ROMAN: Robust Object Map Alignment Anywhere

Mason B. Peterson, Yi Xuan Jia, Yulun Tian, Annika Thomas and Jonathan P. How

***Abstract*—Global localization is essential for long-term, drift-free robot navigation, but existing methods struggle with varying viewpoints or lighting. ROMAN (Robust Object Map Alignment Anywhere) addresses this challenge by using open-set object maps for global localization, offering a more reliable representation than visual features. These sparse maps are built via open-set segmentation, allowing localization in novel environments. ROMAN aligns small object submaps using a graph-theoretic global data association approach, incorporating gravity direction, open-set semantics, and shape similarity for robust global data association. Tested on large-scale outdoor datasets, ROMAN outperforms other methods, with 36% higher recall and 40% lower trajectory error in challenging multi-robot SLAM tasks. More details are at https://acl.mit.edu/ROMAN/.**

I. Introduction

Global localization [1] refers to the task of positioning a robot within a reference map created either in a previous mapping session or by another robot in real-time [2], and is a crucial capability for drift-free navigation. This work focuses on global localization and loop closure using *object-* or *segment-level* representations, which recent studies [3–5] show to be effective in environments with significant sensor noise and viewpoint changes.

The core challenge is *global data association*, which involves finding correspondences between observed and mapped objects without an initial guess. Earlier approaches such as [6–9] rely on geometric verification based on RANSAC [10], which exhibits high computational complexity under high outlier regimes. Newer graph-theoretic approaches offer better accuracy and robustness by forming consistency graphs to solve the correspondence problem [3, 11–14]. However, these methods struggle in challenging regimes such as environments where objects have ambiguous spatial configurations.

To address this, we propose extending graph-theoretic data association beyond mutual geometric consistency by incorporating (i) open-set semantics extracted as semantically meaningful 3D segments [15, 16] with descriptors obtained from [17]; (ii) segment-level geometric attributes like volume and shape; and (iii) prior knowledge of gravity direction from inertial sensors. This unified approach improves the state-of-the-art [11] in terms of both precision and recall metrics.

II. ROMAN

ROMAN, overviewed in Fig. 1, constructs segment-level maps using high-level features from RGB-D images and robot pose estimates to enable global localization in unseen environments. Using Kimera-VIO for odometry [18], FastSAM [16] for image segmentation and averaged CLIP embeddings [17] for semantic descriptors, the system processes and tracks 3D object segments, filtering undesirable objects and merging segments with high voxel overlap [19].

This work is supported in part by the Ford Motor Company, DSTA, ONR, and ARL DCIST under Cooperative Agreement Number W911NF-17-2-0181.

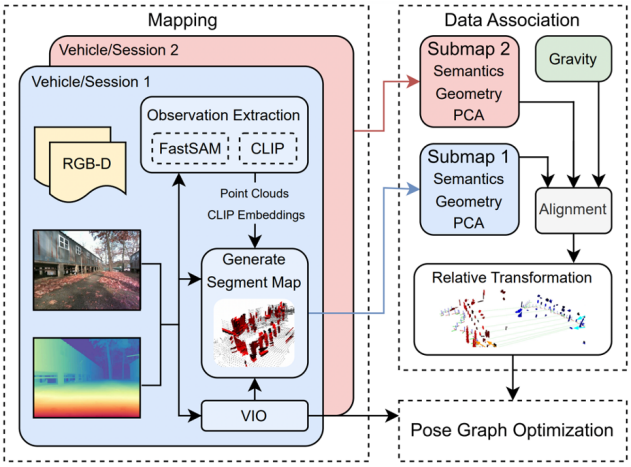

Fig. 1: ROMAN includes three modules including mapping, data association and pose graph optimization. The front end mapping pipeline tracks segments across RGB-D images to generate segment maps. The data association module incorporates semantics, geometry and PCA from submaps along with gravity as a prior into the ROMAN alignment module to return relative transformations between submaps. These transformations are used along with VIO for pose graph optimization.

To perform global localization, we divide each robot's segment map into overlapping submaps and solve a registration problem to align submaps from different robots by associating 3D segments using geometric and semantic attributes. The process involves detecting map overlap and matching segments despite uncertainty and outliers, with the final transformation between frames estimated using Arun's method [20].

We incorporate segment-based loop closures into the robust multi-robot pose graph solver from [21] by registering submaps between robots and within each robot over time, rejecting loop closures with insufficient correspondences, and feeding the valid ones into the robust pose graph optimization of [21] to estimate multi-robot trajectories.

Preliminaries: Robust Data Association Finding object associations between submaps leverages prior work from CILPPER [11]. CLIPPER first constructs a consistency graph, $\mathcal{G}$, where each node in the graph is a putative association $a_p = (p_i, p_j)$ between segments p_i and p_j from robots i and j respectively. From the consistency graph $\mathcal{G}$, a weighted affinity matrix $\mathbf{M}$ is created where $\mathbf{M}_{p,q} = s_a(a_p, a_q)$ and $\mathbf{M}_{p,p} = 1$, and $s_a(a_p, a_q) \in [0, 1]$ scores the geometric consistency (i.e., whether $\|p_i - q_i\|$ is similar to $\|p_j - q_j\|$) between two associations. Inlier associations are determined by (approximately) solving for the densest subset of consistent associations, by maximizing the Rayleigh Quotient of $\mathbf{M}$ and

the binary $\mathbf{u}$ association vector, where u_p is 1 when association a_p is accepted as an inlier and 0 otherwise. See [11] for more details.

Incorporating additional information into association affinity. In its original form, the affinity matrix $\mathbf{M}$ relies solely on distance information between pairs of centroids. We propose a method that directly incorporates shape information via PCA and semantic information as well as the direction of gravity (when available) in the underlying densest subgraph optimization problem using the *geometric mean*,

$$\mathbf{M}_{p,q} = \text{GM}(s_a(a_p, a_q), s_o(a_p), s_o(a_q)), \tag{1}$$

where s_o is object similarity in terms of shape and semantics and s_a is our novel association pair scoring using gravity as a prior. $s_o(a_p) = \text{GM}(s_{o_{\text{semantic}}}, s_{o_{\text{shape}}})$, where we emphasize that this scoring is between pairs of objects p_i and p_j rather than pairs of associations. We make $s_{o_{\text{semantic}}}$ a scaled cosine similarity between CLIP embeddings and $s_{o_{\text{shape}}}$ the geometric mean of the ratio between four indicators of the two objects' shape: volume, linearity, planarity, and 3D-ness, where the last three are computed using principal component analysis. We incorporate prior knowledge from gravity by decoupling the vertical and lateral components of object distances for computing association pair similarity in s_a.

III. Experiments

We evaluated ROMAN's map alignment using the outdoor Kimera-Multi Dataset [22], where each robot creates submaps using Kimera-VIO [18] for odometry. Baselines include RANSAC-100K and RANSAC-1M, on segment centroids with max iteration count of 100,000 and 1 million resepectively, standard CLIPPER, CLIPPER with pruned initial associations, and visual features with VoW descriptors of ORB features.

Results in Fig. 2 including precision-recall, distance error, and heading difference plots, demonstrate ROMAN's ability to correctly align maps while rejecting incorrect ones, even under challenging viewpoint scenarios. ROMAN achieved higher precision/recall, aligned maps with less error, and handled diverse heading angles better than baseline methods, with a maximum recall of 0.67 for same-direction submaps and 0.26 for opposite-direction submaps, outperforming the next best method, RANSAC-1M. ROMAN is also 4.5 times faster than RANSAC-1M, with each submap packet under 250 KB and requiring less than 24 MB for a 1 km trajectory.

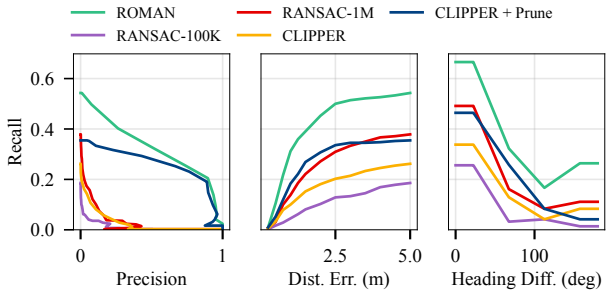

Fig. 2: Performance of ROMAN and baseline methods.

We run the full ROMAN pose graph optimization pipeline on the tunnel, hybrid, and outdoor Kimera-Multi datasets and compare the root-mean squared (RMS) absolute trajectory error (ATE) of the estimated multi-robot trajectories.

ROMAN's ability to get loop closures in challenging visual scenarios allows moderate improvements to the ATE; however, due to robot experiment design, the robot paths are well connected and most loop closure opportunities occur when robots are traveling in the same direction, leading to only opportunity for small improvement.

TABLE I: Multi-robot SLAM Results - ATE RMSE (m)

Dataset	Visual Features	ROMAN	Combined
Tunnel	4.38	4.20	4.12
Hybrid	5.83	5.12	4.77
Outdoor	9.38	8.77	7.77

We further evaluate the proposed method's ability to register segment maps in an outdoor, off-road environment with high visual ambiguity.

In this experiment, data is recorded on Clearpath Jackals using Intel RealSense D455 to capture RGB-D images and Kimera-VIO [18] for odometry. We run the ROMAN-SLAM pipeline on easy (same direction), medium (partial opposite directions) and hard (opposite direction) pairs of robot trajectory data taken in an outdoor, off-road environment. We compare ROMAN to Kimera-Multi [21] using visual features for loop closure in Fig. 3, where ROMAN finds loop closures in cases that Kimera-Multi does not.

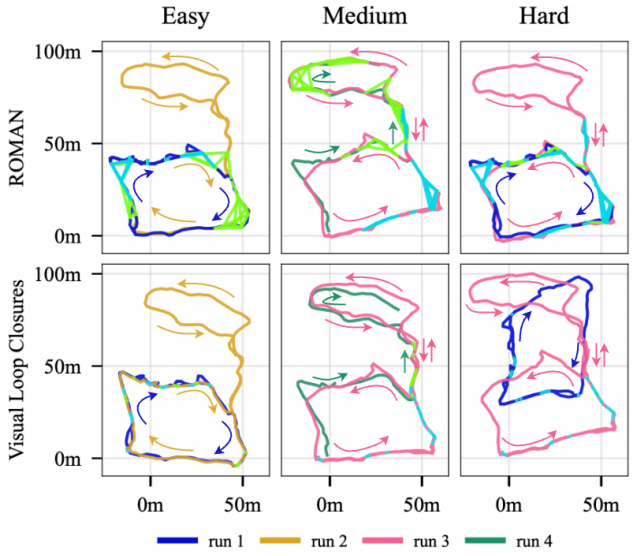

Fig. 3: Off-road qualitative pose graph trajectory estimate. Blue lines (▬) represent single-robot loop closures and green lines (▬) show multi-robot loop closures.

IV. Conclusion

This work presented a method for performing global localization in challenging outdoor environments by robust registration of 3D open-set segment maps. Associations between maps were informed by geometry of 3D segment locations, segment shape attributes, and direction of the gravity vector in segment maps. Future work includes incorporating additional shape information from learned shape descriptors for computing shape similarity.

REFERENCES

[1] H. Yin, X. Xu, S. Lu, X. Chen, R. Xiong, S. Shen, C. Stachniss, and Y. Wang, "A survey on global lidar localization: Challenges, advances and open problems," *arXiv preprint arXiv:2302.07433*, 2023.

[2] P.-Y. Lajoie, B. Ramtoula, F. Wu, and G. Beltrame, "Towards collaborative simultaneous localization and mapping: a survey of the current research landscape," *Field Robotics*, 2022.

[3] J. Yu and S. Shen, "Semanticloop: loop closure with 3d semantic graph matching," *RA-L*, vol. 8, no. 2, pp. 568–575, 2022.

[4] A. Thomas, J. Kinnari, P. Lusk, K. Kondo, and J. P. How, "SOS-Match: segmentation for open-set robust correspondence search and robot localization in unstructured environments," *arXiv:2401.04791*, 2024.

[5] X. Liu, J. Lei, A. Prabhu, Y. Tao, I. Spasojevic, P. Chaudhari, N. Atanasov, and V. Kumar, "Slideslam: Sparse, lightweight, decentralized metric-semantic slam for multi-robot navigation," *arXiv preprint arXiv:2406.17249*, 2024.

[6] R. Dubé, D. Dugas, E. Stumm, J. Nieto, R. Siegwart, and C. Cadena, "Segmatch: segment based place recognition in 3d point clouds," in *ICRA*. IEEE, 2017.

[7] G. Tinchev, S. Nobili, and M. Fallon, "Seeing the wood for the trees: Reliable localization in urban and natural environments," in *IROS*. IEEE, 2018.

[8] R. Dube, A. Cramariuc, D. Dugas, H. Sommer, M. Dymczyk, J. Nieto, R. Siegwart, and C. Cadena, "SegMap: segment-based mapping and localization using data-driven descriptors," *IJRR*, vol. 39, no. 2-3, pp. 339–355, 2020.

[9] A. Cramariuc, F. Tschopp, N. Alatur, S. Benz, T. Falck, M. Brühlmeier, B. Hahn, J. Nieto, and R. Siegwart, "SemSegMap–3D segment-based semantic localization," in *IROS*. IEEE, 2021, pp. 1183–1190.

[10] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[11] P. C. Lusk and J. P. How, "CLIPPER: robust data association without an initial guess," *RA-L*, 2024.

[12] J. G. Mangelson, D. Dominic, R. M. Eustice, and R. Vasudevan, "Pairwise consistent measurement set maximization for robust multi-robot map merging," in *ICRA*. IEEE, 2018, pp. 2916–2923.

[13] J. Shi, H. Yang, and L. Carlone, "Robin: a graph-theoretic approach to reject outliers in robust estimation using invariants," in *ICRA*. IEEE, 2021, pp. 13 820–13 827.

[14] B. Forsgren, M. Kaess, R. Vasudevan, T. W. McLain, and J. G. Mangelson, "Group-k consistent measurement set maximization via maximum clique over k-uniform hypergraphs for robust multi-robot map merging," *IJRR*, p. 02783649241256970, 2023.

[15] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[16] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, "Fast segment anything," *arXiv preprint arXiv:2306.12156*, 2023.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[18] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: an open-source library for real-time metric-semantic localization and mapping," in *ICRA*. IEEE, 2020.

[19] L. Schmid, M. Abate, Y. Chang, and L. Carlone, "Khronos: A unified approach for spatio-temporal metric-semantic slam in dynamic environments," in *Proc. of Robotics: Science and Systems*, 2024.

[20] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *TPAMI*, no. 5, pp. 698–700, 1987.

[21] Y. Tian, Y. Chang, F. H. Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems," *T-RO*, vol. 38, no. 4, 2022.

[22] Y. Tian, Y. Chang, L. Quang, A. Schang, C. Nieto-Granda, J. How, and L. Carlone, "Resilient and distributed multi-robot visual SLAM: Datasets, experiments, and lessons learned," in *IROS*, 2023.