depth space. To mitigate the impact of noise and erroneous measurements in groundtruth, we select depth thresholds ranging from $0.1m$ to $15m$ for the Sea-thru dataset [1] and $0.1m$ to $20m$ for the SQUID dataset [4] in our comparisons. We denote the aligned estimation as $\hat{\mathbf{z}}_i$, $i = 1 \cdots K$, and K is the number of valid pixels. $\mathbf{z}_i$ is the groundtruth depth. We utilize the following evaluation metrics described in [12] in our comparison:

- Mean absolute value of the relative error (AbsRel): $\frac{1}{K}\sum_{i=1}^{K}\frac{|\hat{\mathbf{z}}_i - \mathbf{z}_i|}{\mathbf{z}_i}$
- Squared relative error (Sq. Rel): $\frac{1}{K}\sum_{i=1}^{K}\left(\frac{\hat{\mathbf{z}}_i - \mathbf{z}_i}{\mathbf{z}_i}\right)^2$
- Root mean square error (RMSE): $\sqrt{\frac{1}{K}\sum_{i=1}^{K}(\hat{\mathbf{z}}_i - \mathbf{z}_i)^2}$
- Root mean square logarithmic error (RMSE log): $\sqrt{\frac{1}{K}\sum_{i=1}^{K}\left(\log(\hat{\mathbf{z}}_i) - \log(\mathbf{z}_i)\right)^2}$
- Scale invariant logarithmic error (SILog): $\frac{1}{K}\sum_{i=1}^{K} e_i^2 - \frac{1}{K^2}\left(\sum_{i=1}^{K} e_i\right)^2$, where $e_i = \log(\hat{\mathbf{z}}_i) - \log(\mathbf{z}_i)$
- Accuracy with threshold $\delta = \max\left(\frac{\hat{\mathbf{z}}_i}{\mathbf{z}_i}, \frac{\mathbf{z}_i}{\hat{\mathbf{z}}_i}\right) < \text{thr}$

The comparison results using the D3 and D5 datasets, with a forward-facing perspective, from the Sea-thru [1] is shown in Tab. 1 and Tab. 2. The comparison using the SQUID dataset [4] is shown in Tab. 3. Note that the thr in the comparison metric is set to be 1.05.

The experimental results unequivocally demonstrate the

Bound losses			$\delta < 1.05$	$\delta < 1.05^2$	$\delta < 1.05^3$	AbsRel	Sq. Rel	RMSE	RMSE log	SILog
$\mathcal{L}_{bl1}$	$\mathcal{L}_{bu}$	$\mathcal{L}_{bl2}$								
	✓	✓	0.360	0.610	0.752	0.111	0.029	1.097	0.161	0.028
✓		✓	0.459	0.712	0.846	0.083	0.016	0.757	0.122	0.016
✓	✓		0.498	0.754	0.878	0.076	**0.014**	0.815	0.122	0.016
✓	✓	✓	**0.502**	**0.761**	**0.881**	**0.075**	**0.014**	**0.737**	**0.117**	**0.015**

Table 4. **Ablation study on D5 dataset.** All bound losses contribute to the final performance of PUDE. PUDE trained with all bound losses has the best evaluation results on metrics.

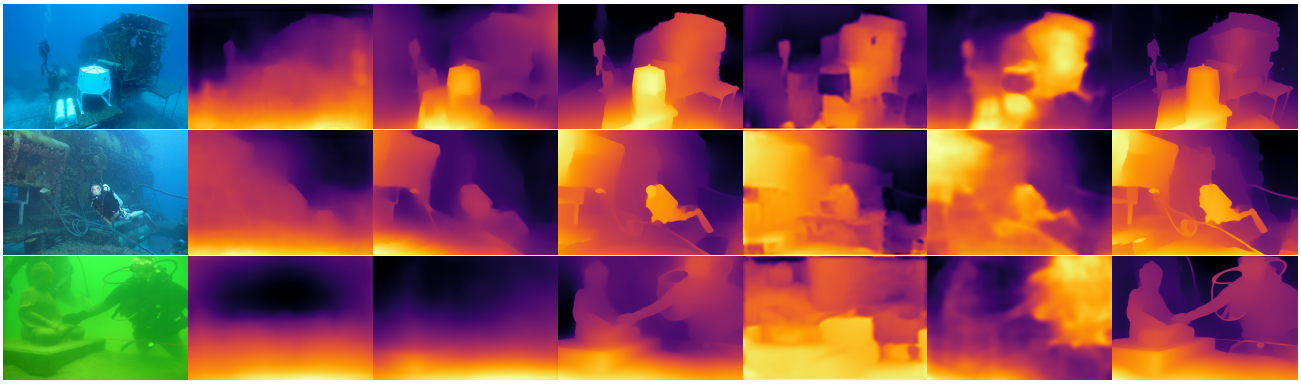


Figure 4. **Qualitative comparison on underwater images.** Qualitative comparison of PUDE with in-air (MegaDepth, MiDaS, DPT) and underwater methods (UWNET, UDepth) on underwater images. PUDE detects clearer background edges and presents more depth details.

superior performance of the PUDE model across all evaluated metrics on three datasets compared to other methods. It achieves lower errors and higher estimation accuracy, substantiating the effectiveness of our proposed approach. Particularly, the D5 and SQUID datasets contain images with a wider range and different levels of blurriness, on which the PUDE model demonstrates clearly improved performance. We point out that, due to the inherent challenges of blurry underwater images, the groundtruth measurements obtained via stereo-matching-based methods are predominantly confined to regions near the camera. This limitation obscures the advantages of our PUDE model in handling blur and background region detection.

5.2.2 Qualitative evaluation

We also provide qualitative evaluation across underwater images collected from different environments. We sample some underwater images that were collected in a turbid environment from the dataset [21] and do the zero-shot evaluation.

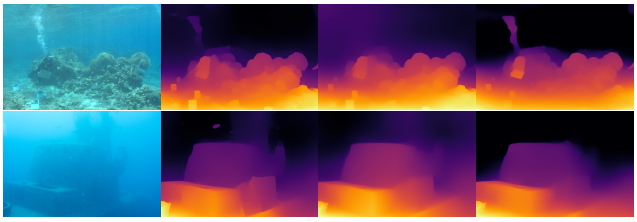


Figure 5. **Study of $\mathcal{L}_{bu}$ and $\mathcal{L}_{bl1}$.** Without $\mathcal{L}_{bu}$, the model cannot distinguish the background area clearly. Without $\mathcal{L}_{bl1}$, the depth on the far regions cannot be well estimated and is losing the details. The PUDE model trained by all bound losses can produce a more detailed depth and a clear background region recognition. Even the outline of the small fish, which is hard to see in human eyes, can be labelled clearly.

The results are shown in Fig. 4. The performance of our PUDE model in underwater depth estimation is evidently superior, particularly in capturing detailed features such as edges. It also effectively identifies background edges, including the fine contours of distant wires and small fishes in water. These outcomes show that PUDE not only inherits robust generalizability from DPT but also incorporates a physics-based understanding of underwater imaging, enabling the PUDE model to discern background regions without requiring labelled training data specific to background areas.

5.3. Ablation studies

To investigate the influence of the bound losses in training, we retrain the model with or without bound losses and do the evaluation on the Sea-thru D5 dataset. The results are shown in Tab. 4. It demonstrates the losses $\mathcal{L}_{bu}$ and $\mathcal{L}_{bl1}$ are the core losses in training, which are consistent with our training goals, solving the issues described in O1 and O2. $\mathcal{L}_{bl2}$ has less effect on the training results but can improve the model's performance. Furthermore, to present the effectiveness of bound losses $\mathcal{L}_{bl1}$ and $\mathcal{L}_{bu}$ visually, we make a comparison using the images from the Sea-thru D5 and SQUID, which is shown in Fig. 5. We can observe that, without $\mathcal{L}_{bl1}$, the estimated depth lacks details in the regions of the coral gullies. Meanwhile, the distant scenes are estimated farther from the camera and not even shown in the depth map. By contrast, without $\mathcal{L}_{bu}$, the detection of the background regions is not clear, with a fogging appearance near the edges.

6. Conclusion

This paper proposes a novel method for transferring the in-air depth estimation model, DPT, to underwater settings. We use the new bound losses to enforce the model to produce depth estimation that follows the underwater imaging formation model. The final quantitative and qualitative evaluation results prove the effectiveness of our method.

References

[1] Derya Akkaynak and Tali Treibitz. Sea-thru: A method for removing water from underwater images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1682–1691, 2019. 2, 3, 4, 7

[2] 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. IEEE, 2023. 2

[3] Codruta O Ancuti, Cosmin Ancuti, Christophe De Vleeschouwer, Laszlo Neumann, and Rafael Garcia. Color transfer for underwater dehazing and depth estimation. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 695–699. IEEE, 2017. 2

[4] Dana Berman, Deborah Levy, Shai Avidan, and Tali Treibitz. Underwater single image color restoration using haze-lines and a new quantitative dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 3, 7

[5] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 1

[6] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1

[7] Levi Cai, Nathan E McGuire, Roger Hanlon, T Aran Mooney, and Yogesh Girdhar. Semi-supervised visual tracking of marine animals using autonomous underwater vehicles. *International Journal of Computer Vision*, 131(6):1406–1427, 2023. 7

[8] Herng-Hua Chang, Chia-Yang Cheng, and Chia-Chi Sung. Single underwater image restoration based on depth estimation and transmission compensation. *IEEE Journal of Oceanic Engineering*, 44(4):1130–1149, 2018. 2

[9] Weifeng Chen, Shengyi Qian, and Jia Deng. Learning single-image depth from videos using quality assessment networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5604–5613, 2019. 2

[10] John Y Chiang and Ying-Ching Chen. Underwater image enhancement by wavelength compensation and dehazing. *IEEE transactions on image processing*, 21(4):1756–1769, 2011. 1, 2

[11] 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. 2, 4, 5

[12] 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. 2, 7

[13] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018. 1, 2

[14] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 2

[15] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 2

[16] Honey Gupta and Kaushik Mitra. Unsupervised single image underwater depth estimation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 624–628. IEEE, 2019. 2

[17] Praful Hambarde, Subrahmanyam Murala, and Abhinav Dhall. Uw-gan: Single-image depth estimation and image enhancement for underwater images. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2021. 2, 6, 7, 8

[18] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 2

[19] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 2

[20] Deborah Levy, Amit Peleg, Naama Pearl, Dan Rosenbaum, Derya Akkaynak, Simon Korman, and Tali Treibitz. Seathru-nerf: Neural radiance fields in scattering media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 56–65, 2023. 2

[21] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, 29:4376–4389, 2019. 7, 8

[22] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 1, 2, 6, 7, 8

[23] Yue Luo, Jimmy Ren, Mude Lin, Jiahao Pang, Wenxiu Sun, Hongsheng Li, and Liang Lin. Single view stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 155–163, 2018. 2

[24] Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, 2021. 1, 2

[25] Yan-Tsung Peng and Pamela C Cosman. Underwater image restoration based on image blurriness and light absorption. *IEEE transactions on image processing*, 26(4):1579–1594, 2017. 2

[26] 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. IEEE, 2015.

[27] Yan-Tsung Peng, Keming Cao, and Pamela C Cosman. Generalization of the dark channel prior for single image restoration. *IEEE Transactions on Image Processing*, 27(6):2856–2868, 2018. 2

[28] Michael Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[29] René 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 and machine intelligence*, 44(3):1623–1637, 2020. 1, 2, 3, 6, 7, 8

[30] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 1, 2, 3, 4, 6, 7, 8

[31] Nisha Varghese, Ashish Kumar, and AN Rajagopalan. Self-supervised monocular underwater depth recovery, image restoration, and a real-sea video dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12248–12258, 2023. 2

[32] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR*, abs/1908.00463, 2019. 3

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

[34] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth perception with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018. 2

[35] Jinghe Yang, Mingming Gong, Girish Nair, Jung Hoon Lee, Jason Monty, and Ye Pu. Knowledge distillation for feature extraction in underwater vslam. *arXiv preprint arXiv:2303.17981*, 2023. 2

[36] Xinchen Ye, Jinyi Zhang, Yazhi Yuan, Rui Xu, Zhihui Wang, and Haojie Li. Underwater depth estimation via stereo adaptation networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1

[37] 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. IEEE, 2023. 6, 7, 8