# Tree-Mamba: A Tree-Aware Mamba for Underwater Monocular Depth Estimation (Supplementary Material)

Peixian Zhuang, *Senior Member, IEEE,* Yijian Wang, Zhenqi Fu, Hongliang Zhang, Sam Kwong, *Fellow, IEEE*, Chongyi Li, *Senior Member, IEEE*

***Abstract*—In the supplementary material, we first revisit the preliminaries of state space models (SSMs), introduce related vision Mamba methods, summarize the meaning of notations in section IV of our main text, and then provide more experimental results and implementation details to complement the main manuscript, including experimental results as follows: 1) visual results using our BlueDepth dataset versus other datasets, 2) visual comparison of real underwater images from Test-FR5691, and 3) visual comparison of different methods on underwater panoramic, hazy, sand-dust, and low-light images.**

TABLE I
DEFINITION OF NOTATIONS IN SECTION IV OF OUR MAIN TEXT.

| Symbol | Definition | Symbol | Definition |
|---|---|---|---|
| $h$ | Latent state | $H$ | Matrixized H |
| $A$ | State matrix | $\bar{A}$ | Discretized $A$ |
| $B$ | Input matrix | $\bar{B}$ | Discretized $B$ |
| $C$ | Output matrix | $\Delta$ | Time-scale parameter |
| $S$ | Parent node matrix | $C$ | Child node matrix |
| $\rho(\cdot)$ | Spectral radius | $\|\cdot\|_\infty$ | Infinity norm |

## I. PRELIMINARIES

State Space Models (SSMs) are proposed for sequence-to-sequence modeling, which maps a 1-dimensional sequence input $x(t) \in \mathbb{R}^{1\times 1}$ to an output $y(t) \in \mathbb{R}^{1\times 1}$ by a latent state $h(t) \in \mathbb{R}^{N\times 1}$, as described by the following linear ordinary differential equations:

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t), \\ y(t) &= Ch(t), \end{aligned} \tag{1}$$

where $t$ and $N$ are the time step and hidden state size, respectively. $h'(t) = \frac{d}{dt}h(t)$. $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times 1}$, and $C \in \mathbb{R}^{1\times N}$ denote state, input, and output matrices, respectively. $A$ determines the influence of previous latent state on current latent state, $B$ decides how much $x(t)$ affects the latent state, and $C$ describes how the latent state is transformed into $y(t)$. To integrate Eq. (1) into deep learning-based architectures, the Zero-Order Hold (ZOH) rule is usually adopted for discretization, since the ZOH can avoid large computational burden caused by calculating integrals. The discretization is defined as follows:

$$\begin{aligned} \bar{A} &= e^{\Delta A}, \\ \bar{B} &= (\Delta A)^{-1}(e^{\Delta A} - I) \cdot \Delta B, \\ h_t &= \bar{A}h_{t-1} + \bar{B}x_t, \\ y_t &= Ch_t, \end{aligned} \tag{2}$$

where $\bar{A}$ and $\bar{B}$ are discrete counterparts. $\Delta$ and $I$ are the time-scale parameter and the identity matrix, respectively.

Since the parameters ($\Delta$, $A$, $B$, and $C$) are randomly initialized and remain invariant to the input $x$, the SSM faces the poor performance in context-dependent learning. To solve the above issue, Dao *et al.* [1] introduce a selective mechanism-based SSM named Mamba. Specifically, the selection mechanism is used to make parameters ($\Delta$, $B$, and $C$) depending on the input $x$:

$$\Delta, B, C = Linear(x), \tag{3}$$

where $Linear$ is a parameterized projection. Such an initialization can effectively improve the performance of the SSM in context-dependent learning.

## II. VISION MAMBA METHODS

Leveraging Mamba's strengths in long-sequence modeling, many Mamba-based models have been proposed for vision tasks. These models flatten 2D images into multiple 1D sequences along different scanning directions, followed by state propagation. Zhu *et al.* [2] proposed the first visual Mamba (ViM) model, which introduces a bidirectional raster scanning strategy to convert 2D images into 1D sequences and learns the visual representation in a sequence modeling manner. Hu *et al.* [3] designed a continuous scanning strategy to preserve spatial dependencies of images and achieve enhanced global context modeling. Shi *et al.* [4] combined four-directional raster scanning with a diagonal scanning strategy to preserve image locality and continuity, but incurred additional computational burden. Li *et al.* [5] introduced a nested S-shape scanning strategy, which divided an image into multiple non-overlapping subregions and performed continuous scanning within each subregion, thus improving local feature modeling capability. Although the above Mamba-based methods achieve promising performance in the image domain, they perform inadequately in underwater monocular depth estimation (UMDE) because their fixed and inflexible scanning strategies fail to effectively model the structural features of underwater images. In contrast, our proposed tree-aware scanning strategy constructs an input-dependent minimum spanning tree and leverages the structural

Peixian Zhuang and Yijian Wang contributed equally to this work.
*Corresponding authors: (Chongyi Li and Peixian Zhuang).

relationships between parent and child nodes to capture the spatial topology of underwater images, thereby enabling multi-scale feature modeling capabilities. Our scanning strategy not only delivers powerful feature representation capabilities but also maintains a high degree of flexibility.

## III. More Experimental Results

### A. Experiment Settings

**Implementation Details.** We implement the proposed Tree-Mamba on the PyTorch 2.1.0 framework with an Intel (R) i9-12900K CPU, 64GB RAM, and an NVIDIA RTX 4090 GPU. We adopt the ADAM optimizer for network optimization and set the initial learning rate to $10^{-4}$. The input underwater images are resized to 256×256. The batch size and training epochs are set to 8 and 50. The hyperparameters, including the learning rate, are adaptively optimized using the Optuna library [6] by minimizing the loss of the training set.

### B. Effects of different UMDE datasets

We investigate the effectiveness of different UMDE datasets in boosting the prediction performance of existing UMDE models. Specifically, four UMDE models (UW-GAN [7], UDepth [8], UW-Depth [9], and our Tree-Mamba) are individually trained on the eight datasets (Sea-Thru [10], NYU-U [11], SQUID [12], FLSea [13], Atlantis [14], SUIM-SDA [15], USOD10K [16], and our BlueDepth). After training for 50 epochs, each UW-GAN [7], UDepth [8], UW-Depth [9], and our Tree-Mamba with the minimum training loss value is retained. Subsequently, a qualitative evaluation is performed on the UIEB dataset [17] to contrast the performance of these trained UMDE models, as shown in Fig. 1. As shown, all models trained on real-labeled datasets (Sea-Thru [10], NYU-U [11], SQUID [12], and FLSea [13]) tend to produce chaotic scene depth distributions, while those trained on pseudo-labeled datasets (Atlantis [14], SUIM-SDA [15], and USOD10K [16]) improve estimation accuracy of scene depth, but their improvements remain limited. In contrast, all models trained with our proposed BlueDepth are able to produce more accurate scene depth, and our proposed Tree-Mamba method yields better depth results than other competitors [7]–[9]. This significant improvement highlights the effectiveness of the proposed BlueDepth baseline for facilitating existing UMDE models to better learn accurate object-depth relationships.

## References


[1] T. Dao and A. Gu, "Transformers are SSMs: generalized models and efficient algorithms through structured state space duality," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 10041–10071, 2024.

[2] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: efficient visual representation learning with bidirectional state space model," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.

[3] V. T. Hu, S. A. Baumann, M. Gui, O. Grebenkova, P. Ma, J. Fischer, and B. Ommer, "ZigMa: A dit-style zigzag mamba diffusion model," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 148–166, 2024.

[4] Y. Shi, B. Xia, X. Jin, X. Wang, T. Zhao, X. Xia, X. Xiao, and W. Yang, "VmambaIR: Visual state space model for image restoration," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 6, pp. 5560–5574, 2025.

[5] B. Li, H. Zhao, W. Wang, P. Hu, Y. Gou, and X. Peng, "MaIR: A locality- and continuity-preserving mamba for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7491–7501, 2025.

[6] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 2623–2631, 2019.

[7] P. Hambarde, S. Murala, and A. Dhall, "UW-GAN: Single-image depth estimation and image enhancement for underwater images," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[8] B. Yu, J. Wu, and M. J. Islam, "UDepth: Fast monocular depth estimation for visually-guided underwater robots," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 3116–3123, 2023.

[9] L. Ebner, G. Billings, and S. Williams, "Metrically scaled monocular depth estimation through sparse priors for underwater robots," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 3751–3757, 2024.

[10] D. Akkaynak and T. Treibitz, "Sea-Thru: A method for removing water from underwater images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1682–1691, 2019.

[11] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognit.*, vol. 98, p. 107038, 2020.

[12] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater single image color restoration using haze-lines and a new quantitative dataset," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2822–2837, 2021.

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

[14] F. Zhang, S. You, Y. Li, and Y. Fu, "Atlantis: Enabling underwater depth estimation with stable diffusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11852–11861, 2024.

[15] 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 Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[16] L. Hong, X. Wang, G. Zhang, and M. Zhao, "USOD10K: A new benchmark dataset for underwater salient object detection," *IEEE Trans. Image Process.*, vol. 34, pp. 1602–1615, 2025.

[17] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 4376–4389, 2020.

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

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

[20] H. Gupta and K. Mitra, "Unsupervised single image underwater depth estimation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pp. 624–628, 2019.

[21] W. Song, Y. Wang, D. Huang, A. Liotta, and C. Perra, "Enhancement of underwater images with statistical model of background light and optimization of transmission map," *IEEE Trans. Broadcast.*, vol. 66, no. 1, pp. 153–169, 2020.

[22] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, 2022.

[23] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-Mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18537–18546, 2023.

[24] 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," *Int. J. Comput. Vis.*, pp. 1–19, 2023.

[25] 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 Trans. Instrum. Meas.*, vol. 74, pp. 1–14, 2025.

[26] N. J. Avanaki, A. Ghildyal, N. Barman, and S. Zadtootaghaj, "LAR-IQA: A lightweight, accurate, and robust no-reference image quality assessment model," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 328–345, 2024.

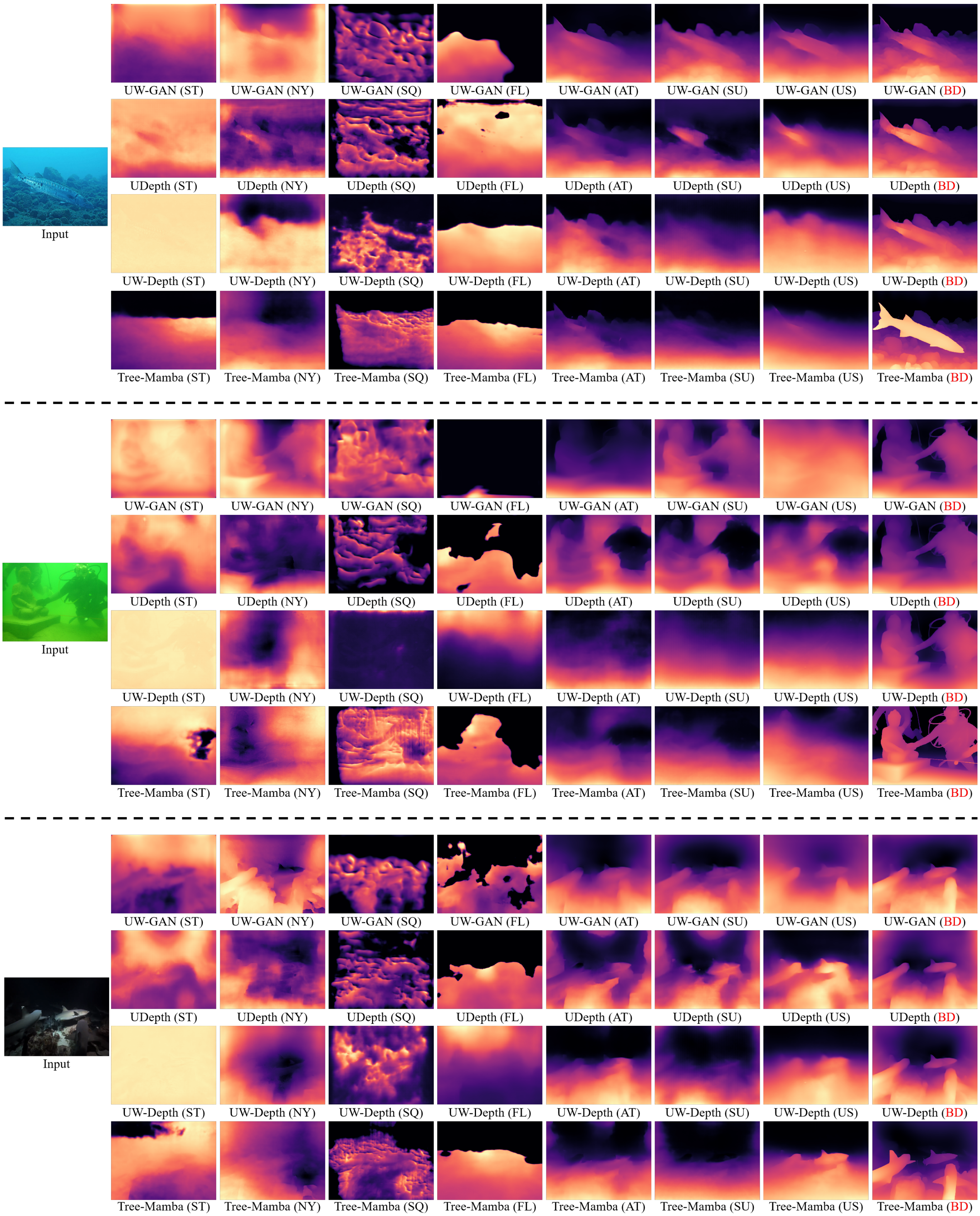



Fig. 1. Visual results using our proposed BlueDepth dataset versus other datasets. ST, NY, SQ, FL, AT, SU, US, and BD denote the model trained on Sea-Thru [10], NYU-U [11], SQUID [12], FLSea [13], Atlantis [14], SUIM-SDA [15], USOD10K [16], and our BlueDepth, respectively. The depth results of UW-GAN [7], UDepth [8], UW-Depth [9], and Tree-Mamba are significantly improved by training on our BlueDepth dataset, meanwhile, our Tree-Mamba method yields better depth results than other competitors.

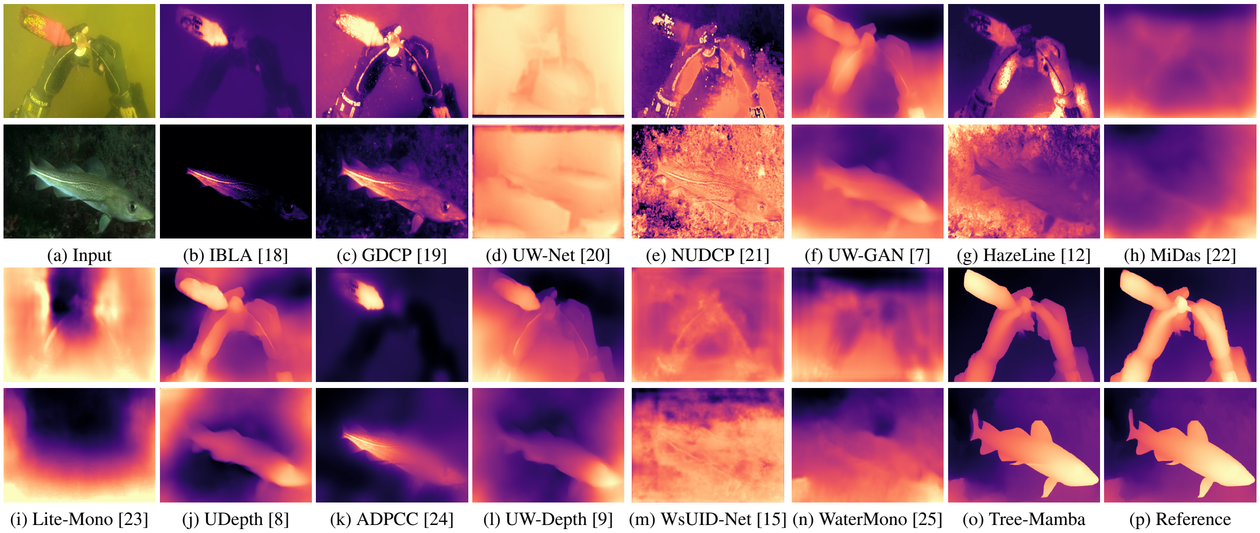



Fig. 2. Visual comparison of different methods on yellowish and low-visibility underwater images from **Test-FR5691**. Compared with other competitors, our Tree-Mamba method yields better depth results on different degraded underwater images, and our depths are closer to those of ground truths.

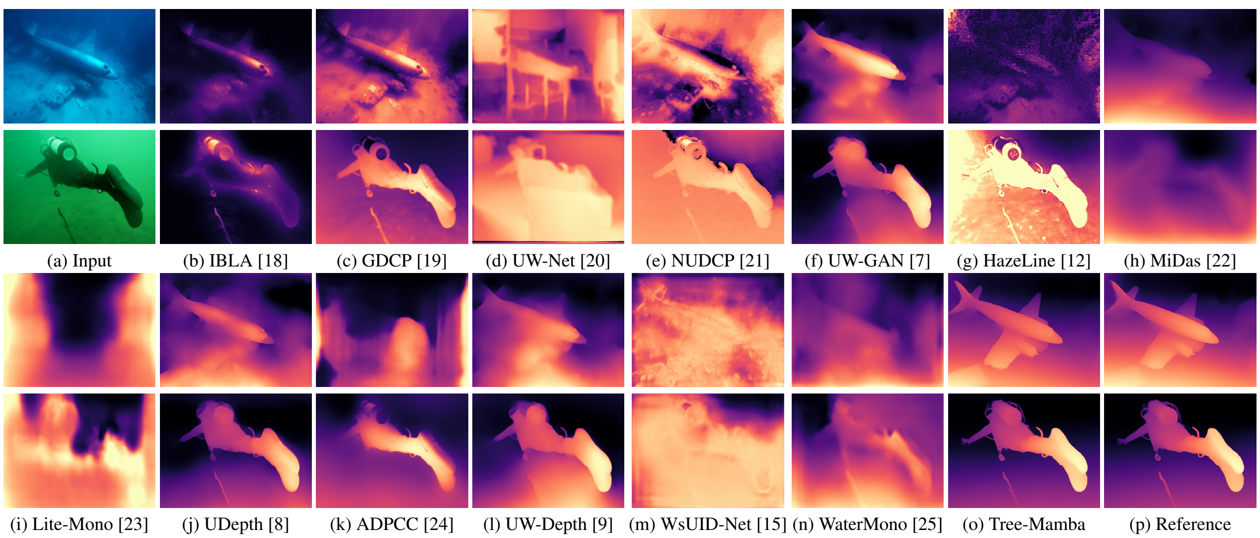



Fig. 3. Visual comparison of different methods on bluish and greenish underwater images from **Test-FR5691**. Compared with other competitors, our Tree-Mamba method yields better depth results on different degraded underwater images, and our depths are closer to those of ground truths.

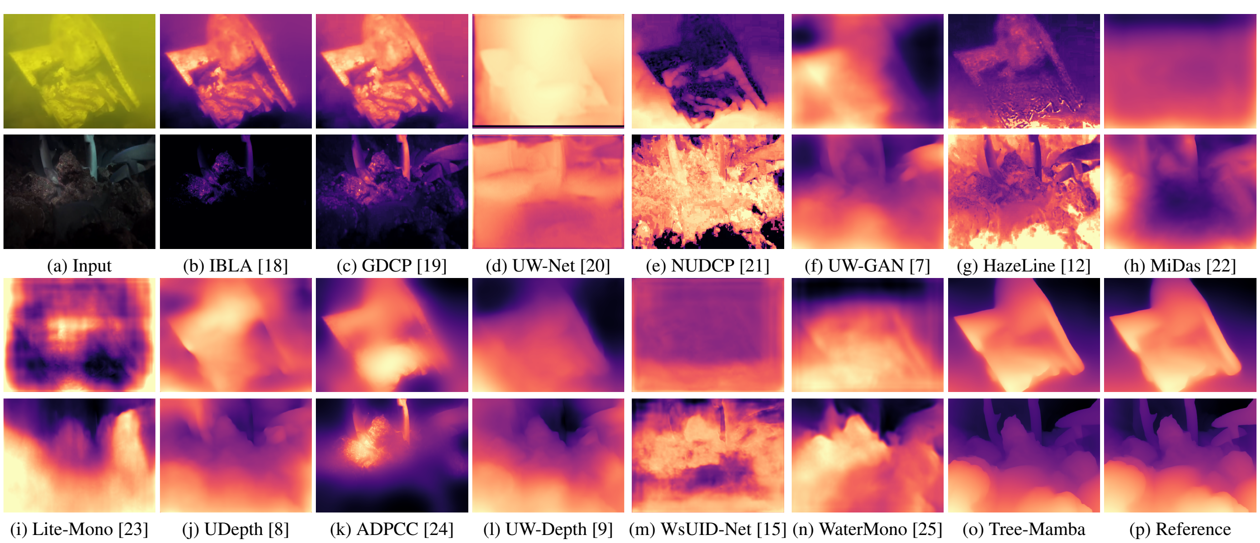
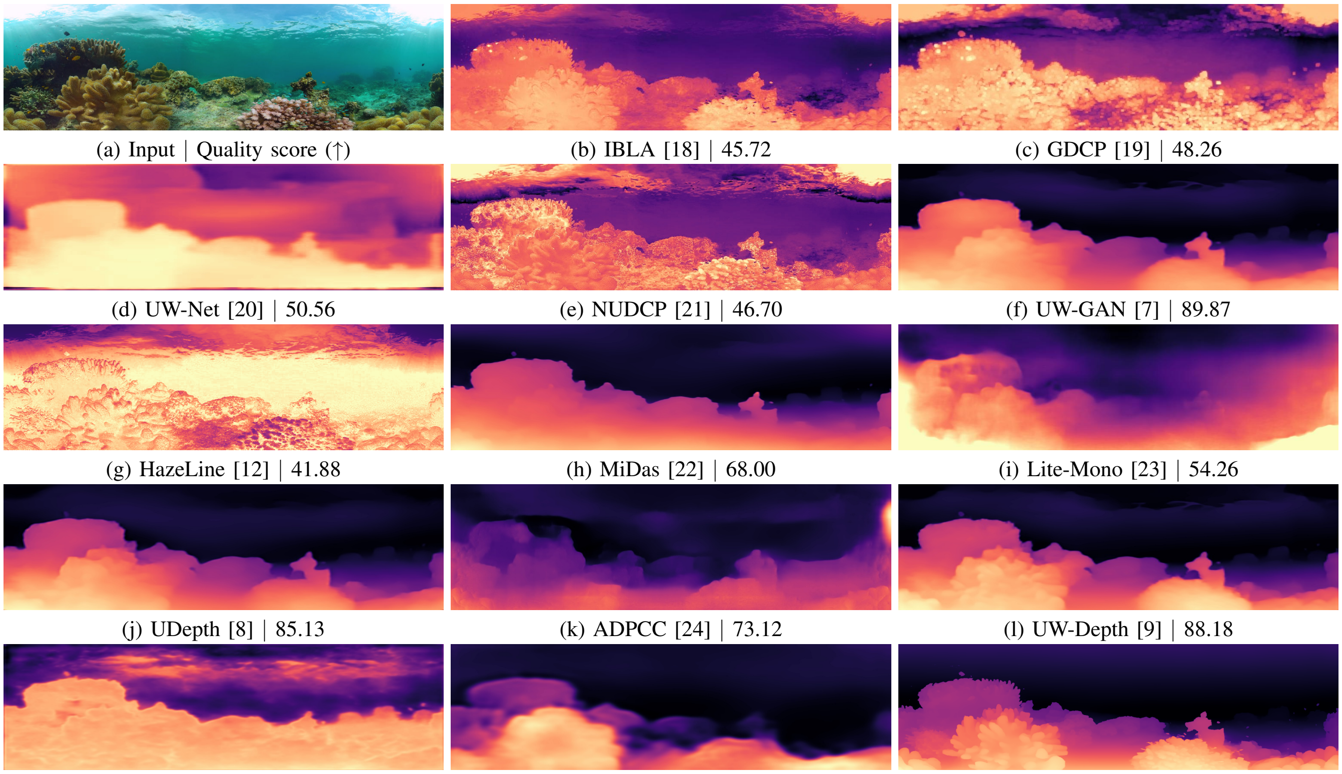
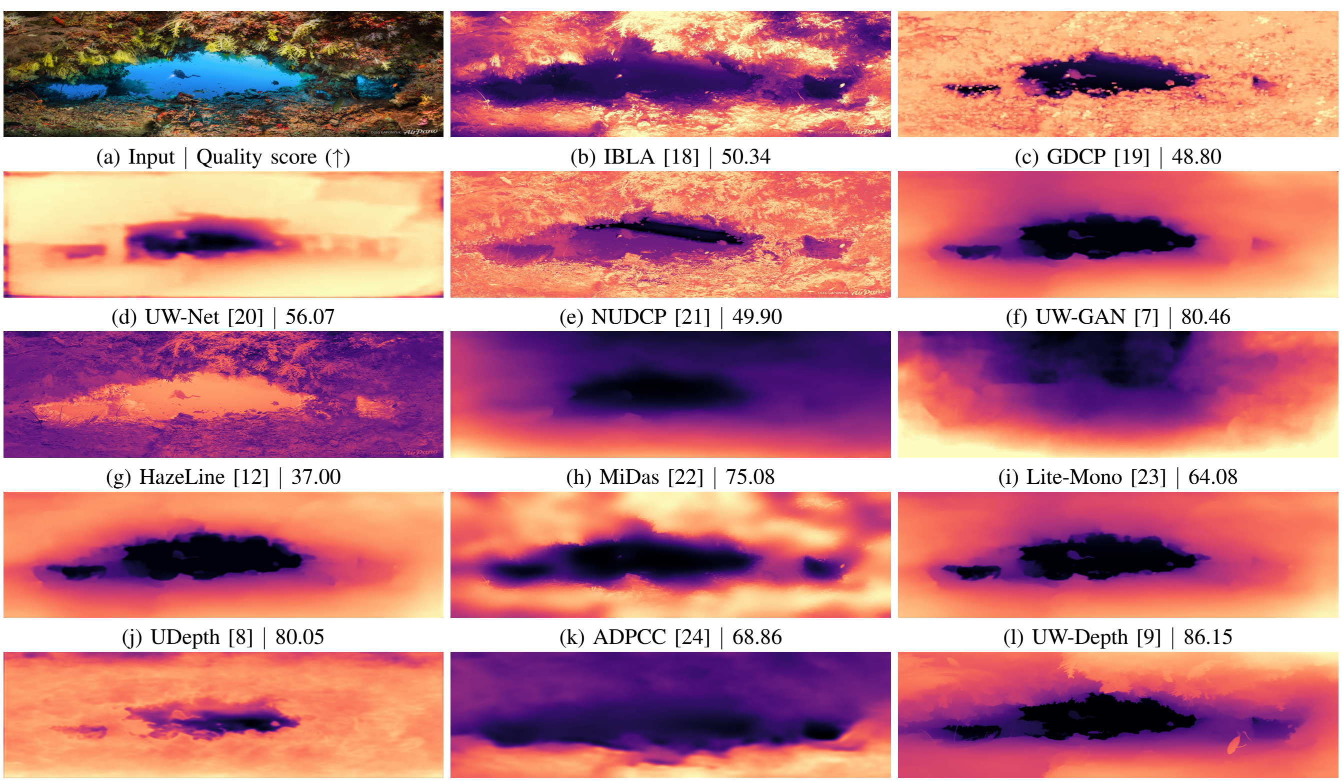



Fig. 4. Visual comparison of different methods on yellowish and low-visibility underwater images from **Test-FR5691**. Compared with other competitors, our Tree-Mamba method yields better depth results on different degraded underwater images, and our depths are closer to those of ground truths.



(a) Input | Quality score (↑)
(b) IBLA [18] | 45.72
(c) GDCP [19] | 48.26
(d) UW-Net [20] | 50.56
(e) NUDCP [21] | 46.70
(f) UW-GAN [7] | 89.87
(g) HazeLine [12] | 41.88
(h) MiDas [22] | 68.00
(i) Lite-Mono [23] | 54.26
(j) UDepth [8] | 85.13
(k) ADPCC [24] | 73.12
(l) UW-Depth [9] | 88.18
(m) WsUID-Net [15] | 50.55
(n) WaterMono [25] | 32.69
(o) Tree-Mamba | 96.52

Fig. 5. Visual comparison of different methods on an underwater panoramic image. The quality score is evaluated by the fine-tuned LAR-IQA model [26]. The best result is marked in red. Compared with other competitors, our Tree-Mamba method yields better results of both panoramic depth and quality score.



(a) Input | Quality score (↑)
(b) IBLA [18] | 50.34
(c) GDCP [19] | 48.80
(d) UW-Net [20] | 56.07
(e) NUDCP [21] | 49.90
(f) UW-GAN [7] | 80.46
(g) HazeLine [12] | 37.00
(h) MiDas [22] | 75.08
(i) Lite-Mono [23] | 64.08
(j) UDepth [8] | 80.05
(k) ADPCC [24] | 68.86
(l) UW-Depth [9] | 86.15
(m) WsUID-Net [15] | 58.66
(n) WaterMono [25] | 52.34
(o) Tree-Mamba | 98.75

Fig. 6. Visual comparison of different methods on an underwater panoramic image. The quality score is evaluated by the fine-tuned LAR-IQA model [26]. The best result is marked in red. Compared with other competitors, our Tree-Mamba method yields better results of both panoramic depth and quality score.

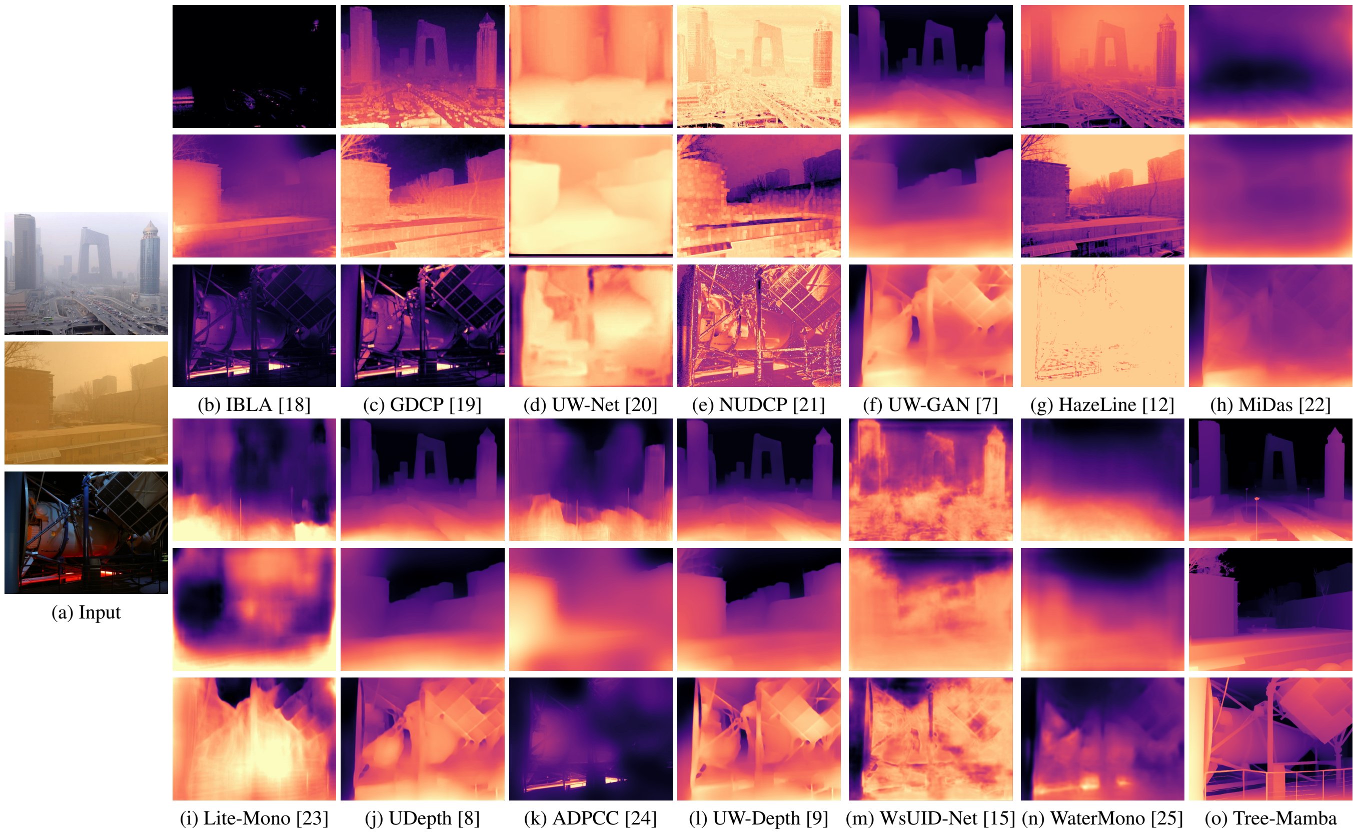



Fig. 7. Visual comparison of different methods on hazy (top), sand-dust (middle), and low-light (bottom) images. Compared with other methods, our Tree-Mamba method yields better depth estimation results on hazy, sand-dust, and low-light images.

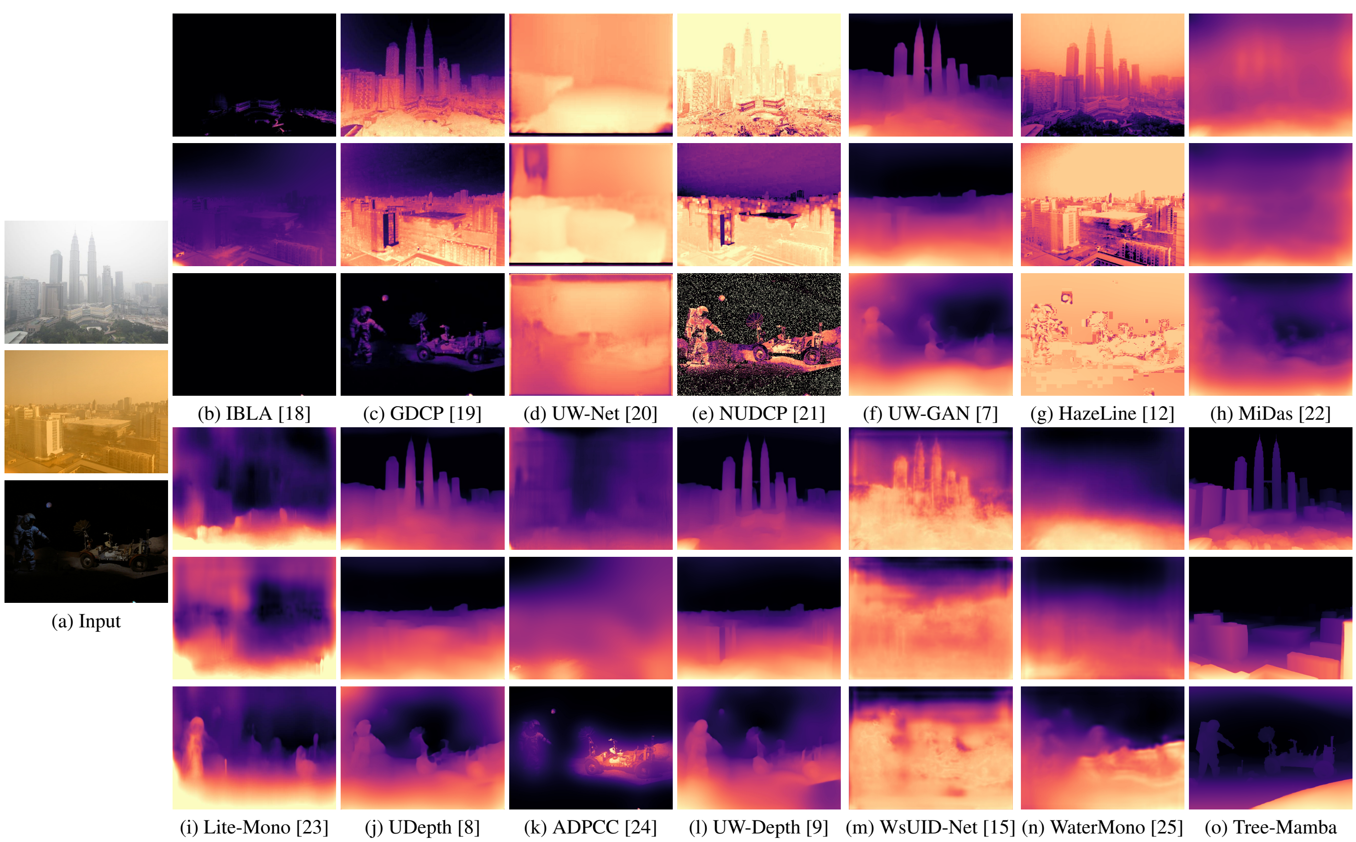



Fig. 8. Visual comparison of different methods on hazy (top), sand-dust (middle), and low-light (bottom) images. Compared with other methods, our Tree-Mamba method yields better depth estimation results on hazy, sand-dust, and low-light images.