Dive into Aerial Remote Sensing Underwater Depth Estimation with Hyperspectral Imagery

Jiahao Qi, Xingyue Liu, Chen Chen, Dehui Zhu, Kangcheng Bin*, Ping Zhong*

National Key Laboratory of Science and Technology on Automatic Target Recognition
National University of Defense Technology, China
{qijiahao1996, liuxingyue18, chenchen21c, dhzhu95, binkc21, zhongping}@nudt.edu.cn

Abstract

Visible spectrum images capture limited information from just three discrete bands, often resulting in suboptimal performance in underwater depth estimation (UDE) due to significant information loss from water absorption. In contrast, hyperspectral imagery, which include hundreds of continuous bands, provide abundant spectral information that offers greater resilience against the adverse effects of water absorption. In this paper, we conduct a comprehensive study to investigate how spectral information can enhance aerial remote sensing UDE through two key aspects: the benchmark dataset and general framework. For the benchmark dataset, we construct a real-world hyperspectral UDE (HUDE) dataset **ATR-HUDE**, comprising approximately 500 synchronized hyperspectral and LiDAR data pairs collected from diverse coastal scenes and flight altitudes. Regarding the general framework, we integrate recent physical imaging models and advances in state space models to design a novel HUDE framework named **HUDEMamba** that estimates underwater depth using both model-driven and data-driven approaches. Experimental results on the constructed benchmark dataset validate the potential of HUDE and the effectiveness of HUDEMamba.

Datasets — https://github.com/qjh1996/ATR-HUDE

Introduction

Aerial remote sensing underwater depth estimation (UDE) focuses on analyzing the depth distribution of seafloor rather than the distance to underwater objects. Recent research (Li et al. 2024; Ye et al. 2023; Yu, Wu, and Islam 2023) employed visible spectrum images as the primary data source for UDE. However, as shown in Figure 1(b), two regions with significantly different depths exhibit identical texture information, indicating that **visible spectrum images struggle to provide discriminative information for remote sensing UDE**. Hyperspectral imagery (HSI), capturing and processing a wide spectrum of light across numerous contiguous spectral bands (Li et al. 2019), is regraded as an alternative to remote sensing UDE. Figure 1(c) demonstrates that the spectral information in HSI can differentiate underwater regions with varying depths, even their visual appearance are nearly identical.

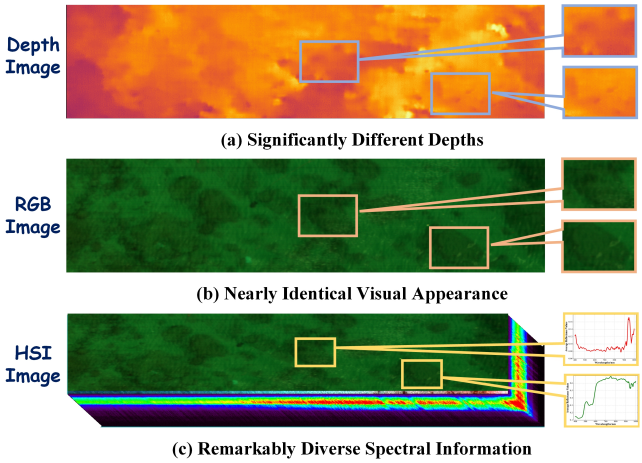


Figure 1: Limitation of visible spectrum imagery and advantage of hyperspectral imagery in underwater depth estimation. (a) Two regions in the depth image with significantly different depths are annotated with blue boxes. (b) The corresponding regions in the RGB image, annotated with pink boxes, have nearly identical visual appearances. (c) The corresponding regions in the HSI image, annotated with yellow boxes, show remarkably diverse spectral information.

Despite this charming characteristic, hyperspectral underwater depth estimation (HUDE) is still under-explored for two main reasons. **(1) Well-Established Dataset Scarcity:** The existing HUDE datasets are limited in scale and often not publicly available (Liu et al. 2020). Additionally, the inconsistency between the data collection platforms of HSI and depth maps can result in image misalignment, thereby undermining the accuracy of depth groundtruth. **(2) Top-Tier Methodology Deficiency:** Most existing methods (Ma et al. 2014; Pan et al. 2015; Alevizos 2020) are specially designed for visible spectrum images and do not consider the unique characteristics of HSI, leading to inferior HUDE performance. Furthermore, they tend to underutilize the spatial correlation among adjacent pixels, which is crucial when visual appearance features are not sufficiently discriminative.

To this end, we establish a new HUDE dataset and propose a novel HUDE framework in this paper. The established dataset, named **ATR-HUDE dataset**, comprises a to-

*Corresponding author.

tal of 500 HSIs that are collocated with unmanned aerial vehicle (UAV) at three distinct coastal scenes. To provide accurate depth groundtruth, we measured the underwater depth information during low tide using a UAV-borne LiDAR sensor to keep the consistency of data collection platform. The LiDAR and HSIs are well-aligned using a commonly employed method (Lee et al. 2015). The proposed framework, termed **HUDEMamba**, consists of a physics-inspired image translator (PIT) and a context-aware depth estimator (CDE). Specifically, the PIT leverages the inherent characteristics of HSI and heuristic rules derived from underwater imaging models to transform the input image into a depth-related image for task-agnostic knowledge distillation. Notably, the PIT is a parameter-free module that greatly improves the computational efficiency. Based on the depth-related image, the CDE designs a hybrid-level state space model (SSM) in combination with a multi-scale bin predictor to yield accurate HUDE results. The hybrid-level SSM adopts a customized multi-level scanning strategy to fully exploit the spatial correlation among adjacent pixels, facilitating efficient spatial-spectral representation learning.

Related Work

Underwater depth estimation

Traditional UDE methods commonly estimate underwater depth based on hand-crafted priors and physical image formation model (PIFM). Peng *et al.* investigated the relationship between depth and blurriness based on the PIFM to estimate underwater depth (Peng, Zhao, and Cosman 2015). Zhang *et al.* employed statistical underwater dark channel priors to perform non-uniform illumination correction for UDE (Zhang et al. 2017). Furthermore, Song *et al.* proposed an underwater light-attenuation prior to estimate underwater depth (Song et al. 2018). However, these methods cannot be applied to HUDE since the utilized priors are derived based on the characteristics of visible spectrum images. Nevertheless, their impressive performances confirm that *combining the PIFM and image properties is a valid solution to obtain depth-related information without any model parameters*. Inspired by this observation, we exploit the spectral analysis theory of HSI and PIFM to establish a physics-inspired image translator for task-agnostic knowledge distillation.

Learning-based UDE methods harness the power of deep neural networks for UDE. Some of these methods (Gupta and Mitra 2019; Ye et al. 2020) model the UDE problem as a linear regression task. Gupta *et al.* established a dense-block based on auto-encoders as the regressor for UDE (Gupta and Mitra 2019). Ye *et al.* developed a U-Net-like depth regression network trained in an adversarial learning manner (Ye et al. 2020). Another line of work formulates UDE as a per-pixel classification task. Yu *et al.* proposed a Transformer-based classifier to transform global image information into UDE results (Yu, Wu, and Islam 2023). Yang *et al.* combined a multilayer perceptron (MLP) head with a fully convolutional network as the depth classifier (Yang et al. 2024). Compared with the traditional methods, the learning-based ones are more competitive due to their superior feature extraction capabilities. However, extending them to HUDE remains challenging due to their inability to fully exploit the spatial correlation among adjacent pixels. For this, we tailor the SSM for UDE to leverage global and local spatial information for HUDE.

Hyperspectral underwater depth estimation

HUDE has not been thoroughly explored until recently (Ma et al. 2014; Pan et al. 2015; Alevizos 2020). Ma *et al.* proposed a band ratio method with minimal tuning parameters for HUDE, which only utilizes a portion of the spectral information (Ma et al. 2014). To improve accuracy, Pan *et al.* introduced support vector regression as an alternative method by using full-band HSI (Pan et al. 2015). Building on this, Alevizos integrated the PIFM with support vector regression to enhance both the accuracy and interpretability of depth estimation results (Alevizos 2020). Considering the limitations of support vector regression in feature extraction, Peng *et al.* proposed a channel-wise spectral attention-based HUDE network to exploit the spatial-spectral information of HSI (Peng et al. 2023). Although significant efforts have been made in HUDE, combining the strengths of PIFM and deep learning methods for HUDE remains largely unexplored. To achieve more accurate results, it is essential to integrate recent advancements in deep learning with insights from PIFM, aiming to develop a more flexible, precise, and generalized HUDE framework. In this paper, we target a novel HUDE framework that incorporates both data-driven and model-driven approaches to leverage the intrinsic structure of HSI and harness the representative capabilities of modern deep neural models simultaneously.

Proposed Dataset

UAV-borne Hyperspectral-LiDAR System

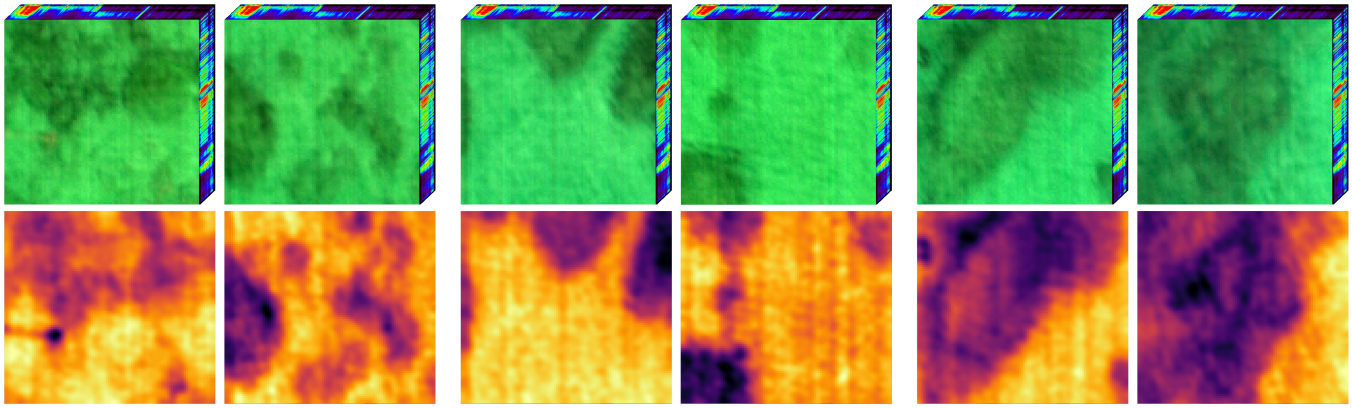
As shown in Figure 2(a), the dataset is acquired using a UAV-borne hyperspectral-LiDAR system, comprising the X20P-LIR sensor and a DJI M300 RTK UAV platform. The X20P-LIR, depicted in Figure 2(b), integrates LiDAR, infrared, and hyperspectral imaging technologies. It provides simultaneous LiDAR and hyperspectral imaging over a 350-1000 nm range, with the LiDAR component offering measurements up to 450 meters at 80% reflectivity. The DJI M300 RTK UAV platform, demonstrated in Figure 2(c), supports a 2.7 kg payload, up to 55 minutes of flight time, and centimeter-level accuracy via RTK/IMU modules.

Data Collection

To acquire precise depth groundtruth, data collection was conducted in two phases. During the first phase at high tide, the hyperspectral sensor captured submerged seafloor images while the LiDAR sensor acquired surface elevation data. In the second phase at low tide, the LiDAR sensor captured depth information of the same seafloor scene without water coverage. Data collection was conducted across diverse coastal scenes in Sanya, China to ensure scene diversity and assess the cross-scene generalization capability. Besides, we also collect data at varying flight altitudes to investigate the impact of image scale on HUDE performance.

(a) Data Collection System (b) X20P-LIR Sensor (c) M300 UAV Platform

Figure 2: Illustration of UAV-borne Hyperspectral-LiDAR System.



(a) Scene1 (b) Scene2 (c) Scene3

Figure 3: Representative samples of the proposed dataset in different scenes.

ATR-HUDE Dataset

Data Processing and Groundtruth Generation. Hyperspectral data preprocessing involved radiometric calibration and geometric correction, whereas LiDAR preprocessing included position calculation and point cloud aggregation. All processing tasks were conducted using software provided by instrument manufacturer. To generate accurate groundtruth, a standard image registration method (Lee et al. 2015) is employed to align the hyperspectral and LiDAR data. Following this registration, the LiDAR data serves as the depth ground truth. Figure 3 illustrates representative samples.

Dataset Configuration. We segmented the processed images into non-overlapping 100×100 blocks, yielding 500 image pairs. We utilized 400 pairs for training and 100 pairs for testing. Image pairs for different tasks were proportionally sampled from diverse scenes and altitudes to minimize data distribution bias.

Method

Motivations

The pipeline of existing UDE methods consists of an image translation module and a depth estimation module (Yu, Wu, and Islam 2023). The former module extracts information highly related to underwater depth and the later one perform UDE based on the depth-related information. It is straightforward to draw on the wisdom of existing UDE methods to develop the HUDE framework. However, there remains significant challenges that need to be addressed: **(1) Heavy computational burden and poor generalization performance**. The image translation module in the existing UDE methods typically employs dense-block structures (such as U-Net) or Transformer-based networks that contain numerous learnable parameters. Unfortunately, the high dimensionality and limited amount of HSIs result in significant computational burdens and severe overfitting issues (Molinier and Kilpi 2019). **(2) Overlook the spatial correlation among adjacent pixels**. The depth estimation modules in the existing UDE methods often assume that each pixel in an image is independent when predicting depth (Bhat, Alhashim, and Wonka 2020). However, this assumption is invalid since pixels corresponding to the same object and located in a compact neighborhood should have similar depth values. Consequently, this pixel-independence assumption leads to the neglect of spatial correlations among adjacent pixels, resulting in inaccurate estimation results.

To address these challenges, our motivations are depicted as follows. Firstly, since the PIFM describes the underwater imaging process that are generalized to any underwater scene, its principles can be exploited to enhance generalization capacity (Wang et al. 2024). Additionally, hyperspectral unmixing (HU) methodology can extract specific information from high-dimensional HSI without any model parame-

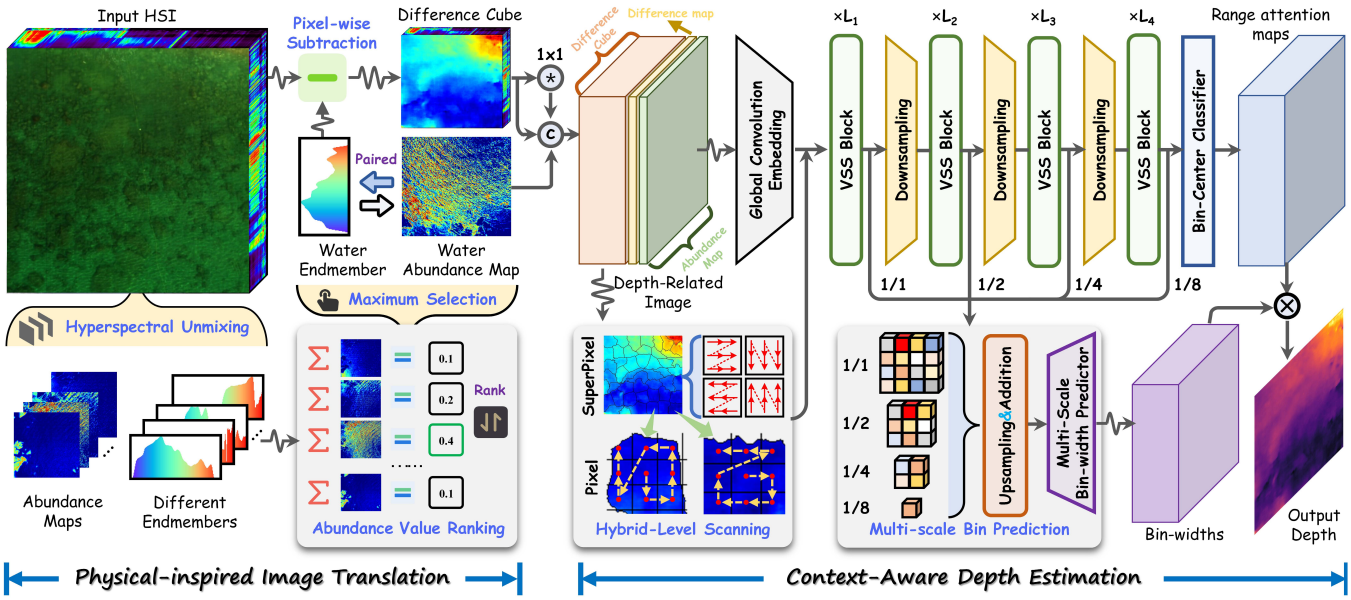


Figure 4: Overall pipeline of HUDEMamba.

ters (Rasti et al. 2024). Motivated by these insights, we leverage the principles derived from the PIFM as priors to guide a parameter-free HU method, conducting a model-driven and parameter-free image translator. Secondly, recent advances of SSM in different computer vision tasks (Liu, Zhang, and Zhang 2024) prove that SSM specializes in spatial correlation modeling. Consequently, we devote to tailoring SSM for HUDE to leverage the spatial correlations among adjacent pixels. Given that superpixel segmentation algorithms can provide pixel adjacency information, we design a superpixel segmentation-guided multi-level scanning method to maximize the effectiveness of SSM for HUDE.

Method Overview. Based on the above motivations, we propose a simple yet effective HUDE framework named **HUDEMamba**, as illustrated in Figure 4. The HUDE process proceeds in two main stages: (a) physics-inspired image translation and (b) context-aware depth estimation. In the following subsections, we will introduce the key modules of HUDEMamba in detail.

Physics-inspired Image Translator

Some research (Yu, Wu, and Islam 2023; Ding et al. 2024) indicate that UDE with depth-related images, rather than the original ones, is more accurate. Therefore, we propose a model-driven and parameter-free image translator to transform the HSI into depth-related images.

a. Physical Image Formation Model. Given an input HSI $\boldsymbol{I}$, the commonly used PIFM (Lee et al. 1998) can be formulated as:

$$\begin{aligned}\boldsymbol{I}(x,y) &= \boldsymbol{I}_w(x,y)\cdot(\mathbf{1}-e^{-k\cdot\boldsymbol{d}(x,y)})+\boldsymbol{I}_b(x,y)\cdot e^{-k\cdot\boldsymbol{d}(x,y)}\\ &= \underbrace{\boldsymbol{I}_w}_{\text{Depth-agnostic term}} + \underbrace{(\boldsymbol{I}_b(x,y)-\boldsymbol{I}_w)\cdot e^{-k\cdot\boldsymbol{d}(x,y)}}_{\text{Depth-related term}}, \quad (1)\end{aligned}$$

where $\boldsymbol{I}$, $\boldsymbol{I}_w$, and $\boldsymbol{I}_b$ are the spectra of the input HSI, water body, and bottom, respectively. Due to minimal variation in water composition within remote sensing underwater scenes, $\boldsymbol{I}_w(x,y)$ can be assumed to be spatially invariant and simplified to $\boldsymbol{I}_w$. k refers to the water attenuation constant, and $\boldsymbol{d}$ is the depth map. According to Equation (1), we can draw the **following principles**:

- $\boldsymbol{I}$ is a linear combination of $\boldsymbol{I}_w$ and $\boldsymbol{I}_b$, with coefficient maps that are functions of $\boldsymbol{d}$. (**Principle 1**)
- $\boldsymbol{I}_w$ is the only depth-agnostic term in $\boldsymbol{I}$. (**Principle 2**)

Based on these principles, the coefficient map $e^{-k\cdot\boldsymbol{d}(x,y)}$ and spectral distance image $\boldsymbol{I}-\boldsymbol{I}_w=(\boldsymbol{I}_b(x,y)-\boldsymbol{I}_w)\cdot e^{-k\cdot\boldsymbol{d}(x,y)}$ are depth-related images. The key to translate $\boldsymbol{I}$ into a depth-related image is driving $\boldsymbol{I}_w$ and $e^{-k\cdot\boldsymbol{d}(x,y)}$.

b. HSI Unmixing and Abundance Ranking. HU can be exploited to decompose $\boldsymbol{I}$ as

$$\begin{aligned}&\{(\boldsymbol{E}_i(x,y),\boldsymbol{A}_i(x,y))\}_i^N=\boldsymbol{HU}(\boldsymbol{I}(x,y)),\\ &s.t.\sum_i^N\boldsymbol{E}_i(x,y)\cdot\boldsymbol{A}_i(x,y)=\boldsymbol{I}(x,y),\end{aligned} \quad (2)$$

where $(\boldsymbol{E}_i(x,y),\boldsymbol{A}_i(x,y))$ denotes the i-th endmember and its corresponding abundance map. N is the hyperparameter that determines the total number of water and bottom endmembers and $\boldsymbol{HU}(\cdot)$ is the HU operation.

$\boldsymbol{A}_i(x,y)$ quantitatively represents the fractional contribution of $\boldsymbol{E}_i(x,y)$ to the spectral signature of $\boldsymbol{I}(x,y)$. Then, abundance value $\boldsymbol{a}_i=\sum_x\sum_y\boldsymbol{A}_i(x,y)$ represents the total contribution of $\boldsymbol{E}_i$ across the entire HSI. Based on the physical meaning of $\boldsymbol{A}_i(x,y)$, Equation (1), and Equation (2), it can be inferred that $\boldsymbol{A}_i(x,y)$ is equal to $\mathbf{1}-e^{-k\cdot\boldsymbol{d}(x,y)}$ if $\boldsymbol{E}_i$ represents $\boldsymbol{I}_w$. Due to the pervasive nature of the water body, the endmember $\boldsymbol{E}$ in $\{\boldsymbol{E}_i(x,y)\}_i^N$ of maximum $\boldsymbol{a}$ refers to

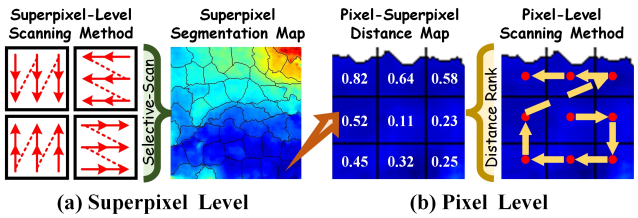


Figure 5: Illustration of superpixel-pixel selective scanning.

$\boldsymbol{I}_w$. To identify $\boldsymbol{I}_w$, the endmember $\boldsymbol{E}_{\max}$ with the highest abundance value is derived using the following strategy:

$$\begin{aligned} \{\boldsymbol{E}_{\max}, \boldsymbol{A}_{\max}\} &= \arg\max_i \boldsymbol{a}_i, \\ \boldsymbol{I}_w = \boldsymbol{E}_{\max}, \;\; e^{-k\cdot\boldsymbol{d}(x,y)} &= \mathbf{1} - \boldsymbol{A}_{\max}. \end{aligned} \tag{3}$$

c. Depth-related Image Translation. After deriving $\boldsymbol{I}_w$ and $e^{-k\cdot\boldsymbol{d}(x,y)}$ using Equation (3), we can translate $\boldsymbol{I}$ into the depth-related image $\boldsymbol{I}_D$ as:

$$\boldsymbol{I}_D = \textbf{Concat}(\boldsymbol{I} - \boldsymbol{E}_{\max}, \textbf{Conv}_{1\times1}(\boldsymbol{I} - \boldsymbol{E}_{\max}), -\ln(\mathbf{1} - \boldsymbol{A}_{\max})) \tag{4}$$

where $\textbf{Concat}(\cdot)$ is the concatenation operation along the spectral dimension, and $\textbf{Conv}_{1\times1}(\cdot)$ represents a 1×1 convolution operation that output dimension is 1.

Context-aware Depth Estimator

As stated in Section 4.1, the spatial correlation among adjacent pixels is critical for HUDE. Consequently, we adapt Mamba (Gu and Dao 2023), an improved variant of SSM, to develop a context-aware depth estimator (CDE).

a. Superpixel-pixel Selective Scanning for HUDE. The sequential nature of the scanning operation in Mamba is not well-suited for non-sequential vision data. Thus, recent efforts proposed several simplistic patch-level scanning methods for input processing (Liu et al. 2024; Zhu et al. 2024). In contrast, we introduce a fine-grained, multi-level scanning approach to adapt Mamba for HUDE, as illustrated in Figure 5. The core idea behind our method is the hierarchical modeling of spatial correlations in a coarse-to-fine manner.

Firstly, the depth-related image $\boldsymbol{I}_D$ is divided into several superpixel blocks using SLIC (Barbato et al. 2022). Then, the ruler of the proposed scan method is described as:

- **Superpixel-Level (coarse)**: Considering the similarity between superpixel blocks and image patches, we apply the SS2D (Liu et al. 2024) to determine the superpixel traversal path.
- **Pixel-Level (fine)**: Within each superpixel block, we calculate the spectral angle distance between each pixel and the superpixel center. The pixel traversal path is then determined by prioritizing pixels with smaller distances.

Finally, position embedding $\boldsymbol{P}$ is achieved by above method.

b. Context-aware Representation Learning. Figure 4 provides an overview of the developed CDE. The input $\boldsymbol{I}_D$ is initially processed by a **global convolution embedding** (GCE) module, which employs 3×3 convolutional kernels with padding to maintain spatial dimensions while reducing dimensionality. The 3×3 convolutional operation in GCE can help to extract the primary spatial-spectral information. Then, the position embedding $\boldsymbol{P}$ transforms the output of GCE $\boldsymbol{I}_G$ into a pixel sequence $\boldsymbol{T}_0$. Finally, several context-aware representation learning stages are employed to yield hierarchical context-aware representations with resolutions $1/1$, $1/2$, $1/4$, $1/8$ based on $\boldsymbol{T}_0$. Each context-aware representation learning stage includes a down-sampling layer (except the first one), followed by a series of Mamba Blocks. The process of the proposed learning strategy is defined as:

$$\boldsymbol{T}_0 = \left[\boldsymbol{C}; \boldsymbol{I}_G^1\boldsymbol{W}; \boldsymbol{I}_G^2\boldsymbol{W}; \cdots; \boldsymbol{I}_G^{\mathrm{M}}\boldsymbol{W}\right] + \boldsymbol{P}. \tag{5}$$

$$\begin{aligned} \boldsymbol{T}_l &= \textbf{MambaBlocks}(\hat{\boldsymbol{T}}_{l-1}) + \hat{\boldsymbol{T}}_{l-1}, \\ \text{where} \quad \hat{\boldsymbol{T}}_{l-1} &= \begin{cases} \boldsymbol{T}_{l-1}, & \text{if } l = 1 \\ \textbf{downsample}(\boldsymbol{T}_{l-1}), & \text{if } l > 1 \end{cases}, \end{aligned} \tag{6}$$

where $\boldsymbol{I}_G^i$ is the i-th pixel in $\boldsymbol{I}_G$, M is the total number of pixels in $\boldsymbol{I}_G$, $\boldsymbol{C}$ is the class token, $\boldsymbol{W}$ is the learnable projection matrix, $\textbf{MambaBlocks}(\cdot)$ are the Mamba blocks, $\textbf{downsample}(\cdot)$ is the downsampling operation, and $\boldsymbol{T}_l$ is the l-th ($l \geq 1$) context-aware representation.

c. Underwater Depth Estimation. As stated in (Bhat, Alhashim, and Wonka 2020), the depth estimation task can be divided into *range-attention map prediction* and *bin-width estimation*. Specifically, bin-width estimation result $\boldsymbol{b}$ is derived from a multi-scale context-aware representation $\boldsymbol{F}_{\textbf{bin}}$:

$$\begin{aligned} \boldsymbol{F}_{\textbf{bin}} &= \boldsymbol{U}\left(\boldsymbol{U}\left(\boldsymbol{U}\left(\boldsymbol{T}_1[1:]\right) + \boldsymbol{T}_2[1:]\right) + \boldsymbol{T}_3[1:]\right) + \boldsymbol{T}_4[1:], \\ \hat{\boldsymbol{b}} &= \textbf{MLP}_{\text{ReLU}}(\boldsymbol{F}_{\textbf{bin}}) = [\hat{\boldsymbol{b}}_1, \hat{\boldsymbol{b}}_2 \cdots, \hat{\boldsymbol{b}}_M], \\ \boldsymbol{b} &= [\textbf{Norm}(\hat{\boldsymbol{b}}_1), \textbf{Norm}(\hat{\boldsymbol{b}}_2), \cdots, \textbf{Norm}(\hat{\boldsymbol{b}}_M)], \\ \textbf{Norm}(\hat{\boldsymbol{b}}_i) &= [\frac{\hat{\boldsymbol{b}}_i^1 + \epsilon}{\sum_{j=1}^{L}\left(\hat{\boldsymbol{b}}_i^j + \epsilon\right)}, \cdots, \frac{\hat{\boldsymbol{b}}_i^L + \epsilon}{\sum_{j=1}^{L}\left(\hat{\boldsymbol{b}}_i^j + \epsilon\right)}]. \end{aligned} \tag{7}$$

where $\boldsymbol{U}(\cdot)$ is the unsampling operation, $\textbf{MLP}_{\text{ReLU}}$ is the MLP head with ReLU activation outputting L-dimensional vector $\hat{\boldsymbol{b}}_i$ (L is also the number of bins), and $\epsilon = 10^{-3}$ ensures each bin-width is strictly positive. Besides, range-attention map prediction result $\boldsymbol{R}$ is derived with highest-level context-aware representation $\boldsymbol{T}_4$:

$$\begin{aligned} \boldsymbol{R} &= \textbf{MLP}_{\text{Softmax}}([\boldsymbol{T}_4[0] * \boldsymbol{T}_4[1], \cdots, \boldsymbol{T}_4[0] * \boldsymbol{T}_4[M]]) \\ &= [\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots, \boldsymbol{R}_M], \end{aligned} \tag{8}$$

where $\textbf{MLP}_{\text{Softmax}}$ is the MLP head with Softmax activation outputting L-dimensional vector $\boldsymbol{R}_i$ and $*$ is the element-wise product. Finally, the HUDE result $\hat{d}_i$ is a linear combination of $\boldsymbol{R}$ and bin-width centers $c(\boldsymbol{b}_i^j)$:

$$\begin{aligned} \hat{d}_i &= \sum_{j=1}^{L} c(\boldsymbol{b}_i^j) \cdot \boldsymbol{R}_i^j, \;\; \{i = 1, 2, \cdots, M\}, \\ c(\boldsymbol{b}_i^j) &= d_{\min} + (d_{\max} - d_{\min})\left(\boldsymbol{b}_i^j/2 + \sum_{j=1}^{L-1}\boldsymbol{b}_i^j\right). \end{aligned} \tag{9}$$

where $[d_{\min}, d_{\max}]$ is the full depth range.

d. Learning objective. We employ the linear combination of two pixel-wise supervised loss functions (Mean Squared

Method	Depth Error (lower, better)					Depth Accuracy (higher, better)		
	$\mathbf{SI}_{log}$	**A.Rel**	**S.Rel**	**RMSE**	$\mathbf{RMSE}_{log}$	$\boldsymbol{\delta}_1$	$\boldsymbol{\delta}_2$	$\boldsymbol{\delta}_3$
UWNet	1.935	0.185	0.095	0.384	0.216	0.694	0.951	0.997
UDepth	0.898	0.117	0.030	0.197	0.127	0.912	0.999	1.000
Joint-ID	0.599	0.066	0.011	0.119	0.077	0.994	1.000	1.000
WaterMono	0.357	0.049	0.006	0.081	0.055	0.999	1.000	1.000
SUTDF	0.324	0.039	0.004	0.065	0.045	1.000	1.000	1.000
HUDEMamba	**0.201**	**0.020**	**0.001**	**0.037**	**0.025**	1.000	1.000	1.000

Table 1: Quantitative results on full-scene testing. The best and second-best results are in **bold** and with underline.

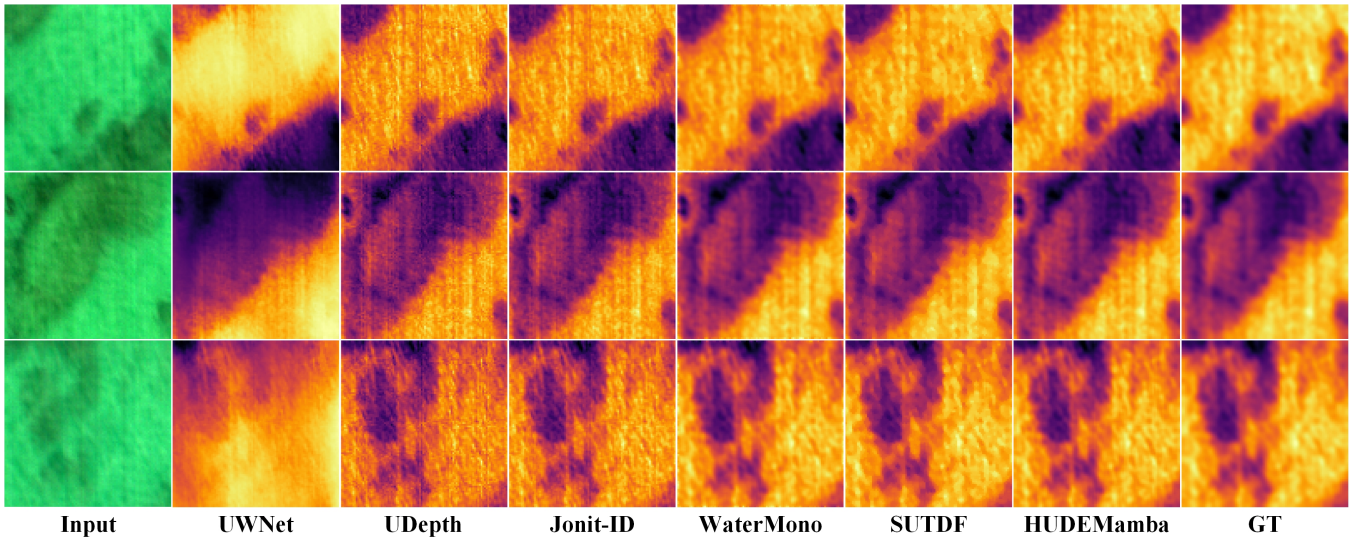


Figure 6: Qualitative results of full-scene testing.

Error loss (MSE) $\mathcal{L}_{MSE}$ and Scale-Invariant Log (SILog) loss $\mathcal{L}_{SILog}$ (Eigen, Puhrsch, and Fergus 2014)) as the learning objective $\mathcal{L}$:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{MSE} + \mathcal{L}_{SILog}, \\ \mathcal{L}_{MSE} &= \mathbb{E}\left(\|d_i - \hat{d}_i\|_2\right), \\ \mathcal{L}_{SILog} &= \alpha \sqrt{\frac{1}{M}\sum_{i=1}^{M} g_i^2 - \frac{\lambda}{M^2}\left(\sum_{i=1}^{M} g_i\right)^2}, \end{aligned} \tag{10}$$

where $g_i = \log(d_i/\hat{d}_i)$ and d_i is the groundtruth. α and λ are hyperparameters that are set as 0.85 and 10.

Experiments

Implementation Details

For the physics-inspired image translation stage, we utilize VCA (Rasti et al. 2024) as the hyperspectral unmixing (HU) method, setting the hyperparameter N in Equation (2) to 5 for all coastal scenes. In the context-aware depth estimation stage, we employ VMamba-T (Liu et al. 2024) as the foundational Mamba model, incorporating the proposed superpixel-pixel selective scanning strategy. Additionally, the full depth range $[d_{min}, d_{max}]$ in Equation (9) is set to $[0.5, 3]$ meters for all compared methods based on the proposed dataset. HUDEMamba is implemented on seven NVIDIA RTX6000 Ada 48GB GPUs running Ubuntu 22.04. We used the Adam optimizer with a batch size of 8. The initial learning rate of 1×10^{-4} was gradually reduced to 1×10^{-6} following a cosine schedule over 300 epochs.

Experimental settings

We compare the proposed method with four UDE methods (*i.e.,* UWNet (Gupta and Mitra 2019), UDepth (Yu, Wu, and Islam 2023), Joint-ID (Yang et al. 2024), and WaterMono (Ding et al. 2024)) and one HUDE method (*i.e.,* SUTDF (Qi et al. 2021)). For a fair comparison, we retrain all the compared methods using our dataset and training settings, along with the hyperparameters specified in their original papers using the proposed HUDE dataset. Following (Ding et al. 2024), the evaluation metrics used for quantitative evaluation include root mean error (**RMSE**) and its log variant ($\mathbf{RMSE}_{log}$), absolute error in log-scale ($\mathbf{Log}_{10}$), absolute (**A.Rel**) and squared (**S.Rel**) mean relative error, the percentage of inlier pixels ($\boldsymbol{\delta}_i$) with threshold 1.25^i, and scale-invariant error in log-scale ($\mathbf{SI}_{log}$).

Main results

To thoroughly evaluate the HUDE performance, we conduct two evaluation modes: full-scene and cross-scene testing.

Method	Scene1			Scene2			Scene3			Average		
	$\mathbf{SI}_{log}\downarrow$	$\mathbf{RMSE}_{log}\downarrow$	$\boldsymbol{\delta}_1\uparrow$	$\mathbf{SI}_{log}\downarrow$	$\mathbf{RMSE}_{log}\downarrow$	$\boldsymbol{\delta}_1\uparrow$	$\mathbf{SI}_{log}\downarrow$	$\mathbf{RMSE}_{log}\downarrow$	$\boldsymbol{\delta}_1\uparrow$	$\mathbf{SI}_{log}\downarrow$	$\mathbf{RMSE}_{log}\downarrow$	$\boldsymbol{\delta}_1\uparrow$
UWNet	3.472	0.446	0.251	3.500	0.448	0.250	3.403	0.439	0.254	3.458	0.444	0.252
UDepth	3.530	0.409	0.302	3.865	0.465	0.248	3.854	0.463	0.250	3.750	0.446	0.267
Joint-ID	3.210	0.396	0.305	3.379	0.421	0.282	3.377	0.419	0.286	3.322	0.412	0.291
WaterMono	3.058	0.371	0.356	3.208	0.397	0.323	3.207	0.396	0.325	3.158	0.388	0.335
SUTDF	3.008	0.342	**0.446**	3.061	0.347	**0.438**	3.072	0.349	**0.436**	3.047	0.346	**0.440**
HUDEMamba	**2.402**	**0.317**	0.403	**2.464**	**0.323**	0.400	**2.450**	**0.323**	0.397	**2.439**	**0.321**	0.400

Table 2: Quantitative results on cross-scene testing. The best and second-best results are in **bold** and with underline.

The full-scene testing, which assesses the learning capacity of the compared methods, *utilizes data from all scenes for both training and testing*. The quantitative results of this evaluation are detailed in Table 1, with corresponding visual results shown in Figure 6. In quantitative comparison, HUDEMamba demonstrates superior performance compared to the SOTA methods, achieving higher accuracy metrics while yielding lower values in certain error metrics. This highlights the capacity of HUDEMamba to leverage hyperspectral information for predicting more accurate HUDE results. The qualitative results in Figure 6 further support this conclusion, demonstrating that outputs of HUDEMamba are closest to the ground truth and accurately reflect the underwater scene geometry. Additionally, the visual results also indicate that HUDEMamba excels in removing background regions and predicting foreground layers with greater precision. **The cross-scene testing** evaluates the generalization performance *by training on one scene and testing on others*. The corresponding results are summarized in Table 2, using $\mathbf{SI}_{log}$, $\mathbf{RMSE}_{log}$, and $\boldsymbol{\delta}_1$ as evaluation metrics. The quantitative recorded in Table 2 indicate that all compared methods experience varying degrees of performance degradation. However, HUDEMamba outperforms the leading approaches by more significant margins, showing improvements of 0.608 in $\mathbf{SI}_{log}$ and 0.025 in $\mathbf{RMSE}_{log}$. These results collectively demonstrate that HUDEMamba exhibits superior generalization performance compared to other SOTA models when adapting to unseen scenes.

Ablation Study

Method	Full Scene		Cross Scene	
	$\mathbf{RMSE}_{log}\downarrow$	$\boldsymbol{\delta}_1\uparrow$	$\mathbf{RMSE}_{log}\downarrow$	$\boldsymbol{\delta}_1\uparrow$
HUDEMamba w/o PIT	0.063	0.984	0.394	0.282
HUDEMamba	0.025	1.000	0.321	0.400

Table 3: Quantitative results on ablation study of physics-inspired image translator.

The benefits of physics-inspired image translator. As detailed in Table 3, HUDEMamba without the physics-inspired image translator (PIT) shows significant performance degradation in both full-scene and cross-scene evaluations. In the full-scene test, performance metrics deteriorate from 0.025/1.000 to 0.063/0.984 in $\mathbf{RMSE}_{log}$ and $\boldsymbol{\delta}_1$, highlighting the critical role of the PIT in capturing essential hyperspectral information for HUDE. In the cross-scene test, performance metrics decline from 0.394/0.282 to 0.321/0.400 in $\mathbf{RMSE}_{log}$ and $\boldsymbol{\delta}_1$, demonstrating the substantial impact of the PIT on the generalization capability of HUDEMamba.

Method	Full Scene		Cross Scene	
	$\mathbf{RMSE}_{log}\downarrow$	$\boldsymbol{\delta}_1\uparrow$	$\mathbf{RMSE}_{log}\downarrow$	$\boldsymbol{\delta}_1\uparrow$
PIT+Transformer	0.042	1.000	0.340	0.342
PIT+vanilla Mamba	0.034	1.000	0.332	0.367
HUDEMamba	0.025	1.000	0.321	0.400

Table 4: Quantitative results on ablation study of context-aware depth estimator.

The benefits of context-aware depth estimator. To assess the effectiveness of the context-aware depth estimator (CDE), we compare it with two depth estimator variants: (1) transformer-based and (2) vanilla Mamba-based. We substitute the depth estimator with these variants, and the quantitative results are presented in Table 4. CDE consistently outperforms both variants, achieving average improvements of 0.01 and 0.02 in $\mathbf{RMSE}_{log}$ and $\boldsymbol{\delta}_1$. This enhancement is attributed to CDE's effective exploitation of spatial correlations among adjacent pixels, facilitated by the Mamba framework and the proposed superpixel-pixel selective scanning strategy. Moreover, the greater performance gap between HUDEMamba and the compared methods in the cross-scene test further highlights the contribution of CDE to the generalization capability. Furthermore, the observed performance disparity between the transformer-based and vanilla Mamba-based estimators indicates that long-sequence modeling is better suited for depth estimation.

Conclusion

In this paper, we built the first open-access real-world hyperspectral underwater depth estimation (HUDE) dataset, collected using a UAV-borne platform equipped with RTK/IMU information. To address the challenges in HUDE, we also propose a novel framework that leverages the intrinsic properties of hyperspectral imagery and the spatial correlations among adjacent pixels to achieve accurate HUDE results. Experimental results on the proposed dataset confirm the effectiveness of our approach. We hope this work can foster further research in HUDE to provide a novel, accurate, and reliable means for underwater environment understanding.

Acknowledgements

This work was supported in part by the Foundation Fund of Science and Technology on Near-Surface Detection Laboratory under Grant 6142414220808, and in part by the National Natural Science Foundation of China 62201586, and in part by China National Postdoctoral Program for Innovative Talents under Grant BX20240492.

References

Alevizos, E. 2020. A Combined Machine Learning and Residual Analysis Approach for Improved Retrieval of Shallow Bathymetry from Hyperspectral Imagery and Sparse Ground Truth Data. *Remote Sensing*, 12(21).

Barbato, M. P.; Napoletano, P.; Piccoli, F.; and Schettini, R. 2022. Unsupervised Segmentation of Hyperspectral Remote Sensing Images with Superpixels. *arXiv: 2204.12296*.

Bhat, S. F.; Alhashim, I.; and Wonka, P. 2020. AdaBins: Depth Estimation using Adaptive Bins. *arXiv: 2011.14141*.

Ding, Y.; Li, K.; Mei, H.; Liu, S.; and Hou, G. 2024. WaterMono: Teacher-Guided Anomaly Masking and Enhancement Boosting for Robust Underwater Self-Supervised Monocular Depth Estimation. *arXiv: 2406.13344*.

Eigen, D.; Puhrsch, C.; and Fergus, R. 2014. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. *CoRR*, abs/1406.2283.

Gu, A.; and Dao, T. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv: 2312.00752*.

Gupta, H.; and Mitra, K. 2019. Unsupervised Single Image Underwater Depth Estimation. In *IEEE International Conference on Image Processing, 2019*, 624–628.

Lee, J.; Cai, X.; Schönlieb, C.; and Coomes, D. 2015. Nonparametric Image Registration of Airborne LiDAR, Hyperspectral and Photographic Imagery of Wooded Landscapes. *IEEE Trans. Geosci. Remote. Sens.*, 53(11): 6073–6084.

Lee, Z.; Mobley, C. D.; Steward, R. G.; and Patch, J. S. 1998. Hyperspectral remote sensing for shallow waters. I. A semianalytical model. *Applied optics*, 37(27): 6329–6338.

Li, K.; Wang, X.; Liu, W.; Qi, Q.; Hou, G.; Zhang, Z.; and Sun, K. 2024. Learning Scribbles for Dense Depth: Weakly Supervised Single Underwater Image Depth Estimation Boosted by Multitask Learning. *IEEE Trans. Geosci. Remote. Sens.*, 62: 1–15.

Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; and Benediktsson, J. A. 2019. Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Trans. Geosci. Remote. Sens.*, 57(9): 6690–6709.

Liu, B.; Liu, Z.; Men, S.; Li, Y.; Ding, Z.; He, J.; and Zhao, Z. 2020. Underwater Hyperspectral Imaging Technology and Its Applications for Detecting and Mapping the Seafloor: A Review. *Sensors*, 20(17).

Liu, X.; Zhang, C.; and Zhang, L. 2024. Vision Mamba: A Comprehensive Survey and Taxonomy. *arXiv: 2405.04404*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; and Liu, Y. 2024. VMamba: Visual State Space Model. *arXiv: 2401.10166*.

Ma, S.; Tao, Z.; Yang, X.; Yu, Y.; Zhou, X.; and Li, Z. 2014. Bathymetry Retrieval From Hyperspectral Remote Sensing Data in Optical-Shallow Water. *IEEE Trans. Geosci. Remote. Sens.*, 52(2): 1205–1212.

Molinier, M.; and Kilpi, J. 2019. Avoiding Overfitting When Applying Spectral-Spatial Deep Learning Methods on Hyperspectral Images with Limited Labels. In *IEEE International Geoscience and Remote Sensing Symposium, 2019*, 5049–5052. IEEE.

Pan, Z.; Glennie, C. L.; Legleiter, C.; and Overstreet, B. 2015. Estimation of Water Depths and Turbidity From Hyperspectral Imagery Using Support Vector Regression. *IEEE Geosci. Remote. Sens. Lett.*, 12(10): 2165–2169.

Peng, D.; Mao, H.; Sun, L.; Li, Q.; and Zhou, M. 2023. Bathymetry Retrieval from Hyperspectral Image Using the Channel-wise Spectral Attention Based Convolutional Neural Network. In *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, 2023*, 1–7. IEEE.

Peng, Y.; Zhao, X.; and Cosman, P. C. 2015. Single underwater image enhancement using depth estimation based on blurriness. In *IEEE International Conference on Image Processing, 2015*, 4952–4956. IEEE.

Qi, J.; Wan, P.; Gong, Z.; Xue, W.; Yao, A.; Liu, X.; and Zhong, P. 2021. A Self-Improving Framework for Joint Depth Estimation and Underwater Target Detection from Hyperspectral Imagery. *Remote. Sens.*, 13(9): 1721.

Rasti, B.; Zouaoui, A.; Mairal, J.; and Chanussot, J. 2024. Image Processing and Machine Learning for Hyperspectral Unmixing: An Overview and the HySUPP Python Package. *IEEE Trans. Geosci. Remote. Sens.*, 62: 1–31.

Song, W.; Wang, Y.; Huang, D.; and Tjondronegoro, D. 2018. A Rapid Scene Depth Estimation Model Based on Underwater Light Attenuation Prior for Underwater Image Restoration. In *Advances in Multimedia Information Processing, 2018*, 678–688.

Wang, C.; Xu, H.; Jiang, G.; Yu, M.; Luo, T.; and Chen, Y. 2024. Underwater Monocular Depth Estimation Based on Physical-Guided Transformer. *IEEE Trans. Geosci. Remote. Sens.*, 62: 1–16.

Yang, G.; Kang, G.; Lee, J.; and Cho, Y. 2024. Joint-ID: Transformer-Based Joint Image Enhancement and Depth Estimation for Underwater Environments. *IEEE Sensors Journal*, 24(3): 3113–3122.

Ye, X.; Li, Z.; Sun, B.; Wang, Z.; Xu, R.; Li, H.; and Fan, X. 2020. Deep Joint Depth Estimation and Color Correction From Monocular Underwater Images Based on Unsupervised Adaptation Networks. *IEEE Trans. Circuits Syst. Video Technol.*, 30(11): 3995–4008.

Ye, X.; Zhang, J.; Yuan, Y.; Xu, R.; Wang, Z.; and Li, H. 2023. Underwater Depth Estimation via Stereo Adaptation Networks. *IEEE Trans. Circuits Syst. Video Technol.*, 33(9): 5089–5101.

Yu, B.; Wu, J.; and Islam, M. J. 2023. UDepth: Fast Monocular Depth Estimation for Visually-guided Underwater Robots. In *IEEE International Conference on Robotics and Automation, 2023*, 3116–3123. IEEE.

Zhang, S.; Gong, X.; Nian, R.; He, B.; Wang, Y.; and Lendasse, A. 2017. A depth estimation model from a single underwater image with non-uniform illumination correction. In *OCEANS 2017*, 1–5.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *arXiv: 2401.09417*.