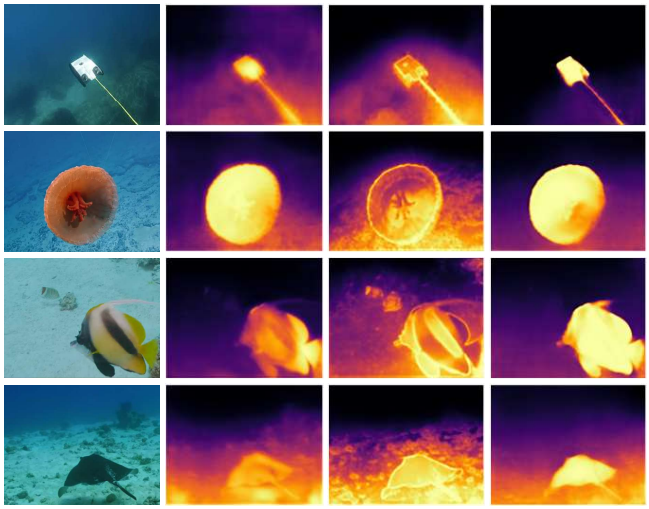
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Fast Monocular Depth Estimation for Underwater Robotics Leveraging Attenuation Differences as Supplementary Information

Hao Wang[1], Liang Lu[1], Yan Dong[1], Bin Han[1] *Senior Member, IEEE*

***Abstract*— Underwater and in-air environments exhibit distinct imaging characteristics, which should be carefully considered and effectively exploited for accurate depth estimation. In this work, we analyze the effectiveness of wavelength-dependent attenuation for underwater depth estimation and show that it is helpful but insufficient to perform depth estimation independently. Therefore, we propose a fast underwater monocular depth estimation network that incorporates underwater light absorption difference (ULAD) as supplementary information. Compared with methods that rely solely on RGB input, the proposed approach provides more accurate depth predictions. In our network, RGB and ULAD features are extracted by MobileNetV4 and fused using FusionMamba, followed by decoding and refinement with a micro Vision Transformer. The network is trained on the USOD10K dataset and evaluated on both its test set and the FLSea dataset. Experimental results demonstrate that our method achieves more accurate depth estimation and higher efficiency compared with other lightweight networks. Furthermore, Compared with existing state-of-the-art fast underwater depth estimation methods, our network further reduces the number of parameters by 10% and improves inference speed by 43%. The source code and pretrained models are available at https://github.com/Sillear/ULAD-Depth**



(a)Original (b)RGB (c)ULAD (d)Our

Fig. 1. Comparison of using only RGB information or Underwater Light Absorption Difference (ULAD) versus our fusion method. Our network achieves better results by leveraging the complementary advantages of RGB and ULAD.

I. INTRODUCTION

Depth information plays a crucial role in high-level vision applications for robotics. However, in underwater environments, the unique properties of the water medium make depth acquisition significantly more challenging than in air. Existing underwater ranging devices, such as LiDAR [1], RGB-D cameras [2], and sonar systems [3], are typically expensive and primarily designed for medium-to-long range measurements. In contrast, efficient and low-cost devices remain scarce for close-range vision in underwater robotics. Given that cameras are fundamental components of robotic perception systems, estimating depth directly from RGB images provides a practical and effective alternative.

Numerous monocular depth estimation methods have been developed for aerial environments [4]–[6]. However, unlike air, the water medium exhibits unique optical properties, such as forward scattering, backscattering, and wavelength-dependent attenuation. These effects lead to image degradation in the form of turbidity, blurring, and color distortion, which significantly hinder the applicability of terrestrial depth estimation methods in underwater scenarios. This limitation has been demonstrated in [7], [8]. Although these unique properties pose significant challenges for depth estimation, we argue that the core difficulty lies in the lack of effective utilization of these properties. Since the differences between underwater and aerial environments stem from these characteristics, a more reasonable approach is to analyze and leverage them to support underwater depth estimation.

For air and underwater environments, wavelength-dependent absorption of light, which behaves differently in water [9], is a critical factor contributing to their differences. We refer to the Underwater Light Absorption Difference as ULAD. Specifically, water absorbs blue and green light less than red light, resulting in an imbalance among the RGB channels in underwater images. This leads to the failure of many depth estimation methods in underwater scenarios, such as the DCP-based method [10], [12], because the prior conflicts with their underlying assumptions. For current mainstream deep learning methods trained on in-air image datasets, such differences prevent the networks from learning the unique characteristics of underwater images, resulting in suboptimal performance in underwater environments, as reported in [7]. Although ULAD leads to the failure of many existing methods, it remains a crucial factor for accurate

*This work was supported in part by the National Nature Science Foundation of China 52375015, in part by the Jing-Jin-Ji Regional Integrated Environmental Improvement-National Science and Technology Major Project 2025ZD1206400, and in part by the Interdiciplinary Research Program of HUST 2024JCYJ037.(Corresponding auther: binhan@hust.edu.cn)

[1]the School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China. wanghao96@hust.edu.cn, liang_lu@hust.edu.cn, dongy2021@outlook.com

underwater depth estimation. The color shift caused by it becomes more severe as depth increases, which is potentially beneficial for underwater depth estimation. Recent studies [13], [14] have also confirmed the effectiveness of utilizing it for this purpose.

Depending on how ULAD is incorporated, existing monocular underwater depth estimation methods can be categorized into two groups. The first group includes methods that rely solely on traditional RGB information, such as [15], [16], where ULAD is either ignored or significantly simplified. The second group consists of methods that rely exclusively on ULAD, such as [9], [14], which discard the original RGB information. It is evident that relying on a single type of information is not reasonable for underwater depth estimation. Numerous monocular depth estimation methods in aerial environments have demonstrated the effectiveness and importance of traditional RGB information. Therefore, we argue that RGB information is essential. Although ULAD has been shown to be beneficial for depth estimation, it still suffers from critical limitations, which we describe in detail in Section III. Therefore, a more appropriate approach is to integrate both traditional RGB information and ULAD information in a complementary manner. To the best of our knowledge, our method is the first work that further explores the fusion of underwater attenuation features and RGB features for underwater depth estimation. The method most similar to ours is [11], which jointly utilizes traditional RGB and ULAD information. However, it simply fuses the depth maps estimated separately from RGB and ULAD, without exploring a deeper integration of the two types of information.

We conduct a detailed analysis of the effectiveness and inherent limitations of ULAD in depth estimation. A special attenuation information space (AIS) is proposed to integrate ULAD information, and an underwater monocular depth estimation network is developed by incorporating RGB information. The network adopts MobileNetV4 as the encoder to independently extract features from RGB images and ULDA data. These features are subsequently fused through the MambaFusion module. The fused representation is then processed by a decoder and mViT to generate the depth estimation output.

Our main contributions are summarized as follows:

(1) To the best of our knowledge, our method is the first work that further explores the fusion of underwater attenuation features and RGB features for underwater depth estimation. We propose a lightweight network that integrates RGB and ULAD information, achieving more accurate underwater depth predictions than existing methods and enabling real-time deployment on embedded robotic platforms.

(2) We design a feature fusion module called MambaFusion, which leverages state space models to effectively integrate heterogeneous features.

(3) We introduce an Attenuation Prior Loss function based on the physical image formation model, which explicitly accounts for the attenuation characteristics of light in underwater environments.

II. Related work

A. Traditional underwater depth estimation

Most traditional underwater depth estimation is typically based on the underwater image formation model [17]. In [12], the depth is estimated by model parameter estimation based on the DCP [18] extended for underwater scenarios. However, the assumptions behind these priors are often too restrictive, leading to limited performance across diverse underwater environments. In addition to the above, many other priors have been explored for underwater depth estimation, including gradient-based method [19], lighting-reflection decomposition approach [20], and methods that integrate multiple priors for improved estimation [21]. Underwater depth estimation benefits from the underwater imaging model. However, the estimation of model parameters is often prone to errors. Moreover, since the ground truth of these parameters is generally unavailable, the accuracy of the imaging model usually cannot be quantitatively assessed. As a result, the quality of the depth estimation is typically judged based on the enhancement outcome, which often leads to suboptimal depth accuracy.

B. Deep learning-based underwater depth estimation

In recent years, the growing popularity of deep learning has led to new applications in underwater monocular depth estimation. With the introduction of new underwater image datasets [24], [25], the successful training of neural networks for this task has become feasible. The physics-based image formation model, as a fundamental mechanism of underwater imaging, has been incorporated into many deep learning approaches. Studies [26], [27] introduced physical priors derived from this model into neural networks, where the estimation of model parameters is typically used to restore underwater images and estimate scene depth. Recently, Vision Transformers (ViT) have provided new perspectives and methodologies for a wide range of visual tasks. This architecture has also demonstrated promising performance in underwater depth estimation [4]. In addition, joint network [28], mutual distillation loss [29], and teacher-student learning strategies [30] have also been introduced into underwater depth estimation. Although the number of underwater datasets continues to grow, their diversity and scale remain significantly limited compared to those of in-air image datasets. To address this limitation, many approaches employ GAN-based architectures or zero-reference networks, such as underwater unsupervised depth estimation method [31].

In summary, although numerous effective deep learning methods have been proposed, most primarily focus on RGB information or partially adapt terrestrial priors to underwater environments. The potential of ULAD information has not been fully explored. Encouragingly, a growing number of researchers have recently begun to recognize its importance in underwater depth estimation.

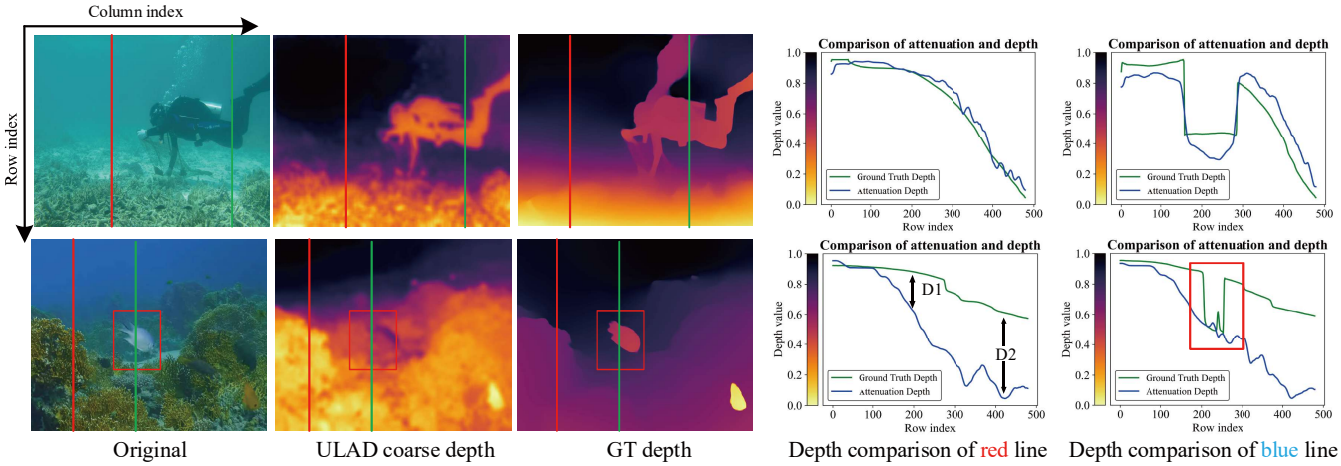


Fig. 2. Visualization of the correlation between ULAD and underwater depth estimation. The first row demonstrates its effectiveness, while the second row illustrates its limitations. The red and blue lines indicate the positions where depth is estimated, corresponding to the curves shown on the right. The red box in the image corresponds to the red box values in the curve on the right.

C. ULAD-based underwater depth estimation

The mechanism of ULAD was discovered long ago [34] and has been widely used in underwater image enhancement for many years, but its benefits for underwater depth estimation have only recently gained attention. A method for underwater depth estimation using underwater attenuation prior is proposed in [9] formally, which has become the foundation for many recent underwater monocular depth estimation approaches. In [13], the method proposed an RMI space and utilized a lightweight network to leverage attenuation information for depth estimation. The experimental results confirmed the importance of ULAD information in underwater depth estimation. This conclusion was further validated in their subsequent work [14]. Besides, the works of [35] employed the RMI space for depth estimation, while the IMT space proposed in [11] also serves as a similar information space, achieving promising results in underwater depth estimation. In [7], the authors studied the issues arising from directly applying the terrestrial depth estimation method DPT [4] to underwater environments. Although ULAD was not directly used for depth estimation, their observation of nonlinear trends of depth serves as an indirect validation of the effectiveness of it.

In summary, ULAD has been increasingly used in underwater depth estimation, with numerous studies demonstrating its effectiveness. As a key differentiating feature between underwater and terrestrial images, it should be fully leveraged in underwater depth estimation.

III. Method

A. ULAD for Depth Estimation

The feasibility of employing ULAD for depth estimation can be justified within the framework of underwater imaging models. At present, the following simplified model is commonly used:

$$\mathbf{I}(x) = \mathbf{J}(x)\,\mathbf{t}(x) + B\,(1 - \mathbf{t}(x)) \tag{1}$$

where,$\mathbf{I}$ denotes the degraded image, $\mathbf{J}$ the original image, $\mathbf{t}$ transmission, and B the background light. The transmission is defined as follows:

$$\mathbf{t}^c(x) = e^{-\beta^c d(x)}, c \in \{R, G, B\} \tag{2}$$

where β_c denotes the attenuation coefficient of light at a specific wavelength c in the underwater environment, and $\mathbf{d}(x)$ represents the scene depth. From this formulation, the concept of ULAD can be readily understood. The Dmip proposed [36] is a widely adopted method for representing ULAD. But it gives up the information of G or B channel. To fully exploit the information contained in each channel and to avoid large depth errors caused by extreme cases in individual channels, we employ the following formulation for ULAD-based depth computation:

$$AD(x) = \sqrt{\left(I_\Omega^r(x) - I_\Omega^g(x)\right)^2 + \left(I_\Omega^r(x) - I_\Omega^b(x)\right)^2} \tag{3}$$

where $AD(x)$ represents the attenuation degree. $I_\Omega^c(x) = \min_{y \in \Omega(x)} I^c(x), c \in \{r, g, b\}$. we use this formulation to compute depth and demonstrate the effectiveness of ULAD-based depth estimation, as illustrated in Fig. 2

Fig. 2 illustrates both the effectiveness and the limitations of ULAD-based depth estimation. In summary, ULAD is an effective mechanism for depth estimationm, as shown in the first row of Fig. 2. However, it still suffers from several limitations, as outlined below:

(1)ULAD is incapable of handling near-range scenes or underwater scenes where color shifts are negligible. $D2$ illustrated in Fig. 2 being significantly larger than $D1$ demonstrates this limitation.

(2)ULAD cannot effectively estimate the depth of special objects such as white or black and vividly colored objects. This limitation is illustrated by the red box in Fig. 2.

To address the aforementioned issues, we propose the Attenuation Information Space (AIS) based on ULAD information, and incorporate it as complementary information to conventional RGB data. The AIS is inspired by the RMI space

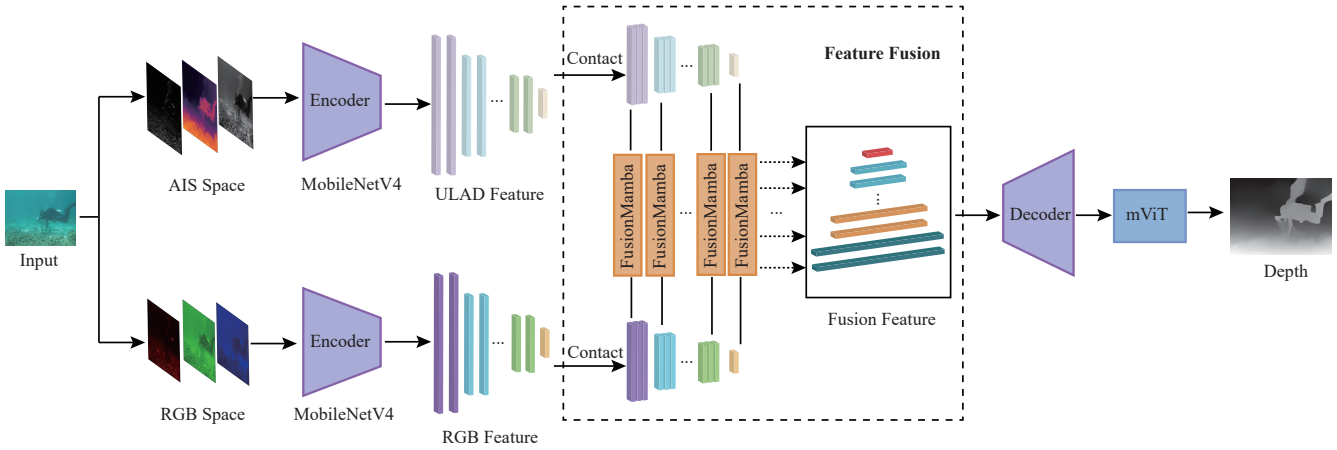


Fig. 3. Diagram of our network architecture. The input RGB images are mapped into both the AIS space and the normal RGB space. Then ULAD and RGB features are then extracted by the MobileNetV4 backbone. These two feature sets are fused using the FusionMamba layers to obtain fusion features. Finally, the fusion features are decoded by our decoder and refined by the Micro Vision Transformer (mViT) to generate the single-channel depth prediction. The corresponding loss functions are described in Section III-C.

[13] and the IMT space [11]. It consists of three components: *Attenuation Section*($AD\,(x)$ in Eq.(3)), *Intensity Section*(the intensity image) and *Gradient Section*(the gradient magnitude). Specifically, the luminance map is obtained from the L channel of the LAB color space. The gradient component is computed using the Sobel operator and the normalized result is used as the Gradient Section. The intensity map is employed to progress the misestimation of white and black objects, while vividly colored objects is still handled by RGB information. The Gradient Section is used to enhance object edges. The features extracted from AIS are fused with those from RGB for more effective depth estimation. CD-Depth [11] employs a confidence matrix consisting only of 0 and 1 to fuse the two depth maps from attenuation and RGB. The approach merely selects the estimate with the smaller error between two independent predictions without exploiting the complementary advantages of RGB and ULAD information. In contrast, we propose a more reasonable approach that adopts a fundamental fusion strategy to integrate both ULAD and RGB information simultaneously, thereby enabling mutual correction and yielding more accurate depth estimation.

B. Framework of Net

1) Architecture description: The pipeline of our network is shown in Fig. 3. We use only RGB images as input. After preprocessing, the images are transformed into both the AIS space and the RGB space and are then fed into their respective encoding layers. The encoder, MobileNetV4, incorporates the latest Universal Inverted Bottleneck (UIB), which includes two optional depthwise convolutions. In addition, combined with the optimized Mobile MQA module and a two-phase neural architecture search (NAS) strategy, MobileNetV4 achieves highly efficient performance, making it well suited for underwater robotic applications. The selected RGB and AIS feature maps are then fused by FusionMamba to generate the integrated representation. Details of the FusionMamba layer are provided in next Section.

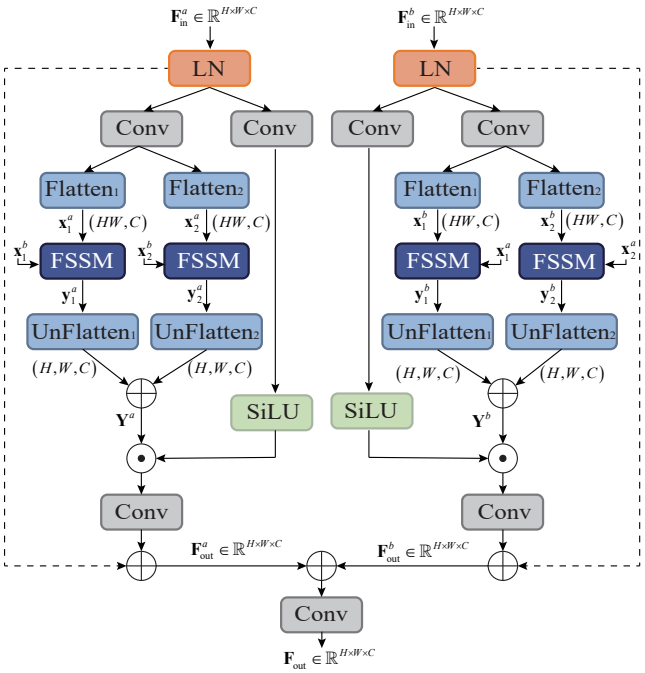


Fig. 4. Diagram of the FusionMamba layer architecture. FSSM denotes the Fusion State Space Model, and SiLU denotes the Sigmoid Linear Unit activation function. $(\cdot)_1$ and $(\cdot)_2$ in the figure represent two directions of four-directional Mamba block [37]. The two inputs correspond to ULAD and RGB features. The final output is the fusion features, which have the same resolution as the input features.

We employ a cascaded multi-layer upsampling module as the decoder, where features of different resolutions are fed into different upsampling layers via skip connections. Inspired by [38], we reformulate underwater monocular depth estimation as a classification problem at the final stage, in which the depth estimation result is expressed as a linear combination of bin centers discretized within the depth

range.

To leverage both local structural features and global information, which are beneficial for depth estimation, the decoder output is further processed by a ViT-based module. A lightweight mViT is adopted to maintain network efficiency. The decoder output is first passed through a 1×1 convolution and then flattened into patch embeddings. Subsequently, a multilayer perceptron (MLP) head with rectified linear unit (ReLU) activation is used to generate the bin centers $f(b)$. Meanwhile, the range attention map is processed through a 1×1 convolution to produce N channels, combining adaptive global information with local pixel-level features extracted from the CNN. This output is then subjected to a softmax operation to obtain p_k. Finally, the estimated depth value at each pixel is computed as the linear combination of p_k and $f(b)$.

2) FusionMamba: The effective fusion of RGB feature information and AIS feature information is critical for depth estimation. Existing feature fusion strategies mainly include conventional CNN-based fusion and transformer-based fusion. While CNNs are computationally efficient, their limited receptive fields constrain their ability to capture global context. Transformers, on the other hand, excel at modeling global dependencies but are computationally expensive [37].

To effectively integrate heterogeneous information, we adopt the fusion state space model (FSSM) to combine different types of features. The detailed architecture of the FusionMamba layer is shown in Fig. 4. For different feature input $\mathbf{F}_{\text{in}}^{a}$ and $\mathbf{F}_{\text{in}}^{b}$, both are processed through identical modules before fusion, and the output dimension remains unchanged. In our fusion structure, we observe that the flattening direction has little impact on the fusion of RGB and AIS features. Therefore, to improve network efficiency, we reduce the number of flattening directions from four to two, corresponding to the first and third directions in [37].

C. Loss Function

Pixel-wise Losses: We employ two classical pixel-wise supervised loss functions to evaluate the discrepancy between the estimated depth and the ground-truth depth. The L2 loss denoted as $\mathcal{L}_{\text{L2}}$ and Scale-Invariant Log (SILog) loss denoted as $\mathcal{L}_{\text{SILog}}$ [**?**].

They are defined as follows:

$$\mathcal{L}_{\text{L2}} = \sum_{x} \left[\left\| \hat{\mathbf{d}}(x) - \mathbf{d}(x) \right\|_2 \right] \tag{4}$$

$$\mathcal{L}_{\text{SILog}} = \alpha \sqrt{\frac{1}{\text{K}} \sum_{x} g(x)^2 - \frac{\lambda}{\text{K}^2} \left(\sum_{x} g(x) \right)^2} \tag{5}$$

where $g(x) = \log \hat{\mathbf{d}}(x) - \log \mathbf{d}(x)$, and K denotes the number of pixels having valid ground truth values. The α and λ are set to 10 and 0.85 following [38].

Attenuation Prior Loss: As shown in Eq. (2), different channels c have distinct attenuation behaviors. By taking the ratio between two channels, we obtain:

$$\frac{\mathbf{t}^{\text{c1}}}{\mathbf{t}^{\text{c2}}} = \frac{\left(\mathbf{I}^{\text{c1}} - B^{\text{c1}}\right) \cdot \left(\mathbf{J}^{\text{c2}} - B^{\text{c2}}\right)}{\left(\mathbf{J}^{\text{c1}} - B^{\text{c1}}\right) \cdot \left(\mathbf{I}^{\text{c2}} - B^{\text{c2}}\right)} \tag{6}$$

After discarding the darkest 0.05% pixels of $\mathbf{I}$, we select the remaining 0.1% darkest pixels as M_{dark}, on which $\mathbf{J}^{ci} \approx 0$. By substituting Eq.2 into Eq.8, we obtain:

$$\mathbf{d}(x)\left(\beta^{\text{c2}} - \beta^{\text{c1}}\right) = \ln\left(\frac{B^{\text{c2}}\left(\mathbf{I}^{\text{c1}}(x) - B^{\text{c1}}\right)}{B^{\text{c1}}\left(\mathbf{I}^{\text{c2}}(x) - B^{\text{c2}}\right)}\right), x \in M_{\text{dark}} \tag{7}$$

According to ULAD, for $c1 \in \{G, B\}$ and $c2 \in \{R\}$, we have $\beta^{c1} - \beta^{c2} < 0$. Hence, Eq.9 can be rewritten as:

$$\boldsymbol{\beta}^{\text{c1,c2}} = \ln\left(\frac{B^{\text{c1}}\left(\mathbf{I}^{\text{c2}} - B^{\text{c2}}\right)}{B^{\text{c2}}\left(\mathbf{I}^{\text{c1}} - B^{\text{c1}}\right)}\right) < 0 \tag{8}$$

where $B^{c} = \frac{1}{\text{N}_2} \sum_{i=0}^{\text{N}_2} \mathbf{I}^{c}(x_i), x_{i \in} M_{\text{far}}$ denotes the farthest 0.1% pixels of $\hat{\mathbf{d}}(x)$, and the number of pixels is denoted as N_2. The bold symbol $\boldsymbol{\beta}^{c}$ represents the attenuation-coefficient loss map. By computing it for the R,G and R,B channels, we obtain $\mathcal{L}_{\text{ULAD}}$, which is defined as follows:

$$\mathcal{L}_{\text{ULAD}} = \sum_{i=0}^{\text{N}_2} f\left(\boldsymbol{\beta}^{\text{R,G}}(x_i)\right) + f\left(\boldsymbol{\beta}^{\text{R,B}}(x_i)\right), x_i \in M_{\text{far}} \tag{9}$$

where $f(\cdot)$ denotes the ReLU function.

Edge-based Loss: To enhance the edge discernibility of small objects and fine details in the depth map, the edge-based loss [**?**] is adopted.

$$\mathcal{L}_{E\text{dge}} = \frac{\sum_{x=1}^{WH} \mathcal{G}(\mathbf{I}(x)) \left[\left\| \hat{\mathbf{d}}(x) - \mathbf{d}(x) \right\|_1 \right]}{WH} \tag{10}$$

where $\mathcal{G}(\cdot)$ denotes the approach for extracting edge information, which includes gradient computation and morphological erosion operations. W and H correspond to the width and height of the image, respectively.

In conclusion, the complete loss function is formulated as follow:

$$\mathcal{L} = \mu_{\text{L2}}\mathcal{L}_{\text{L2}} + \mu_{\text{S}}\mathcal{L}_{\text{SILog}} + \mu_{\text{D}}\mathcal{L}_{\text{ULAD}} + \mu_{\text{E}}\mathcal{L}_{\text{Edge}} \tag{11}$$

where μ_{L2}, μ_{S}, μ_{D} and μ_{E} denote the weighting factors of the corresponding loss terms, which are set to 0.2, 0.6, 0.1 and 0.1 during training, respectively.

IV. Experiment

A. Datasets and Implementation Details

1) Datasets: We conduct training on the USOD10K [24] dataset. Approximately 10,000 underwater RGB images and their corresponding depth maps are contained in the dataset with a resolution of 480 × 640. Additionally, we conduct generalization experiments on the FLSea [25] dataset. To ensure a consistent scale, the image size of this dataset is also resized to 480 * 640 during testing. The Canyons-Flatiron and RedSea-Big dice loop subsets are primarily used for our evaluation.

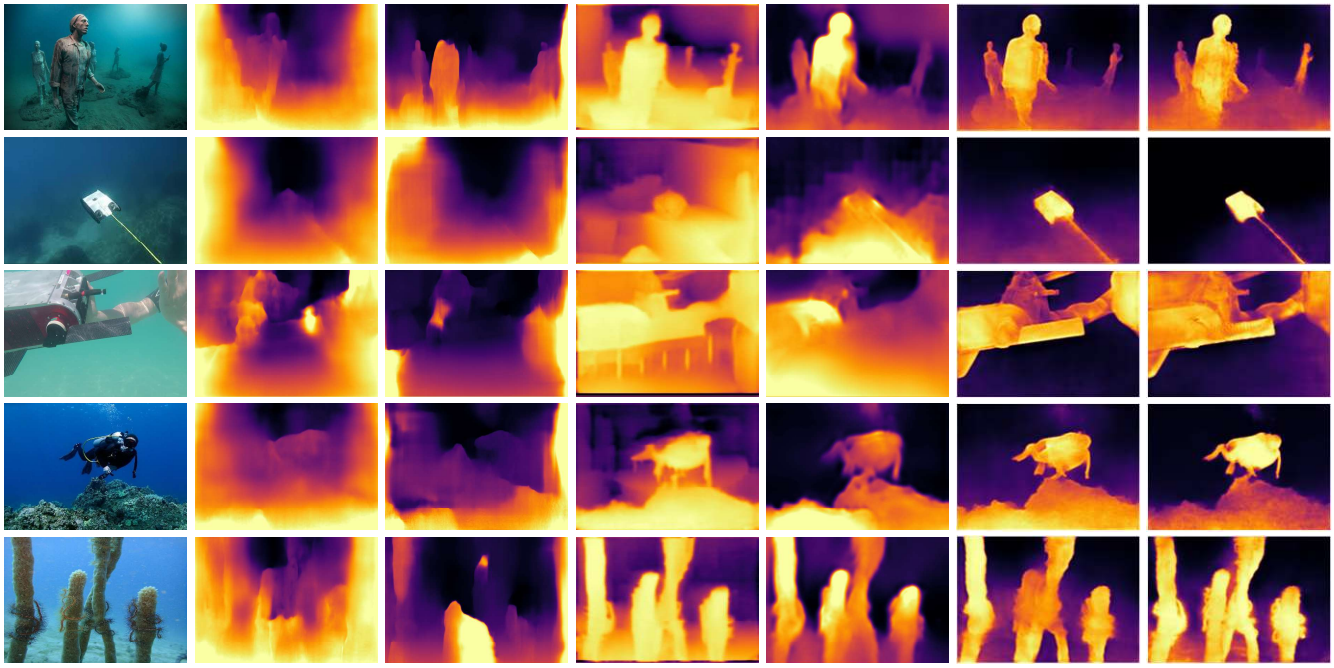


Fig. 5. Qualitative comparison results of underwater depth estimation. The original images are selected from the USOD10K [24] and FLSea [25] datasets. The compared methods include LiteMono [5], MonoDepth2 [23], UW-Net [31], WaterMono [30], UDepth [13], and our method.

TABLE I

QUANTITATIVE RESULTS OF DEPTH ESTIMATION ON THE USOD10K AND FLSEA-VI DATASETS.

Methods	Year	Dataset	Depth Error(↓)				Depth Accuracy(↑)			Param(↓) (M)	Mem(↓) (M)	FPS(↑)
			Abs Rel	Sq Rel	RSME	RSME log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$			
LiteMono	2023	USOD	0.885	0.379	0.285	0.687	0.313	0.517	0.664	3.08	11.70	11.10
		FLSea	0.914	0.302	0.262	0.767	0.300	0.579	0.734			
Monodepth2	2019	USOD	0.857	0.384	0.307	0.730	0.318	0.586	0.760	14.33	54.70	9.10
		FLSea	1.263	0.687	0.323	0.932	0.244	0.494	0.605			
UWNet	2019	USOD	1.142	0.651	0.373	0.866	0.260	0.336	0.674	29.60	112.92	5.80
		FLSea	1.080	0.885	0.475	1.089	0.116	0.354	0.595			
WaterMono	2025	USOD	0.689	0.257	0.246	0.645	0.345	0.566	0.910	**3.07**	**11.70**	12.50
		FLSea	0.913	0.236	0.217	0.692	0.376	0.649	0.886			
UDepth	2022	USOD	0.369	0.326	0.148	0.630	0.550	0.773	0.925	14.26	54.40	32.30
		FLSea	0.945	0.381	0.276	0.893	0.314	0.564	0.793			
Our	2025	USOD	**0.287**	**0.157**	**0.140**	**0.373**	**0.653**	**0.899**	**0.973**	13.74	52.40	**46.50**
		FLSea	**0.635**	**0.253**	**0.215**	**0.670**	**0.494**	**0.744**	**0.908**			

The **best** results are marked in black bold type. The input image resolution is set to 480*640.

2) Implementation Details: In our network, the RGB image is the only input, while the depth map from the dataset serves as the ground truth (GT). The proposed method is implemented using PyTorch and trained on a single NVIDIA GeForce RTX 4090 GPU. The input resolution is set to 480 * 640 with a batch size of 8. Under this setting, training one epoch takes less than 5 minutes. The model is trained for 80 epochs on the USOD10K dataset. The AdamW optimizer is employed with a weight decay of 5e-3. The initial learning rate is set to 1e-4, following a cosine learning rate schedule.

B. Qualitative and Quantitative Evaluation

For baseline performance comparison, we selected five advanced underwater depth estimation networks: LiteMono, MonoDepth2, UW-Net, WaterMono, and UDepth. Most of these models are lightweight. We adopted the recommended settings of these networks in testing and directly used their pretrained models whenever available. For LiteMono and MonoDepth2 which are primarily designed for terrestrial scenes, we did not perform additional training specifically for the underwater setting. Instead, we directly adopted the original recommended configurations. This choice was made to assess, to some extent, the feasibility of transferring

terrestrial monocular depth estimation methods to underwater environments based on their resulting performance, and to further demonstrate the necessity of dedicated underwater depth estimation approaches.

In the quantitative evaluation, we employ standard error metrics and accuracy metrics [**?**]. The error metrics include mean absolute relative error (Abs Rel), squared relative error (Sq Rel), root mean squared error (RMSE), and root mean squared logarithmic error (RMSE log). The accuracy metrics include accuracy under thresholds $\delta < 1.25$, $\delta < 1.25^2$, and $\delta < 1.25^3$. Considering that the GT depth does not cover all pixels in some images, only the pixels with valid depth labels are used for evaluation.

Qualitative comparison results are shown in Fig. 5, and quantitative evaluation results are presented in Table IV-A.2. It can be observed that LiteMono and MonoDepth2 produce relatively poor results, failing to accurately estimate scene depth. This is attributed to the weak generalization ability of their pretrained models in underwater scenarios. UW-Net generates more accurate depth maps, but still contains many misestimated regions. WaterMono demonstrates severe edge information lossing, resulting in blurred object boundaries. Among these comparative methods, UDepth achieves the best performance as an underwater depth estimation network, with fewer parameters and higher real-time efficiency. Compared with these methods, our network achieves superior depth estimation accuracy and real-time performance with fewer parameters. From Table IV-A.2, it is evident that compared with the already lightweight UDepth network, our method further reduces the number of parameters by **10%**. By fusing AIS and RGB information, it also produces more accurate depth estimation results. In real-time testing, with the input resolution fixed at 480 × 640, our network achieves 46.5 FPS on the RTX 4090 platform.

C. Ablation Studies

To verify the effectiveness of fusing AIS and RGB information, we conduct ablation studies. We use the classical networks VGG and MobileNetV2 for experimental validation, since they facilitate modifications in the fusion part of the network structure. The experimental results are shown in Fig. 6, and the quantitative results are presented in Table III and II, corresponding to the USOD10K and FLSea datasets, respectively. The results show that using only the AIS space as input enables depth estimation, but the accuracy is lower than that of using only the RGB space. Its results emphasize edges and details, which is attributed to the inclusion of gradient information in the AIS space. When AIS and RGB are fused, the depth estimation error decreases, and the edges become more accurate and clearer. As shown in Table III and II, the fusion consistently outperforms the results of using either space alone. These ablation studies demonstrate that combining RGB and AIS space information is more effective than relying on a single source of information.

TABLE II

QUANTITATIVE RESULTS OF ABLATION STUDY ON USOD10K

Method	Abs Rel	Sq Rel	RSME	RSME log
our	**0.287**	**0.057**	**0.140**	**0.373**
our_RGB	0.500	0.149	0.199	0.520
our_RMI	0.613	0.208	0.230	0.570
VGG	**0.337**	**0.069**	**0.167**	**0.487**
VGG_RGB	0.386	0.090	0.170	0.496
VGG_RMI	0.405	0.098	0.176	0.497
MobileNetV2	**0.424**	**0.108**	**0.178**	**0.510**
MobileNetV2_RGB	0.443	0.122	0.181	0.568
MobileNetV2_RMI	0.452	0.123	0.187	0.594

TABLE III

QUANTITATIVE RESULTS OF ABLATION STUDY ON FLSEA-VI

Method	Abs Rel	Sq Rel	RSME	RSME log
our	**1.230**	**0.520**	**0.310**	**0.860**
our_RGB	1.340	0.630	0.380	1.080
our_RMI	1.530	0.770	0.400	1.080
VGG	**1.290**	**0.570**	**0.370**	**1.070**
VGG_RGB	1.330	0.610	0.390	1.170
VGG_RMI	1.400	0.650	0.380	1.090
MobileNetV2	**1.440**	**0.700**	**0.390**	**1.090**
MobileNetV2_RGB	1.460	0.730	0.400	1.090
MobileNetV2_RMI	1.510	0.750	0.400	1.120

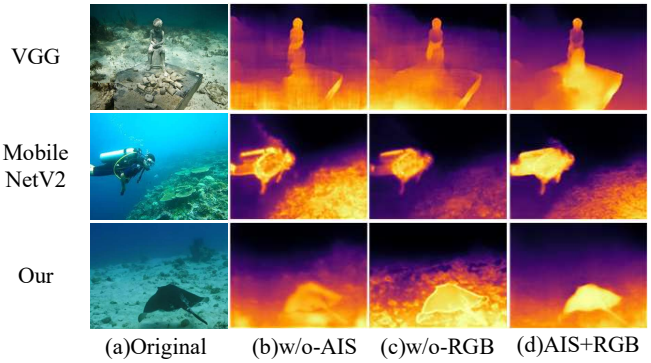


Fig. 6. Ablation studies of AIS and RGB spaces. The first, second, and third rows show the results of VGG, MobileNetV2, and our network, respectively. The cases w/o-AIS and w/o-RGB are obtained by modifying network through removing the feature fusion module and the corresponding encoder layer

V. CONCLUSIONS

In this paper, we analyze the effectiveness and limitations of ULAD in underwater depth estimation. Based on this analysis, we propose a monocular depth estimation method that incorporates ULAD as supplementary information. By fully exploiting RGB image information while leveraging ULAD to bridge the gap between in-air and underwater environments, our network achieves more accurate depth estimation. Extensive comparative experiments and ablation studies conducted on the USOD10K and FLSea datasets demonstrate that our method achieves superior depth estimation accuracy compared with existing approaches. The ablation studies further confirm the effectiveness of ULAD as supplementary information. In real-time testing, our method outperforms current state-of-the-art lightweight depth estimation networks, reducing parameters by 10% and improving frame rates by 43%. In future work, we plan to further apply ULAD to underwater visual enhancement, where the esti-

mated depth and enhanced images can be jointly optimized in an iterative manner.

REFERENCES

[1] G. Zhou, C. Li, D. Zhang, D. Liu, X. Zhou, and J. Zhan, "Overview of underwater transmission characteristics of oceanic lidar," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8144–8159, 2021.

[2] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. S. Kweon, "High-quality depth map upsampling and completion for rgb-d cameras," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5559–5572, 2014.

[3] J. Zhang, F. Han, D. Han, J. Yang, W. Zhao, and H. Li, "Integration of sonar and visual–inertial systems for slam in underwater environments," *IEEE Sensors Journal*, vol. 24, no. 10, pp. 16 792–16 804, 2024.

[4] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[5] 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.

[6] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21 875–21 911, 2024.

[7] J. Yang, M. Gong, and Y. Pu, "Physics-informed knowledge transfer for underwater monocular depth estimation," in *European Conference on Computer Vision*. Springer, 2024, pp. 449–465.

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

[9] 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 *Pacific rim conference on multimedia*. Springer, 2018, pp. 678–688.

[10] J. Zhou, Q. Liu, Q. Jiang, W. Ren, K.-M. Lam, and W. Zhang, "Underwater camera: Improving visual perception via adaptive dark pixel prior and color correction," *International Journal of Computer Vision*, pp. 1–19, 2023.

[11] J. Guo, J. Ma, F. Sun, Z. Gao, Á. F. García-Fernández, H.-N. Liang, X. Zhu, and W. Ding, "Cd-udepth: Complementary dual-source information fusion for underwater monocular depth estimation," *Information Fusion*, vol. 118, p. 102961, 2025.

[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*, vol. 62, pp. 1–16, 2024.

[13] B. Yu, J. Wu, and M. J. Islam, "Udepth: Fast monocular depth estimation for visually-guided underwater robots," *arXiv preprint arXiv:2209.12358*, 2022.

[14] M. A. B. Siddique, J. Wu, I. Rekleitis, and M. J. Islam, "Aquafuse: Waterbody fusion for physics-guided view synthesis of underwater scenes," *IEEE Robotics and Automation Letters*, 2025.

[15] L. Ebner, G. Billings, and S. Williams, "Metrically scaled monocular depth estimation through sparse priors for underwater robots," in *2024 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2024, pp. 3751–3757.

[16] X. Ye, Y. Chang, R. Xu, and H. Li, "Uw-adapter: Adapting monocular depth estimation model in underwater scenes," *IEEE Transactions on Multimedia*, 2025.

[17] D. Akkaynak, T. Treibitz, T. Shlesinger, Y. Loya, R. Tamir, and D. Iluz, "What is the space of attenuation coefficients in underwater computer vision?" in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4931–4940.

[18] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.

[19] Y.-Z. Hsieh and M.-C. Chang, "Underwater image enhancement and attenuation restoration based on depth and backscatter estimation," *IEEE Transactions on Computational Imaging*, 2025.

[20] M. Yang, A. Sowmya, Z. Wei, and B. Zheng, "Offshore underwater image restoration using reflection-decomposition-based transmission map estimation," *IEEE Journal of Oceanic Engineering*, vol. 45, no. 2, pp. 521–533, 2019.

[21] M. Roznere and A. Q. Li, "Real-time model-based image color correction for underwater robots," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 7191–7196.

[22] S. Amitai, I. Klein, and T. Treibitz, "Self-supervised monocular depth underwater," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1098–1104.

[23] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[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*, vol. 34, pp. 1602–1615, 2023.

[25] Y. Randall, "Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets," Master's thesis, University of Haifa (Israel), 2023.

[26] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 8, pp. 1362–1376, 2009.

[27] Z. Huang, X. Wang, C. Xu, J. Li, and L. Feng, "Underwater variable zoom: Depth-guided perception network for underwater image enhancement," *Expert Systems with Applications*, vol. 259, p. 125350, 2025.

[28] G. Yang, G. Kang, J. Lee, and Y. Cho, "Joint-id: Transformer-based joint image enhancement and depth estimation for underwater environments," *IEEE Sensors Journal*, vol. 24, no. 3, pp. 3113–3122, 2023.

[29] J. Baek, G. Kim, and S. Kim, "Semi-supervised learning with mutual distillation for monocular depth estimation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4562–4569.

[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] H. Gupta and K. Mitra, "Unsupervised single image underwater depth estimation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 624–628.

[32] Y. Huang, F. Yuan, F. Xiao, J. Lu, and E. Cheng, "Underwater image enhancement based on zero-reference deep network," *IEEE Journal of Oceanic Engineering*, vol. 48, no. 3, pp. 903–924, 2023.

[33] F. Zhang, S. You, Y. Li, and Y. Fu, "Atlantis: Enabling underwater depth estimation with stable diffusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 11 852–11 861.

[34] S. Q. Duntley, "Light in the sea," *Journal of the optical society of America*, vol. 53, no. 2, pp. 214–233, 1963.

[35] W. Gu and L. Qi, "Dense geometry supervision for underwater depth estimation," in *Advanced Fiber Laser Conference (AFL 2024)*, vol. 13544. SPIE, 2025, pp. 25–32.

[36] 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.

[37] S. Peng, X. Zhu, H. Deng, L.-J. Deng, and Z. Lei, "Fusionmamba: Efficient remote sensing image fusion with state space model," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[38] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4009–4018.