Underwater Visual SLAM with Depth Uncertainty and Medium Modeling

Supplementary Material

This supplementary document provides additional experimental results and extended discussions of our approach, which are organized as follows:

- *Additional Results (§A)*
- *Dataset Details (§B)*
- *Model Details (§C)*
- *Discussion (§D)*

A. Additional Results

Failure Cases. Fig. C1 demonstrates failure cases in (a) rendering quality and (b) tracking trajectories. Although our method achieves successful convergence on most sequences, failures occasionally occur in particularly challenging long-duration trajectories, *e.g.*, the length of trajectory shown in (b) exceeds three times that of typical sequences. Moreover, rendering artifacts appear in difficult scenarios characterized by severe photometric variations and poor illumination. In future work, we plan to explore advanced approaches, such as integrating underwater-specific image enhancement techniques [3] and multi-sensor fusion methods [7], to further enhance robustness and tracking accuracy in challenging underwater scenarios.

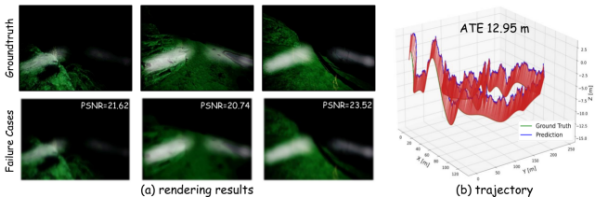

Figure C1. Failure cases of (a) rendering and (b) tracking trajectory in MIMIR-UW [1] (§A).

B. Dataset Details

Data Processing. All underwater datasets are reorganized following the input format of the TUM dataset [9].

Dataset Details. FLSea [8] consists of two primary subsets, *Canyons* and *Red Sea*, each capturing diverse underwater environments with varying dynamics and visual characteristics. The *Canyons* subset includes four sequences designed to evaluate SLAM robustness in underwater scenarios: U Canyon (UC) with 2,895 frames, Flatiron (Fla) with 2,475 frames, Horse Canyon (HC) with 2,230 frames, and Tiny Canyon (TC) with 1,012 frames. The *Red Sea* subset consists of eight distinct sequences covering diverse paths and loops: Northeast Path (NP, 2,593 images), Landward Path (LP, 1,204 images), Dice Path (DP, 1,428 images), Pier Path (PP, 1,695 images), Coral Table Loop (CTL, 1,017 images), Cross Pyramid Loop (CPL, 1,652 images), Big Dice Loop (BDL, 3,159 images), and Sub Pier (SP, 1,091 images).

MIMIR-UW [1] consists of four distinct underwater scenarios: *SeaFloor*, *SeaFloor Algae*, *OceanFloor*, and *SandPipe*. The *SeaFloor* scenario contains three sequences: track0 (t0) with 2,847 frames, track1 (t1) with 2,030 frames, and track2 (t2) with 2,537 frames. The *SeaFloor Algae* scenario expands this with three additional sequences: track0 (t0) having 2,934 frames, track1 (t1) having 2,076 frames, and track2 (t2) having 2,489 frames. The *OceanFloor* scenario specifically introduces illumination-related challenges through three subsets: track0-light (t0-l, 2,421 frames), track0-dark (t0-d, 2,404 frames), and track1-light (t1-l, 6,263 frames), highlighting photometric degradation effects across identical paths. Lastly, the *SandPipe* scenario addresses feature-scarce environments, offering two sequences: track0-dark (t0-d, 2,741 frames) and track0-light (t0-l, 2,605 frames), depicting sandy substrates with low-contrast lighting conditions.

Tartanair [11] is a large-scale synthetic benchmark widely used for visual SLAM research. For evaluating underwater SLAM methods, we adopt its *Ocean* scenarios, which include 12 Easy-level and 10 Hard-level sequences. The Easy-level sequences simulate relatively stable marine environments featuring moderate turbulence and consistent illumination, whereas the Hard-level sequences introduce significant challenges such as severe photometric variations and transient occlusion events.

C. Model Details

Tracking and Mapping. For tracking, we initialize dense bundle adjustment layers using pre-trained weights [10] derived from large-scale monocular videos, while the mapping module is trained from scratch. In the mapping loss function (Eq. 13), we balance $\mathcal{L}_1$ and $\mathcal{L}_{\text{DSSIM}}$ terms in $\mathcal{L}_{\text{rgb}}$ with weighting factors of 0.7 and 0.3, respectively.

D. Discussion

Terms of use, Privacy, and License. The datasets and algorithms described in this work are made available exclusively for academic research purposes. For privacy protection, no personally identifiable information was recorded during underwater vehicle operation, and any human-annotated data utilized herein (if applicable) was anonymized through cryptographic hashing of operator identifiers. Third-party datasets referenced in this study, such as FLSea [8], MIMIR-UW [1], and TartanAir [11], maintain their respective Creative Commons Attribution-

NonCommercial 4.0 International (CCBY-NC4.0) licenses.

Limitations. (1) Our tracking module utilizes weights pre-trained on large-scale synthetic datasets for dense bundle adjustment, achieving favorable performance on certain synthetic sequences, *e.g.*, TartanAir-Ocean. To further enhance real-world applicability, future work could involve collecting a substantial corpus of real-world underwater sequences pairs for practical applications. (2) Although our approach demonstrates robust reconstruction capabilities across multiple datasets, it currently does not explicitly model dynamic underwater scenes, potentially limiting its effectiveness in highly dynamic environments. Integrating methods like 4D Gaussian Splatting for dynamic scene rendering [13] could address this limitation in future studies.

Future Direction. (1) Unlike terrestrial or aerial scenarios, sonar sensors play an indispensable role underwater. Optical sensors (*i.e.*, cameras) are prone to performance degradation under challenging illumination conditions, such as dark marine environments. In contrast, sonar sensors provide enhanced robustness for underwater mapping tasks. Therefore, multimodal fusion (*e.g.*, cameras and sonar) will be a crucial research direction for navigation [2, 4–6, 12]. (2) Visual SLAM methods inevitably accumulate errors over long distances, undermining reliability and robustness in large-scale underwater applications. Addressing error accumulation and drift correction remains a longstanding challenge, making error correction methods a central focus of future research. (3) Real-world underwater conditions (*e.g.*, illumination variability, hydrodynamic disturbances, and marine animal activities) constantly fluctuate. Future efforts will involve collecting a broader range of diverse underwater data or introducing controllable synthetic data generation [14–17], further improving robustness and generalization capabilities across underwater environments.

Broader Impacts. The ocean hosts abundant life and plays a vital role in the global carbon cycle. The motivation behind our DUV-SLAM approach is to advance fully autonomous underwater vehicles, facilitating deeper and more efficient exploration of one of Earth's largest and most diverse ecosystems. Additionally, we encourage further research efforts toward autonomous biological detection, tracking, monitoring, and environmental management in complex marine habitats.

References

[1] Olaya Álvarez-Tuñón, Hemanth Kanner, Luiza Ribeiro Marnet, Huy Xuan Pham, Jonas le Fevre Sejersen, Yury Brodskiy, and Erdal Kayacan. Mimir-uw: A multipurpose synthetic dataset for underwater navigation and inspection. In *IROS*, 2023. 1

[2] Jianzhe Gao, Rui Liu, and Wenguan Wang. 3d gaussian map with open-set semantic grouping for vision-language navigation. In *ICCV*, 2025. 2

[3] Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE TIP*, 29: 4376–4389, 2019. 1

[4] Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *ICCV*, 2023. 2

[5] Rui Liu, Wenguan Wang, and Yi Yang. Vision-language navigation with energy-based policy. In *NeurIPS*, 2024.

[6] Rui Liu, Wenguan Wang, and Yi Yang. Volumetric environment representation for vision-language navigation. In *CVPR*, 2024. 2

[7] Sharmin Rahman, Alberto Quattrini Li, and Ioannis Rekleitis. Svin2: An underwater slam system using sonar, visual, inertial, and depth sensor. In *IROS*, 2019. 1

[8] Yelena Randall and Tali Treibitz. Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets. *arXiv preprint arXiv:2302.12772*, 2023. 1

[9] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IROS*, pages 573–580, 2012. 1

[10] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In *NeurIPS*, 2021. 1

[11] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *IROS*, 2020. 1

[12] Aming Wu, Rui Liu, Yahong Han, Linchao Zhu, and Yi Yang. Vector-decomposed disentanglement for domain-invariant object detection. In *ICCV*, 2021. 2

[13] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024. 2

[14] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. Migc++: Advanced multi-instance generation controller for image synthesis. *IEEE TPAMI*, 2024. 2

[15] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *CVPR*, pages 6818–6828, 2024.

[16] Dewei Zhou, Ji Xie, Zongxin Yang, and Yi Yang. 3dis: Depth-driven decoupled instance synthesis for text-to-image generation. *arXiv preprint arXiv:2410.12669*, 2024.

[17] Dewei Zhou, Mingwei Li, Zongxin Yang, and Yi Yang. Dreamrenderer: Taming multi-instance attribute control in large-scale text-to-image models. In *ICCV*, 2025. 2