WaterSplat-SLAM: Photorealistic Monocular SLAM in Underwater Environment

Kangxu Wang
Department of Electronic Engineering
Tsinghua University
Beijing, China
kx-wang23@mails.tsinghua.edu.cn

Shaofeng Zou
Key Laboratory of Marine Robotics
University of Chinese Academy of Sciences
Beijing, China
zoushaofeng23@mails.ucas.ac.cn

Chenxing Jiang
Department of Electronic and Computer Engineering
HKUST
Hong Kong, China
cjiangan@connect.ust.hk

Yixiang Dai
Department of Electronic Engineering
Tsinghua University
Beijing, China
daiyx23@mails.tsinghua.edu.cn

Siang Chen
Department of Electronic Engineering
Tsinghua University
Beijing, China
csa21@mails.tsinghua.edu.cn

Shaojie Shen
Department of Electronic and Computer Engineering
HKUST
Hong Kong, China
eeshaojie@ust.hk

Guijin Wang
Department of Electronic Engineering
Tsinghua University
Beijing, China
wangguijin@tsinghua.edu.cn

***Abstract*—Underwater monocular SLAM is a challenging problem with wide applications. However, existing methods struggle to produce maps with high-fidelity rendering. In this paper, we propose WaterSplat-SLAM, a novel monocular underwater SLAM system that achieves robust pose estimation and photorealistic dense mapping. Specifically, we couple semantic medium filtering into two-view 3D reconstruction prior to enable underwater-adapted camera tracking and depth estimation. Moreover, we present a semantic-guided rendering and adaptive map management strategy with an online medium-aware Gaussian map, modeling underwater environment in a photorealistic and compact manner. Experiments on multiple underwater datasets show that WaterSplat-SLAM achieves robust camera tracking and high-fidelity rendering in underwater environments.**

***Index Terms*—SLAM; Mapping; Marine Robotics**

I. Introduction

Online photorealistic dense mapping is crucial for underwater robotic applications. Prior visual underwater SLAM methods rely on handcrafted features [1]–[3], making them fragile in turbid, low-light scenes and prone to producing sparse maps. Some dense SLAM approaches [4]–[7] fuse structured light, sonar, or laser to build dense point clouds or occupancy maps, but these representations are not suited for photorealistic rendering. Recent 3DGS-SLAM methods [8]–[12] enable photorealistic mapping, but ignore underwater attenuation and scattering. These methods fail to model water as a light-scattering medium, degrading tracking stability and limiting downstream underwater tasks. To overcome these challenges, we propose WaterSplat-SLAM, achieving robust camera tracking and high-quality underwater scene modeling by RGB input. As shown in Fig. 1, our method extracts semantic medium information from images to suppress the adverse effects of water on camera tracking. We then incorporate semantic information into two-view 3D reconstruction prior to generating multi-view consistent depth maps and globally consistent camera poses, providing a high-quality initialization for the Gaussian map. To jointly model the underwater objects and the surrounding medium, we propose a semantic-guided rendering approach with an online medium-aware Gaussian map, avoiding misrepresenting the volumetric medium with Gaussian primitives. To ensure global consistency, we also adaptively adjust Gaussian primitives in response to loop closure and keyframe updates, and merge overlapping primitives across loop-closure frames to reduce redundancy. Across extensive experiments, it exceeds the NeRF-based and 3DGS-based SLAM methods in both tracking and mapping accuracy.

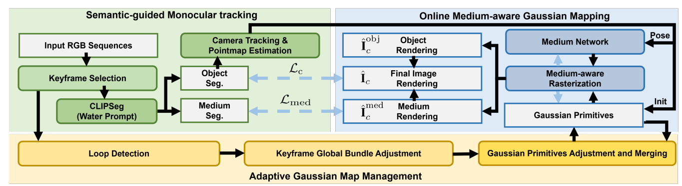

Fig. 1. System overview of WaterSplat-SLAM

II. Methods

A. Semantic-guided Monocular Tracking

We use MASt3R [13], which takes an image pair $\mathbf{I}_i, \mathbf{I}_j \in \mathbb{R}^{H\times W\times 3}$ as input and predicts pointmaps $\mathbf{X}_i^i, \mathbf{X}_i^j \in \mathbb{R}^{H\times W\times 3}$ together with confidence maps $\mathbf{Q}_i^i, \mathbf{Q}_i^j \in \mathbb{R}^{H\times W\times 1}$. Here, $\mathbf{X}_i^j$ denotes the pointmap of image i expressed in camera j's coordinate frame, and the same convention applies to $\mathbf{Q}$. To prevent water regions from affecting camera tracking, we use CLIPSeg [14] with a water prompt to obtain masks and set the confidences of masked pixels to zero. Given the last keyframe pointmap estimate $\tilde{\mathbf{X}}_k^k$, the current-frame pointmap estimate $\mathbf{X}_f^f$, and their correspondences $\mathbf{m}_{f,k}$, we estimate the camera pose $\mathbf{T}_{kf} \in \mathbf{Sim}(3)$ by minimizing $\sum_{(m,n)\in\mathbf{m}_{f,k}} \left\| q_{m,n}\left(\tilde{\mathbf{X}}_{k,n}^k - \mathbf{T}_{kf}\mathbf{X}_{f,m}^f\right)\right\|_\rho$, where $\|\cdot\|_\rho$ is the Huber norm and $q_{m,n} = \sqrt{\mathbf{Q}_{f,m}^f \mathbf{Q}_{f,n}^k}$, following [15]. This confidence weighting prevents low-texture or backscatter-dominated water areas from dominating pose estimation, while using monocular geometry priors for Gaussian initialization.

B. Online Medium-aware Gaussian Mapping

As shown in the left image of Fig. 2, objects are represented by anisotropic 3D Gaussian primitives [16], which are initialized from the tracking module. For a ray

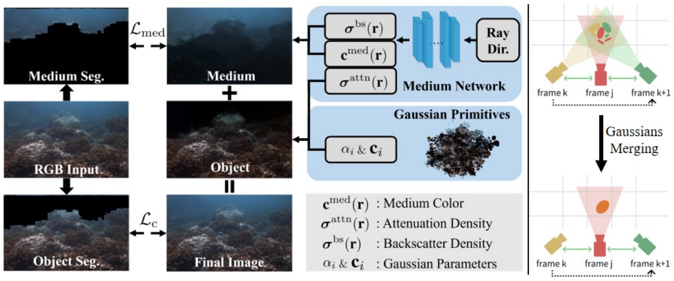

Fig. 2. Left: Online medium-aware gaussian mapping; Right: Gaussian primitives merging pipeline

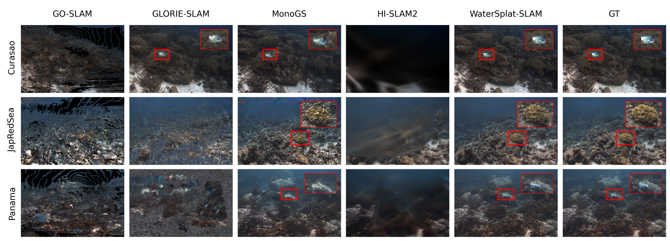

Fig. 3. Comparison of reconstruction results on SeaThru-NeRF dataset.

Fig. 4. Comparison of reconstruction results on WaterSplat-SLAM dataset.

$\mathbf{r}$, 3DGS renders the clear object color as $\hat{\mathbf{C}}^{\text{clr}}(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \mathbf{c}_i$, $T_i = \prod_{j<i}(1-\alpha_j)$, where α_i denotes the opacity obtained by projecting each Gaussian onto the image plane. We render an underwater image by separating scene objects from the underwater scattering medium. A Medium Network predicts the ray-dependent attenuation density $\boldsymbol{\sigma}^{\text{attn}}(\mathbf{r})$, backscatter density $\boldsymbol{\sigma}^{\text{bs}}(\mathbf{r})$, and medium color $\mathbf{c}^{\text{med}}(\mathbf{r})$. Accordingly, object rendering is given by $\hat{\mathbf{C}}^{\text{obj}}(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \exp(-\boldsymbol{\sigma}^{\text{attn}}(\mathbf{r}) t_i)\mathbf{c}_i$, medium rendering is given by $\hat{\mathbf{C}}^{\text{med}}(\mathbf{r}) = \mathbf{c}^{\text{med}}(\mathbf{r}) \sum_{i=1}^{N} T_i \big[\exp(-\boldsymbol{\sigma}^{\text{bs}}(\mathbf{r}) t_{i-1}) - \exp(-\boldsymbol{\sigma}^{\text{bs}}(\mathbf{r}) t_i)\big]$, and the final rendered image is obtained as $\hat{\mathbf{C}}^{\text{render}}(\mathbf{r}) = \hat{\mathbf{C}}^{\text{obj}}(\mathbf{r}) + \hat{\mathbf{C}}^{\text{med}}(\mathbf{r})$. New Gaussians are inserted only in non-water regions with insufficient coverage: $\mathbf{X}_k^{\text{add}} = \mathbf{X}_k \odot (\mathbf{1} - \mathbf{M}_k) \odot (\mathbf{A} < \tau_A)$, where $\mathbf{A}$ denotes the cumulative opacity. Given the water mask $\mathbf{M}_c$, the semantic photometric loss is defined as $\mathcal{L}_{\text{sempho}} = (1-\lambda_{\text{ssim}})(\mathcal{L}_{\text{med}} + \mathcal{L}_c) + \lambda_{\text{ssim}}(1 - \text{SSIM}(\hat{\mathbf{I}}_c, \mathbf{I}_c))$, where $\mathcal{L}_{\text{med}} = |\mathbf{M}_c \odot (\hat{\mathbf{I}}_c^{\text{med}} - \mathbf{I}_c)|_1$ and $\mathcal{L}_c = |(\mathbf{1} - \mathbf{M}_c) \odot (\hat{\mathbf{I}}_c - \mathbf{I}_c)|_1$. With visible primitives VIS_c, we regularize the shape using $\mathcal{L}_s = \sum_{i \in \text{VIS}_c} |\mathbf{S}_i - \bar{\mathbf{S}}_i|$, and the final mapping loss is $\mathcal{L} = \lambda_{\text{sempho}} \mathcal{L}_{\text{sempho}} + \lambda_s \mathcal{L}_s$.

C. Adaptive Gaussian Map Management

After loop closure and global BA, the keyframe poses are updated from $\mathbf{T}_k = [s_k \mathbf{R}_k, \mathbf{t}_k]$ to $\mathbf{T}'_k = [s'_k \mathbf{R}'_k, \mathbf{t}'_k]$. For each Gaussian anchored to keyframe k, we update its geometry as $\mathbf{S}'_i = \frac{s'_k}{s_k}\mathbf{S}_i, \mathbf{R}'_i = \mathbf{R}'_k \mathbf{R}_k^{-1} \mathbf{R}_i, \boldsymbol{\mu}'_i = \mathbf{T}'_k \mathbf{T}_k^{-1} \boldsymbol{\mu}_i$. A keyframe is re-rendered if its pose change satisfies $\Delta d > d_{\text{th}}$ or $\Delta\theta > \theta_{\text{th}}$. If the ratio of re-rendered keyframes exceeds $GLOMAP_{\text{th}}$, global mapping is triggered. For loop-closure pairs, we merge redundant primitives by voxelizing all Gaussians anchored to the current and matched historical frames using voxel size v. As shown in the right image of Fig. 2, all primitives within the same non-empty voxel are replaced by a single representative Gaussian, whose color, scale, orientation, position, and opacity are averaged, while its anchor is randomly selected from the merged anchors. This procedure keeps the map globally consistent and compact while preserving rendering quality.

TABLE I
AVERAGE RENDERING QUALITY ON TWO DATASETS.

Method	Type	SeaThru-NeRF Avg			WaterSplat-SLAM Avg		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GO-SLAM [18]	NeRF	15.74	0.267	0.831	14.56	0.293	0.714
GLORIE-SLAM [19]		18.18	0.470	0.538	20.51	0.658	0.557
MonoGS [8]	3DGS	18.04	0.558	0.584	23.32	0.827	0.648
HI-SLAM2 [11]		13.42	0.295	0.830	23.90	0.762	0.539
OpenGS-SLAM [20]		18.35	0.613	0.646	23.89	0.816	0.578
S3PO-GS [21]		19.06	0.593	0.503	26.68	0.789	0.552
WaterSplat-SLAM		**24.14**	**0.696**	**0.335**	**30.19**	**0.835**	**0.269**

TABLE II
TRACKING ACCURACY (ATE(M) ↓) ON WATERSPLAT-SLAM DATASET.

Method	Big_gate	Pipe_local	Pool_up	Pool_up2	Pool_loop	Avg
ORB-SLAM3 [22]	0.091	2.768	Fail	Fail	Fail	×
GO-SLAM [18]	0.041	0.292	2.532	0.205	3.231	1.260
GLORIE-SLAM [19]	0.036	0.236	Fail	**0.086**	Fail	×
MonoGS [8]	1.729	3.472	2.899	2.578	3.240	2.784
HI-SLAM2 [11]	**0.034**	**0.135**	2.533	1.213	2.343	1.252
OpenGS-SLAM [20]	Fail	Fail	2.063	2.249	Fail	×
S3PO-GS [21]	0.114	0.592	1.593	0.834	Fail	×
WaterSplat-SLAM	0.066	0.164	**0.289**	0.224	**0.353**	**0.219**

TABLE III
ABLATION STUDY ON *Pool_loop* OF WATERSPLAT-SLAM DATASET.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	ATE ↓
	w/o Water Mask	27.21	0.797	0.353	0.375
Pool_loop	w/o Primitives Adjustment	26.56	0.790	0.376	/
	WaterSplat-SLAM	**27.69**	**0.808**	**0.326**	**0.353**

III. EXPERIMENTS

We evaluate on the SeaThru-NeRF [17] and the self-collected WaterSplat-SLAM datasets against NeRF-SLAM methods [18], [19], Gaussian-SLAM methods [8], [11], [20], [21], and ORB-SLAM3 [22]. Our method achieves the best average rendering quality and tracking accuracy, as shown in Tabs. I and II, with ablation studies in Tab. III. Moreover, Figs. 3 and 4 show qualitative comparisons. On an RTX 4090/i9-13900K desktop, our method runs at 3 FPS, matching HI-SLAM2 [11] while producing higher-quality renderings.

IV. CONCLUSION

We propose WaterSplat-SLAM, a novel monocular underwater SLAM system based on medium-aware 3DGS, achieving robust camera tracking and photorealistic rendering in underwater scenes. The semantic-guided water rendering method achieves high-quality reconstruction of both objects and scattering media. The voxel-based merging of Gaussian primitives upon loop closure reduces redundancy between associated frames, making our medium-aware Gaussian map more efficient. Above all, WaterSplat-SLAM establishes a new foundation for aquatic spatial intelligence systems.

REFERENCES

[1] J. Drupt, C. Dune, A. I. Comport, and V. Hugel, “Qualitative evaluation of state-of-the-art dso and orb-slam-based monocular visual slam algorithms for underwater applications,” in *OCEANS 2023 - Limerick*, 2023, pp. 1–7.

[2] F. Hidalgo, C. Kahlefendt, and T. Bräunl, “Monocular orb-slam application in underwater scenarios,” in *2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*, 2018, pp. 1–4.

[3] H. Zhao, R. Zheng, M. Liu, and S. Zhang, “Detecting loop closure using enhanced image for underwater vins-mono,” in *Global Oceans 2020: Singapore–US Gulf Coast*. IEEE, 2020, pp. 1–6.

[4] Y. Wang, Y. Ji, H. Woo, Y. Tamura, H. Tsuchiya, A. Yamashita, and H. Asama, “Acoustic camera-based pose graph slam for dense 3-d mapping in underwater environments,” *IEEE Journal of Oceanic Engineering*, vol. 46, no. 3, pp. 829–847, 2020.

[5] Y. Ou, J. Fan, C. Zhou, P. Zhang, and Z.-G. Hou, “Hybrid-vins: Underwater tightly coupled hybrid visual inertial dense slam for auv,” *IEEE Transactions on Industrial Electronics*, vol. 72, no. 3, pp. 2821–2831, 2025.

[6] Y. Ou, J. Fan, C. Zhou, P. Zhang, Z. Shen, Y. Fu, X. Liu, and Z. Hou, “An underwater, fault-tolerant, laser-aided robotic multi-modal dense slam system for continuous underwater in-situ observation,” *arXiv preprint arXiv:2504.21826*, 2025.

[7] Y. Ou, J. Fan, C. Zhou, S. Kang, Z. Zhang, Z.-G. Hou, and M. Tan, “Structured light-based underwater collision-free navigation and dense mapping system for refined exploration in unknown dark environments,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 55, no. 1, pp. 110–123, 2024.

[8] H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison, “Gaussian splatting slam,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 18 039–18 048.

[9] V. Yugay, Y. Li, T. Gevers, and M. R. Oswald, “Gaussian-slam: Photo-realistic dense slam with gaussian splatting,” 2024. [Online]. Available: https://arxiv.org/abs/2312.10070

[10] H. Huang, L. Li, H. Cheng, and S.-K. Yeung, “Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular stereo and rgb-d cameras,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 584–21 593.

[11] W. Zhang, Q. Cheng, D. Skuddis, N. Zeller, D. Cremers, and N. Haala, “Hi-slam2: Geometry-aware gaussian slam for fast monocular scene reconstruction,” *arXiv preprint arXiv:2411.17982*, 2024.

[12] B. Lee, J. Park, K. T. Giang, S. Jo, and S. Song, “Mvs-gs: High-quality 3d gaussian splatting mapping via online multi-view stereo,” 2024. [Online]. Available: https://arxiv.org/abs/2412.19130

[13] V. Leroy, Y. Cabon, and J. Revaud, “Grounding image matching in 3d with mast3r,” in *European Conference on Computer Vision*. Springer, 2024, pp. 71–91.

[14] T. Lüddecke and A. Ecker, “Image segmentation using text and image prompts,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 7086–7096.

[15] B. Duisterhof, L. Zust, P. Weinzaepfel, V. Leroy, Y. Cabon, and J. Revaud, “Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion,” *arXiv preprint arXiv:2409.19152*, 2024.

[16] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, “3d gaussian splatting for real-time radiance field rendering.” *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[17] D. Levy, A. Peleg, N. Pearl, D. Rosenbaum, D. Akkaynak, S. Korman, and T. Treibitz, “Seathru-nerf: Neural radiance fields in scattering media,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 56–65.

[18] Y. Zhang, F. Tosi, S. Mattoccia, and M. Poggi, “Go-slam: Global optimization for consistent 3d instant reconstruction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 3727–3737.

[19] G. Zhang, E. Sandström, Y. Zhang, M. Patel, L. Van Gool, and M. R. Oswald, “Glorie-slam: Globally optimized rgb-only implicit encoding point cloud slam,” *arXiv preprint arXiv:2403.19549*, 2024.

[20] S. Yu, C. Cheng, Y. Zhou, X. Yang, and H. Wang, “Rgb-only gaussian splatting slam for unbounded outdoor scenes,” *arXiv preprint arXiv:2502.15633*, 2025.

[21] C. Cheng, S. Yu, Z. Wang, Y. Zhou, and H. Wang, “Outdoor monocular slam with global scale-consistent 3d gaussian pointmaps,” *arXiv preprint arXiv:2507.03737*, 2025.

[22] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, “Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam,” *IEEE transactions on robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.