Physics-informed Knowledge Transfer for Underwater Monocular Depth Estimation

Anonymous ECCV 2024 Submission

Paper ID #9034

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 small underwater image set without groundtruth measurement, guided by a physical model for underwater image formation. We propose novel bound losses based on the physical model to rectify the depth estimations to align with actual underwater physical properties. Finally, in the evaluations across multiple datasets, we compare our Physics-informed Underwater Depth Estimation (**PUDE**) model with other existing in-air and underwater methods. The results reveal that the **PUDE** model excels in both quantitative and qualitative comparisons.

Keywords: Underwater depth · Physics-informed · Knowledge transfer

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 [23]. 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 [5,6,12,21,27–29]. Benefiting from the massive and various training datasets, some of in-air monocular depth estimation models, such as MiDaS [29] and DPT [28], 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 image formation [36]. For example, we applied the in-air trained models, e.g., DPT in [28], 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, 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 [9]. Using the newly proposed bound losses, which use the physics information to correct the inaccurate depth on the distant regions and background regions, our training method enforces the model to produce estimates that align with the extracted physical properties. Notably, this adaptation only requires a low amount of 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 [28], to underwater images. Inspired by these observations, we propose a novel training approach to adapt in-air pre-trained depth estimation models for underwater environments. 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 [28], to tackle the challenge of data shortage; 2) the integration of the underwater image formation model proposed in [9] 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 [11, 12, 18, 22]. 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 [14] or calibrated stereo image pairs [13] presents an attractive avenue for exploration but performs less compared to supervised learning [23], 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 [8,21,34]. Building further on that, MiDaS [29] extract depth data from 3D movies to enhance dataset diversity and improve generalization. DPT [28] 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 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 [9] has been demonstrated to be effective for solving problems in computer vision for underwater scenarios, for example, feature extraction [35], neural radiance fields (NeRFs) [19] and depth estimation [10]. The traditional methods utilize the physics-based image formation model to estimate the depth directly from the underwater images [3, 4, 7, 24–26]. As the foundation of these methods, Dark Channel Prior (DCP) [17], 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 [10], 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 [15] and UW-GAN [16]. They use the GAN method through color restoration to estimate depth. The recent works [2, 30] 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 [9] 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:

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

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

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 water particles. 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 [28] and MiDaS [29], generate the inverse relative depth estimation from a single image. 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:

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

Here, s and h are two unknown scale and shift coefficients. In practice, they can be off-line estimated for specific scenes using sequence images. The overall

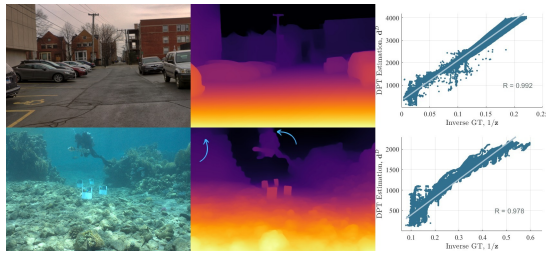


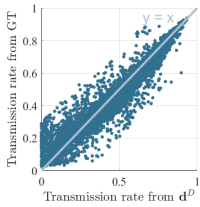
Fig. 1: DPT estimation of the underwater image has a curve tendency, with overestimation for distant regions. Meanwhile, the estimated depth has the issue of unclear background detection (blue).

mapping relationship between the inverse relative depth $\mathbf{d}_i$ and the inverse of metric depth $\mathbf{z}_i$ is linear on the in-air images [32]. An example can be seen in the first row of Fig. 1, which is the depth estimation of DPT on an outdoor in-air image from DIODE [31], we can observe that DPT estimation on the in-air image maintains a good linear relationship (the correlation coefficient $r = 0.992$). However, when we apply DPT directly to underwater images from the dataset [1, 4], in which the depth groundtruth is obtained through the Structure-from-motion and stereo matching, 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 and present a curve tendency with a weaker correlation, especially when $\mathbf{d}_i^D$ is small. An example is shown in the second row of Fig. 1 (More in Supple.).

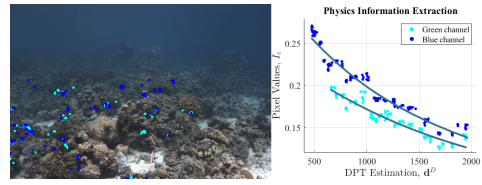
To observe the nonlinear trend caused by image blurriness, we applied the Sea-thru method [1] to calculate the medium transmission rate t_i in Eq. (2) on the grayscale images. We then compared these rates with those obtained from the DPT depth (aligned with groundtruth depth). An example of this procedure on the Sea-thru D3 dataset is shown in Fig. 2a, where we randomly sample 100 points from each image. We observe that the estimated transmission rate from DPT is mostly lower than that calculated by groundtruth depth when the transmission rate is small (when $t_i < 0.2$). The results reveal that the DPT underestimates the transmission rate for the small transmission rate regions, which means overestimating the depth in the distant regions. Furthermore, the background edges on the underwater depth image of DPT, such as the blue arrow pointed regions around corals, are less clear, exhibiting fogging compared with the estimation on the in-air image. 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.

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.



(a) t_i in Eq. (2) (D3 dataset)



(b) Point selection and model fitting in Eq. (1), Eq. (3)

Fig. 2: (a) The comparison between the calculated transmission rate t_i from groundtruth and the estimated t_i from DPT depth, presents an underestimation of the t_i for small values of DPT. (b) The left image shows the points selected in the refined estimation. The right shows the fitting of the estimation.

For these issues, we propose two bound losses (Sec. 4.3) in our training.

4 Method

4.1 Framework overview

Inspired by the observations in Sec. 3, we propose a physics-informed knowledge transfer method for 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 transition from in-air to underwater environments. Specifically, we adopt the pre-trained DPT-Hybrid model [28], 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.

Most daily images are *non-linear images*, which are automatically post-processed by nonlinear algorithms integrated into cameras, for example, gamma correction, to approach the non-linear perception of human eyes. By contrast, *linear images* are images in which pixel values are directly captured by the camera without nonlinear post-processing or only with ratio processing. To ensure the PUDE model's generalizability in reality and keep consistency in training and testing, we utilize post-processed non-linear underwater images as inputs. In contrast, for precise physical information extraction, we utilize the linear underwater images in the estimation of the physics parameters in Eq. (1) and Eq. (3).

Our training framework consists of two branches shown in Fig. 3. 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 (in Sec. 4.2). The second branch is the physics-informed knowledge transfer.

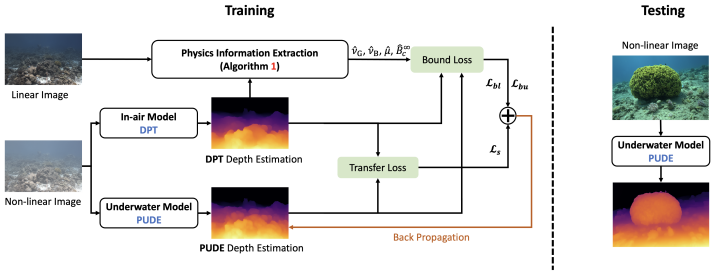


Fig. 3: Overall training framework for PUDE. The overall training framework comprises two components: the direct knowledge transfer from the DPT model and the physic-informed knowledge transfer based on the DPT estimation. Physical information is extracted from DPT's output to establish the bound losses.

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 [10]. Based on that, we propose two 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 the 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 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:

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

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 estimation 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 overestimate the depth in distant and blurry regions but offer more trustable estimation for the clearer regions, showing a good linear relationship described in Eq. (3). Therefore, we extract the physics information harnessed from DPT's depth estimations for clear regions.

The estimation process utilises the DPT depth estimation (on nonlinear post-processed images) and the linear images, consisting of two steps, the rough and refined estimation. The first step is the rough estimation using the whole image to obtain a robust estimation of the background light $\hat{B}_c^\infty$. The rationale for extracting the darkest points in the Sea-thru method 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. Therefore, we evenly select the N_1 darkest points, on which $J_{c,i} \approx 0$, from the whole image 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 suggestions in UDCP [10], we only use blue and green channels in this process. We denote the selected point set 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:

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

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

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

Utilizing the preliminary estimates $\hat{\nu}_G$, $\hat{\nu}_B$, and $\hat{\mu}$, we calculate the medium transmission rate $t_{c,i}$ in Eq. (2) for each pixel. Observations from Fig. 2a suggest that depth overestimation of DPT primarily occurs in regions with low transmission rates t_i. Therefore, in our implementation, in the refined estimation step, we exclude the regions where DPT's estimation is inaccurate while maximizing the utilization of the largest point sampling regions, in which the transmission rate is smaller than 0.2. Similar to rough estimation, in refined estimation, we select N_2 darkest points from the selected regions to construct the set M_{c2} and address a least-square problem. However, we streamline the process by utilizing the estimated $\hat{B}_c^\infty$ in the first step. The error ϵ_{c2} in this is:

$$\epsilon_{c2} := \sum_{i \in M_{c2}} ||I_{c,i} - \hat{B}_c^\infty (1 - e^{-\frac{\nu_c}{\mathbf{d}_i^D + \mu}})||_2^2, \tag{7}$$

and the least-square problem is:

$$(\hat{\nu}_G, \hat{\nu}_B, \hat{\mu}) = \underset{\nu_G, \nu_B, \mu}{\operatorname{argmin}} \sum_{c \in \{G,B\}} \epsilon_{c2}, \tag{8}$$

(a) RGB (b) DPT depth (c) $\mathcal{L}_{bu}$ (d) $\mathcal{L}_{bl}$ (e) local patches

Fig. 4: Loss visualization: Depth estimation from DPT loses details due to blurriness. However, our designed bound losses have differences across pixels according to the pixel intensity thereby enhancing the depth estimation details in training.

In our implementation, N_1 and N_2 is set to be 500 and 200 respectively. An implementation example of the refined estimation is shown in Fig. 2b.

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

Lower bound: The first bound is a lower bound on the estimated medium transmission rate $t^P_{c,i}$, $c \in \{G, B\}$. 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 the parameters estimation step, and using the PUDE depth estimation, we get the estimated $\hat{t}^P_{c,i}$:

$$\hat{t}^P_{c,i} = e^{-\frac{\hat{\nu}_c}{\mathbf{d}^P_i + \hat{\mu}}}. \tag{9}$$

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

$$I_{c,i} \geq (1 - t_{c,i}) B^{\infty}_c, \tag{10}$$

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

$$\hat{t}^P_{c,i} \geq 1 - \frac{I_{c,i}}{\hat{B}^{\infty}_c}. \tag{11}$$

This lower bound on $\hat{t}^P_{c,i}$ 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. (9). Consequently, the captured pixel value $I_{c,i}$ should not be lower than the estimated backscattering, as shown in Eq. (10).

Upper bound: The second bound is an upper bound on the estimated medium transmission rate $t^P_{c,i}$, $c \in \{G, B\}$ 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:

$$I_{c,i} \leq t_{c,i} + (1 - t_{c,i}) B^{\infty}_c. \tag{12}$$

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 $\hat{t}^D_{c,i}$, which is defined as:

$$\hat{t}^D_{c,i} = e^{-\frac{\hat{\nu}_c}{\mathbf{d}^D_i + \hat{\mu}}}. \tag{13}$$

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:

$$M_b = \{i \mid \frac{\hat{B}^\infty_c(1-\hat{t}^D_{c,i})}{I_{c,i}} \geq \gamma\}. \tag{14}$$

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, $\hat{t}^D_{c,i}$ is greater than the true value. Then based on the Eq. (12), we have,

$$I_{c,i} < \hat{t}^D_{c,i} + (1-t_{c,i})B^\infty_c, \quad i \in M_b, \tag{15}$$

substituting the estimated parameters, we have the bound:

$$\hat{t}^P_{c,i} < \frac{-I_{c,i} + \hat{t}^D_{c,i} + \hat{B}^\infty_c}{\hat{B}^\infty_c}, \quad i \in M_b. \tag{16}$$

This upper bound on $t^P_{c,i}$ in Eq. (16) is to adjust the depth estimation on the background regions. For the pixels on the background regions, the original DPT produce lower depth estimations. 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. (15).

Bound losses: We set the bound losses to enforce the estimated transmission rate $\hat{t}^P_{c,i}$ to satisfy the listed two 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 $f(x) = \max(0, x)$. The number of elements in M_b denoted as N_b. According to the bounds in Eq. (11), Eq. (16), the bound losses are:

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

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

Table 1: Quantitative results: The test datasets, Sea-thru D3, D5, and SQUID, comprise 43, 48, and 57 samples, respectively, collected from open ocean environment.

Datasets	Methods	$\delta < 1.05 \uparrow$	$\delta < 1.05^2 \uparrow$	$\delta < 1.05^3 \uparrow$	AbsRel ↓	Sq. Rel↓	RMSE ↓	RMSE log↓	SILog↓
D3	MegaDepth [21]	0.275	0.500	0.659	0.171	0.112	1.019	0.225	0.054
	MiDaS [29]	0.228	0.429	0.570	0.408	0.953	1.635	0.425	0.230
	DPT [28]	0.421	0.673	0.790	0.123	0.085	0.895	0.187	0.038
	UWNET [16]	0.097	0.192	0.290	0.774	2.829	2.027	0.650	0.451
	UDepth [37]	0.315	0.554	0.707	0.158	0.094	1.001	0.213	0.051
	Ours (PUDE)	**0.443**	**0.709**	**0.826**	**0.101**	**0.059**	**0.863**	**0.161**	**0.027**
D5	MegaDepth [21]	0.352	0.580	0.722	0.130	0.043	1.467	0.202	0.043
	MiDaS [29]	0.340	0.573	0.714	0.131	0.043	1.305	0.197	0.044
	DPT [28]	0.374	0.638	0.774	0.124	0.051	1.479	0.205	0.045
	UWNET [16]	0.075	0.158	0.240	0.395	0.334	2.576	0.409	0.184
	UDepth [37]	0.376	0.628	0.776	0.117	0.038	1.343	0.192	0.039
	Ours (PUDE)	**0.435**	**0.678**	**0.803**	**0.105**	**0.034**	**1.343**	**0.186**	**0.037**
SQUID	MegaDepth [21]	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 [28]	0.345	0.599	0.743	0.136	0.068	1.657	0.182	0.040
	UWNET [16]	0.082	0.167	0.255	0.510	0.741	3.363	0.478	0.278
	UDepth [37]	0.274	0.464	0.606	0.213	0.163	1.883	0.236	0.083
	Ours (PUDE)	**0.387**	**0.632**	**0.772**	**0.119**	**0.043**	**1.524**	**0.163**	**0.032**

These two bound losses are related to the pixel-wise intensity on color channel. Especially for the dark and distant regions, the difference in the pixel intensity can reflect the depth differences across pixels, the pixel intensity is mostly decided by the depth-related backscattering. The differences in bound losses caused by pixel intensity can encourage the model to learn the depth difference across pixels in small local blurry regions. In other words, the two bound losses guide the model to learn the nonlinear mechanism of backscattering in water, namely "how blurriness affects the pixel intensity". Fig. 4 shows an example. Therefore, the overall performance of the model in terms of detail detection gets improved.

Final loss: We use the parameters β_0, β_1 and β_2 to adjust the weights between the similarity loss and bound losses, which are set to 1, 10 and 10, respectively. The weights are decided by the metric and visual results on the validation set during training. The total loss $\mathcal{L}_{total}$ in training is:

$$\mathcal{L}_{total} = \beta_0 \mathcal{L}_s + \beta_1 \mathcal{L}_{bl} + \beta_2 \mathcal{L}_{bu}. \tag{19}$$

5 Experiments

5.1 Implementation details

The training dataset for the PUDE method contains 76 white-balanced linear underwater images collected in an open ocean environment with continuous depth variation, sourced from the SeaThru-NeRF [19] dataset. This scene structure is beneficial for the accurate estimation of the parameters in the underwater image formation model Eq. (1). We apply gamma correction to these images to

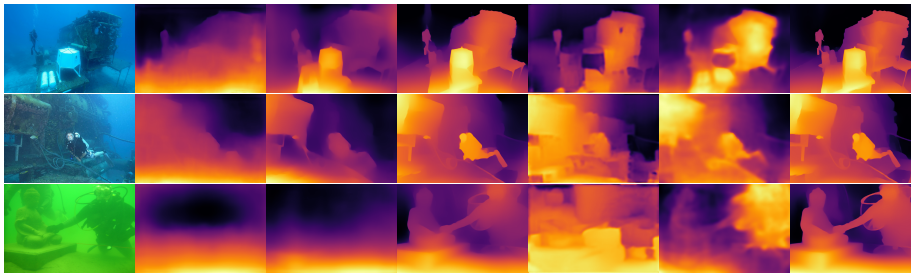


Fig. 5: Qualitative comparison on underwater images. Qualitative comparison of PUDE with in-air (MegaDepth, MiDaS, DPT) and underwater methods (UWNET, UDepth) on underwater images. PUDE detects clearer background edges and presents more depth details. (More in Supplementary)

generate corresponding nonlinear images for training. We randomly sample 20 images with groundtruth from the Sea-thru D3 datasets to construct the validation dataset. During training, the nonlinear images are downsampled to the dimension of 576×384 and fed into the neural network. PUDE is initialized with the DPT-Hybrid [28]. The training runs 3 epochs, employing the Adam optimizer with a learning rate of 10^{-5}. Notably, an early stopping is applied to prevent overfitting by looking at the the validation results.

In testing, we compare the performance of the PUDE model with the underwater depth estimation methods, UDepth [37], UW-Net [16], and in-air methods with good generalization properties MiDaS V2.1 [29], MegaDepth [21] and DPT-Hybrid [28]. The comparison encompasses both quantitative and qualitative evaluations. The quantitative evaluation uses the remaining images from the Sea-thru D3 and the whole D5 datasets [1]. Another zero-shot testing is performed on the SQUID dataset from [4], including ground truth measurements obtained through stereo-matching. Given the availability of RAW format images in these datasets, we applied gamma correction and brightness enhancement to the images. In the zero-shot qualitative evaluation, the images are sampled from the real underwater dataset [20].

5.2 Evaluations

Quantitative evaluation 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 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 utilize the evaluation metrics described in [11] in our

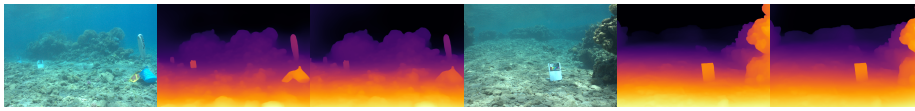


Fig. 6: Ablation study: (a) Without using $\mathcal{L}_{bu}$, the model cannot clearly detect the background edges near the rocks and has a fogging appearance. (b) Without using $\mathcal{L}_{bl}$, the distant rocks cannot be detected, depth of that region is overestimated.

comparison: Mean absolute value of the relative error (AbsRel), Squared relative error (Sq. Rel), Root mean square error (RMSE), Root mean square logarithmic error (RMSE log), Scale-invariant logarithmic error (SILog) and Accuracy with threshold ($\delta < thr$). Note that the thr in the comparison metric is set to be 1.05. The results are shown in Tab. 1. The experimental results demonstrate the 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, PUDE demonstrates promising performance in zero-shot testing.

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 [20] and do the zero-shot evaluation. The results are shown in Fig. 5. The performance of our PUDE model in underwater depth estimation is superior, particularly in capturing detailed features such as edges. It also effectively identifies background edges, including the fine contours of distant wires. 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.

5.3 Ablation studies

To investigate the influence of the bound losses in training, we retrain the model with and without bound losses and evaluate them on the Sea-thru D5 dataset. The results, shown in Tab. 2, demonstrate both the losses $\mathcal{L}_{bu}$ and $\mathcal{L}_{bl}$ contribute to training, which is consistent with our training goals, solving the issues

Table 2: Ablation study on D5 dataset. All bound losses contribute to the final performance of PUDE. PUDE has the best evaluation results on metrics.

$\mathcal{L}_{bu}$	$\mathcal{L}_{bl}$	$\delta < 1.05$	$\delta < 1.05^2$	$\delta < 1.05^3$	AbsRel	Sq. Rel	RMSE	RMSE log	SILog
		0.374	0.638	0.774	0.124	0.051	1.479	0.205	0.045
✓		0.426	0.672	0.798	0.116	0.048	1.456	0.199	0.042
	✓	0.398	0.657	0.795	0.109	0.034	1.383	0.192	0.039
✓	✓	**0.435**	**0.678**	**0.803**	**0.105**	**0.034**	**1.343**	**0.186**	**0.037**

Fig. 7: Method generalization: PUDE(M) also has improved performance in edge detection and detail detection compared with the in-air model, MiDaS.

described in O1 and O2. Moreover, to visually demonstrate the effectiveness of bound losses $\mathcal{L}_{bl}$ and $\mathcal{L}_{bu}$, we conducted a comparative analysis using images from the Sea-thru D5 dataset, as presented in Fig. 6. Without $\mathcal{L}_{bl}$, the depth in distant regions is overestimated. By contrast, without $\mathcal{L}_{bu}$, the detection of the background edges is not clear, resulting in a foggy appearance near the edges.

5.4 Method generalization on other in-air models

To further explore our method's generalization on different in-air models, we have conducted an additional experiment replacing DPT by MiDaS v2.1 [29] in the training loop, which has the same non-linearity described in Sec. 3 and demonstrate the second-best performance when applied to underwater images directly, and obtained **PUDE(M)**. The quantitative and qualitative results are shown in Fig. 7 and Tab. 3. The results confirm that our method can also enhance the in-water performance of the non-SOTA model, MiDaS. However, as our method relies on knowledge transfer, the efficacy is contingent upon the capability of the original in-air model. A better model yields better transfer results.

6 Conclusion

This paper proposes a novel method for transferring the in-air depth estimation model, DPT, to underwater settings. We observe DPT has issues with overestimating depth for distant regions and unclear background edge detection on underwater images. To tackle these, we use the novel 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. Additionally, the further experiment shows our method's generalization capabilities on other in-air models.

Table 3: PUDE(M) on SQUID dataset: PUDE(M) shows improved performance on the metrics compared to MiDaS, but PUDE has the best performance.

Models	$\delta < 1.05$	$\delta < 1.05^2$	$\delta < 1.05^3$	AbsRel	Sq. Rel	RMSE	RMSE log	SILog
MiDaS	0.196	0.379	0.507	0.296	0.347	2.444	0.309	0.139
PUDE(M)	0.279	0.463	0.601	0.188	0.115	1.869	0.221	0.067
PUDE	**0.435**	**0.678**	**0.803**	**0.105**	**0.034**	**1.343**	**0.186**	**0.037**

References

1. Akkaynak, D., Treibitz, T.: Sea-thru: A method for removing water from underwater images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1682–1691 (2019) 3, 4, 5, 7, 12
2. Amitai, S., Klein, I., Treibitz, T.: Self-supervised monocular depth underwater. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 1098–1104. IEEE (2023) 4
3. Ancuti, C.O., Ancuti, C., De Vleeschouwer, C., Neumann, L., Garcia, R.: Color transfer for underwater dehazing and depth estimation. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 695–699. IEEE (2017) 3
4. Berman, D., Levy, D., Avidan, S., Treibitz, T.: Underwater single image color restoration using haze-lines and a new quantitative dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 3, 5, 12
5. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4009–4018 (2021) 1
6. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023) 1
7. Chang, H.H., Cheng, C.Y., Sung, C.C.: Single underwater image restoration based on depth estimation and transmission compensation. IEEE Journal of Oceanic Engineering **44**(4), 1130–1149 (2018) 3
8. Chen, W., Qian, S., Deng, J.: Learning single-image depth from videos using quality assessment networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5604–5613 (2019) 3
9. Chiang, J.Y., Chen, Y.C.: Underwater image enhancement by wavelength compensation and dehazing. IEEE transactions on image processing **21**(4), 1756–1769 (2011) 2, 3, 4
10. Drews, P.L., Nascimento, E.R., Botelho, S.S., Campos, M.F.M.: Underwater depth estimation and image restoration based on single images. IEEE computer graphics and applications **36**(2), 24–35 (2016) 3, 7, 8
11. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014) 3, 12
12. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2002–2011 (2018) 1, 3
13. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017) 3
14. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019) 3
15. Gupta, H., Mitra, K.: Unsupervised single image underwater depth estimation. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 624–628. IEEE (2019) 4
16. Hambarde, P., Murala, S., Dhall, A.: Uw-gan: Single-image depth estimation and image enhancement for underwater images. IEEE Transactions on Instrumentation and Measurement **70**, 1–12 (2021) 4, 11, 12

17. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence **33**(12), 2341–2353 (2010) 3
18. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth international conference on 3D vision (3DV). pp. 239–248. IEEE (2016) 3
19. Levy, D., Peleg, A., Pearl, N., Rosenbaum, D., Akkaynak, D., Korman, S., Treibitz, T.: Seathru-nerf: Neural radiance fields in scattering media. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 56–65 (2023) 3, 11
20. Li, C., Guo, C., Ren, W., Cong, R., Hou, J., Kwong, S., Tao, D.: An underwater image enhancement benchmark dataset and beyond. IEEE Transactions on Image Processing **29**, 4376–4389 (2019) 12, 13
21. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2041–2050 (2018) 1, 3, 11, 12
22. Luo, Y., Ren, J., Lin, M., Pang, J., Sun, W., Li, H., Lin, L.: Single view stereo matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 155–163 (2018) 3
23. Ming, Y., Meng, X., Fan, C., Yu, H.: Deep learning for monocular depth estimation: A review. Neurocomputing **438**, 14–33 (2021) 1, 3
24. Peng, Y.T., Cao, K., Cosman, P.C.: Generalization of the dark channel prior for single image restoration. IEEE Transactions on Image Processing **27**(6), 2856–2868 (2018) 3
25. Peng, Y.T., Cosman, P.C.: Underwater image restoration based on image blurriness and light absorption. IEEE transactions on image processing **26**(4), 1579–1594 (2017) 3
26. Peng, Y.T., Zhao, X., Cosman, P.C.: Single underwater image enhancement using depth estimation based on blurriness. In: 2015 IEEE International Conference on Image Processing (ICIP). pp. 4952–4956. IEEE (2015) 3
27. Ramamonjisoa, M., Lepetit, V.: Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019) 1
28. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021) 1, 2, 3, 4, 6, 11, 12
29. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: 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, 3, 4, 11, 12, 14
30. Varghese, N., Kumar, A., Rajagopalan, A.: 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. pp. 12248–12258 (2023) 4
31. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., Shakhnarovich, G.: DIODE: A Dense Indoor and Outdoor DEpth Dataset. CoRR **abs/1908.00463** (2019), http://arxiv.org/abs/1908.00463 5
32. Wu, C.Y., Wang, J., Hall, M., Neumann, U., Su, S.: Toward practical monocular indoor depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3814–3824 (2022) 5

33. Wu, Y., Zhou, Y., Chen, S., Ma, Y., Li, Q.: Defect inspection for underwater structures based on line-structured light and binocular vision. Applied Optics **60**(25), 7754–7764 (2021) 2
34. Xian, K., Shen, C., Cao, Z., Lu, H., Xiao, Y., Li, R., Luo, Z.: Monocular relative depth perception with web stereo data supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 311–320 (2018) 3
35. Yang, J., Gong, M., Nair, G., Lee, J.H., Monty, J., Pu, Y.: Knowledge distillation for feature extraction in underwater vslam. arXiv preprint arXiv:2303.17981 (2023) 3
36. Ye, X., Zhang, J., Yuan, Y., Xu, R., Wang, Z., Li, H.: Underwater depth estimation via stereo adaptation networks. IEEE Transactions on Circuits and Systems for Video Technology (2023) 1
37. Yu, B., Wu, J., Islam, M.J.: Udepth: Fast monocular depth estimation for visually-guided underwater robots. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 3116–3123. IEEE (2023) 11, 12