OPENNAVMAP: Multi-Session Structure-Free Topometric Mapping for Scalable Visual Navigation

Jianhao Jiao[1], Changkun Liu[2], Jingwen Yu[3], Dimitrios Kanoulas[1,4]

***Abstract*—This paper proposes OPENNAVMAP, a multi-session mapping system designed for scalable visual navigation. Rather than relying on the 3D structure-based representation of the environment, OPENNAVMAP adopts a robust collaborative localization strategy to facilitate map merging, taking only 2D images as input. The resulting topometric map is thus lightweight and structure-free, composed of three layered graphs: odometry, covisibility, and traversability. This design enables autonomous visual navigation without the need for prior structure-based maps. Experiments on map merging demonstrate that OPENNAVMAP achieves high accuracy ($< 3m$ ATE over $15km$) and strong robustness to challenging conditions such as day-night transitions and large viewpoint changes. The system has been successfully deployed on a quadruped robot using only monocular RGB inputs for image-goal visual navigation. A video is provided to explain the methodology and experimental results[1].**

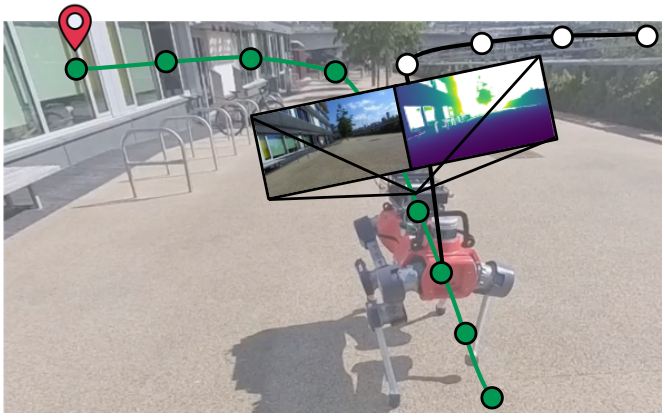


Fig. 1. Visual navigation with the proposed structure-free topometric map. Green dots show the path to the goal, white dots other nodes.

I. INTRODUCTION

A. Motivation

Robots navigating daily tasks require map representations that are expressive, scalable, and easy to maintain. Traditional dense 3D maps are computationally heavy, difficult to update, and brittle under environmental changes. To address this, we propose a mapping framework that incrementally integrates maps from multiple sessions [1], emphasizing scalability and robustness for long-term visual navigation.

B. Challenges

Constructing scalable maps across diverse environments remains challenging.

1) Map Representation: Conventional methods for visual navigation, such as SfM [2], SLAM [3], [4], or volumetric mapping [5], generate accurate but storage-heavy maps or lightweight ones with reduced precision. Both are hard to maintain as environments evolve. Inspired by the map-free benchmark [6], we adopt a sparse, structure-free representation that avoids explicit 3D priors and improves adaptability.

2) Collaborative Localization: Merging submaps from multiple sessions presents a significant challenge, particularly due to the limited overlap inherent in crowd-sourced data. This limitation stems from the lack of a unified collection strategy, which yields sparse and non-uniform environmental coverage, thereby restricting the common geometric features necessary for robust map registration. This raises the need for robust mechanisms to handle: *1) Spatial Scalability*: adapting to diverse environments and arbitrary viewpoints; and *2) Temporal Adaptability*: ensuring robustness against environmental dynamics, such as movable objects and diurnal variations [7].

3) Scalable Mapping Data Acquisition: High-end LiDARs and expert calibration [2], [8] limit scalability in mapping data acquisition. In contrast, consumer devices (*e.g.,* smartphones, AR/VR headsets) enable *crowd-sourced mapping* [9], [10], though they introduce sensor diversity and perceptual noise. Unlike OSM or Google Street View [11], robotic navigation requires sub-meter accuracy in GNSS-denied settings.

C. Contributions

We present OPENNAVMAP, a collaborative multi-session mapping system for visual relocalization (VLoc) and navigation (VNav). The term "Open", reflects our commitment to collaborative mapping construction inspired by platforms such as OpenStreetMap [12], where crowd-sourced data aggregation enables extensible environmental representation. OPENNAVMAP uses a lightweight *topometric map* organized as three layers: odometry, covisibility, and traversability. Each node stores 6-DoF poses, while edges capture spatial and semantic constraints. Unlike 3D point clouds or voxel grids [2]–[4], our structure-free map avoids heavy storage. Using a SoTA stereo reconstruction model [13], geometry

[1]Department of Computer Science, University College London, Gower Street, WC1E 6BT, London, UK. {ucacjji, d.kanoulas}@ucl.ac.uk.

[2]Department of Computer Science and Engineering, HKUST, China.

[3]Department of Electronic and Computer Engineering, HKUST, China.

[4]Dimitrios Kanoulas is also with the AI Centre, Department of Computer Science, University College London, Gower Street, WC1E 6BT, London, UK and Archimedes/Athena RC, Greece.

This work was supported by the UKRI Future Leaders Fellowship [MR/V025333/1] (RoboHike).

[1]https://drive.google.com/file/d/1bFKZstoTOoO_OOAB6hvKeVK5O_q2e-zq/view?usp=drive_link

is reconstructed on demand, retaining metric precision while remaining lightweight and editable. This enables efficient updates under dynamic changes. OpenNavMap is multi-session mapping system with a structure-free representation, explicitly designed for scalable VLoc and VNav. It requires only monocular images and scale-aware poses, achieving below $3m$ ATE across $15km$ of real-world data collected over 10 months. The system also supports cross-device localization and has been deployed on simulated and real robots (Fig. 1), completing multiple autonomous image-goal navigation tasks under significant perceptual variation.

II. Methodology

A. Scene Representation

The resulting topometric map $\mathcal{M}^W = \{\mathcal{G}_C^W, \mathcal{G}_O^W, \mathcal{G}_T^W\}$ represents scenes using three distinct graphs in a global world frame $\{\}^W$: **Covisibility Graph** *(CvG)*, **Odometry Graph** *(OdG)*, and **Traversability Graph** *(TrG)*, as illustrated in Fig. 2. Each graph serves a specific function. CvG supports VLoc. It connects nodes based on visual co-visibility and stores global descriptors for VPR, along with raw images and their associated poses. OdG functions as a factor graph, encoding odometry measurements and loop closures. Two nodes are connected if their relative transformations are available either from odometry data or from relative pose estimation. TrG is utilized for motion planning and connects nodes that are mutually traversable. Nodes store their poses, while edges store traversability costs. Importantly, even if two nodes have pose constraints and are visually co-visible, they are not necessarily traversable if an obstacle obstructs their direct path.

We describe our collaborative localization approach, which integrates submaps generated independently by multiple devices. Modeling the OdG as a factor graph [14], the problem is addressed in three steps: *1)* submap construction, *2)* relative pose estimation across submaps, and *3)* pose graph optimization (PGO) for global consistency.

B. Submap Construction

Each submap consists of CvG, OdG, and TrG graphs containing nodes and edges. Devices generate submaps during capture sessions using VIO, which fuses camera imagery and IMUs for locally consistent, scale-aware poses. This setup is compatible with smartphones, AR glasses [15], multi-sensor SLAM robots [4], 360° cameras, or vehicles with wheel odometry [16]. Submaps can be derived from: *1)* SfM/SLAM pipelines [17], *2)* odometry-based robots, or *3)* geo-referenced image repositories. This flexibility supports long-term, scalable mapping across diverse sources.

C. Topological-Level Localization

Without GPS, topological localization relies on visual place recognition (VPR). Robustness requires rejecting false matches (high precision) while retaining correct but rare matches (high recall). Our approach integrates: *1)* **descriptor extraction**: global descriptors from images using CosPlace [18] (ResNet-18 backbone, 256-D output); *2)* **sequence-based matching**: align descriptor sequences via a DP-based shortest-path search through the difference matrix,

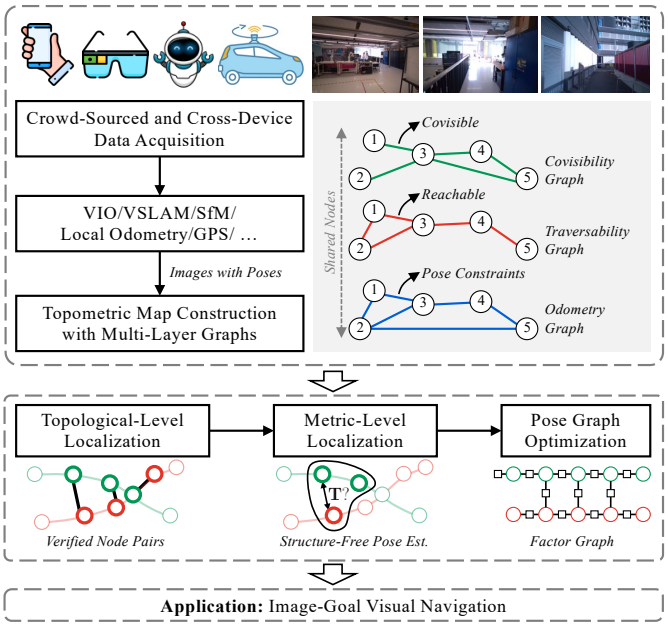


Fig. 2. Block diagram illustrating the pipeline of the proposed OpenNavMap system. The system builds on the topometric map with multiple layers for different utilities: covisbility, traversibility, and odometry. a) Individual submaps can be constructed from data collected by various types of devices. b) The collaborative localization consists of two steps to compute the relative transformation between the reference and query map, and then perform the PGO to jointly estimate their transformation. c) The resulting topometric map is deployed for image-goal navigation.

enabling flexible in-sequence and jump operations [19], [20]; *3)* **geometric verification**: validate correspondences via RANSAC-based fundamental matrix estimation, producing high-confidence pairs $\mathcal{P}_{GV}$.

D. Metric-Level Localization

We refine inter-submap transformations using a stereo reconstruction network [13] to predict geometry and poses from image pairs. Multiple references are integrated via global optimization of 3D pointmaps and camera parameters, improving accuracy even with sparse overlaps. Confidence maps are calibrated into the optimization process, weighting residuals to downplay unreliable predictions. High-confidence pairs $\mathcal{P}_{ML}$ are retained as loop-closure constraints.

E. Pose Graph Optimization

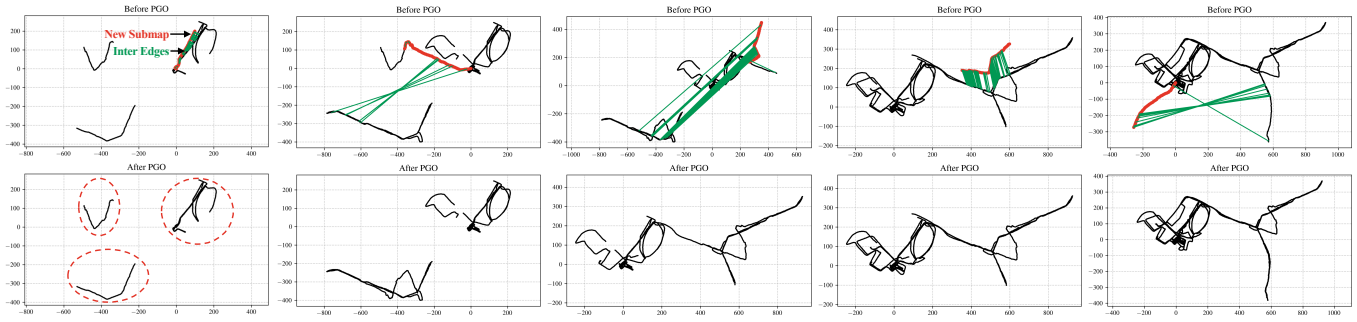
Finally, loop-closure constraints from $\mathcal{P}_{ML}$ are incorporated into the odometry graph $\mathcal{G}_O^R$. We solve a robust nonlinear least-squares PGO:

$$\{\mathbf{T}_i^R\}^* = \arg\min_{\mathbf{T}_i^R} \sum_{\mathbf{e}_{O_{i,j}}^R \in \mathcal{E}_O^R} \rho\Big(\|\log[(\bar{\mathbf{T}}_j^i)^{-1}(\mathbf{T}_i^R)^{-1}\mathbf{T}_j^R]\|_{\mathbf{\Sigma}_j^i}^2\Big), \tag{1}$$

using GTSAM [14] for efficient optimization. The result is a unified, globally consistent map.

III. Datasets

This section describes our experimental setup and datasets. The collaborative mapping and localization system is implemented in Python. All experiments, except the real-world



(a) 10 submaps (b) 22 submaps (c) 27 submaps (d) 36 submaps (e) 55 submaps

Fig. 3. The incremental map merging process is performed with the incoming submaps, where the order of submaps are randomly shuffled and they may present non-overlapping at the beginning, as seen in (a), where disconnected submaps present, but not affect the pose graph optimization. The data used is S2-R4 for the example. Green lines indicate the reliable loop factors.

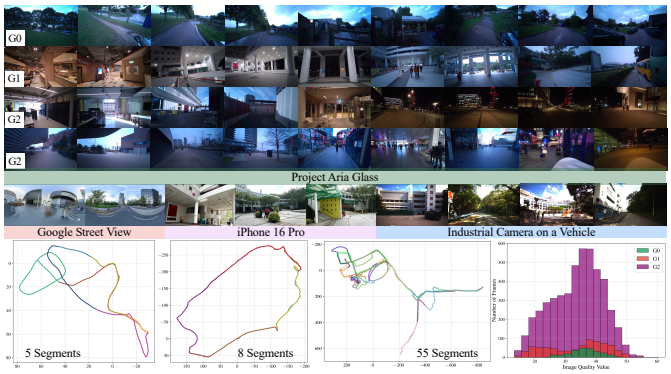


Fig. 4. Overview of our dataset collected with Aria glasses across offices, campuses, shopping centers, and vineyards in two countries, spanning 9 months, 37 sequences, and $19km$ of trajectories.

VNav deployment, are run on a desktop with an Intel i9 CPU and NVIDIA RTX 4090 GPU. Models are used without additional fine-tuning. We collected data using Meta Project Aria glasses [15], equipped with grayscale cameras, a front RGB camera, IMU, and GPS. Data span two countries, 9 months, 37 sequences, and $19km$ of trajectories across offices, campuses, shopping centers, parks, and vineyards. Sequences are grouped as: G0: a vineyard session; G1: campus recordings (day and night); G2: large-scale urban areas recorded over three months, capturing long-term variations.

Five users with varied behaviors contributed, introducing challenges such as wide viewpoint changes and irregular motion. Ground-truth poses were generated using Meta's cloud-based SLAM, and images were anonymized. A dataset overview is shown in Fig. 4. For map merging, we use all G0-G2 sequences. Each sequence is divided into segments of up to $300m$, with keyframes sampled by a distance threshold[2], yielding 68 segments. For each, we provide intrinsics, extrinsics, timestamps, VIO poses, and GT poses.

IV. Experimental Results

We evaluate our system in two parts: *1)* map merging across different devices and *2)* integration into a closed-loop VNav system tested in simulation and real robots. Overall, results show that even with a structure-free map, our approach achieves robust localization and scalable navigation.

[2]Translation and rotation thresholds: $3.9m$ and $60°$.

TABLE I
Theoretical map size per image and relative ratio of baselines compared to ours.

Methods	Parameters	Map Size $[MB]$
Ours	$H = 512, W = 288, C \in (0, 1]$	$0.423CN$
Hloc (DISK+LG)	$M = 5000, D = 128$	$1.22N$ ($\frac{2.89}{C}\times$)
Hloc (SP+LG)	$M = 4096, D = 256$	$2N$ ($\frac{4.66}{C}\times$)

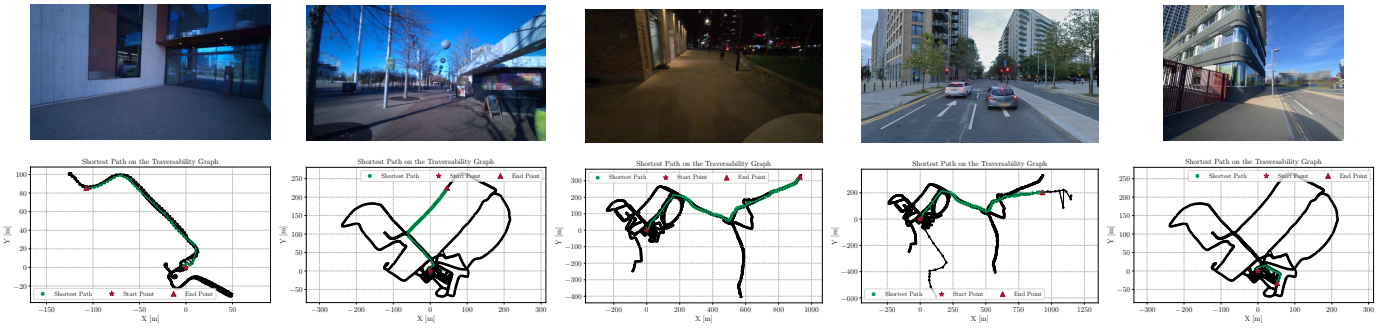
*N: Number of reference images for a map.

A. Map Size

The primary storage concern for our mapping solution is the co-visibility graph and its associated keyframe images. We evaluate the **Theoretical Map Size** to demonstrate the lightweight nature of our approach compared to SfM-based methods like Hloc [21], a state-of-the-art feature-based visual localization system. SfM-based systems, such as Hloc, must store a $3D$ point cloud and per-$3D$ point feature descriptors (along with co-visibility and visual dictionary information), where the feature descriptors constitute the largest data volume. Specifically, the Hloc map size is estimated as $N \times 2DM$ bytes, where N is the number of reference images, $D = 256$ is the descriptor dimension (half precision, 2 bytes per entry), and M is the number of features per image. In contrast, ours only store compressed reference images. The map size for these methods is given by $N \times 3WH \times C$ bytes, where W and H are the image dimensions, and C is the average JPEG compression ratio (empirically, $C \approx 0.14$ in our experiments). Table I provides a detailed comparison of the storage requirements. For instance, our map shown in Fig. 3 achieves a total storage size under 255MB for 4273 images.

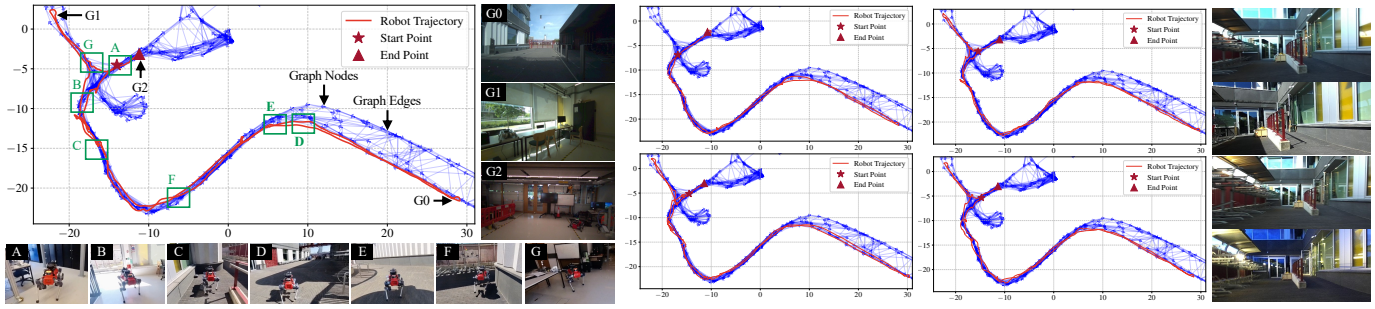
B. Map Merging

Map merging incrementally aligns multiple submaps represented as pose graphs, using our collaborative localization pipeline and PGO. Data from three regions (G0-G2) include challenges such as low texture, illumination changes, and viewpoint variation. To test robustness, we randomly shuffled submap order (Gi-Rj), simulating unstructured crowdsourced



(a) 2024/09/04-2024/11/28 (b) 2024/09/04-2024/12/04 (c) 2024/09/04-2024/12/23 (d) 2012/04/01-2024/12/23 (e) 2024/09/04-2024/12/04

Fig. 5. Global path planning on the proposed topometric map. Subfigures (a)-(e) represent different stages of the mapping process. As more data are collected, the map incrementally expands to cover a larger area, enabling the robot to navigate to a broader set of goals via the shortest paths. The goal images in (a)-(c) are captured using Aria glasses, while (d) uses a Google Street View image. As demonstrated in (e), our framework also supports input captured by heterogeneous devices such as smartphones, highlighting its generalization capability.



(a) First run (b) Multiple runs

Fig. 6. Real-world experiment (R0) with a quadruped robot. (a) Navigation through indoor and outdoor areas using goal images. (b) Multiple runs across day/night show robustness under lighting variations.

TABLE II
ABSOLUTE TRAJECTORY ERROR (ATE) MEASUREMENTS ACROSS DIFFERENT SCENARIOS.

Data	Dist./Dura.	Shuffle	Trans. ATE $[m]$	Rot. ATE $[deg]$
G0-InOrder	$0.6km$ $6mins$	N	0.47	0.70
G0-R0		Y	0.65	0.86
G0-R1		Y	0.27	0.50
G1-InOrder	$3.2km$ $18hours$	N	1.18	0.44
G1-R0		Y	2.39	1.15
G1-R1		Y	1.27	0.41
G2-InOrder	$15.3km$ $110days$	N	1.51	1.71
G2-R0		Y	2.30	1.10
G2-R1		Y	2.87	1.52
G2-R2		Y	2.30	1.56
G2-R3		Y	1.93	2.06
G2-R4		Y	1.78	0.98
G2-R5		Y	2.05	1.52
G2-R6		Y	1.80	1.35
G2-R7		Y	1.72	0.79
G2-R8		Y	2.58	1.00

input. Key observations: *1)* Shuffled submaps may lack overlap, but mismatches are filtered via geometric verification and confidence weighting. *2)* The final map achieves $<$ 3m translational and $< 2.1°$ rotational ATE over 15.3km, demonstrating robustness. *3)* Node-level matching enables independent submap integration.

C. Closed-Loop Visual Navigation

We evaluate image-goal navigation, where the target is a goal image, by integrating our topometric map into a closed-loop VNav system [22]. This interaction is intuitive, avoids coordinate specification, and can extend to object- or language-based goals. We test two tasks: *1)* global path planning on the merged map, and *2)* closed-loop navigation on a real robot.

1) Global Path Planning: The map supports metric-based shortest path planning with newly added submaps, ensuring globally consistent trajectories. It also handles in-the-wild goal images from mobile phones. As shown in Fig. 5, our method consistently selects correct shortest paths, even when user intuition suggests longer routes.

2) Real-World Experiments: We validate in three scenarios: R0 (indoor and outdoor lab), R1 (outdoor with bridge crossing), and R2 (building perimeter with turns). In R0, the robot follows three image goals over 160m in 312s at 0.5m/s, robustly handling dynamic objects. Repeated runs confirm consistent trajectories across lighting changes. In R1, the robot successfully navigates sparse visual features; in R2, it manages complex turns where odometry alone would drift (Fig. 7).

V. CONCLUSION

We presented OPENNAVMAP, a collaborative localization and multi-session mapping system for scalable robot navigation. The method employs a lightweight, structure-free topometric map that reduces complexity and storage compared

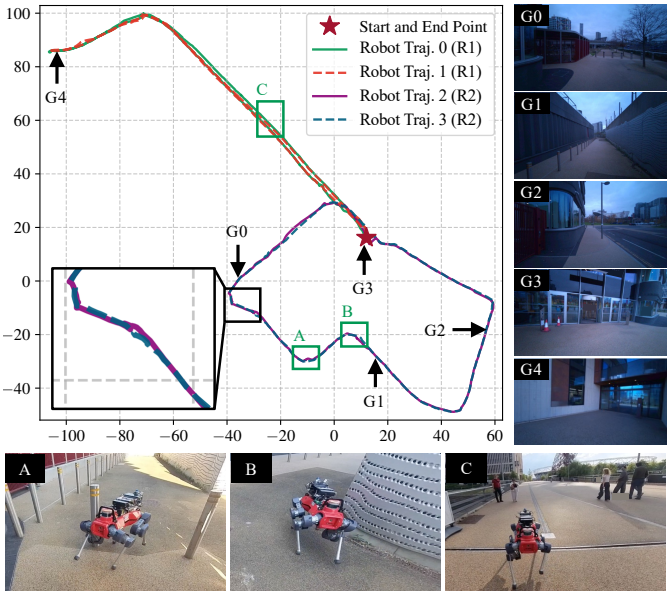


Fig. 7. Experiments R1 and R2. The robot navigates challenging terrains, including a bridge and confined paths, using only image goals.

to structure-based maps. Extensive experiments across $19km$ of trajectories in varied environments show that our approach achieves sub-$3m$ ATE over $15km$ using only monocular RGB inputs. The system was also deployed on a real quadruped robot, completing multiple autonomous navigation tasks and demonstrating practical scalability. To support future research, we will release our code and datasets.

REFERENCES

[1] M. Fernandez-Cortizas, H. Bavle, D. Perez-Saura, J. L. Sanchez-Lopez, P. Campoy, and H. Voos, "Multi s-graphs: an efficient distributed semantic-relational collaborative slam," *IEEE Robotics and Automation Letters*, 2024.

[2] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] R. Murai, E. Dexheimer, and A. J. Davison, "MASt3R-SLAM: Real-time dense slam with 3d reconstruction priors," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 16 695–16 705.

[4] C. Zheng, W. Xu, Z. Zou, T. Hua, C. Yuan, D. He, B. Zhou, Z. Liu, J. Lin, F. Zhu *et al.*, "Fast-livo2: Fast, direct lidar-inertial-visual odometry," *IEEE Transactions on Robotics*, 2024.

[5] H. Oleynikova, Z. Taylor, M. Fehr, R. Siegwart, and J. Nieto, "Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1366–1373.

[6] E. Arnold, J. Wynn, S. Vicente, G. Garcia-Hernando, A. Monszpart, V. Prisacariu, D. Turmukhambetov, and E. Brachmann, "Map-free visual relocalization: Metric pose relative to a single image," in *Springer European Conference on Computer Vision*, 2022, pp. 690–708.

[7] P. Yin, J. Jiao, S. Zhao, L. Xu, G. Huang, H. Choset, S. Scherer, and J. Han, "General place recognition survey: Towards real-world autonomy," *IEEE Transactions on Robotics*, 2025.

[8] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[9] T. Qin, C. Li, H. Ye, S. Wan, M. Li, H. Liu, and M. Yang, "Crowd-sourced nerf: Collecting data from production vehicles for 3d street view reconstruction," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[10] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "Nerf in the wild: Neural radiance fields for unconstrained photo collections," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7210–7219.

[11] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver, "Google street view: Capturing the world at street level," *Computer*, vol. 43, no. 6, pp. 32–38, 2010.

[12] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org, 2017.

[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] F. Dellaert and G. Contributors, "borglab/gtsam," May 2022. [Online]. Available: https://github.com/borglab/gtsam)

[15] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith *et al.*, "Project Aria: A new tool for egocentric multi-modal ai research," *arXiv preprint arXiv:2308.13561*, 2023.

[16] H. Wei, J. Jiao, X. Hu, J. Yu, X. Xie, J. Wu, Y. Zhu, Y. Liu, L. Wang, and M. Liu, "Fusionportablev2: A unified multi-sensor dataset for generalized slam across diverse platforms and scalable environments," *The International Journal of Robotics Research*, p. 02783649241303525, 2024.

[17] H. Xu, P. Liu, X. Chen, and S. Shen, "D2SLAM: Decentralized and distributed collaborative visual-inertial slam system for aerial swarm," *IEEE Transactions on Robotics*, 2024.

[18] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4878–4888.

[19] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *2012 IEEE international conference on robotics and automation*. IEEE, 2012, pp. 1643–1649.

[20] O. Vysotska and C. Stachniss, "Lazy data association for image sequences matching under substantial appearance changes," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 213–220, 2015.

[21] E. Brachmann, T. Cavallari, and V. A. Prisacariu, "Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5044–5053, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258841110

[22] J. Jiao, J. He, C. Liu, S. Aegidius, X. Hu, T. Braud, and D. Kanoulas, "LiteVLoc: Map-lite visual localization for image goal navigation," *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025.