Outdoor Global Localization via Robust Registration of 3D Open-Set Segments

Mason B. Peterson, Yi Xuan Jia, Yulun Tian and Jonathan P. How
LIDS, MIT
Emails: masonbp@mit.edu, yixuany@mit.edu, yulun@mit.edu, jhow@mit.edu

***Abstract*—Global localization is a fundamental capability enabling long-term and drift-free robot navigation. This paper presents an outdoor global localization method based on robust registration of open-set segment maps. Our method detects and tracks segments using open-set image segmentation models that enable direct generalization to unseen environments. To perform global localization under the high outlier regimes that are typical of natural/outdoor environments, we formulate a registration problem between small submaps of 3D segments and solve for the correspondences using a graph-theoretic global data association approach. Further, to guide registration in highly noisy or ambiguous scenarios, we propose novel ways of incorporating additional information (*e.g.*, segment attributes and known gravity direction) within the global data association formulation. The proposed method is evaluated on outdoor datasets recorded by multiple robots and shown to outperform existing methods in terms of precision-recall metrics and localization accuracy.**

I. Introduction

Global localization [1] refers to the task of localizing a robot in a reference map or another robot's local map (*i.e.*, inter-robot loop closure) using onboard measurements and without any initial guess. It is a cornerstone capability for drift-free navigation in GPS-denied scenarios including many natural environments such as forests and underground caves. However, these natural environments are often characterized by unstructured geometries and self-similar visual appearances, which lead to significant noise and outliers in onboard perception and make the association of perceived features in global localization challenging.

Compared to conventional keypoints extracted from visual or lidar observations, *objects* or *segment-level* representations offer many advantages for global localization. First, they are more stable against sensor noise and viewpoint changes: in comparison, methods based on visual features quickly fail when scenes are viewed from very different viewpoints [2]. Second, objects and segments are more lightweight and thus much more efficient to transmit compared to local visual features in multi-robot applications. SegMatch [3] is an earlier work that extracts 3D segments from dense lidar maps, and solves global localization using eigenvalue/shape-based matching and RANSAC [4]. Subsequent works improved on this framework using learned descriptors [5], semantic information [6, 7], and an improved feature extraction process [8]. Ankenbauer *et al.* [9] present an object-based global localization method that leverages graph-theoretic data association [10] as the back-end solver, which is shown to achieve superior robustness against outlier putative associations. While these prior works show promising results, their performance in novel natural environments are limited by their dependence on domain-specific segmentation or object detection methods that require manual training and/or careful parameter tuning. To address this issue, recent work [11] proposes to use pre-trained, open-set foundation models for zero-shot segmentation in novel environments. However, this approach uses a simple object representation given by their 3D centroids, which limits global localization performance in more challenging regimes when using centroids alone fails, *e.g.*, due to the ambiguity caused by a small number of objects in the scene or the symmetry in their spatial configurations.

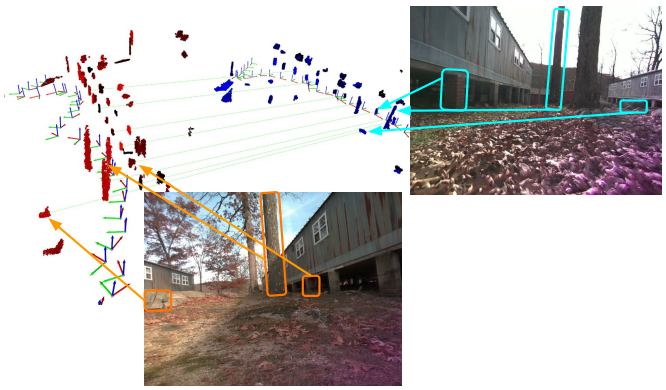

Fig. 1: Pair of segment submaps matched by two robots traveling in *opposite* directions in an off-road environment. The submaps are shown in different colors (red and blue) together with robot trajectories. Green edges denote segment-level associations found by the proposed method, and a subset of these associated segments are annotated in the corresponding image observations.

Contributions. We present a method for performing global localization in natural environments with high visual ambiguity resulting from different viewpoints and visually similar scenes (e.g. as shown in Fig 1). To accomplish this, we present the following contributions:

1) A pipeline for creating open-set object-based maps using a single onboard RGB-D camera and FastSAM [12] for open-set image segmentation in previously unseen environments. These maps compactly summarize the detailed RGB-D point clouds into sparse representations consisting of segment location and shape attributes, which enable efficient and robust global localization.
2) A method extending the graph-theoretic global data association method of [10] to incorporate 3D-segment-level similarity information (*e.g.*, based on shape and volume) and a gravity-direction prior. Our method implicitly guides the solver to correct 3D segment-to-segment associations in challenging regimes when object centroids alone are not sufficient for identifying the correct data association (*e.g.*, due to repetitive geometric structures or scenes with few distinct objects).
3) Experimental evaluation of the proposed method us-

ing real-world multi-robot navigation datasets in natural/outdoor environments, demonstrating that our method outperforms baseline methods in regimes with opposite viewpoints and challenging visual similarity.

II. Proposed Approach

A. Open-set object-level mapping

To enable global localization in previously unseen natural environments, the proposed method constructs segment-level maps based on high-level features detected by recent zero-shot open-set segmentation models. The inputs consist of RGB-D images and robot pose estimates (e.g., provided by a visual-inertial SLAM system). The images are downsampled and processed with FastSAM [12] to extract high-level image segments for each robot. We track the detected segments by computing the Intersection over Union (IoU) between segments from consecutive keyframes. For each track of the associated 2D segments, we create a corresponding 3D segment in the robot map by merging the RGB-D point cloud observations in the local reference frame. We also implement segment merging by computing IoU in both the projected 2D masks and 3D volumes to handle cases when the same object gets segmented into multiple parts. Fig. 1 visualizes portions of two segment maps constructed by two robots traveling in opposite directions, together with the correspondences detected by the proposed approach (shown as green edges).

B. Submap Alignment

To perform global localization, we divide each robot's onboard segment map into multiple, potentially overlapping submaps. We then consider the problem of aligning robot i's local submap $\mathcal{M}_i$ in robot i's local frame $\mathcal{F}_i$ with robot j's map $\mathcal{M}_j$ in $\mathcal{F}_j$. We formulate this as a registration problem where each 3D segment is represented by a 3D point and feature vector (*e.g.*, volume and shape attributes). Successful global localization requires that segments in $\mathcal{M}_i$ are correctly associated with segments in $\mathcal{M}_j$, which is a challenging task in the presence of uncertainty, outliers, and geometric ambiguity. To this end, we extend the graph-theoretic data association framework, CLIPPER [10], to tackle segment map registration. In the following, we first present a brief review of the approach behind CLIPPER and then describe the proposed extension. Then, once correspondences between $\mathcal{M}_i$ and $\mathcal{M}_j$ have been determined, the relative transformation from $\mathcal{F}_j$ to $\mathcal{F}_i$, $\hat{\mathbf{T}}^i_j$, can be found using the closed-form Arun's method [13].

Preliminaries: Graph-Theoretic Global Data Association. 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 a segment p_i in $\mathcal{M}_i$ and a segment p_j in $\mathcal{M}_j$. Edges are created between nodes when associations are geometrically consistent with each other. Specifically, given two putative correspondences $a_p = (p_i, p_j)$ and $a_q = (q_i, q_j)$, CLIPPER declares that a_p and a_q are consistent if the distance between segment centroids in the same map is preserved, i.e., if $d(a_p, a_q) \triangleq |\,\|p_i - q_i\| - \|p_j - q_j\|\,|$ is small. Then, a weighted edge $\mathcal{E}_{p,q} = s_a(a_p, a_q)$ is added to the graph according to $s_a(a_p, a_q) = \exp\left(-\frac{1}{2}\frac{d(a_p,a_q)^2}{\sigma^2}\right)$ if $d(a_p, a_q) \leq \epsilon$, where $s_a(a_p, a_q) \in [0, 1]$ scores the similarity between two associations and ϵ and σ are tuneable parameters expressing bounded noise in the segment point representation.

Given 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 CLIPPER determines inlier associations by (approximately) solving for the densest subset of consistent associations, formulated as the following optimization problem,

$$\max_{\mathbf{u}\in\{0,1\}^n} \frac{\mathbf{u}^\top \mathbf{M}\mathbf{u}}{\mathbf{u}^\top\mathbf{u}}. \quad \text{subject to} \quad u_p u_q = 0 \text{ if } \mathbf{M}_{p,q} = 0,\ \forall_{p,q}, \tag{1}$$

where u_p is 1 when association a_p is accepted as an inlier and 0 otherwise. See [10] for more details.

When point registration methods such as CLIPPER are applied on segment maps, unique challenges are introduced that are often not faced in other point registration problems (e.g., lidar point cloud registration), including dealing with greater noise in segment centroids (*e.g.*, due to partial observation) and few inlier segments (*e.g.*, often less than 10) mapped in both $\mathcal{M}_i$ and $\mathcal{M}_j$. These problems combined can lead to ambiguity when performing segment submap registration, and we show in Sec. III that CLIPPER often struggles to determine correct associations when aligning segment maps. To address these problems, other works have proposed pre-processing or post-processing methods that leverage additional information to filter incorrect global localization results. For instance, in [11] and [14], a putative association is removed if the difference in size of the associated segments is above a threshold. Additionally, in the event when the two submaps share gravity direction (so the true relative transform only involves rotation around the gravity direction), [11] checks the roll/pitch angles of the estimated transform $\hat{\mathbf{T}}^i_j$ in a post-processing step to remove incorrect registration solutions.

Leveraging Additional Information within CLIPPER. In comparison to works that explicitly use prior information in pre-processing or post-processing steps, we propose a method that directly incorporates this information in the underlying optimization problem (1). The key to our approach is to extend the original similarity metric to (i) account for segment-to-segment attribute similarity and (ii) use knowledge of the gravity direction when available.

First, we address the task of incorporating information from segment attributes. For a putative associate $a_p = (p_i, p_j)$, let $s_o(a_p)$ denote a similarity measure between the segments p_i and p_j. While [10] suggests to set the diagonal entries of $\mathbf{M}$ to reflect this information, *e.g.*, by setting $\mathbf{M}_{p,p} = s_o(a_p)$, this similarity measure has a limited impact as the objective function value tends to be dominated by association-to-association similarity terms (off-diagonals of $\mathbf{M}$). Another solution from [7] proposes multiplying the association affinity score by $s_o(\cdot)$ so that $\mathbf{M}_{p,q} = s_a(a_p, a_q)s_o(a_p)s_o(a_q)$. We find that although this approach allows segment-to-segment similarity to play a significant role in the registration problem, the elements of $\mathbf{M}$ are skewed to be much smaller resulting in many fewer accepted inlier associations. To incorporate segment-to-segment similarity without significantly diminishing the magnitudes of the entries of $\mathbf{M}$, we instead propose using the weighted *geometric mean*,

$$\mathbf{M}_{p,q} = \left(s_a(a_p, a_q)^w s_o(a_p)s_o(a_q)\right)^{\frac{1}{w+2}} \tag{2}$$

where w can be used to balance the association similarity score with the segment-to-segment similarity score. In our

implementation, we set $w = 2$ to balance the impacts of the s_a and s_o terms. For the segment-to-segment similarity, we use

$$s_o(a_p) = \left(\prod_{k=1}^{K} \frac{\min\left(f_k(p_i), f_k(p_j)\right)}{\max\left(f_k(p_i), f_k(p_j)\right)} \right)^{1/K}, \quad (3)$$

where $f_k(p_i)$ represents the k-th shape attribute of p_i. In this work, three shape attributes are used, which include the segment volume together with the minimum and maximum lengths of the oriented bounding box.

Next, we address implicitly incorporating knowledge of the gravity direction in the global data association formulation. Due to the geometric-invariant formulation of eq. (1), the solver naturally considers registering object maps as a 6-DOF problem. Often in robotics, with an onboard IMU the direction of the gravity vector is well defined, and we are instead interested in considering transformations with only x, y, z, and yaw components. In this work, we propose a method to leverage knowledge of the gravity direction *within* the data association step, which guides the solver to select associations that are consistent with the direction of the gravity vector.

To accomplish this, we propose a geometric invariant that inherently represents prior knowledge of the gravity vector by decoupling computations in the x-y plane and along the z axis:

$$s_a(a_p, a_q) = \exp\left(-\frac{1}{2} \left(\frac{d_{xy}^2(a_p, a_q)}{\frac{2}{3}\sigma^2} + \frac{d_z^2(a_p, a_q)}{\frac{1}{3}\sigma^2} \right) \right), \quad (4)$$

where

$$\begin{aligned} d_{xy}(a_p, a_q) &= \left| \, \|p_{i,xy} - q_{i,xy}\| - \|p_{j,xy} - q_{j,xy}\| \, \right| \\ d_z(a_p, a_q) &= \left| \, (p_{i,z} - q_{i,z}) - (p_{j,z} - q_{j,z}) \, \right|. \end{aligned}$$

It is important to note that we use the *difference* in the z-axis since we have directional information from the gravity vector while we only use *distance* in the x-y plane. The directional information helps further disambiguate correspondence selection as opposed to merely decoupling x-y distance from z.

III. Experiment

We evaluate our method for performing global localization in challenging multi-robot scenarios including robots traversing off-road paths in natural environments and in opposite directions. We compare the `proposed` method of using shape attribute similarities and the gravity vector against the following baselines. `RANSAC-100K` and `RANSAC-1M` apply RANSAC [4], as implemented in [15], on segment centroids with a max iteration count of 100,000 and 1 million. `CLIPPER` runs standard CLIPPER [10] on segment centroids, and `CLIPPER + Prune` further prunes registration results using volume and gravity information (so it has access to similar information as the proposed method). Additionally, we introduce the following variants of our method for the purpose of ablation. In `Proposed w/o Shape`, segment shape similarity (*i.e.*, s_o) is not used. `Product` and `Arithmetic Mean` replace the geometric mean in eq. (2) with direct multiplication and weighted average, respectively.

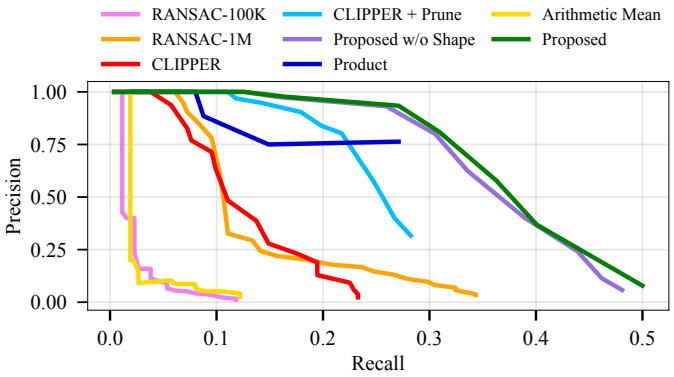

Fig. 2: Precision vs recall plot comparing global localization methods on submaps created from MIT campus robot data. The parameter for minimum number of associated objects is swept to generate different levels of precision and recall.

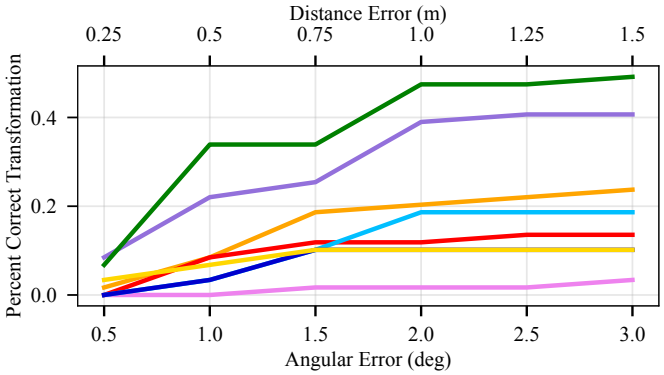

Fig. 3: Percentage of correctly aligned submaps from 59 challenging overlap cases. Along the x-axis, maximum allowed registration error in both distance (top x-axis) and angle (bottom x-axis) are varied. The proposed method computes the correct $\hat{\mathbf{T}}_j^i$ for 49% of the challenging overlapping submaps with an error of less than $1.5\,\mathrm{m}$ and $3\,\mathrm{deg}$.

A. MIT Campus Global Localization

We first evaluate our method for aligning segment submaps using the outdoor Kimera Multi Dataset [2] recorded at the MIT campus. Maps are created using our segment mapping pipeline along ground truth robot trajectories. The resulting segments are grouped into submaps, where a submap includes all 3D segments within a $20\,\mathrm{m}$ radius and a new submap is created every $10\,\mathrm{m}$. To demonstrate our global localization algorithm, we select six events where robots paths overlap, including three instances where robots travel in the same direction and three challenging cases where robots cross paths perpendicularly or travel in opposite directions to each other.

We compare our method for aligning 3D segment submaps against baseline methods. Fig. 2 shows the combined precision-recall performance for all six overlap events. We require that the registration error of the computed $\hat{\mathbf{T}}_j^i$ must be below $1.5\,\mathrm{m}$ and $3\,\mathrm{deg}$ for the global localization estimate to be accepted as correct. To generate the precision-recall curves, we vary the threshold on the number of objects required for each algorithm to accept a candidate alignment. The proposed method achieves significantly higher precision-recall results than baseline methods and does so with similar computation times as seen in Table I.

To provide further insights on the registration accuracy, we evaluate the ability of segment registration methods to

accurately estimate $\mathbf{T}_j^i$ given two overlapping submaps from one of the three challenging configurations with opposite or perpendicular viewpoints. Fig. 3 shows the total percentage of correct $\hat{\mathbf{T}}_j^i$ from 59 evaluated pairs of submaps. Results show that the proposed method computes accurate transformations for 49% of overlapping submaps in these challenging scenarios, where in comparison `RANSAC-1M` only correctly estimate 24% of these transformations with a runtime that is three times slower. We additionally note that on average, each submap can be represented with only 1.6 KB of data which is much more efficient to communicate than visual feature descriptors.

TABLE I: Mean Submap Registration Timing Analysis

Method	Time (ms)	Method	Time (ms)
RANSAC-100K	37.6	Proposed w/o Shape	42.6
RANSAC-1M	187.6	Product	45.4
CLIPPER	57.8	Arithmetic Mean	819.6
CLIPPER + Prune	22.5	Proposed	56.7

B. Off-road Global Localization

We further evaluate the proposed method's ability to register segment maps in an outdoor, off-road environment with high visual ambiguity. We select overlapping submaps which were created by two robots traveling in opposite directions using Kimera-VIO [16] for ego localization. Fig. 1 shows two submaps, separated for visualization, selected for qualitative evaluation of our global localization method. In Fig. 4, we qualitatively compare the computed $\hat{\mathbf{T}}_j^i$ from CLIPPER and the proposed method by overlaying the submap created by robot 1 and the submap created by robot 2 transformed by $\hat{\mathbf{T}}_j^i$. Due to considerable ambiguity in the scene, CLIPPER's $\hat{\mathbf{T}}_j^i$ is yawed about $180\,\text{deg}$ from the true $\mathbf{T}_j^i$ while our method harnesses the information from segment shape similarities and the gravity vector to compute an accurate $\hat{\mathbf{T}}_j^i$.

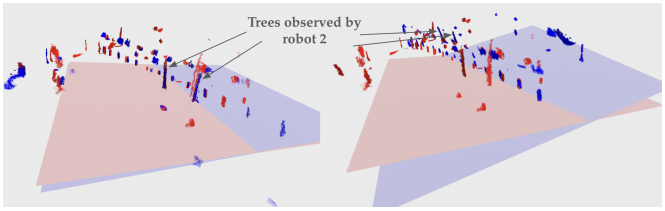

Fig. 4: Comparison of our proposed method (left) versus CLIPPER (right). The submap created by robot 2 (marked in blue) was transformed using the estimate $\hat{\mathbf{T}}_j^i$. The x-y planes of each robot's frames are shown to demonstrate the quality of computed transformations.

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.

Acknowledgments

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

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] 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 *IEEE/RSJ IROS*, 2023.

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

[4] 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.

[5] 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," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 339–355, 2020.

[6] 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 *2021 IEEE/RSJ IROS*, pp. 1183–1190, IEEE, 2021.

[7] J. Yu and S. Shen, "Semanticloop: loop closure with 3d semantic graph matching," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 568–575, 2022.

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

[9] J. Ankenbauer, P. C. Lusk, A. Thomas, and J. P. How, "Global localization in unstructured environments using semantic object maps built from various viewpoints," in *2023 IEEE/RSJ IROS*, pp. 1358–1365, IEEE, 2023.

[10] P. C. Lusk and J. P. How, "CLIPPER: robust data association without an initial guess," *IEEE Robotics and Automation Letters*, 2024.

[11] 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.

[12] 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.

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

[14] M. B. Peterson, P. C. Lusk, A. Avila, and J. P. How, "MOTLEE: collaborative multi-object tracking using temporal consistency for neighboring robot frame alignment," *arXiv preprint arXiv:2405.05210*, 2024.

[15] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," *arXiv preprint arXiv:1801.09847*, 2018.

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