Loggerhead: Grounding Feed-forward Transformers with Multimodal Odometry for Metric-scale Reef Reconstructions

Daniel Yang[1,2], Jungseok Hong[1], John J. Leonard[1], and Yogesh Girdhar[2]

***Abstract*— Accurate underwater SLAM remains challenging due to the absence of GPS, limited visibility, and drift accumulation from dead-reckoning. Advances in feed-forward 3D reconstruction provide dense geometric pointmaps with relative pose estimates from long trajectories of monocular imagery, but their estimates are up to scale and drift over long horizons. In this work, we present Loggerhead, a graph-based trajectory optimization approach that grounds these approaches with metric, underwater vehicle sensor data. We integrate into a unified optimization framework the odometry estimated from multimodal sensor inputs as well as visual loop closures with constraints derived from the chunk-based processing of long horizon feed-forward 3D reconstruction models. We evaluate the proposed approach on three underwater reef trajectories collected in the U.S. Virgin Islands and demonstrate more accurate trajectory estimation compared to both odometry and feedforward reconstruction alone while providing a unified, coarse pointmap representation of the scene.**

I. Introduction

Advances in underwater vehicles have enabled their deployment in challenging underwater environments for tasks including environmental monitoring and exploration [1]–[5]. However, accurate state estimation is still challenging in these GPS-denied environments. While acoustic positioning systems, such as USBL, are available, they require complex deployment of additional instrumentation in the environment. As a result, localization often relies on SLAM methods that use combinations of acoustic (e.g. Doppler Velocity Log (DVL)), inertial, visual, and barometric sensors [6], [7]. While these sensors provide locally accurate pose estimates, drift accumulates over long trajectories.

Advances in foundation models for visual reconstruction have enabled estimation of camera poses and scene geometry from monocular images [8]–[11]. These models are computationally expensive, requiring high amounts of memory per image and scaling quadratically with pairwise token comparisons. More recent work addresses this issue for long sequences by dividing the input into more computationally tractable windows and aligning these with optimization [12], [13] or by introducing architectural changes to maintain continuity [14]. However, they still rely on the backbone of the model to perform high quality feature matching and be robust on the input data distribution and still can accumulate drift across long trajectories. These issues are exacerbated with underwater data, which is subject to visual degradation through medium and other environmental effects.

In this paper, we present Loggerhead, coined after the sea turtle, which integrates classical graph-based optimization

[1]Massachusetts Institute of Technology [2]Woods Hole Oceanographic Institution

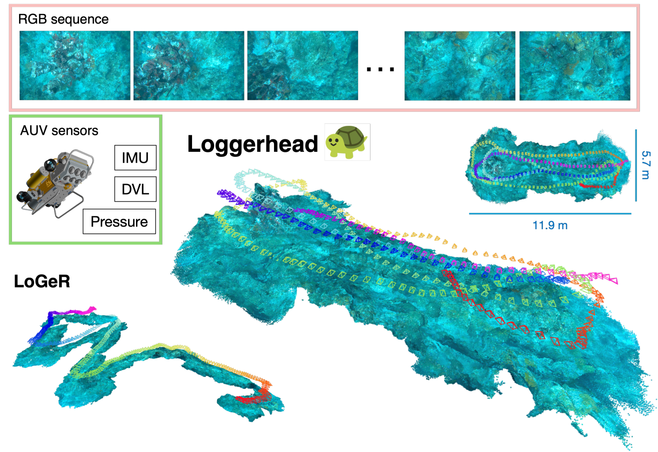

Fig. 1: **Loggerhead** grounds feed-forward transformers with metric odometry derived from multimodal sensors common on underwater vehicles. This enables rapid processing of long sequences of RGB video with minimal drift and pose error accumulation. Output of Loggerhead and LoGeR [14] shown on the same DendrogyraLoop trajectory.

with a feed-forward 3D reconstruction model, LoGeR [14], to achieve grounded, metric underwater SLAM. Instead of blindly relying on a feed forward model's predictions, we derive geometric constraints for our pose graph grounded in odometry data obtained from acoustic, inertial, and pressure sensors typically found on underwater vehicles. We further incorporate loop closure information estimated from image retrieval and local reconstruction with these feed-forward models, enabling global trajectory improvements.

Our contributions are:

1) Loggerhead, a pose-graph formulation that incorporates scale-aware odometry with relative poses derived from a feed-forward 3D reconstruction model.
2) Loop closure detection combining visual place recognition, feature matching, and local feed-forward reconstruction.
3) Experimental validation on three real-world trajectories collected in the U.S. Virgin Islands.

II. Method

A. Odometry Estimation

We fuse inertial, velocity, and depth measurements through an extended Kalman filter (EKF) [15], thus providing the metric backbone of our pose graph. We also run off-the-shelf LoGeR [14] across our full image sequence to obtain an unscaled estimate of our trajectory, from which we can directly utilize the rotation estimates but model the translation with an unknown scale factor, shared with all other relative translations from a window of input, that can be estimated and optimized later, $t_{ij} = s\hat{t}_{ij}$ where $\hat{t}_{ij}$. EKF

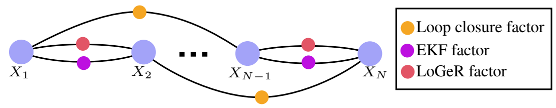

Fig. 2: We integrate odometry estimated from multimodal sensor data with loop closure and LoGeR-derived constraints into a factor-graph to estimate vehicle state as well as the scale needed to utilize LoGeR pointmaps.

and LoGeR estimates are incorporated into the pose graph as $f_{\text{EKF}}(X_i, X_j)$ and $f_{\text{LoGeR}}(X_i, X_j)$, respectively.

B. Loop Closure Detection

Similar to other approaches designed to handle long trajectories [13], we utilize SALAD [16] features to detect loop closure candidates for minimizing long-term drift. Candidate image pairs are subsequently verified through geometric feature matching to reject false positives. For each verified loop closure pair, temporally neighboring frames are selected and processed, together as a single window by LoGeR independently of the main temporally sequential chain of input. This provides a relative pose constraint, similar to the above except the scale factor is only associated with the edge on which we add a constraint instead of a whole window of poses, on temporally distant but spatially neighboring portions of the trajectory. The resulting loop closure constraints are added to the graph as $f_{\text{Loop}}(X_i, X_j)$.

C. Pose-graph optimization

We formulate the pose estimation problem as a pose graph optimization problem, in which nodes represent camera poses and edges encode relative pose constraints. The factor graph is composed of three factor types: $\mathcal{F}_{\text{LoGeR}} \cup \mathcal{F}_{\text{EKF}} \cup \mathcal{F}_{\text{Loop}}$. Given the set of measurements $\mathcal{Z}$, the camera pose $\mathcal{X}$ and per-window scale $\mathcal{S}$ is estimated via Maximum a Posteriori (MAP) inference.

$$\mathcal{X}^\star, \mathcal{S}^\star = \underset{\mathcal{X} \in SE(3), \mathcal{S} \in \mathbb{R}}{\arg\max} \; p(\mathcal{X}, \mathcal{S} \mid \mathcal{Z}). \tag{1}$$

III. Experiments

A. Dataset collection

We evaluate Loggerhead on three trajectories of coral reef benthic survey data collected in the US Virgin Islands. These sites are biologically active and contain a patchy mix of sand, seagrass, and hard and soft corals [17]. Our sensor suite includes a Nortek Nucleus 1000 DVL, which includes an IMU and pressure sensor, and a downward-facing ZED X Mini camera with a flat viewport. For this work we utilize only the monocular output of the stereo camera. We perform in-water calibration with Kalibr [18] and extract rectified imagery at 5 fps at 1920 x 1200 resolution. We refer to our trajectories as **Lawnmower** (7029 frames over 520.0 m), **DendrogyraLoop** (923 frames over 69.6 m), and **TagLoop** (513 frames over 42.2 m) as shown in Figs. 1 and 3.

B. Evaluation metrics

We report absolute trajectory error (ATE) on each trajectory. We approximate the ground-truth reference using Metashape [19], scaled based on fiducial markers in the scene, and use evo [20] to align to the reference before computing the ATE RMSE in meters. For comparison with LoGeR [14], we allow for a scale correction, but not with our method and the EKF.

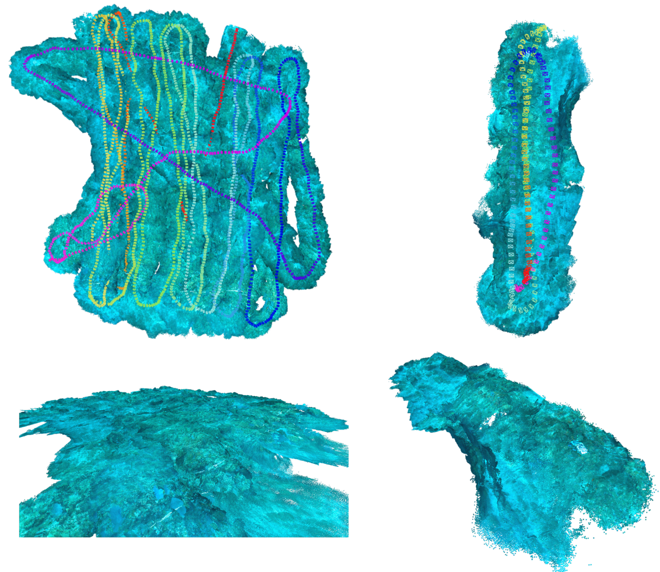

Fig. 3: **Qualitative evaluation** Output of Loggerhead on the Lawnmower and TagLoop trajectories, showing the top down view with trajectory overlaid as well as off-angle views to see the 3D information, derived from the LoGeR [14] pointmaps, present.

Method \Trajectory num frames, scale	Lawnmower 7029, 553.8 m	DendrogyraLoop 923, 69.6 m	TagLoop 513, 44.1 m
EKF	8.95	0.42	0.27
LoGeR	8.73	2.99	0.61
Loggerhead	**3.41**	**0.19**	**0.26**

TABLE I: **Trajectory estimation metrics.** ATE RMSE (lower is better ↓) is calculated in meters. First-place results are **bolded**.

IV. Results

As seen in Tab. I, Loggerhead achieves lower ATE RMSE than both the raw EKF data alone as well as post-scaled and aligned LoGeR [14]. As seen in Figs. 1 and 3, we recover the desired shape of the trajectories (a lawnmower or looping transects) and can leverage the pointmap output of LoGeR to provide a crude proxy for geometric information. Specifically, visual seams of incoherence between different pointmaps as well as the lack of precise depth from downward-facing imagery can be seen, highlighting a space for future improvement. The failure mode of trajectory drift of feed forward transformer approaches is shown in Fig. 1

V. Conclusions

We present Loggerhead , a pose graph-based underwater SLAM approach that integrates feed-forward model outputs, odometry estimates from multimodal sensor fusion, and loop closure constraints obtained through visual place recognition models. Experiments on challenging real-world underwater datasets demonstrate that combining learned geometric, even on out-of-distribution scenarios, priors with grounded, classical pose graph optimization significantly reduces drift. Future work will focus on more robust loop closures and improving geometric reconstruction quality.

References

[1] Y. Girdhar, N. McGuire, L. Cai, S. Jamieson, S. McCammon, B. Claus, J. E. S. Soucie, J. E. Todd, and T. A. Mooney, "Curee: A curious underwater robot for ecosystem exploration," in *2023 IEEE International Conference on Robotics and Automation*, 2023, pp. 11 411–11 417.

[2] B. Joshi, M. Xanthidis, M. Roznere, N. J. Burgdorfer, P. Mordohai, A. Q. Li, and I. Rekleitis, "Underwater exploration and mapping," in *2022 IEEE/OES Autonomous Underwater Vehicles Symposium (AUV)*. IEEE, 2022, pp. 1–7.

[3] M. Johnson-Roberson, M. Bryson, A. Friedman, O. Pizarro, G. Troni, P. Ozog, and J. C. Henderson, "High-resolution underwater robotic vision-based mapping and three-dimensional reconstruction for archaeology," *Journal of Field Robotics*, vol. 34, no. 4, pp. 625–643, 2017.

[4] F. F. R. Merveille, B. Jia, Z. Xu, and B. Fred, "Advancements in sensor fusion for underwater slam: A review on enhanced navigation and environmental perception," *Sensors (Basel, Switzerland)*, vol. 24, no. 23, p. 7490, 2024.

[5] S. Williams and I. Mahon, "Simultaneous localisation and mapping on the great barrier reef," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, vol. 2, 2004, pp. 1771–1776 Vol.2.

[6] S. Rahman, A. Q. Li, and I. Rekleitis, "Svin2: An underwater slam system using sonar, visual, inertial, and depth sensor," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1861–1868.

[7] J. Song, O. Bagoren, R. Andigani, A. Sethuraman, and K. A. Skinner, "Turtlmap: Real-time localization and dense mapping of low-texture underwater environments with a low-cost unmanned underwater vehicle," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 1191–1198.

[8] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 697–20 709.

[9] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 5294–5306.

[10] N. Keetha, N. Müller, J. Schönberger, L. Porzi, Y. Zhang, T. Fischer, A. Knapitsch, D. Zauss, E. Weber, N. Antunes, J. Luiten, M. Lopez-Antequera, S. R. Bulò, C. Richardt, D. Ramanan, S. Scherer, and P. Kontschieder, "MapAnything: Universal feed-forward metric 3D reconstruction," in *International Conference on 3D Vision (3DV)*. IEEE, 2026.

[11] Y. Wang, J. Zhou, H. Zhu, W. Chang, Y. Zhou, Z. Li, J. Chen, J. Pang, C. Shen, and T. He, "pi3: Permutation-equivariant visual geometry learning," *arXiv preprint arXiv:2507.13347*, 2025.

[12] D. Maggio, H. Lim, and L. Carlone, "Vggt-slam: Dense rgb slam optimized on the sl (4) manifold," *Advances in Neural Information Processing Systems*, vol. 39, 2025.

[13] K. Deng, Z. Ti, J. Xu, J. Yang, and J. Xie, "Vggt-long: Chunk it, loop it, align it – pushing vggt's limits on kilometer-scale long rgb sequences," 2025. [Online]. Available: https://arxiv.org/abs/2507.16443

[14] J. Zhang, C. Herrmann, J. Hur, C. Sun, M.-H. Yang, F. Cole, T. Darrell, and D. Sun, "Loger: Long-context geometric reconstruction with hybrid memory," *arXiv preprint arXiv:2603.03269*, 2026.

[15] T. Moore and D. Stouch, "A generalized extended kalman filter implementation for the robot operating system," in *Intelligent Autonomous Systems 13: Proceedings of the 13th International Conference IAS-13*. Springer, 2016, pp. 335–348.

[16] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

[17] S. McCammon, S. Jamieson, T. A. Mooney, and Y. Girdhar, "Discovering biological hotspots with a passively listening auv," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 3789–3795.

[18] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1280–1286.

[19] "Agisoft Metashape: Professional Edition." [Online]. Available: https://www.agisoft.com/features/professional-edition/

[20] M. Grupp, "evo: Python package for the evaluation of odometry and slam." https://github.com/MichaelGrupp/evo, 2017.