IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Self-supervised Underwater Monocular Depth Estimation Informed by Multi-physics Processes

Fengqi Xiao, Juntian Qu*, Member, IEEE

***Abstract*—Depth information is crucial for underwater robotic detection and navigation tasks. However, the underwater imaging environment is complex and variable. The images captured by robots are typically sequences or videos with uniform scene content, and the ground-truth of depth is difficult to obtain. This challenge hinders the generalization of existing self-supervised monocular depth estimation (SMDE) schemes for practical underwater detection applications. To address this issue, we propose an SMDE method for underwater images informed by the physical process of optical degradation. Specifically, we developed a further degradation process for underwater images, which can constrain the image restoration process to solve the attenuation coefficient and depth map, and then combine it with the ego-motion based framework to form a self-supervised learning closed loop. Guided by inherent optical properties, this closed-loop can learn depth cues from the underwater image formation model and the geometric relationships involved in view transformation. Experiments demonstrate that the proposed method is reduced by about 9.1% in RMSE index and improved by about 3.5% in threshold accuracy compared with the SOTA method and can adapt to various underwater robot detection scenarios.**

***Index Terms*—RGB-D Perception; Computer Vision for Transportation; Deep Learning for Visual Perception; Underwater Image; Underwater Physical Model.**

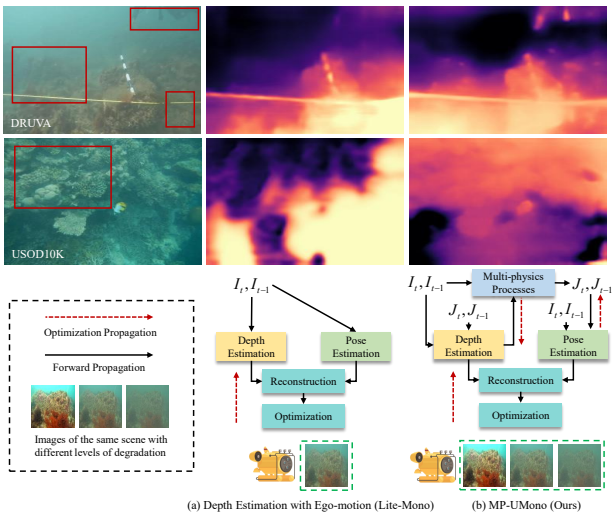

Fig. 1. The depth estimation results of the self-supervised method Lite-Mono [4] and the proposed method. I_{t-1} and I_t represent the previous frame and current frame, respectively. J represents restoration image. We train the two networks on the DRUVA dataset [5]. As illustrated, both methods demonstrate excellent performance on the DRUVA, while the proposed method has a more obvious advantage over Lite-Mono on the USOD10K.

I. Introduction

UNDERWATER robots have become the essential tools for human exploration and development of the ocean. Effective posture correction, positioning, and navigation of of underwater robots during exploration are highly dependent on accurate scene depth estimation technology. On land, various visual systems, such as binocular cameras [1], ToF cameras [2], and LiDAR [3], can capture depth information. However, the underwater environment is complex and variable, with ocean currents significantly disrupting the navigation of underwater robots. Additionally, the limited space and energy resources of underwater vehicles render these expensive, large, and precise depth acquisition devices impractical for underwater use. In contrast, monocular depth estimation technology, which requires only a single image as input without special settings, is particularly well-suited for underwater robot applications.

Monocular depth estimation technology is well-established on land, with numerous methods demonstrating excellent performance. However, these methods face challenges when applied directly to underwater environments. For supervised methods, obtaining depth ground-truth (GT) underwater is challenging, which impedes the effectiveness of supervised end-to-end training. Consequently, self-supervised monocular depth estimation (SMDE) has emerged as a more practical solution for underwater scenarios. Although there are several ego-motion based self-supervised methods on land, they often struggle to generalize to underwater images. This difficulty arises because underwater video sequences typically feature uniform scenes and simple content, combined with unique light attenuation effects, deep networks are prone to overfitting and may not handle diverse underwater scenes effectively. As illustrated in Fig.1, directly transferring land-based methods to underwater environments results in insufficient model generalization performance. The pose information between sample frames can no longer satisfy the self-supervised depth estimation learning in this scenario. Therefore, how to obtain knowledge specific to underwater scenes other than samples

Manuscript received: March 31, 2025; Revised June, 23, 2025; Accepted July, 17, 2025.

This paper was recommended for publication by Editor Abhinav Valada upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the National Key R&D Program of China (Grant No. 2024YFC2815000), the Shenzhen "Pengcheng Peacock Program," the Beijing "Youth Talent Promotion Project," the Tsinghua SIGS Cross-disciplinary Research and Innovation Fund (Grant No. JC2022002), the Shenzhen Science and Technology Program (Grant No. WDZC20231128114452001 and JCYJ20240813112107010), the Tsinghua SIGS Overseas Research Cooperation Fund (Grant No. HW2023001), the Tsinghua SIGS Scientific Research Startup Fund (Grant No. QD2022021C), the Jianghuai Dream Fund (Grant No. 2023-ZM01Z006), and the Shenzhen Key Laboratory of Advanced Technology for Marine Ecology (Grant No. ZDSYS20230626091459009). (*Corresponding Author: Juntian Qu.)

Fengqi Xiao and Juntian Qu are with Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. (Email: xiaofengqi@sz.tsinghua.edu.cn; juntian.qu@sz.tsinghua.edu.cn.)

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

to improve the generalization ability of the self-supervised network is the focus of this paper.

To address the aforementioned challenges, this paper proposes an underwater SMDE method named MP-UMono. Specifically, we integrate underwater image restoration and further degradation processes within an ego-motion self-supervised framework. Initially, we employ a neural network to estimate the attenuation coefficient of underwater images guided by inherent optical properties (IOP). This coefficient is then combined with depth information to solve for scene radiation according to the underwater image formation model. Subsequently, the underwater image is further degraded, with the attenuation coefficient in this process adhering to IOP and serving as a perturbation term in the transition from scene radiation to the further degraded image. This setup facilitates the derivation of the relationship between the attenuation coefficients of the original and further degraded images, thereby constraining the network output and establishing a closed loop for self-supervised learning. Consequently, the framework effectively learns depth information from various physical processes, including ego-motion, image restoration, and further degradation, significantly enhancing the network's generalization capabilities. The contributions of this paper are summarized as follows:

- This paper proposes an underwater SMDE framework, capable of learning depth information from various underwater physical processes.
- We designed the underwater image restoration and further degradation processes that can establish a self-supervised learning closed loop.
- The IOP are used to drive the physical process, simplifying the solution of physical parameters and improves the method's performance.

II. Related Work

A. Monocular Depth Estimation for Underwater Robotics

Monocular depth estimation methods for underwater robotics can be broadly classified into traditional and learning-based methods. Traditional methods typically rely on statistical priors derived from degraded underwater images to estimate depth maps, including maximum intensity prior [6], red channel prior [7], underwater light attenuation prior [8], and dark channel prior [9], among others. However, these methods are computationally intensive and time-consuming, often producing significant errors in depth estimation under severe light attenuation or mismatched prior conditions. Recently, deep learning techniques have demonstrated superior performance in underwater depth estimation tasks. These methods can be trained under supervised or self-supervised conditions. Yu et al. [10] proposed a rapid monocular depth estimation method that integrates domain knowledge of natural underwater scene features to train an end-to-end network. Wang et al. [11] incorporated underwater image formation model features into an end-to-end learning framework, effectively utilizing both local and global features of underwater images. Wang [12] et al. introduced an encoder that combines Transformer and inverted transmission map (TM) for end-to-end encoding and decoding, addressing the issue of non-uniform degradation in underwater images. Although these supervised learning methods achieve high accuracy, they require a large number of training samples. Given the challenging underwater imaging environment and the difficulty of obtaining depth GT, there is a growing need for depth map estimation methods under self-supervised conditions.

B. Self-supervised learning for Monocular Depth Estimation

SMDE methods rely on learning geometric projection relationships from video sequences to estimate depth information. These methods have evolved from depth estimation techniques initially developed for images captured in air. Zhou et al. [13] introduced an unsupervised learning framework to estimate dense 3D geometry and camera motion from unstructured video sequences. Zhang et al. [4] proposed an effective combination of CNNs and Transformers for self-supervised depth map estimation, significantly reducing the number of parameters and enhancing computational efficiency. In the underwater scene, Yang et al. [14] developed a self-supervised depth estimation network that, guided by multiple constraints based on underwater feature synthesis, learns depth trends from attenuated information in monocular underwater videos. Varghese et al. [5] combined reprojection loss with the underwater image formation model, creating a self-supervised underwater depth estimation network. These self-supervised methods mitigate the challenge of limited underwater paired samples and provide a foundational basis for the methods proposed in this paper.

III. Motivation

A. Depth in Underwater Image Formation Model

The Jaffe-McGlamery underwater image formation model [15] describes the degradation of light as it propagates underwater, accounting for absorption and scattering effects before it reaches the optical camera. According to this model, an underwater degraded image I_c captured by the camera can be represented as:

$$I_c(x) = J_c(x)T_c(x) + A_\infty(1 - T_c(x)), c \in \{R, G, B\}, \quad (1)$$

where $J(x)$ is the undegraded image, $T(x)$ is the TM, x represents the image pixel, and A_∞ is the background light (BL). According to the Beer-Lamber law [16], the TM can be further expressed using the attenuation coefficient term as $T_c(x) = e^{-\beta_c d(x)}$. Consequently, the model can be rewritten as:

$$I_c(x) = J_c(x)e^{-\beta_c d(x)} + A_\infty(1 - e^{-\beta_c d(x)}). \quad (2)$$

Among them, β_c is the attenuation coefficient, and $d(x)$ denotes the distance from the scene to the camera. The equation also indicates that the degradation of underwater optical images is highly correlated with $d(x)$, providing a crucial clue for the monocular depth estimation method proposed in this paper.

B. Depth in Image Further Degradation

As discussed in Section II.A, depth information can be derived not only from the robot's ego-motion underwater

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

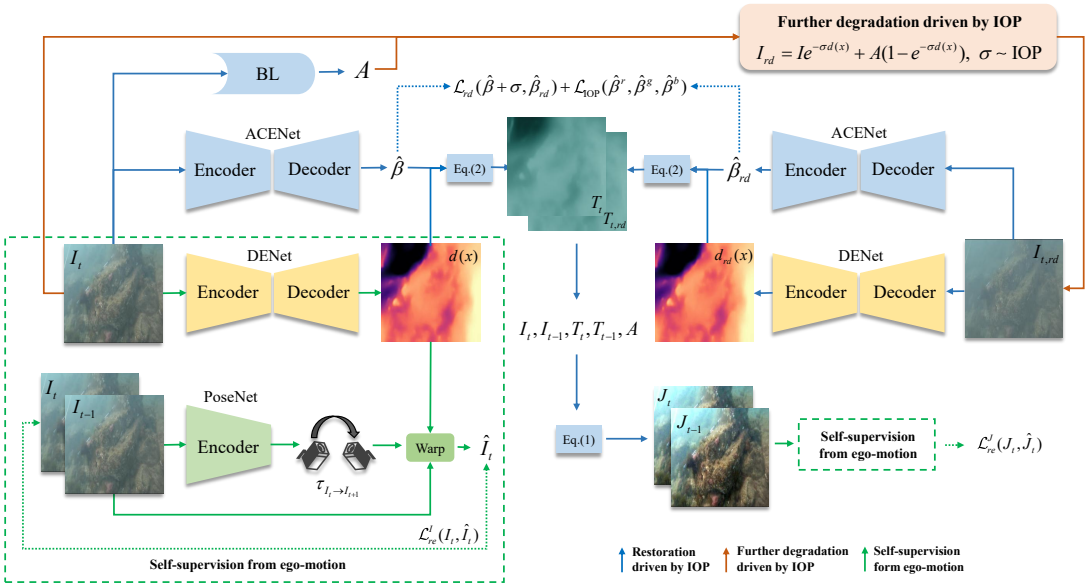

Fig. 2. The overall framework of the proposed method. It consists of three physical processes: the ego-motion process (indicated by the green line), the image restoration process (indicated by the blue line), and the further degradation process (indicated by the orange line.) ACENet, DENet, and PoseNet were all trained from scratch.

but also from the restoration or degradation processes of underwater images. Inspired by the work of Kar et al. [17], we propose a further degradation process driven by the IOP. To simplify the expression, we rewrite Eq.(2),

$$I_1 = Je^{-\beta_1 d_1} + A_1(1 - e^{-\beta_1 d_1}), \tag{3}$$

where J represents the undegraded image, and I_1 is the underwater image degraded from the undegraded image J using physical parameters such as attenuation coefficient β_1, propagation distance d_1, and BL A_1. The parameter A_1 can be estimated by the method in [18], and d_1 can be initially obtained from ego-motion. Therefore, among the remaining two variables J and β_1, it is sufficient to estimate only one in order to solve for the other. Subsequently, we further degrade I_1 based on the underwater image formation model:

$$I_2 = I_1 e^{-\beta_2 d_2} + A_2(1 - e^{-\beta_2 d_2}). \tag{4}$$

Substituting Eq.(3) into Eq.(4), we can get:

$$\begin{aligned} I_2 = Je^{-\beta_1 d_1} e^{-\beta_2 d_2} + A_1 e^{-\beta_2 d_2} \\ - A_1 e^{-\beta_1 d_1} e^{-\beta_2 d_2} + A_2 - A_2 e^{-\beta_2 d_2}. \end{aligned} \tag{5}$$

Here, we assume that the BL is independent to the degree of degradation, meaning $A_1 = A_2 = A$, that the depth information for the same scene remains constant, that is, $d_1 = d_2 = d$. Therefore, Eq.(5) can be written as:

$$I_2 = Je^{-(\beta_1+\beta_2)d} + A(1 - e^{-(\beta_1+\beta_2)d}). \tag{6}$$

It is evident from the above equation that the further degraded image I_2 also satisfies the underwater image formation model and can be represented as the result of additional degradation, following the J-M model with the undegraded image J as the input. Moreover, its attenuation coefficient can be regarded as a perturbation β_2 added to β_1.

To utilize the further degradation process for a self-supervised training loop, we designed an optimization strategy, where $\beta' = \beta_1$ is the parameter to be estimated. Let $\sigma = \beta_2$ denote an attenuation coefficient interference, so that I_1 degenerates to I_2, then the following three equations can be obtained from Eq. (4), (3), and (6):

$$I_2 = I_1 e^{-\sigma d} + A(1 - e^{-\sigma d}), \tag{7}$$

$$I_1 = Je^{-\beta' d} + A(1 - e^{-\beta' d}), \tag{8}$$

$$I_2 = Je^{-(\beta'+\sigma)d} + A(1 - e^{-(\beta'+\sigma)d}). \tag{9}$$

Therefore, we can further degrade the underwater image I_1 to obtain I_2 through Eq.(7), that is, further degrade I_t to obtain $I_{t,rd}$ in Fig.2. If an appropriate estimator to estimate the attenuation coefficient $\hat{\beta}$ from I_1, then the attenuation coefficient estimated from I_2 using this estimator will definitely be close to $\hat{\beta} + \sigma$, as indicated in Eq.(8) and (9). Let β_{rd} represent the estimated value from I_2. The optimization goal of the loop, comprising image restoration and further degradation, is to minimize the dissimilarity between $\hat{\beta} + \sigma$ and $\hat{\beta}_{rd}$.

C. Underwater Inherent Optical Properties

The attenuation of light during underwater propagation is wavelength-dependent, which implies that for color underwater images, the attenuation coefficients and perturbations discussed previously should exhibit three-channel variations. To simplify the solution process, we introduced the IOP to guide the further degradation process, thereby making the self-supervised learning approach more physically interpretable.

In the visible light spectrum, the absorption coefficient varies irregularly with wavelength [19], whereas the scattering coefficient, though slightly variable, remains relatively constant in certain types of water [20]. Richard et al. developed a model to describe spectral variations in attenuation coefficients [21], [22], which approximates the wavelength-dependent behavior of these coefficients through linear relationships. According to this model, the ratio of attenuation coefficients

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

across different color channels can be calculated as follows:

$$\frac{\beta^k}{\beta^r} = \frac{A^r_\infty(m\lambda^k + i)}{A^k_\infty(m\lambda^r + i)}, k \in (g, b). \tag{10}$$

Among them, λ^c, $c \in (r, g, b)$ represent the wavelengths of the red, green and blue channels respectively. For this study, these wavelengths are set to 620 nm, 540 nm and 450 nm respectively, with empirical parameters $m = -0.00113$ and $i = 1.62517$.

If the attenuation coefficient is $\beta = [\beta^r, \beta^g, \beta^b]$ and the perturbation is $\sigma = [\sigma^r, \sigma^g, \sigma^b]$, then both are subject to the above IOP. By setting the value of one channel, the values of the remaining channels can be obtained according to Eq.(10). Additionally, according to the geometric principles, the further degraded attenuation coefficient $[\beta^r + \sigma^r, \beta^g + \sigma^g, \beta^b + \sigma^b]$ still obeys IOP. Therefore, it is appropriate to use IOP to guide the further degradation process.

IV. Proposed Method

A. Overall Framework

To learn depth information from the physical process of light degradation, this paper proposes an underwater SMDE network named MP-UMono. It consists of three interconnected components: underwater image restoration process, further degradation process and ego-motion process, as shown in Fig.2. The restoration and further degradation processes are driven by the underwater IOP. During the self-supervised learning phase based on ego-motion, image frame I_t in a video sequence is used as the input of the depth map estimation network (DENet), and the previous frame I_{t-1} and current frame I_t are used as the input of the pose estimation network (PoseNet). The outputs of both will participate in the warp of the source frame and optimize the reconstruction with the target frame. In the underwater image restoration process, the BL is estimated using the method from [18]. Then, an attenuation coefficient estimation network (ACENet) is used to estimate the attenuation coefficient of the current frame. The coefficient is constrained by the IOP and can be combined with the estimated depth map and the original image according to Eq.(2) to obtain the restored image J. Finally, the IOP is used to construct a further degradation process to obtain I_{rd}. The restoration process is repeated, and the attenuation coefficients of the further degraded image and the current frame form a perturbation-related constraint. This results in a self-supervised constraint loop integrating all three processes.

B. Loss Functions

1) Loss from Ego-motion: To learn depth information from ego-motion, we follow [4] to build a self-supervised framework which utilizes the reprojection relationship between frames. In an image sequence, the target frame I_t can be reconstructed by sampling the pixels in the source frame I_s using the depth map D_t and relative pose estimated from the target frame. The projection of the coordinates in the target frame on the coordinates in the source frame p_s can be calculated:

$$p_s \sim K\hat{T}_{t\to s}\hat{D}_t\left(p_t\right)K^{-1}p_t. \tag{11}$$

Here K is the intrinsic camera matrix calculated from the registration process. Therefore, each pixel in the reprojected image $I_t(p_t)$ can be filled with the value of $I_s(p_s)$. According to the reprojection relationship described, the objective is to minimize the reconstruction loss between the reconstructed image $\hat{I}_t$ and the target image I_t. Following [4], we employ the combination of two similarity metrics L_1 and SSIM [23] to quantify the loss:

$$\mathcal{L}^I_{\text{re}} = \alpha||I_t, \hat{I}_t||_1 + (1-\alpha)\,\text{SSIM}\left(I_t, \hat{I}_t\right). \tag{12}$$

Similarly, we perform the same projection process on the restored image J_t. Therefore, the complete reconstruction loss will be expressed as:

$$\mathcal{L}^J_{\text{re}} = \alpha||J_t, \hat{J}_t||_1 + (1-\alpha)\,\text{SSIM}\left(J_t, \hat{J}_t\right). \tag{13}$$

So the reconstruction loss can be expressed as the sum of the above two losses: $\mathcal{L}_{\text{re}} = \mathcal{L}^I_{\text{re}} + \mathcal{L}^J_{\text{re}}$.

To smooth the generated disparity, an edge-aware smoothing loss is calculated:

$$\mathcal{L}_{\text{smooth}} = |\partial_x d^*_t|\, e^{-|\partial_x I_t|} + |\partial_x d^*_t|\, e^{-|\partial_y I_t|}, \tag{14}$$

where d^*_t represent the average normalized disparity.

2) Loss from Physical Processes: To accurately estimate the attenuation coefficient of the image, we use the underwater IOP to constrain the relationship between different channels of the estimated attenuation coefficient. Additionally, the attenuation coefficient estimated from the underwater image and the attenuation coefficient estimated from the further degraded image need to satisfy the perturbation constraint. Therefore, according to Eq.(7), (8), (9), and (10), the loss of the attenuation coefficient can be expressed as the sum of the above two constraints:

$$\mathcal{L}_{\text{ac}} = \mathcal{L}_{\text{rd}} + \mathcal{L}_{\text{IOP}} = ||(\hat{\beta}+\sigma) - \hat{\beta}_{rd}||_2 + \sum_{k\in g,b} ||\eta^k\hat{\beta}^r - \hat{\beta}^k||_2, \tag{15}$$

where η represents the ratio between channels in the IOP, as shown in Eq.(10). In addition, to maintain the consistency of the depth map in different physical processes, the depth maps are constrained as follow:

$$\mathcal{L}_{\text{depth}} = ||d - d_{rd}||_1. \tag{16}$$

Here, d and d_{rd} denote the depth maps estimated from the underwater image and the further degraded image, respectively.

Finally, to ensure that the restored image more realistic, it needs to satisfy the grayscale world assumption, so a color constancy loss is introduced:

$$\mathcal{L}_{cc} = \sum_{c\in\Omega}\left(\mu(J^c) - 0.5\right)^2, \tag{17}$$

where $\Omega = \{R, G, B\}$, J^c denotes a channel of the enhanced image, and $\mu(\cdot)$ denotes the operation of taking the mean value of the image. In summary, the overall loss of the proposed method is:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{re}} + \lambda_1\mathcal{L}_{\text{smooth}} + \lambda_2\mathcal{L}_{\text{ac}} + \lambda_3\mathcal{L}_{\text{depth}} + \lambda_4\mathcal{L}_{cc}. \tag{18}$$

Among them, λ is the weight for balancing each loss function.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

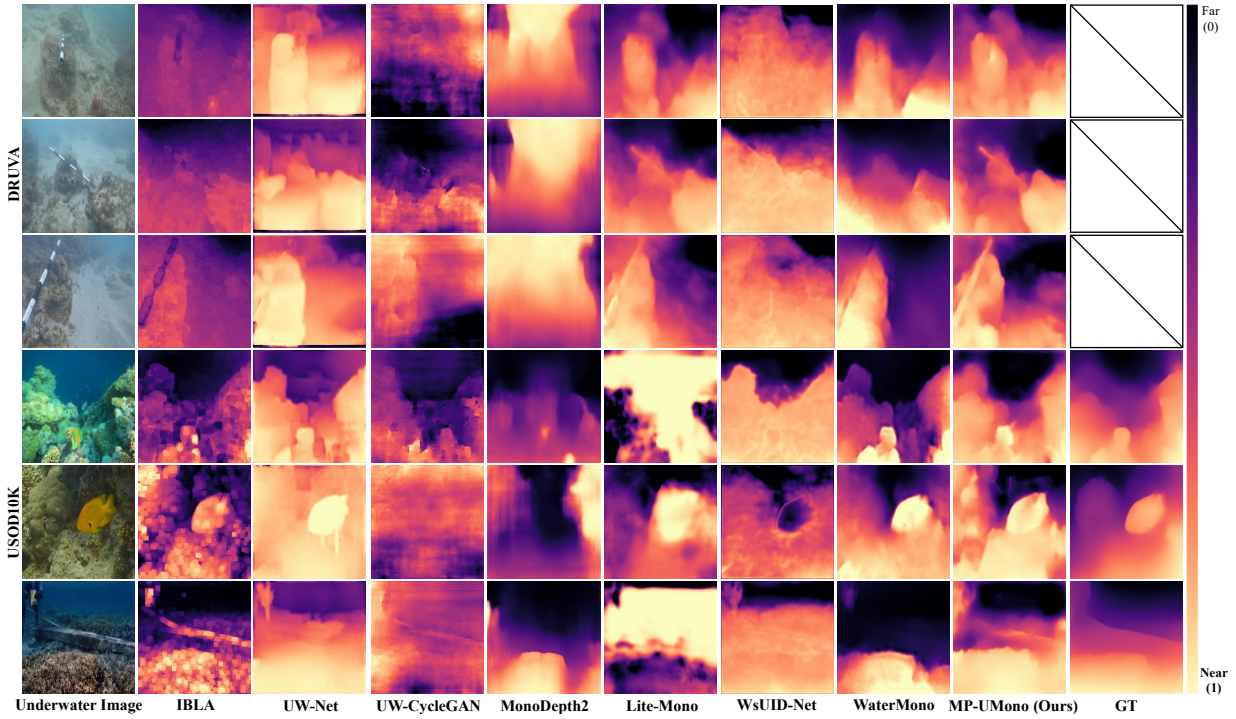

Fig. 3. Comparison of results across various methods. The proposed method demonstrates strong generalization performance across different datasets. Note that the DRUVA dataset is not equipped with depth GT.

V. Experiments

A. Experimental Settings

1) Datasets and Evaluation Metrics: Different from supervised methods, our method adopts a self-supervised approach and does not require paired data. We employ 19 sequences from DRUVA [5] with a total of 30,063 images for training. The performance of the trained model is evaluated on DRUVA, USOD10K [24], and UIEBD [25]. The comparative evaluation includes methods such as Lite-Mono [4], Monodepth2 [26], IBLA [7], UW-net [27], UW-CycleGAN (denoted as UW-C) [28], WsUID-Net [29], and WaterMono [30]. We use metrics described in [31] including AbsRel, SqRel, RMSE, RMSElog, as well as threshold accuracy $\delta \leq 1.25^n (n = 1, 2, 3)$ to assess depth estimation accuracy. SSIM [23], PSNR [32], UIQM [33] and FDUM [34] are used to measure the image quality.

2) Implementation Details: The proposed method employs three networks to estimate the attenuation coefficient, depth map and pose, respectively. The ACENet is based on an encoder-decoder architecture with skip connections. The DENet integrates CNNs and Transformers, as described in [4]. The PoseNet is based on Resnet-18 [35]. ACENet, DENet, and PoseNet were all trained from scratch. The proposed method is implemented using Pytorch framework on a workstation with a GeForce RTX 3090 Ti GPU. The size of the images in the training and testing is set to $3\times256\times256$. The training is performed for 30 epochs with a learning rate of 1e-4 and a batch size of 16. The weights λ_1, λ_2, λ_3 and λ_4 are set to 0.001. The perturbation σ is set to 0.1. The depth range is constrained between 0.5 and 4 meters, while the attenuation coefficient is limited between 0.5 and 1.

TABLE I
Quantitative evaluation of depth map estimation. The top two results are marked with bold and underline respectively.

Methods	USOD10K						
	AbsRel↓	SqRel↓	RMSE↓	RMSElog↓	$\delta \leq 1.25$↑	$\delta \leq 1.25^2$↑	$\delta \leq 1.25^3$↑
IBLA	0.6212	0.3729	0.3714	0.2605	<u>0.4183</u>	**0.6317**	<u>0.7630</u>
UW-net	0.7939	0.3757	0.3477	0.2758	0.3076	0.4687	0.5996
UW-C	0.6765	0.2629	0.3198	0.2698	0.2292	0.4208	0.6140
Monodepth2	0.7372	0.3440	0.3579	0.2907	0.2540	0.4275	0.5946
Lite-Mono	0.7794	0.4056	0.3461	0.2675	0.3360	0.5207	0.6590
WsUID-Net	0.6079	0.2464	0.3394	0.2944	0.2305	0.3738	0.5357
WaterMono	<u>0.5301</u>	<u>0.2419</u>	<u>0.2944</u>	<u>0.2457</u>	0.4110	0.6184	0.7412
Ours	**0.5069**	**0.2035**	**0.2677**	**0.2128**	**0.4320**	<u>0.6222</u>	**0.7677**

B. Evaluation on Depth Estimation

Fig.3 presents the results of the proposed method and the comparison method on DRUVA and USOD10K. The traditional IBLA method performs suboptimally for depth estimation of local details. The deep learning methods such as UW-Net, MonoDepth2, UW-CycleGAN, and Lite-Mono struggle to maintain consistent performance across different datasets. In contrast, the proposed method demonstrates good generalization across various data sets. To quantitatively

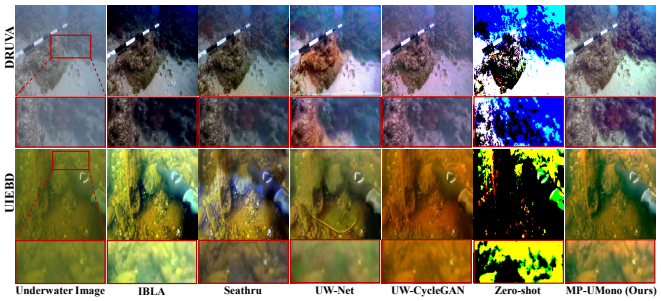

Fig. 4. Comparison results of underwater image restoration.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

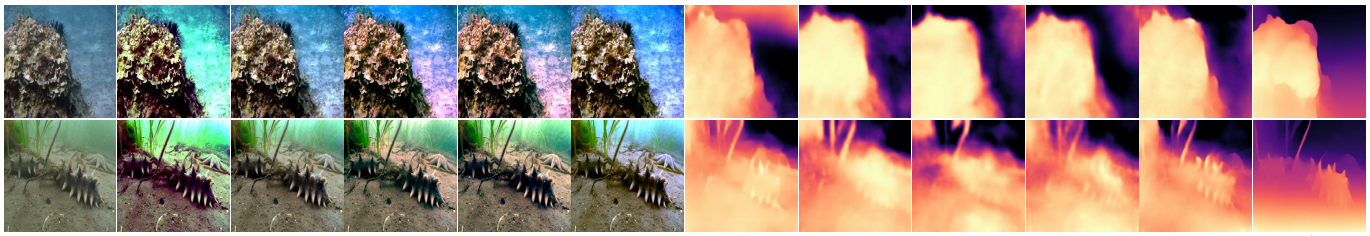

Fig. 5. Results of image restoration and depth estimation in ablation studies.

evaluate the proposed method, we employ seven metrics to measure the results of each methods as detailed in Table I. The table indicates that the proposed method outperforms other methods in objective metrics. Among them, the proposed method achieved the best and suboptimal results in all metrics, especially the RMSE, which was improved by 9.1% compared with the suboptimal.

C. Evaluation on Underwater Image Restoration

The proposed method enhances depth estimation by leveraging the restoration and further degradation processes of underwater images. Consequently, the accuracy of image restoration is crucial for improving depth estimation performance. We evaluated the proposed method against several similar methods and the baseline Zero-shot [17] on the DRUVA and UIEBD for image visual quality, as depicted in Fig.4. The results indicate that, compared to other methods, the proposed method excels in both global visual quality and local detail texture. Table II presents the objective visual quality scores. On the DRUVA, the proposed method ranks second in all non-reference metrics. On the UIEBD, it achieves the highest score in full-reference metrics, with non-reference metrics ranking second and fourth, respectively. Overall, the proposed method demonstrates relatively stable performance in underwater image restoration.

TABLE II
QUANTITATIVE EVALUATION OF VISUAL QUALITY. THE TOP TWO RESULTS ARE MARKED WITH BOLD AND UNDERLINE RESPECTIVELY.

Datasets	Metrics	IBLA	UW-Net	Seathru	UW-C	Zero-shot	**Ours**
DRUVA	UIQM↑	3.9068	4.0421	**4.4774**	4.0334	2.4711	<u>4.1324</u>
	FDUM↑	0.5564	0.5652	**0.61**	0.5688	NaN	<u>0.5722</u>
UIEBD	SSIM↑	0.6528	0.195	0.7218	0.7086	0.1009	**0.7251**
	PSNR↑	16.707	10.292	<u>18.568</u>	18.164	8.1638	**18.665**
	UIQM↑	3.2647	3.556	4.0032	<u>3.6467</u>	**9.0870**	3.4308
	FDUM↑	0.7114	0.6783	0.8007	**0.8285**	NaN	<u>0.803</u>

D. Ablation Study

To evaluate the effectiveness of each loss in the proposed method, we established four control groups, denoted as A, B, C and D, and present their results and evaluation metrics in Fig.5 and Table III. As illustrated, model A lacks the ego-motion guidance of the restored image due to the absence of $\mathcal{L}_{\text{re}}^{J}$, leading to poor performance in depth estimation. Model B uses random numbers instead of IOP to generate perturbations and omits the $\mathcal{L}_{\text{IOP}}$, which affects the restoration process and causes errors in depth estimation in the foreground area. Model C and D do not employ the further degradation process (without $\mathcal{L}_{\text{rd}}$ and $\mathcal{L}_{\text{depth}}$) and $\mathcal{L}_{cc}$, respectively. Both show serious color casts in the restoration results. From the table, the depth estimation performance of model C is inferior to that of model B and D, which also reflects the significance of the further degradation process for the proposed method. The proposed method demonstrates good performance in both image restoration and depth estimation with complete losses.

TABLE III
QUALITATIVE EVALUATION OF ABLATION STUY. THE TOP TWO RESULTS ARE MARKED WITH BOLD AND UNDERLINE RESPECTIVELY.

Models	Losses					Depth Estimation		Image Restoration	
	$\mathcal{L}_{\text{re}}^{J}$	IOP	$\mathcal{L}_{\text{rd}}$	$\mathcal{L}_{\text{depth}}$	$\mathcal{L}_{\text{cc}}$	AbsRel↓	$\delta \leq 1.25^3$↑	SSIM↑	PSNR↑
A	×	✓	✓	✓	✓	0.6488	0.6782	0.6987	17.852
B	✓	×	✓	✓	✓	<u>0.5396</u>	<u>0.7463</u>	0.7185	18.382
C	✓	✓	×	×	✓	0.6956	0.6762	0.6821	17.516
D	✓	✓	✓	✓	×	0.5740	0.7391	<u>0.7192</u>	<u>18.433</u>
Ours	✓	✓	✓	✓	✓	**0.5069**	**0.7677**	**0.7251**	**18.665**

E. Evaluation on Reliability

Fig.6 visualizes each component of underwater images with varying BLs during image restoration. The figure demonstrates that the proposed method effectively estimate attenuation coefficient without GT supervision. This is achieved through the constraints imposed by the IOP and the further degradation process. Consequently, the method integrates the estimated depth map to derive the TM and accurately reconstructs scene radiance.

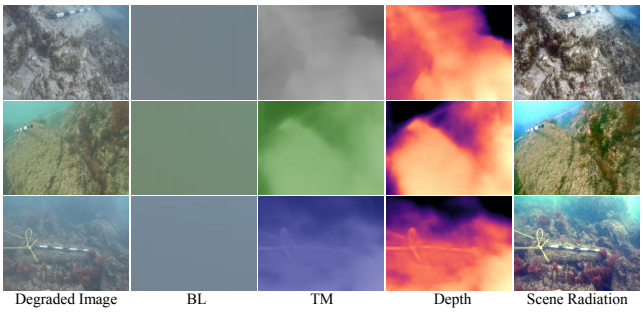

Fig. 6. Visualization of components in physical processes.

TABLE IV
EVALUATION OF COMPUTATIONAL EFFICIENCY AND COST. THE TOP TWO RESULTS ARE MARKED WITH BOLD AND UNDERLINE RESPECTIVELY.

Methods	Image Size 3×256×256				Platforms
	Parameters↓	FLOPS (G)↓	Runtime(s)↓	FPS↑	
IBLA	-	-	5.5703	0.1795	CPU
UW-net	9,866,697	281.6473	0.1601	6.2484	GPU
UW-C	15,909,910	<u>1.1155</u>	0.0779	12.8369	GPU
Monodepth2	14,329,236	209.32	1.1774	0.8493	GPU
Lite-Mono	**3,068,579**	**2.6837**	**0.01563**	**63.9795**	GPU
WsUID-Net	<u>7,434,310</u>	147.905	0.0673	14.8588	GPU
WaterMono	**3,068,579**	**2.6837**	0.01606	62.2665	GPU
Ours	**3,068,579**	**2.6837**	<u>0.01598</u>	<u>62.5782</u>	GPU

F. Evaluation on Computational Efficiency and Cost

Depth estimation is the core technology for underwater robots to complete navigation and detection, and its compu-

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

tational overhead is the key to whether it can participate in the actual application of robots. Therefore, this paper finally evaluates the proposed method and the comparative method in terms of parameters, floating-point operations per second (FLOPS), runtime, and frame rate (FPS). As shown in Table IV, compared with other methods, the performance of the proposed method in each indicator has obtained the best or suboptimal scores, which is competitive and is expected to promote the actual deployment and application of the algorithm on underwater robots.

VI. Conclusion

This paper proposes a SMDE method for underwater robots, which learns depth information from multi-physics Processes, and can overcome the overfitting problem of the self-supervised framework based on ego motion in underwater image scenes. We propose to use IOP to drive the restoration and further degradation process of the image, and establish a self-supervised learning loop with the framework based on ego motion. Experiments show that, this method shows better generalization ability in depth estimation, and shows stable and competitive performance in image restoration tasks, computational costs, etc. The effective fusion of real-world physical priors and data samples in the proposed method can promote the environmental adaptability of underwater robots and is expected to provide new ideas for the practical application of autonomous navigation and positioning of underwater robots.

References

[1] Y. Hu, W. Zhen, and S. Scherer, "Deep-learning assisted high-resolution binocular stereo depth reconstruction," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8637–8643.

[2] J. Noraky and V. Sze, "Low power depth estimation of rigid objects for time-of-flight imaging," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1524–1534, 2019.

[3] J. Choe, K. Joo, T. Imtiaz, and I. S. Kweon, "Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4672–4679, 2021.

[4] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 537–18 546.

[5] N. Varghese, A. Kumar, and A. 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*, 2023, pp. 12 248–12 258.

[6] N. Carlevaris-Bianco, A. Mohan, and R. M. Eustice, "Initial results in underwater single image dehazing," in *Oceans 2010 Mts/IEEE Seattle*. IEEE, 2010, pp. 1–8.

[7] Y.-T. Peng and P. C. Cosman, "Underwater image restoration based on image blurriness and light absorption," *IEEE transactions on image processing*, vol. 26, no. 4, pp. 1579–1594, 2017.

[8] W. Song, Y. Wang, D. Huang, and D. Tjondronegoro, "A rapid scene depth estimation model based on underwater light attenuation prior for underwater image restoration," in *Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part I 19*. Springer, 2018, pp. 678–688.

[9] Y.-T. Peng, K. Cao, and P. C. Cosman, "Generalization of the dark channel prior for single image restoration," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 2856–2868, 2018.

[10] B. Yu, J. Wu, and M. J. Islam, "Udepth: Fast monocular depth estimation for visually-guided underwater robots," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3116–3123.

[11] J. Wang, J. Wang, S. Rong, and B. He, "Umono: Physical model informed hybrid cnn-transformer framework for underwater monocular depth estimation," *arXiv preprint arXiv:2407.17838*, 2024.

[12] C. Wang, H. Xu, G. Jiang, M. Yu, T. Luo, and Y. Chen, "Underwater monocular depth estimation based on physical-guided transformer," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[13] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[14] X. Yang, X. Zhang, N. Wang, G. Xin, and W. Hu, "Underwater self-supervised depth estimation," *Neurocomputing*, vol. 514, pp. 362–373, 2022.

[15] J. S. Jaffe, "Computer modeling and the design of optimal underwater imaging systems," *IEEE Journal of Oceanic Engineering*, vol. 15, no. 2, pp. 101–111, 1990.

[16] D. F. Swinehart, "The beer-lambert law," *Journal of chemical education*, vol. 39, no. 7, p. 333, 1962.

[17] A. Kar, S. K. Dhara, D. Sen, and P. K. Biswas, "Zero-shot single image restoration through controlled perturbation of koschmieder's model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 205–16 215.

[18] Z. Fu, H. Lin, Y. Yang, S. Chai, L. Sun, Y. Huang, and X. Ding, "Unsupervised underwater image restoration: From a homology perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 643–651.

[19] R. C. Smith and K. S. Baker, "Optical properties of the clearest natural waters (200–800 nm)," *Applied optics*, vol. 20, no. 2, pp. 177–184, 1981.

[20] A. H. Barnard, W. S. Pegau, and J. R. V. Zaneveld, "Global relationships of the inherent optical properties of the oceans," *Journal of Geophysical Research: Oceans*, vol. 103, no. C11, pp. 24 955–24 968, 1998.

[21] R. W. Gould, R. A. Arnone, and P. M. Martinolich, "Spectral dependence of the scattering coefficient in case 1 and case 2 waters," *Applied Optics*, vol. 38, no. 12, pp. 2377–2383, 1999.

[22] X. Zhao, T. Jin, and S. Qu, "Deriving inherent optical properties from background color and underwater image enhancement," *Ocean Engineering*, vol. 94, pp. 163–172, 2015.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[24] L. Hong, X. Wang, G. Zhang, and M. Zhao, "Usod10k: a new benchmark dataset for underwater salient object detection," *IEEE transactions on image processing*, 2023.

[25] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Transactions on Image Processing*, vol. 29, pp. 4376–4389, 2019.

[26] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth prediction," October 2019.

[27] H. Gupta and K. Mitra, "Unsupervised single image underwater depth estimation," *arXiv preprint arXiv:1905.10595*, 2019.

[28] H. Yan, Z. Zhang, J. Xu, T. Wang, P. An, A. Wang, and Y. Duan, "Uw-cyclegan: Model-driven cyclegan for underwater image restoration," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[29] K. Li, X. Wang, W. Liu, Q. Qi, G. Hou, Z. Zhang, and K. Sun, "Learning scribbles for dense depth: Weakly supervised single underwater image depth estimation boosted by multitask learning," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[30] Y. Ding, K. Li, H. Mei, S. Liu, and G. Hou, "Watermono: Teacher-guided anomaly masking and enhancement boosting for robust underwater self-supervised monocular depth estimation," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[31] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[32] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369.

[33] K. Panetta, C. Gao, and S. Agaian, "Human-visual-system-inspired underwater image quality measures," *IEEE Journal of Oceanic Engineering*, vol. 41, no. 3, pp. 541–551, 2015.

[34] N. Yang, Q. Zhong, K. Li, R. Cong, Y. Zhao, and S. Kwong, "A reference-free underwater image quality assessment metric in frequency domain," *Signal Processing: Image Communication*, vol. 94, p. 116218, 2021.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.