# ROMAN: Open-Set Object Map Alignment for Robust View-Invariant Global Localization

Mason B. Peterson<sup>1</sup>, Yi Xuan Jia<sup>1</sup>, Yulun Tian<sup>2</sup>, Annika Thomas<sup>1</sup>, and Jonathan P. How<sup>1</sup>

Abstract—Global localization is a fundamental capability required for long-term and drift-free robot navigation. However, current methods fail to relocalize when faced with significantly different viewpoints. We present ROMAN (Robust Object Map Alignment Anywhere), a robust global localization method capable of localizing in challenging and diverse environments based on creating and aligning maps of open-set and viewinvariant objects. To address localization difficulties caused by feature-sparse or perceptually aliased environments, ROMAN formulates and solves a registration problem between object submaps using a unified graph-theoretic global data association approach that simultaneously accounts for object shape and semantic similarities and a prior on gravity direction. Through a set of challenging large-scale multi-robot or multi-session SLAM experiments in indoor, urban and unstructured/forested environments, we demonstrate that ROMAN achieves a maximum recall 36% higher than other object-based map alignment methods and an absolute trajectory error that is 37% lower than using visual features for loop closures. Our project page can be found at https://acl.mit.edu/ROMAN/.

Index Terms—Localization, Mapping, Visual-Inertial SLAM, Multi-Robot SLAM

## I. INTRODUCTION

Global localization [1] refers to the task of localizing a robot in a reference map produced in a prior mapping session or by another robot in real-time, *i.e.*, inter-robot loop closures in collaborative SLAM [2]. It is a cornerstone capability for drift-free navigation in GPS-denied scenarios. In this letter, we consider global localization using *object-* or *segment-level* representations, which have been shown by recent works [3–6] to hold great promise in challenging domains that involve drastic changes in viewpoint, appearance, and lighting.

At the heart of object-level localization is a *global data* association problem, which requires finding correspondences between observed objects and existing ones in the map without an initial guess. Earlier approaches such as [7–10] rely on geometric verification based on RANSAC [11], which exhibits intractable computational complexity under high outlier regimes. Recently, graph-theoretic approaches [4, 12–15] have emerged as a powerful alternative that demonstrates superior accuracy and robustness when solving the correspondence problem. In particular, methods based on consistency graphs [12–15] formulate a graph where nodes denote putative object correspondences and edges denote their geometric consistencies. The data association problem is then solved by extracting

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

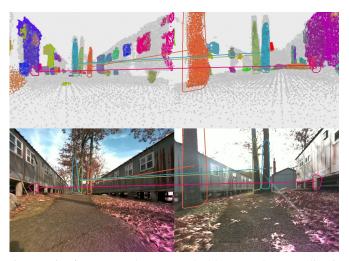


Fig. 1: Pair of segment submaps matched by two robots traveling in *opposite* directions in an off-road environment. Associated segments found by the proposed method are connected by lines and projected onto the image plane. (**Top**) Each pair of associated segments is drawn with the same color. The remaining, unmatched segments are shown in random colors and all other background points are shown in gray. (**Bottom**) The same associated segments and their convex hulls are visualized in the original image observations.

large and densely connected subsets of nodes yielding the desired set of *mutually consistent* correspondences. While these methods represent the current state of the art, their performance is severely limited in challenging regimes where mutual geometric consistency is not sufficient. This is the case, for example, when the environment contains few objects or when the objects' spatial configurations are highly ambiguous.

In this letter, we address the aforementioned technical gap by extending graph-theoretic data association to use information beyond mutual (pairwise) geometric consistency. We develop a unified graph-theoretic formulation that incorporates: (i) *open-set semantics*, extracted as semantically meaningful 3D segments [16, 17] with descriptors obtained from vision-language foundation model, CLIP [18]; (ii) *segment-level geometric attributes*, such as the volume and 3D shapes of segments that provide additional discriminative power; and (iii) an *additional prior* about gravity direction that is readily available from onboard inertial sensors. By fusing this information in the data association formulation, our approach significantly improves the state-of-the-art [12] in terms of both precision and recall metrics.

**Contributions.** We present ROMAN (Robust Object Map Alignment Anywhere), a robust global localization method in challenging unseen environments. In detail, ROMAN consists of the following contributions:

<sup>&</sup>lt;sup>1</sup>Massachusetts Institute of Technology, Cambridge, MA 02139, USA. {masonbp, yixuany, annikat, jhow}@mit.edu.

<sup>&</sup>lt;sup>2</sup>University of California San Diego, San Diego, CA 92093, USA. yut034@ucsd.edu

- A pipeline for creating open-set 3D segment maps from a single onboard RGB-D camera, using FastSAM [17] for open-set image segmentation and CLIP [18] for computing open-set feature descriptors. These maps compactly summarize the detailed RGB-D point clouds into sparse and view-invariant representations consisting of segment locations and metric-semantic attributes, which enable efficient and robust global localization.
- 2) An extension of the graph-theoretic global data association method of [12] to incorporate segment-level similarities computed using CLIP descriptors and geometric attributes based on shape and volume. When gravity direction is known, a gravity-direction prior is also utilized. Our method implicitly guides the solver to correct 3D segment-to-segment associations in challenging regimes when object centroids alone are insufficient for identifying correct associations (*e.g.*, due to repetitive geometric structures or scenes with few distinct objects).
- 3) Extensive experimental evaluation of the proposed method using real-world datasets that involve both urban and off-road scenarios (see Fig. 1). Our approach improves global localization recall by 36% in challenging problem instances involving large viewpoint changes. When using ROMAN rather than visual features for inter-robot loop closures, our method reduces the overall localization error by 7.6% on large-scale collaborative SLAM problems involving 6-8 robots and by 37% on a subset of particularly challenging sequences.

## II. RELATED WORKS

Object-based maps are lightweight environment representations that enable robots to match perceived objects with previously built object maps using object geometry or semantic labels as cues for object-to-object data association. Compared to conventional keypoints extracted from visual or lidar observations, *object-* or *segment-level* representations are more stable against sensor noise and viewpoint, lighting, or appearance changes, which often cause visual feature-based methods to fail [19]. Furthermore, these representations are lightweight and efficient to transmit, an important criterion for multi-robot systems. In this section, we review related methods for using object maps for global localization and SLAM.

**Object SLAM.** To incorporate discrete objects into SLAM, sparse maps of objects are described with geometric primitives such as points [20], cuboids [21] or quadrics [22]. SLAM++ [3] trains domain-specific object detectors for objects like tables and chairs. Choudhary *et al.* [23] use objects as landmarks for localization, providing a database of discovered objects. Lin *et al.* [24] showed that semantic descriptors can improve frame-to-frame object data association. Recent works [6, 25] further leverage *open-set* semantics from pre-trained models. Other methods [26, 27] combine the use of coarse objects for high-level semantic information with fine features for high accuracy in spatial localization. Object-level mapping also conveniently handles dynamic parts of an environment which can be naturally described at an object level [28, 29].

Random sampling for object-based global localization. Object-level place recognition may be performed by an initial coarse scene matching procedure (e.g., matching bag-of-words descriptors for scenes [30]) but is commonly solved in conjunction with the object-to-object data association by attempting to associate objects and accepting localization estimates when object matches are good [5, 31]. Object-to-object data association may be solved by sampling potential rotation and translation pairs between maps [6] or object associations [7–10] using RANSAC [11]. Random sampling methods often require significant computation for satisfactory results and the probability of finding correct inlier associations diminishes exponentially as the number of outliers grows [32].

Graph matching for object-based global localization. Recently, graph-based methods have emerged as a fast and accurate alternative for object data association. Objects are represented as nodes in a graph with graph edges encoding distance between objects [4, 31, 33]. Data association can be performed by matching small, local target graphs with the prior map graph using graph-matching techniques.

Maximal consistency for object-based global localization. Different from graph-matching methods, consistency graph algorithms use nodes to represent potential associations between two objects in different datasets, and edges to encode consistency between pairs of associations. Data associations are found by selecting large subsets of mutually consistent nodes (associations), which can be formulated as either a maximum clique [13–15] or densest subgraph [12] problem. Ankenbauer et al. [34] leverage graph-theoretic data association [12] as the back-end association solver to perform global localization in challenging outdoor scenarios. Matsuzaki et al. [35] use semantic similarity between a camera image and a predicted image to evaluate pairwise consistency. Thomas et al. [5] use pre-trained, open-set foundation models for zero-shot segmentation in novel environments for open-set object map alignment. Our method extends these prior works by incorporating object-to-object similarity and an additional pairwise association prior used to guide the optimization to correct associations.

Inter-Robot Loop Closures for Collaborative SLAM. In the context of multi-robot collaborative SLAM (CSLAM), our approach serves to detect inter-robot loop closures that fuses individual robots' trajectories and maps. State-of-theart CSLAM systems [36-40] commonly adopt a two-stage loop closure pipeline, where a place recognition stage finds candidate loop closures by comparing global descriptors and a geometric verification stage finds the relative pose by registering the two keyframes. To improve loop closure robustness, Mangelson et al. [13] proposes pairwise consistency maximization (PCM) which extracts inlier loop closures from candidate loop closures by solving a maximum clique problem. Do et al. [41] extends PCM [13] by incorporating loop closure confidence and weighted pairwise consistency. Choudhary et al. [42] performs inter-robot loop closure via object-level data association; however, a database of 3D object templates is required. Hydra-Multi [43] employs hierarchical inter-robot loop closure that includes places, objects, and visual features summarized in a scene graph.

## III. ROMAN

This section presents an overview of the full ROMAN system, shown in the system diagram in Fig. 2. We first introduce our pipeline for creating open-set segment-level maps in Section III-B. Section III-C presents a high-level overview of our global localization method which aligns small submaps of mapped objects. The details of our object-level data association method are deferred until Section IV. Finally, we discuss the incorporation of our object-based loop closures in pose graph optimization in Section III-D.

## A. Notation

We use boldfaced lowercase and uppercase letters to denote vectors and matrices, respectively. We define  $[n] = \{1,2,...,n\}$ . For any  $n \in \mathbb{N}$  and  $x_1,...,x_n \in \mathbb{R}$ , we use  $GM(x_1,...,x_n) \triangleq (\prod_{i=1}^n x_i)^{\frac{1}{n}}$  to denote the geometric mean of  $x_1,...,x_n$ , and  $GM(\mathbf{x})$  to denote the geometric mean of the elements of the vector  $\mathbf{x}$ . For any vectors  $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ , their cosine similarity is denoted as  $\cos_{\sin(\mathbf{x},\mathbf{y})} \triangleq \frac{\langle \mathbf{x},\mathbf{y} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$ . We define the element-wise operation  $\operatorname{ratio}(\mathbf{x},\mathbf{y}) \triangleq \min(\frac{\mathbf{x}}{\mathbf{y}},\frac{\mathbf{y}}{\mathbf{x}})$ , where  $\min$  and  $\frac{\mathbf{x}}{\mathbf{y}}$  are also performed element-wise.

# B. Open-set object-level mapping

To enable global localization in previously unseen environments, the proposed method constructs segment-level maps based on high-level features detected by recent zero-shot openset segmentation models. The inputs consist of RGB-D images and robot pose estimates (*e.g.*, provided by a visual-inertial odometry system). Pedestrians are filtered out using YOLO-V7 [44], and a CLIP [18] embedding vector is computed for each remaining segment using its minimal bounding box. The 2D segments are projected to 3D voxels using depth images, and 3D voxels and CLIP embeddings are then fed into a frame-to-frame data association and tracking module.

Data association is performed between existing 3D segment tracks and incoming 3D observations by computing the gridaligned voxel-based IOU between pairs of tracks and observations with 3D voxel overlap [29]. We use a global nearest neighbor approach [45] to assign observations to existing object tracks and create new tracks for any unassociated observation. Semantic descriptors of the associated segments are fused by taking the mean of all of the normalized CLIP embeddings. Because FastSAM may segment objects differently depending on the view, we create a merging mechanism to avoid duplications of the same object. Specifically, 3D segments are merged based on high grid-aligned voxel IOU or when a projection of the two segments onto the image plane results in a high 2D IOU. The result of our mapping pipeline is a set of open-set 3D objects with an abstractable representation. While performing mapping, objects are represented by dense voxels helping the frame-to-frame data association and object merging. However, our global localization only uses a low-data representation of segments consisting of centroid position, shape attributes, and mean semantic embedding. This enables robots to store and communicate maps efficiently.

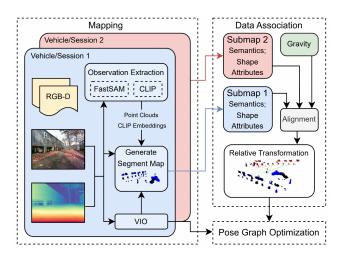


Fig. 2: ROMAN has three modules: mapping, data association, and pose graph optimization. The front-end mapping pipeline tracks segments across RGB-D images to generate segment maps. The data association module incorporates semantics and shape geometry attributes from submaps along with gravity as a prior into the ROMAN alignment module to align maps and detect loop closures. These loop closures and VIO are then used for pose graph optimization.

## C. Submap Alignment

To perform global localization, we divide each robot's on-board segment map into a sequence of potentially overlapping submaps. A submap  $\mathcal{M}_i$  is associated with a center pose  $\mathbf{T}^i_{\mathcal{M}_i}$  and center time  $t_{\mathcal{M}_i}$ , where  $\mathcal{M}_i$  consists of a set of objects within some radius r whose centroid positions are expressed in the map frame  $\mathcal{F}_{\mathcal{M}_i}$ . To keep submap computation controlled, we use a maximum submap size N and remove objects (starting at objects farthest from the center) so that the submap size is  $\leq N$ . After a robot has traveled a distance  $c_d$  from the last submap center, a new submap is created.

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 containing geometric and semantic attributes. Successful global localization requires that submap overlap is correctly detected and 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. Once correspondences between  $\mathcal{M}_i$  and  $\mathcal{M}_j$  have been determined, the relative transformation from  $\mathcal{F}_j$  to  $\mathcal{F}_i$ ,  $\hat{\mathbf{T}}_j^i$ , can be found using the closed-form Arun's method [46].

# D. Pose Graph Optimization

We incorporate our segment-based loop closures into the robust multi-robot pose graph solver from [36]. To do this, we attempt to register each submap from robot i with each submap from robot j. Additionally, intra-robot loop closures are found by registering every pair of submaps from robot i provided enough time has passed. Loop closures are rejected when the number of correspondences found by the data association module  $<\tau$ , where  $\tau$  is a parameter that can be used to reject unlikely loop closures. Finally, loop closure and odometry keyframes are fed into the robust pose graph optimization of [36] to estimate multi-robot trajectories.

## IV. ROBUST OBJECT MAP DATA ASSOCIATION

In this section, we present in depth our proposed data association framework. We will first briefly review the approach behind CLIPPER and then describe the proposed extension.

## A. 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  $d(a_p, a_q) \triangleq |\|\mathbf{c}(p_i) - \mathbf{c}(q_i)\| - \|\mathbf{c}(p_j) - \mathbf{c}(q_j)\||$  is less than a threshold  $\epsilon$ , where  $\mathbf{c}(\cdot) \in \mathbb{R}^3$  is centroid position of a segment. In this case, a weighted edge between  $a_p$  and  $a_q$  is created with weight  $s_a(a_p, a_q) \triangleq \exp\left(-\frac{1}{2}\frac{d(a_p, a_q)^2}{\sigma^2}\right)$ . Intuitively,  $s_a(a_p, a_q) \in [0, 1]$  scores the consistency between two associations, and  $\epsilon$  and  $\sigma$  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,

$$\max_{\mathbf{u} \in \{0,1\}^n} \frac{\mathbf{u}^{\top} \mathbf{M} \mathbf{u}}{\mathbf{u}^{\top} \mathbf{u}}.$$
subject to  $\mathbf{u}_p \mathbf{u}_q = 0$  if  $\mathbf{M}_{p,q} = 0, \ \forall_{p,q}$ ,

where  $\mathbf{u}_p$  is 1 when association  $a_p$  is accepted as an inlier and 0 otherwise. See [12] for more details.

# B. Improving affinity metrics M: general strategies

In its original form, the affinity matrix M in Equation (1) relies solely on distance information between pairs of centroids. However, when applied to 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 mapped in both  $\mathcal{M}_i$  and  $\mathcal{M}_j$ , which can lead to ambiguity when performing segment submap registration. To address these problems, other works [5, 47] have proposed pre-processing or post-processing methods that leverage additional information such as segment size and gravity direction to filter incorrect object associations.

In comparison to works that use prior information in preprocessing or post-processing steps, ROMAN directly incorporates this information in the underlying optimization problem in Equation (1). The key to our approach is to extend the original similarity metric to (i) use additional geometric (e.g., volume, spatial extent) and semantic (e.g., CLIP embeddings) attributes to disambiguate segments, and (ii) directly incorporate knowledge of the gravity direction (when available) to guide the data association solver. Consider the putative association  $a_p = (p_i, p_j)$ . Intuitively, if objects  $p_i$  and  $p_j$  are dissimilar, then the association  $a_p$  is less likely to be correct, which should be represented in the data association optimization formulation of Equation (1). Given a segment similarity score  $s_o(a_p)$  comparing objects  $p_i$  and  $p_j$ , [12] and [48] suggest setting the diagonal entries of M to reflect object similarity information, e.g., by setting  $\mathbf{M}_{p,p} = s_o(a_p)$ ; however, expanding the numerator of Equation (1) shows that this approach has limited impact,

$$\mathbf{u}^{\top} \mathbf{M} \mathbf{u} = \Sigma_{p \in [n]} \left( \mathbf{M}_{p,p} \mathbf{u}_{p}^{2} + \Sigma_{q \in [n], q \neq p} \left( \mathbf{M}_{p,q} \mathbf{u}_{p} \mathbf{u}_{q} \right) \right). \tag{2}$$

As the dimension of M increases, the number of off-diagonal terms (pairwise association affinity terms) increases quadratically and will quickly dominate the overall objective function. Alternatively, [41] and [4] propose multiplying the association affinity score by  $s_o(\cdot)$  so that  $\mathbf{M}_{p,q} = s_a(a_p, a_q)s_o(a_p)s_o(a_q)$ . While this gives segment-to-segment similarity a significant role in the registration problem, the elements of 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 M, we instead propose using the *geometric mean*,

$$\mathbf{M}_{p,q} = GM(s_a(a_p, a_q), s_o(a_p), s_o(a_q)),$$
 (3)

In this work, prior information is incorporated into the optimization problem (1) through careful designs of  $s_a(\cdot, \cdot)$  and  $s_o(\cdot)$ , which will be explained in the subsequent subsections.

# C. Improving affinity metrics M: incorporating metricsemantic segment attributes

In this subsection, we design the segment-to-segment similarity score  $s_o(\cdot)$  by comparing geometric and semantic attributes of the mapped segments. From the relatively dense point-cloud representation created for online mapping, a low-data shape descriptor and the averaged semantic feature descriptor are extracted for each 3D segment. These descriptors are compared using a shape similarity scoring function  $s_{\text{shape}}(\cdot)$  and a semantic similarity score  $s_{\text{semantic}}(\cdot)$ , which we present next. The final segment-to-segment similarity score  $s_o(\cdot)$  is set to be the geometric mean of those two scores.

- a) Semantic similarity metric: To incorporate semantic information, we define the segment-to-segment semantic similarity score by taking the cosine similarity of their CLIP descriptors:  $s_{\text{semantic}}(a_p) = \cos_{\text{sim}}(\text{CLIP}(p_i), \text{CLIP}(p_j))$ . We observe that the cosine similarity score of pairs of CLIP embeddings from images is usually higher than 0.7, which does not allow semantic similarity to play a significant role in determining data associations in Equation (1). We propose to rescale the cosine similarity score using hyperparameters  $\phi_{\text{min}}$  and  $\phi_{\text{max}}$ , so that scores less than  $\phi_{\text{min}}$  are set to 0, scores larger than  $\phi_{\text{max}}$  are set to 1.0, and scores between  $\phi_{\text{min}}$  and  $\phi_{\text{max}}$  are scaled linearly so that they range from 0 to 1.
- b) Shape similarity metric: To incorporate segment shape attributes, we define a segment-to-segment shape similarity score:

$$s_{\text{shape}}(a_p) = \text{GM}\left(\text{ratio}(\mathbf{f}(p_i), \mathbf{f}(p_i))\right),$$
 (4)

where  $\mathbf{f}(p)$  returns a four-dimensional vector of the shape attributes of p and is defined as follows. For each segment p,  $\mathbf{f}_1(p)$  is the volume of the bounding box created from the point cloud of segment p, and  $\mathbf{f}_2(p)$ ,  $\mathbf{f}_3(p)$ , and  $\mathbf{f}_4(p)$  denote the linearity, planarity, and scattering attributes of the 3D points computed via principle component analysis (PCA). The interested reader is referred to [49] for details. The scoring function  $s_{\text{shape}}(\cdot) \in [0,1]$  allows direct feature element-to-element scale comparison. Intuitively, if one element is much larger than the other, the score will be near 0, while if the element is very similar in scale,  $s_{\text{shape}}$  will be close to 1.

## D. Improving affinity metrics M: incorporating gravity prior

We additionally address implicitly incorporating knowledge of the gravity direction in the global data association formulation. Due to the geometric-invariant formulation of Equation (1), the solver naturally considers registering object maps as a 6-DOF problem. Often in robotics, an onboard IMU makes the direction of the gravity vector well-defined, so we are only interested in transformations with x, y, z, and yaw components. Because the optimization variable of Equation (1) is a set of associations rather than a set of transformations, it is not immediately clear how to leverage this information within the optimization problem, motivating the postprocessing rejection step from [5]. In this work, we propose a method to leverage this extra knowledge within the data association step by re-designing  $s_a(\cdot,\cdot)$  to guide the solver to select associations that are consistent with the direction of the gravity vector. Specifically, we represent this prior knowledge of the gravity vector by decoupling computations in the x-yplane and along the z axis:

$$s_a(a_p, a_q) = \exp\left(-\frac{1}{2}\left(\frac{d_{xy}^2(a_p, a_q)}{\frac{2}{3}\sigma^2} + \frac{d_z^2(a_p, a_q)}{\frac{1}{3}\sigma^2}\right)\right), (5)$$

where

$$d_{xy}(a_p, a_q) = ||\mathbf{c}_{xy}(p_i) - \mathbf{c}_{xy}(q_i)|| - ||\mathbf{c}_{xy}(p_j) - \mathbf{c}_{xy}(q_j)|| |$$

$$d_z(a_p, a_q) = ||\mathbf{c}_z(p_i) - \mathbf{c}_z(q_i)| - ||\mathbf{c}_z(p_j) - \mathbf{c}_z(q_j)||.$$

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 in scenarios where distance information is insufficient.

## V. EXPERIMENTS

In this section, we evaluate ROMAN in an extensive series of diverse, real-world experiments. Our evaluation settings consist of urban domains from the large-scale Kimera-Multi datasets [19], off-road domains in an unstructured, natural environment, and ground-aerial localization in a manually constructed, cluttered indoor environment. Experimental results demonstrate that ROMAN achieves superior performance compared to existing baseline methods, obtaining up to 36% improvement in overall recall and 37% improvement in final trajectory estimation error in a subset of particularly challenging sequences from the Kimera-Multi datasets.

Baselines. We compare the alignment performance of ROMAN against the following baselines. RANSAC-100K and RANSAC-1M apply RANSAC [11], as implemented in [50], on segment centroids with a max iteration count of 100,000 and 1 million respectively. CLIPPER runs standard CLIPPER [12] on segment centroids, and CLIPPER + Prune prunes initial putative associations using volume and rejects incorrect registration results using gravity information (so it has access to similar information as the proposed method). Finally, Visual Features uses BoW descriptors of ORB features for global localization as implemented in Kimera-Multi [36].

Performance metrics. We use the following metrics for comparing global localization and full SLAM results. Recall is evaluated by giving a global localization method a set of submaps from two robots, where some submaps may overlap yielding opportunities for global localization. A global localization prediction is determined to be correct if an algorithm decides that a pair of submaps overlaps and correctly computes the transformation with less than 5 m translation and  $10\deg$  rotation error. We evaluate the success rate on submaps with an area of overlap  $> \frac{2}{3}$  the area of a single submap. Precision is defined as the number of correct global localization instances divided by the number of hypothesized overlapping submaps, and recall is defined as the number of correct global localization instances divided by the number of true overlapping submaps. Pose graph optimization is evaluated using root mean squared (RMS) absolute trajectory error (ATE) between the multi-robot registered estimated and ground truth trajectories. We use open-source evo [51] to compute ATE.

**Hyperparameters.** For global localization, we use the parameter values outlined in Table I. In pose graph optimization, we use odometry covariances with uncorrelated rotation and translation noise parameters. We use standard deviations of 0.1 m and 0.5 deg for odometry and 1.0 m and 2.0 deg for loop closures.

TABLE I: Parameters

Parameter	Value	Description		
σ/ε	0.4 m / 0.6 m	Pairwise consistency noise parameters		
$r / c_d$	$15\mathrm{m}$ / $10\mathrm{m}$	Submap radius and spacing distance		
$\phi_{\min}$ / $\phi_{\max}$	0.85 / 0.95	Cosine similarity scaling values		
$\tau$ / $N$	4 / 40	Association threshold and max submap size		

## A. MIT Campus Global Localization

We first evaluate ROMAN's map alignment using the outdoor Kimera-Multi Dataset [19] recorded on MIT campus. Each robot creates a set of submaps using Kimera-VIO [52] for odometry and we attempt to align each submap with every other submap to search for inter and intra robot loop closures as described in Section III. This amounts to over 120,000 pairs of submaps that are given as inputs to the methods where 420 of those pairs have  $\geq \frac{2}{3}$  overlap in terms of area covered.

In Fig. 3, we show combined plots of recall as a function of precision, distance error in estimated transformation, and heading difference when observing the pair of submaps. The precision-recall plot is created by varying tau and gives

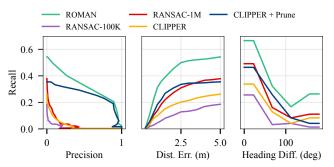


Fig. 3: Performance comparison between ROMAN and baseline methods. ROMAN achieves better precision and recall than baselines, aligns maps with less error, and has greater alignment success in challenging opposite-view alignment cases than baselines.

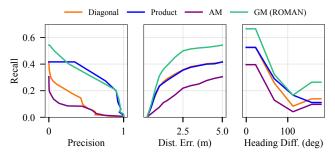


Fig. 4: Comparison of fusion methods. We show that our method of using the geometric mean (GM) to fuse  $s_o$  and  $s_a$  outperforms fusing these methods with arithmetic mean (AM), by taking the product, or by altering only the diagonal elements of the affinity matrix. Other than geometric mean, the most competitive fusion method is taking the product, which is also able to boost precision since, like geometric mean, it strongly penalizes inconsistent  $s_o$  scores. However, it tends to over-penalize (since including  $s_o$  only hurts the overall similarity score) resulting in a strong cutoff at recall of only 0.417 compared to the 0.543 max recall of geometric mean.

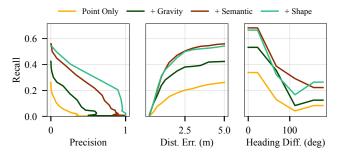


Fig. 5: Cumulative contribution of each similarity component of ROMAN. Using gravity to guide pairwise scores  $s_a$  and semantics in object-to-object scores  $s_o$  improves both precision and recall results. While adding shape similarity does not have a significant benefit in maximum recall, the boost in precision improves ROMAN's alignment results to be used in our full SLAM system.

intuition on a method's ability to determine a correct map alignment while rejecting incorrect alignments. Additionally, we compare recall as we vary the distance error in the alignment to demonstrate the accuracy of the map alignments. Finally, we compare recall at different relative true heading angles to show how methods compare in more challenging viewpoint scenarios. When the heading difference is small, alignment is comparatively easier. While opposite viewpoints create the most challenging scenario in terms of difference in the appearance of objects, crossing paths perpendicularly means fewer common objects are seen compared to traveling in opposite directions, making heading difference near  $90\deg$  the most difficult scenarios to align correctly.

Fig. 3 shows that ROMAN achieves higher precision/recall, aligns maps with less error, and aligns maps from scenes viewed from different heading angles more successfully than baseline methods. For same-direction matchings (with heading difference around zero), ROMAN has a max recall of 0.67. For opposite-direction matchings (with heading difference around 180 deg), ROMAN achieves max recall of 0.26, much higher than the next best method RANSAC-1M which only achieves a recall of 0.11. Furthermore, ROMAN is 4.5 times faster than RANSAC-1M as shown in Table II. In terms of communication and submap storage size, each object includes a 3D centroid, a four-dimensional shape descriptor and a 768-dimensional semantic descriptor. With each submap consisting of at most N=40 objects, a submap packet size is strictly less than 250 KB. For a trajectory of length 1 km, the entire map would require less than 24 MB of storage.

TABLE II: Mean Submap Registration Computation Time (ms)

RANSAC-100K	RANSAC-1M	CLIPPER	CLIPPER+Prune	ROMAN
75.9	488.4	48.2	19.3	108.3

## B. Global Localization Ablation Study

In addition to comparing submap alignment results between different object data association methods, we compare different methods for fusing object similarity scores  $s_o$  with pairwise scores  $s_a$  in Fig. 4. Compared fusion methods include geometric mean ROMAN), product [4, 41], arithmetic mean, and setting the diagonal elements of the affinity matrix  $M_{pp} = \mathrm{GM}(s_o(a_p)s_o(a_q))$  [12, 48]. Fusing scores using the arithmetic mean has fewer zeroed-out elements of the affinity matrix which results in the optimization problem becoming less well constrained. Fusing via the product of scores does make a significant difference in recall, but ultimately gets a maximum recall of only 41.7 compared to the 54.3% max recall of ROMAN's geometric mean method.

Fig. 5 shows an ablation demonstrating the performance boost given from each of the gravity, semantic, and geometric components that are incorporated into ROMAN. While adding gravity and semantic components both improve the recall of our method, adding object shape similarity benefits precision, which is important for integrating global localization into a full SLAM system.

## C. Pose Graph SLAM Results

We run our full SLAM pipeline on the tunnel, hybrid, and outdoor Kimera-Multi datasets and compare the RMSE ATE of the estimated multi-robot trajectories, as reported in Table III. The results show that ROMAN's ability to get loop closures in challenging visual scenarios results in moderate improvements to the overall ATE. This is because by the experiment design, the robot paths are well connected and most loop closure opportunities occur when robots are

TABLE III: 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

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

Robot Subset	Visual Features	ROMAN	Combined
Hybrid: robots 1, 2, 3	10.34	6.91	5.17
Hybrid: robots 4, 5	6.11	2.80	3.43
Outdoor: 1, 3	10.12	7.67	7.97

traveling in the same direction, leading to limited opportunity for significant improvement. However, Table IV shows that a subset of robot trajectories that contain difficult instances for visual loop closures (*e.g.*, perpendicular path crossing and scenes with high visual aliasing) and show that ROMAN has a significantly lower ATE in these challenging scenarios.

# D. Evaluation in 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 (Fig. 1). In this experiment, data is recorded on Clearpath Jackals using Intel RealSense D455 to capture RGB-D images and Kimera-VIO [52] for odometry. We run the ROMAN pipeline on three different pairs of robot trajectories. We compare ROMAN to Visual Features loop closures in Fig. 6. The three pairs consists of an easy, medium, and hard case. The easy case involves two robots that traverse the same loop in the same direction (with one robot who leaves the loop and later returns). In the medium difficulty case, the robots travel in opposite directions except for a short section in the middle where both robots briefly view the scene from the same direction. Finally, in the difficult case, robots travel in a large loop in opposite directions. While ground-truth pose is not available for this data, Fig. 6 shows qualitatively shows that ROMAN successfully detects loop closures in all three cases. More importantly, ROMAN successfully closes loops in opposite-direction traversals, while loop closures from Visual Features only work reliably in same-direction traversals and fail to align trajectories in the hard case.

## E. Evaluation in Ground-Aerial Cross View Localization

Finally, we evaluate ROMAN's robustness to view changes by conducting indoor localization experiments where segment maps created from ground views are aligned with segment maps created from aerial views. Snapshots of the setup from both views are shown in Fig. 7. We test on 20 ground-aerial pairs of trajectories through the constructed environment. One segment map is created for each trajectory and localization is performed between segment maps from the aerial and ground views. We report maximum recall and data association accuracy in Table V. Association accuracy is calculated by computing the ratio between the number of correctly associated objects and the number of all associated objects for each method, averaged over all the segment map pairs. We show that ROMAN maintains an advantages over

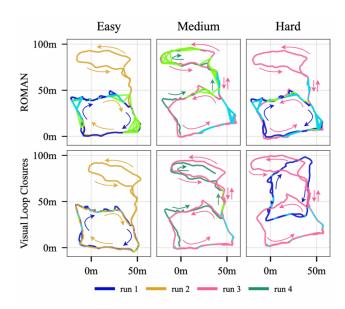


Fig. 6: Off-road qualitative pose graph trajectory estimate. Easy, medium, and hard cases comparing using ROMAN and visual feature for loop closures. Blue lines (—) represent single-robot loop closures and green lines (—) show multi-robot loop closures. Easy is all same direction, medium is all opposite direction except for the small connecting neck, and hard is all opposite direction. Using visual features, no loop closures are found when going opposite directions.



Fig. 7: Setup of the indoor localization task from both ground view (left) and aerial view (right).

TABLE V: Analysis of the Indoor Localization Task

	RANSAC-100K	RANSAC-1M	CLIPPER	CLIPPER+Prune	ROMAN
Assoc. Accuracy	0.276	0.556	0.347	0.277	0.712
Recall	0.1	0.5	0.3	0.2	0.7

other baselines, demonstrating its global localization capability in aerial-ground setups.

## VI. CONCLUSION

This work presented ROMAN, 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, object shape and semantic attributes, and the direction of the gravity vector in object 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] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. Kelly, and A. J. Davison, "Slam++: Simultaneous localisation and mapping at the level of objects," in CVPR, 2013, pp. 1352–1359.
- [4] J. Yu and S. Shen, "Semanticloop: loop closure with 3d semantic graph matching," RA-L, vol. 8, no. 2, pp. 568–575, 2022.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] G. Tinchev, S. Nobili, and M. Fallon, "Seeing the wood for the trees: Reliable localization in urban and natural environments," in *IROS*. IEEE, 2018.
- [9] 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.
- [10] 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.
- [11] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [12] P. C. Lusk and J. P. How, "CLIPPER: robust data association without an initial guess," RA-L, 2024.
- [13] J. G. Mangelson, D. Dominic, R. M. Eustice, and R. Vasudevan, "Pairwise consistent measurement set maximization for robust multirobot map merging," in *ICRA*. IEEE, 2018, pp. 2916–2923.
- [14] 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. 13820–13827.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *ICRA*. IEEE, 2017, pp. 1722–1729.
- [21] S. Yang and S. Scherer, "Cubeslam: Monocular 3-d object slam," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 925–938, 2019.
- [22] L. Nicholson, M. Milford, and N. Sünderhauf, "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam," *RA-L*, vol. 4, no. 1, pp. 1–8, 2018.
- [23] S. Choudhary, A. J. Trevor, H. I. Christensen, and F. Dellaert, "Slam with object discovery, modeling and mapping," in *IROS*. IEEE, 2014, pp. 1018–1025.
- [24] S. Lin, J. Wang, M. Xu, H. Zhao, and Z. Chen, "Topology aware object-level semantic mapping towards more robust loop closure," *RA-L*, vol. 6, no. 4, pp. 7041–7048, 2021.
- [25] D. Maggio, Y. Chang, N. Hughes, M. Trang, D. Griffith, C. Dougherty, E. Cristofalo, L. Schmid, and L. Carlone, "Clio: Real-time task-driven open-set 3d scene graphs," arXiv preprint arXiv:2404.13696, 2024.
- [26] Y. Wang, C. Jiang, and X. Chen, "Voom: Robust visual object odometry and mapping using hierarchical landmarks," arXiv preprint arXiv:2402.13609, 2024.
- [27] M. Zins, G. Simon, and M.-O. Berger, "Oa-slam: Leveraging objects for camera relocalization in visual slam," in *ISMAR*. IEEE, 2022, pp. 720–728
- [28] R. Tian, Y. Zhang, Z. Cao, J. Zhang, L. Yang, S. Coleman, D. Kerr,

- and K. Li, "Object slam with robust quadric initialization and mapping for dynamic outdoors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 10, pp. 11080–11095, 2023.
- [29] 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.
- [30] N. Hughes, Y. Chang, S. Hu, R. Talak, R. Abdulhai, J. Strader, and L. Carlone, "Foundations of spatial perception for robotics: Hierarchical representations and real-time systems," *IJRR*, 2024.
- [31] A. Gawel, C. Del Don, R. Siegwart, J. Nieto, and C. Cadena, "X-view: Graph-based semantic multi-view localization," *RA-L*, vol. 3, no. 3, pp. 1687–1694, 2018.
- [32] R. Raguram, J.-M. Frahm, and M. Pollefeys, "A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus," in ECCV. Springer, 2008, pp. 500–513.
- [33] Y. Wang, C. Jiang, and X. Chen, "Goreloc: Graph-based object-level relocalization for visual slam," RA-L, 2024.
- [34] 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 *IROS*. IEEE, 2023, pp. 1358–1365.
- [35] S. Matsuzaki, K. Koide, S. Oishi, M. Yokozuka, and A. Banno, "Single-shot global localization via graph-theoretic correspondence matching," *Advanced Robotics*, vol. 38, no. 3, pp. 168–181, 2024.
- [36] 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.
- [37] P. Schmuck, T. Ziegler, M. Karrer, J. Perraudin, and M. Chli, "Covins: Visual-inertial slam for centralized collaboration," in *ISMAR*. IEEE, 2021, pp. 171–176.
- [38] P.-Y. Lajoie and G. Beltrame, "Swarm-slam: Sparse decentralized collaborative simultaneous localization and mapping framework for multirobot systems," RA-L, vol. 9, no. 1, pp. 475–482, 2023.
- [39] Y. Huang, T. Shan, F. Chen, and B. Englot, "Disco-slam: Distributed scan context-enabled multi-robot lidar slam with two-stage global-local graph optimization," *RA-L*, vol. 7, no. 2, pp. 1150–1157, 2021.
- [40] Y. Chang, K. Ebadi, C. E. Denniston, M. F. Ginting, A. Rosinol, A. Reinke, M. Palieri, J. Shi, A. Chatterjee, B. Morrell *et al.*, "Lamp 2.0: A robust multi-robot slam system for operation in challenging largescale underground environments," *RA-L*, vol. 7, no. 4, pp. 9175–9182, 2022
- [41] H. Do, S. Hong, and J. Kim, "Robust loop closure method for multirobot map fusion by integration of consistency and data similarity," RA-L, vol. 5, no. 4, pp. 5701–5708, 2020.
- [42] S. Choudhary, L. Carlone, C. Nieto, J. Rogers, H. I. Christensen, and F. Dellaert, "Distributed mapping with privacy and communication constraints: Lightweight algorithms and object-based models," *IJRR*, vol. 36, no. 12, pp. 1286–1311, 2017.
- [43] Y. Chang, N. Hughes, A. Ray, and L. Carlone, "Hydra-multi: Collaborative online construction of 3d scene graphs with multi-robot teams," in *IROS*. IEEE, 2023, pp. 10995–11002.
- [44] Y. Shi, N. Wang, and X. Guo, "Yolov: Making still image object detectors great at video object detection," arXiv preprint arXiv:2208.09686, 2022.
- [45] H. W. Kuhn, "The hungarian method for the assignment problem," Naval research logistics quarterly, vol. 2, no. 1-2, pp. 83–97, 1955.
- [46] 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.
- [47] 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.
- [48] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *ICCV*, vol. 2. IEEE, 2005, pp. 1482–1489.
- [49] M. Weinmann, B. Jutzi, and C. Mallet, "Semantic 3d scene interpretation: A framework combining optimal neighborhood size selection with relevant features," ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 2, pp. 181–188, 2014.
- [50] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," arXiv preprint arXiv:1801.09847, 2018.
- [51] M. Grupp, "evo: Python package for the evaluation of odometry and slam." https://github.com/MichaelGrupp/evo, 2017.
- [52] 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.