Robotics: Science and Systems 2025
Los Angeles, CA, USA, June 21–25, 2025

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

Mason B. Peterson[1], Yixuan Jia[1], Yulun Tian[2], Annika Thomas[1], and Jonathan P. How[1]

***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 global localization method capable of localizing in challenging and diverse environments by creating and aligning maps of *open-set* and *view-invariant* objects. ROMAN formulates and solves a registration problem between object submaps using a unified graph-theoretic global data association approach with a novel incorporation of a gravity direction prior and object shape and semantic similarity. This work's open-set object mapping and information-rich object association algorithm enables global localization, even in instances when maps are created from robots traveling in *opposite* directions. Through a set of challenging global localization experiments in indoor, urban, and unstructured/forested environments, we demonstrate that ROMAN achieves higher relative pose estimation accuracy than other image-based pose estimation methods or segment-based registration methods. Additionally, we evaluate ROMAN as a loop closure module in large-scale multi-robot SLAM and show a 35% improvement in trajectory estimation error compared to standard SLAM systems using visual features for loop closures. Code and videos can be found at https://acl.mit.edu/roman.**

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 paper, 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, 12–16] have emerged as a powerful alternative that demonstrates superior accuracy and robustness when solving the correspondence problem. In particular, methods based on consistency graphs [12–16] formulate a graph where nodes denote putative object correspondences and edges denote their geometric consistencies. The data association problem is then solved by extracting large and densely connected subsets of nodes yielding the desired set of *mutually consistent* correspondences. While segment-based matching has become an established strategy for loop closures, prior approaches were largely demonstrated in indoor/structured settings [17], with limited object variations, or with accurate lidar sensing [9, 16, 18]. In contrast, we focus on unseen environments (*i.e.*, we do not make assumptions about the type of environment in which we operate), noisy segmentations, extreme viewpoint changes (Fig. 1), and RGB-D only sensing. Our key claim is that the proposed work is the only method that performs reliably in such extreme regimes and clearly outperforms state-of-the-art segment-based [11, 12, 19] and visual-feature-based [20, 21] methods in global localization tasks.

Performance in these challenging scenarios is made possible by extending graph-theoretic data association to use information beyond mutual (pairwise) geometric consistency. We enhance the representational richness of association affinity

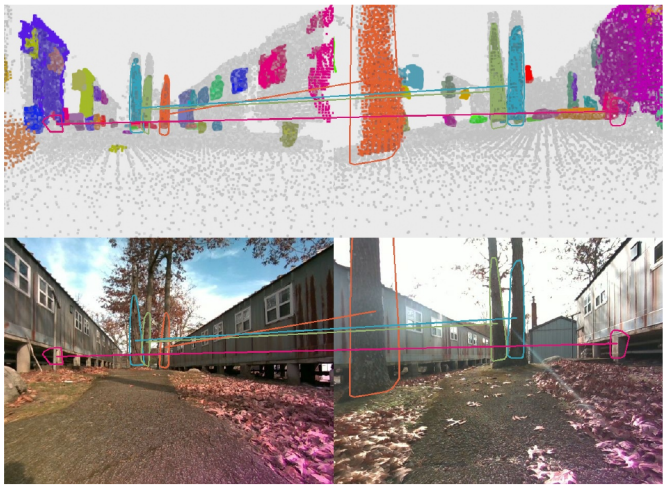

Fig. 1: Pair of segment submaps matched by two robots traveling in *opposite* directions in an off-road environment. Associated segments found by the proposed method are connected by lines and projected onto the image plane. **(Top)** Each pair of associated segments is drawn with the same color. The 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. Further visualization is shown in the supplementary video.

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

[1]Massachusetts Institute of Technology, Cambridge, MA 02139, USA. {masonbp, yixuany, annikat, jhow}@mit.edu.

[2]University of California San Diego, San Diego, CA 92093, USA. yut034@ucsd.edu

*We use *object* and *segment* interchangeably.

metrics by developing a unified formulation that incorporates: (i) *open-set semantics*, extracted as semantically meaningful 3D segments [22, 23] with descriptors obtained from vision-language foundation model, CLIP [24]; (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.

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:

1) A graph-theoretic data association formulation with a novel method 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)
2) A pipeline for creating open-set 3D segment maps from a single onboard RGB-D camera, using FastSAM [23] for open-set image segmentation and CLIP [24] 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.
3) Extensive experimental evaluation of the proposed method using real-world datasets (see Fig. 1) that involve urban, off-road, and ground-aerial scenarios. Our approach improves pose estimation accuracy by 45% in challenging, opposite-view global localization problems. When using ROMAN rather than visual features for inter-robot loop closures in multi-robot SLAM, our method reduces the overall localization error by 8% on large-scale collaborative SLAM problems involving 6-8 robots and by 35% 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 [25]. 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 [26], cuboids [27] or quadrics [28]. SLAM++ [3] trains domain-specific object detectors for objects like tables and chairs. Choudhary *et al.* [29] use objects as landmarks for localization, providing a database of discovered objects. Lin *et al.* [30] showed that semantic descriptors can improve frame-to-frame object data association. Recent works [6, 31] further leverage *open-set* semantics from pre-trained models. Other methods [32, 33] 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 [34, 35].

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 [36]) 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, 37]. 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 [38].

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, 37, 39]. 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 [13–16] or densest subgraph [12] problem. The work by Dubé *et al.* [16] is one of the early works that performs global localization by finding maximum cliques of consistency graphs. Ankenbauer *et al.* [40] leverage graph-theoretic data association [12] as the back-end association solver to perform global localization in challenging outdoor scenarios. Matsuzaki *et al.* [41] 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-the-art CSLAM systems [42–46] 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.* [47] extends PCM [13] by incorporating loop closure confidence and weighted pairwise consistency. Choudhary *et al.* [48] performs inter-robot loop closure via object-level data association; however, a database of 3D object templates is required. Hydra-Multi [49] employs hierarchical inter-robot loop closure that includes places, objects, and visual features summarized in a scene graph.

III. ROMAN

We now give an overview of the ROMAN global localization method. The core idea behind this work is that small, local maps of objects near a robot give rich, global information about the robot's pose in a previously mapped area. To leverage this information, ROMAN uses a mapping module to create object submaps and a robust data association module to associate objects in the robot's local map with objects seen by another robot or mapping session (see Fig. 3).

Our mapping pipeline begins with open-set image segmentation to extract initial observations of objects. Then, object observations are aggregated into an abstract object map. While we initially represent mapped objects with a dense point cloud, once the robot has moved on from an area, objects are abstracted to a single point and a feature descriptor, making our world representation communication- and storage-efficient. A submap centered around a robot's pose and containing nearby sparse, abstract objects is then created and used for global localization. Using local 3D segments, global localization can be achieved by matching objects in a local submap with objects from another robot or session. This is accomplished using our robust object data association method that leverages segment geometry, semantic information, and the direction of gravity to correctly associate objects. Our view-invariant global localization formulation enables global localization even in cases when maps were created by robots traveling in opposite directions. We first describe ROMAN's object data association method in Section IV and then present our approach for creating open-set object maps in Section V.

A. Notation

We use boldfaced lowercase and uppercase letters to denote vectors and matrices, respectively. We define $[n] = \{1, 2, \ldots, n\}$. For any $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathbb{R}$, we use $\text{GM}(x_1, \ldots, x_n) \triangleq (\Pi_{i=1}^{n} x_i)^{\frac{1}{n}}$ to denote the geometric mean of $x_1, \ldots, x_n$, and $\text{GM}(\mathbf{x})$ 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 $\text{cos\_sim}(\mathbf{x}, \mathbf{y}) \triangleq \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$. We define the element-wise operation $\text{ratio}(\mathbf{x}, \mathbf{y}) \triangleq \min(\frac{\mathbf{x}}{\mathbf{y}}, \frac{\mathbf{y}}{\mathbf{x}})$, where $\min$ and $\frac{\mathbf{x}}{\mathbf{y}}$ are also performed element-wise. We use $\mathbf{T}_b^a \in \text{SE}(3)$ to denote the pose of frame $\mathcal{F}_b$ with respect to frame $\mathcal{F}_a$.

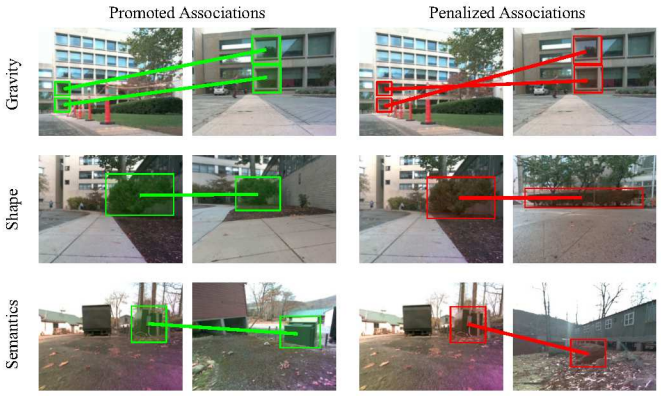

Fig. 2: Visualization of improved affinity metrics. The gravity-based distance score, s_{gravity} promotes pairs of associations that are consistent with the direction of gravity, while s_{shape} and s_{semantic} are used to encourage individual associations to be consistent in terms of geometric shape and semantics respectively.

IV. Robust Object Data Association

While our data association method can be used for general point cloud registration, we focus on the problem of associating objects between two local object submaps for global localization. We first detail submap alignment for global localization in Section IV-A then briefly review fundamentals in graph-theoretic data association in Section IV-B before describing the proposed affinity metrics for object association in Sections IV-C to IV-E.

A. Submap Alignment

We consider a pair of submaps $\mathcal{M}_i$ and $\mathcal{M}_j$ which are associated with gravity-aligned poses $\mathbf{T}^i_{\mathcal{M}_i}$ and $\mathbf{T}^j_{\mathcal{M}_j}$. Each submap $\mathcal{M}_i = \{p_1, \ldots, p_{m_i}\}$ where p_k is a 3D segment, represented by a 3D point in the gravity-aligned map frame $\mathcal{F}_{\mathcal{M}_i}$ and a feature vector containing shape and semantic attributes (object feature descriptors are discussed in greater detail in Section IV-D). We formulate global localization as the problem of estimating the transformation $\hat{\mathbf{T}}^i_j$ which relates the two local frames $\mathcal{F}_i$ and $\mathcal{F}_j$. To accomplish this, we attempt to associate objects in $\mathcal{M}_i$ with objects in $\mathcal{M}_j$. After finding these associations, $\hat{\mathbf{T}}^{\mathcal{M}_i}_{\mathcal{M}_j}$ can be computed using the closed-form Arun's method [50], enabling the relation between frames $\mathcal{F}_i$ and $\mathcal{F}_j$ given that $\hat{\mathbf{T}}^i_j = \mathbf{T}^i_{\mathcal{M}_i} \hat{\mathbf{T}}^{\mathcal{M}_i}_{\mathcal{M}_j} \left(\mathbf{T}^j_{\mathcal{M}_j}\right)^{-1}$. Thus, the core challenge in this global localization setup is to correctly associate segments, a challenging task in the presence of uncertainty, outliers, and geometric ambiguity. To this end, we construct a novel map-to-map object association method leveraging a graph-theoretic formulation incorporating the direction of gravity within maps and object shape and semantic attributes.

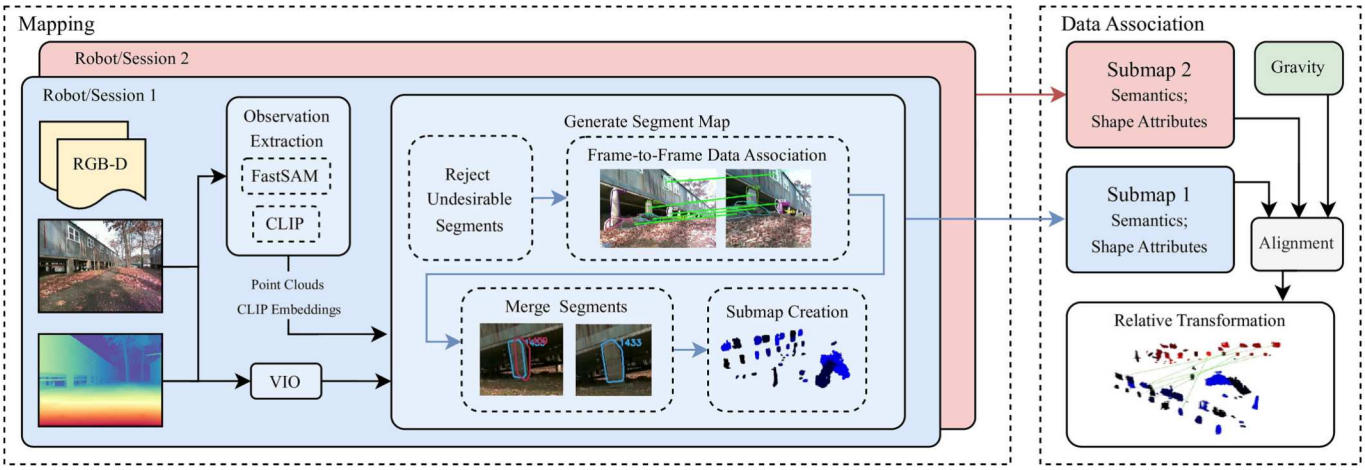

Fig. 3: ROMAN employs a front-end mapping module to create maps of open-set objects, representing each object with its centroid and feature descriptor. Local collections of objects are grouped into submaps and used for global localization by matching objects between two submaps. Accurate data association is achieved using a graph-theoretic formulation which leverages object shape and semantic similarity and a gravity prior.

B. Preliminaries: Graph-Theoretic Global Data Association

We follow the formulation used by CLIPPER [12] by first constructing 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 ϵ, 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 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 $\mathbf{M}_{p,p} = 1$. CLIPPER determines inlier associations by (approximately) solving for the densest subset of consistent associations, formulated as the following optimization problem,

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

where $\mathbf{u}_p$ is 1 when association a_p is accepted as an inlier and 0 otherwise. In the following sections, we describe methods to improve affinity metrics. Given our construction of $\mathbf{M}$, we then use CLIPPER's solver to find inlier associations $\mathbf{u}$. See [12] for more details.

C. Improving affinity metrics: general strategies

In its original form, the affinity matrix $\mathbf{M}$ 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, 51] have proposed pre-processing or post-processing methods that leverage additional information such as segment size and gravity direction to filter incorrect object associations or reject returned inlier associations if they result in an estimated $\hat{\mathbf{T}}_j^i$ that is inconsistent with gravity.

In comparison to works that use prior information in pre-processing or post-processing steps which may discard valuable information, ROMAN directly incorporates gravity and object similarity into 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 [52] suggest setting the diagonal entries of $\mathbf{M}$ 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 $\mathbf{M}$ increases, the number of off-diagonal terms (pairwise association affinity terms) increases quadrati-

cally and will quickly dominate the overall objective function. Alternatively, [47] 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 $\mathbf{M}$ 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 *geometric mean*,

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

The use of geometric mean in merging scores of potentially different scales is well-studied in the field of operation research [53]. It was shown that, under reasonable assumptions, the geometric mean is the only averaging function that merges scores correctly [54, 55]. With this insight in mind, we incorporate additional information 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. An ablation study on fusion methods is presented in Section VI-F.

D. Improving affinity metrics: incorporating metric-semantic 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 (visualized in Fig. 2). 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: $s_{\text{semantic}}(a_p) = \text{cos\_sim}(\text{CLIP}(p_i), \text{CLIP}(p_j))$. 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 ϕ_{min} and ϕ_{max}, so that scores less than ϕ_{min} are set to 0, scores larger than ϕ_{max} are set to 1, and scores between ϕ_{min} and ϕ_{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:

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

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 [56] 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_{shape} will be close to 1.

E. Improving affinity metrics: 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 post-processing rejection step from [5]. In this work, we propose a method to leverage this extra knowledge *within* the data association step by replacing $s_a(\cdot,\cdot)$ with a redesigned pairwise score metric, $s_{\text{gravity}}(\cdot,\cdot)$, to guide the solver to select pairs of 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-y plane and along the z axis:

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

where

$$\begin{aligned} d_{xy}(a_p, a_q) &= \left|\, \|\mathbf{c}_{xy}(p_i) - \mathbf{c}_{xy}(q_i)\| - \|\mathbf{c}_{xy}(p_j) - \mathbf{c}_{xy}(q_j)\| \,\right| \\ d_z(a_p, a_q) &= \left|\, (\mathbf{c}_z(p_i) - \mathbf{c}_z(q_i)) - (\mathbf{c}_z(p_j) - \mathbf{c}_z(q_j)) \,\right|. \end{aligned}$$

In effect, this prohibits selecting pairs of associations where the vertical distances between objects within the same submap are dissimilar, as visualized in Fig. 2. 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. Open-Set Object Mapping

This section describes ROMAN's approach to creating open-set object maps used for global localization in diverse environments. A map containing accurate and concise metric-semantic, object-level information is important for accurate object-based global localization. However, creating such a map has historically been difficult due to the need for an object classifier. Using recent zero-shot open-set segmentation, object-level environment information can easily be extracted from each image, but aggregating this information is difficult due objects or groups of objects being segmented inconsistently between views, occluded object observations, and drift in robot odometry. To overcome these difficulties, we

propose the following open-set object mapping pipeline, which is visualized in Fig. 3.

A. Mapping

The inputs to ROMAN's mapping module consist of RGB-D images and robot pose estimates (*e.g.*, provided by a visual-inertial odometry system). Per image object observations are made by segmenting a color image using FastSAM [23] and applying a series of preprocessing steps to filter out undesirable segments. Distinct and stationary objects are most likely to be segmented consistently across different views, so our segment filtering aims to capture only such segments. We use YOLO-V7 [57] to reject segments containing people. Additionally, we project segments into 3D using the depth image and remove large planar segments which are often large ground regions or non-distinct walls which cannot be represented well as an object. Each of the remaining segments is fed into CLIP [24] to compute a semantic descriptor. Observations, made up of CLIP embeddings and 3D voxels, are then sent to 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 grid-aligned voxel-based IOU between pairs of tracks and observations with 3D voxel overlap [35]. We use a global nearest neighbor approach [58] to assign observations to existing object tracks and create new tracks for any unassociated observation. Semantic descriptors of the associated segments are merged by taking a weighted average of descriptors of the existing segment and the incoming segment as in [59]. 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, which enables efficient map communication and storage.

B. Submap Creation

As a robot travels, submaps are periodically created. After a robot's odometry estimate reaches a distance greater than c_d from the previous submap pose, a new submap is instantiated. The new submap is assigned the current robot's pose with pitch and roll components removed using the IMU's gravity direction estimate, which ensures that objects are represented in a gravity-aligned frame for data association. All objects within a radius r of the submap center are added, and objects continue to be added until the robot's distance from the submap center is greater than r. The submap is then saved, after using a maximum submap size N to remove objects (starting at objects farthest from the center) so that the submap size $m_i \leq N$ thus limiting submap alignment computation. Finally, a newly created submap is fed to the global data association module and ROMAN attempts to align the current submap with previous submaps (e.g., from earlier in the run or from another robot or session). Resulting $\hat{\mathbf{T}}_j^i$ estimates from the submap object data association and alignment are used for global localization if the number of associated objects is greater than a threshold, τ.

VI. 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 [25], 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 45% improvement in relative pose estimation accuracy in opposite directions and 35% improvement in final trajectory estimation error in a subset of particularly challenging sequences from the Kimera-Multi datasets. The experiments were run on a laptop with a 4090 Mobile GPU and a 32-thread i9 CPU.

A. Experimental Setup

Baselines. We compare the alignment performance of `ROMAN` against the following baselines. `RANSAC-100K` and `RANSAC-1M` apply RANSAC [11], as implemented in [60], 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 semantic and shape attributes and rejects incorrect registration results using gravity information (so it has access to similar information as the proposed method). `TEASER++ / Prune` runs the robust registration of [19] using the same pruning mechanism as `CLIPPER / Prune`. `Binary Top-K`, which mimics the association method of SegMap [9], takes the top-k most similar segments (in terms of the semantic and shape descriptors) and constructs a binary affinity matrix that we use for finding associations with solver from [12]. We also compare against recent image-based pose estimation methods. `MASt3R` and `MASt3R (GT Scale)` use the learned 3D reconstruction model of [21] to estimate relative camera poses with the model's estimated translation scale and the ground truth translation scale respectively. `SuperGlue (GT Scale)` similarly estimates relative camera poses using [20] to match SuperPoint features [61]. Additionally, we incorporate ROMAN as a loop closure detection module in single-robot and multi-robot SLAM and compare against `KM` (Kimera-Multi [42]) and `ORB3` (ORB-SLAM3 [62]) which both use BoW descriptors of ORB features for loop closures.

Performance metrics. We use the following metrics for comparing segment-based place recognition, submap alignment (equivalent to relative pose estimation for image-based

methods),and full SLAM results. For place recognition, each algorithm is given a query submap and a database composed of submaps from every other robot run. Submap registration is performed on the query submap and every submap from the database. The database submap with the highest number of associations is returned and success is achieved if the query and returned submaps overlap. We vary the threshold on number of required object association τ to generate precision-recall curves, and following [63], precision-recall area under the curve (AUC) is reported.

To evaluate alignment success rate, an algorithm is given a pair of submaps whose center poses are within $10\,\mathrm{m}$ of each other. We evaluate the image-based methods by giving an algorithm the two images corresponding to the two submap center poses. To avoid giving segment-based methods an unfair advantage, we do not include submaps whose camera fields of view (FOVs) do not overlap. Following [64], alignment (*i.e.*, pose estimation) success is determined when the transformation error is less than $1\,\mathrm{m}$ and $5\,\mathrm{deg}$, with respect to ground truth.

Full SLAM results are evaluated using root mean squared (RMS) absolute trajectory error (ATE) between the registered estimated and ground truth multi-robot trajectories. We use open-source evo [65] to compute ATE.

Parameters. For global localization, we use the parameter values outlined in Table I. We additionally include results for two larger variants of our work: `ROMAN-L`, which uses $r = 25, N = 60$, and `ROMAN-XL`, which uses $r = 30, N = 80$. 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 sparse odometry and 1.0 m and 2.0 deg for loop closures.

TABLE I: Parameters

Parameter	Value	Description
σ / ϵ	$0.4\,\mathrm{m}$ / $0.6\,\mathrm{m}$	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
τ / N	4 / 40	Association threshold and max submap size

B. MIT Campus Global Localization

We first evaluate ROMAN's map alignment using the outdoor Kimera-Multi Dataset [25] recorded on MIT campus. Each robot creates a set of submaps using Kimera-VIO [66] for odometry and our ROMAN mapping pipeline. We use these submaps to evaluate segment-based place recognition and submap alignment for global localization, as described in Section VI-A. We evaluate methods on all multi-robot submap pairs from this dataset that are within $10\,\mathrm{m}$ of each other and whose corresponding camera FOVs overlap. In Table II, we show place recognition and submap alignment results. To highlight performance across different viewpoints, we bin the alignment tests into three different ground-truth relative heading groups: $\theta \leq 60\,\mathrm{deg}$ (same direction), $60\,\mathrm{deg} < \theta \leq 120\,\mathrm{deg}$ (perpendicular), and $\theta > 120\,\mathrm{deg}$. When the heading difference is small, alignment is comparatively easier. Aligning submaps from opposite views or from paths that cross perpendicularly, presents the hardest cases for global localization.

TABLE II: Kimera-Multi Outdoor Data Alignment Success Rate

	Method	Place Recognition (AUC)	Pose Estimation Success Rate ($\leq 5^\circ$, 1 m error) 0–60°	60–120°	120–180°	Mean	Runtime (ms)
Segment-Based	`RANSAC-100K`	0.106	0.141	0.000	0.000	0.047	75.9
	`RANSAC-1M`	0.160	0.341	0.065	0.054	0.153	488
	`CLIPPER`	0.145	0.235	0.043	0.000	0.093	48.2
	`CLIPPER/Prune`	0.531	0.429	0.109	0.108	0.215	<u>23.9</u>
	`TEASER++/Prune`	0.426	0.441	0.125	0.083	0.216	498.6
	`Binary Top-K`	0.307	0.377	0.130	0.054	0.187	**21.3**
Visual	`SuperGlue` (GT Scale)	–	0.685	0.043	0.000	0.243	87.4
	`MASt3R` (GT Scale)	–	**0.775**	0.152	0.297	0.408	2950
	`MASt3R`	–	0.211	0.043	0.000	0.085	2950
Ours	`ROMAN`	0.552	0.521	0.152	0.189	0.287	28.9
	`ROMAN-L`	**0.663**	0.723	<u>0.370</u>	**0.432**	<u>0.508</u>	92.9
	`ROMAN-XL`	<u>0.654</u>	<u>0.745</u>	**0.457**	<u>0.405</u>	**0.536**	213

Table II shows that the `ROMAN` outperforms other segment-based methods in terms of place recognition and alignment success rate in all heading intervals while operating at a similar runtime. In opposite directions, `ROMAN` achieves a pose estimation success rate 75% higher than the next-best segment-based method, `CLIPPER/Prune`. Compared to image-based methods, the ROMAN variant with more objects, `ROMAN-XL` outperforms the next-best method, `MASt3R` (which is given ground truth scale), in every case except for in similar direction scenarios, all while running 10 times faster. In particular, `ROMAN-XL` achieves a pose estimation success rate in opposite directions that is 45% better than `MASt3R` (GT Scale) and 31% better when averaged across the different headings.

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 $1\,\mathrm{km}$, the entire map could be represented with less than 25 MB of data.

C. Loop Closures in Visual SLAM

We integrate ROMAN as a loop closure detection module for single and multi-robot pose-graph SLAM and compare the trajectory estimation results here and in Section VI-D. We use Kimera-VIO [66] for front-end odometry when creating initial ROMAN submaps. Then, we attempt to register each new submap with all existing submaps from the ego robot and other robots. Loop closures are reported when the number of associations found is at least τ. Then, sparsified Kimera-VIO odometry and ROMAN loop closures are fed into the robust pose graph optimization of Kimera-Multi [42] to estimate multi-robot trajectories. Root-mean-squared (RMS) absolute trajectory errors (ATE) in the tunnel, hybrid, and outdoor Kimera-Multi datasets are reported in Table III. We compare SLAM with ROMAN loop closures against a centralized Kimera-Multi (`KM`) [42] and multi-session ORB-SLAM3 (`ORB3`) [62]. Note that in the single-robot case, the baselines are essentially a single-robot version of Kimera and

TABLE III: Kimera-Multi Data [25] SLAM Comparison Against Various Loop Closure Methods (RMS ATE m)

Dataset	Num. Robots	Total Dist. (m)	ORB3 [62]	KM [42]	ROMAN
Easy: Single Robot Tunnels					
Tunnel 0	1	635	**2.08**	4.20	4.16
Tunnel 1	1	780	26.19	**1.61**	2.15
Tunnel 2	1	854	9.53	**5.29**	6.12
Tunnel 3	1	845	16.61	5.29	**3.90**
Mean			13.60	4.10	**4.08**
Medium: Full Multi-Robot Datasets					
Tunnel All	8	6753	–	4.38	**4.20**
Hybrid All	8	7785	–	5.83	**5.12**
Outdoor All	6	6044	–	9.38	**8.77**
Mean			–	6.53	**6.03**
Difficult: Challenging Multi-Robot Combinations					
Hybrid 1, 2, 3	3	3551	–	10.34	**6.91**
Hybrid 4, 5	2	1896	28.09	6.11	**2.80**
Outdoor 1, 2	2	2011	11.93	10.12	**7.67**
Mean			–	8.86	**5.79**

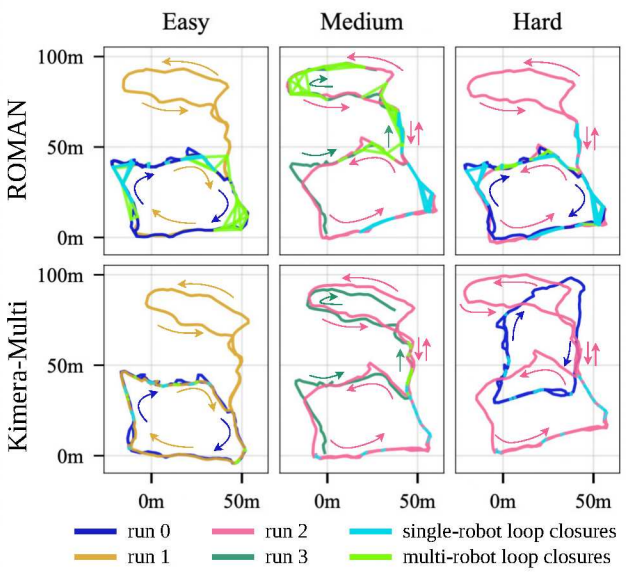

Fig. 4: Off-road qualitative pose graph trajectory estimate. Easy, medium, and hard cases comparing using ROMAN and KM for loop closures. Different combinations were paired together to make easy, medium, and hard cases. In the easy case, robots travel the same direction; in the medium case, the two runs go in opposite direction except for the small connecting neck; and in the hard case, robots only cross paths going opposite directions. Only ROMAN successfully finds loop closures between robots running in opposite directions.

single-robot ORB-SLAM3, where a deeper comparison was made in [67]. Similar to [67], we found that ORB-SLAM3 fails to find reasonable trajectory estimates in some robot configurations, and this is represented with a dash in Table III.

Estimation errors show that, on average, in easier single-robot tunnel runs, ROMAN loop closures result in lower trajectory errors than ORB-SLAM3 and errors comparable to Kimera-Multi. The full, large-scale multi-robot runs show that ROMAN's ability to detect loop closures in challenging visual scenarios results in moderate gains compared to Kimera-Multi's trajectory errors. Improvement is somewhat limited due to the high-connectivity of robot paths and the fact that most robot trajectory overlap occurs when robots are traveling in the same direction, which are loop closure opportunities in which visual-feature-based methods already perform well. However, when SLAM results are compared on a subset of robot trajectories that contain difficult instances for visual loop closures (*e.g.*, perpendicular path crossing and scenes with high visual aliasing), results show that ROMAN has a significantly lower ATE in these challenging scenarios. The trend is that as loop closure scenarios become increasingly difficult, ROMAN demonstrates more significant improvements over state-of-the-art methods.

D. Loop Closures in Off-Road Environment

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 a Clearpath Jackal using Intel RealSense D455 to capture RGB-D images and Kimera-VIO [66] is used for odometry. The robot is teleoperated across four runs, following similar trajectories but with different runs traversing the same area while traveling in different directions. We run the ROMAN pipeline on three different pairs of robot trajectories. We compare ROMAN to KM loop closures in Fig. 4. The three pairs consist 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 that 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. 4 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 KM only work reliably in same-direction traversals and fail to find any loop closures in the hard case.

E. Ground-Aerial Cross-View Localization

We also 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. 5. We test object map alignment on 20 ground-aerial pairs of traverses through the environment, and report alignment success rate in Table IV. We show that ROMAN maintains an advantage over other baselines, demonstrating its global localization capability in a small-scale aerial-ground cross-view localization demonstration.

F. Ablation Study

Finally, we perform an extensive set of ablation studies examining the contribution of different affinity metric improvements, fusion methods, and other algorithmic elements.

Fig. 5: Environmental setup used in the ground-aerial cross-view localization experiment as seen from both ground view (left) and aerial view (right).

TABLE IV: Ground-Aerial Cross-View Localization Results

Method	Recall	Method	Recall
RANSAC-100K	0.25	RANSAC-1M	0.45
CLIPPER	0.3	CLIPPER/Prune	0.35
TEASER++/Prune	0.55	ROMAN	0.60

TABLE V: Ablations Results

	Ablations	Place Recognition (AUC)	Pose Estimation Success Rate ($\leq 5^\circ$, 1 m error) 0–60°	60–120°	120–180°	Mean
	ROMAN	0.552	0.521	0.152	0.189	0.287
Exclude	Gravity	0.522	0.474	0.109	0.081	0.221
	Semantics	0.497	0.500	0.146	0.167	0.271
	Shape	0.530	0.532	0.152	0.108	0.264
Fusion	Arithmetic Mean	0.517	0.199	0.000	0.054	0.085
	Product	0.505	0.388	0.087	0.027	0.167
	Diagonal	0.336	0.388	0.109	0.027	0.175
(σ, ϵ)	$(0.2, 0.3)$	0.504	0.433	0.229	0.028	0.230
	$(0.6, 0.9)$	0.543	0.522	0.104	0.139	0.255
	$(0.8, 1.2)$	0.535	0.548	0.125	0.194	0.289
Image Size	128×128	0.378	0.296	0.149	0.108	0.184
	512×512	0.525	0.535	0.085	0.135	0.252

Fusion methods. Here, we compare different methods for fusing object similarity scores s_o with pairwise scores s_a in Table V. We investigate fusing scores with geometric mean (ROMAN), product [4, 47], arithmetic mean, and setting the diagonal elements of the affinity matrix $M_{pp} = \text{GM}(s_o(a_p)s_o(a_q))$ [12, 52]. Table V shows that fusing scores using the geometric mean results in a much higher alignment success rate compared to other fusion methods. Intuitively, 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 improves the alignment success, but tends to over-penalize, since in this case, including s_o can only lower the overall similarity score. Changing only the diagonal elements also improves over standard CLIPPER [12], but is limited in impact as described in Section IV-C.

Affinity component contributions. In Table V, we additionally examine the effect of using ROMAN for map alignment while excluding the following individual affinity metric components: the gravity-guided pairwise score s_a, the shape similarity score s_{shape}, and the semantic similarity score s_{semantic}. While each component helps ROMAN achieve higher alignment success, the gravity prior makes the most significant difference and the semantic similarity score makes the least. However, in terms of place recognition, semantics makes the largest difference.

Robustness to segmentation errors. As a small experiment, we change the input image size from 256 (the default value for which ROMAN is tuned) to obtain degraded segmentation (128) and over-segmentation (512). On average, FastSAM [23] returns 4.0 segments at image size 128, 11.3 at 256, and 18.7 at 512. As shown in Table V, in the case of over-segmentation, we report only 12% performance decrease in terms of mean pose estimation success rate. With severe under-segmentation, ROMAN achieves 0.184 mean success, which is slightly lower than the best segment-based baselines shown in Table II, however some of the effects of under-segmentation could be mitigated by including segments in a larger radius r.

Robustness to dynamic objects. The ROMAN pipeline deliberately filters out pedestrians, and the robust data association effectively rejects other dynamic objects. To demonstrate the effect of dynamic objects, we disable the pedestrian filter and report that ROMAN achieves a mean alignment success rate of 0.251, which is still better than other segment-based baselines in Table II.

Hyperparameter sensitivity. Table II shows the effect of varying ROMAN submap sizes, controlled by N (maximum submap size) and r (submap radius). We vary N from the default value 40 to 80 with r increasing from 15 m to 30 m correspondingly. The results show that these two submap size parameters can be effectively altered to achieve a trade-off between alignment success rate and runtime. An ablation over the segment noise parameters σ and ϵ are recorded in Table V. We note that the lowest mean recall over all pairs is still higher than the mean recall of any other segment-based method in Table II.

Scalability. Our mapping pipeline runs at 9.6 Hz when computing CLIP [24] embeddings and at 17.9 Hz without running CLIP on the outdoor Kimera-Multi Dataset [25] As shown in Table V, alignment success rate only drops 12% without CLIP embeddings which could be used for running ROMAN on a more compute-constrained platform. Removing CLIP embeddings also reduces map size by 100 times.

VII. Limitations

One of the fundamental challenges with using open-set segmentation like FastSAM [23] for object mapping is determining what constitutes a discrete object. ROMAN's filtering and merging steps significantly improve the quality of resulting object maps; however, inconsistent segmentations may sometimes still result in duplicate representations of objects (e.g., a car and each of its doors may be represented as distinct 3D segments).

Additionally, ROMAN seeks to reject non-object-like segments (*e.g.* ground, walls, etc.) because they do not fit well into the centroid-focused object data association. This does not exploit the information present in non-object segments, *e.g.* roads, walls, and buildings. Our object registration could additionally be improved by employing a coarse-to-fine tech-

nique for using more precise information than object centroids for submap registration.

Finally, while ROMAN runs fast enough for the scale of experiments shown in this paper (*i.e.* up to eight $1000\,\mathrm{m}$ long robot trajectories), trajectories longer than this would require significant computation to register the growing number of submaps as robots continue mapping. A faster place recognition stage could improve scalability.

VIII. 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, which enabled global localization even in instances of robots viewing scenes from opposite directions.

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," *International Journal of Computer Vision*, pp. 1–33, 2024.

[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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1352–1359.

[4] 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.

[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 *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 5266–5272.

[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 International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 8239–8246.

[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," *The International Journal of Robotics Research*, 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 *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (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," *Communications of the 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," *IEEE Robotics and Automation Letters*, 2024.

[13] J. G. Mangelson, D. Dominic, R. M. Eustice, and R. Vasudevan, "Pairwise consistent measurement set maximization for robust multi-robot map merging," in *2018 IEEE international conference on robotics and automation (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 *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 820–13 827.

[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," *The International Journal of Robotics Research*, vol. 43, no. 14, pp. 2245–2273, 2024.

[16] R. Dubé, M. G. Gollub, H. Sommer, I. Gilitschenski, R. Siegwart, C. Cadena, and J. Nieto, "Incremental-segment-based localization in 3-d point clouds," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1832–1839, 2018.

[17] S. D. Sarkar, O. Miksik, M. Pollefeys, D. Barath, and I. Armeni, "Sgaligner: 3d scene alignment with scene graphs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 927–21 937.

[18] L. Li, X. Kong, X. Zhao, W. Li, F. Wen, H. Zhang, and Y. Liu, "Sa-loam: Semantic-aided lidar slam with loop closure," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 7627–7634.

[19] H. Yang, J. Shi, and L. Carlone, "Teaser: Fast and certifiable point cloud registration," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 314–333, 2020.

[20] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[21] 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.

[22] 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.

[23] 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.

[24] 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 *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[25] Y. Tian, Y. Chang, L. Quang, A. Schang, C. Nieto-Granda, J. P. How, and L. Carlone, "Resilient and distributed multi-robot visual slam: Datasets, experiments, and lessons learned," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 11 027–11 034.

[26] S. L. Bowman, N. Atanasov, K. Daniilidis, and G. J. Pappas, "Probabilistic data association for semantic slam," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1722–1729.

[27] S. Yang and S. Scherer, "Cubeslam: Monocular 3-d object slam," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 925–938, 2019.

[28] L. Nicholson, M. Milford, and N. Sünderhauf, "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 1–8, 2018.

[29] S. Choudhary, A. J. Trevor, H. I. Christensen, and F. Dellaert, "Slam with object discovery, modeling and mapping," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 1018–1025.

[30] S. Lin, J. Wang, M. Xu, H. Zhao, and Z. Chen, "Topology aware object-level semantic mapping towards more robust loop closure," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7041–7048, 2021.

[31] 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.

[32] Y. Wang, C. Jiang, and X. Chen, "Voom: Robust visual object odometry and mapping using hierarchical landmarks," *arXiv preprint arXiv:2402.13609*, 2024.

[33] M. Zins, G. Simon, and M.-O. Berger, "Oa-slam: Leveraging objects for camera relocalization in visual slam," in *2022 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 2022, pp. 720–728.

[34] 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. 11 080–11 095, 2023.
[35] 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.
[36] 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," *The International Journal of Robotics Research*, p. 02783649241229725, 2024.
[37] A. Gawel, C. Del Don, R. Siegwart, J. Nieto, and C. Cadena, "X-view: Graph-based semantic multi-view localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1687–1694, 2018.
[38] R. Raguram, J.-M. Frahm, and M. Pollefeys, "A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus," in *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II 10*. Springer, 2008, pp. 500–513.
[39] Y. Wang, C. Jiang, and X. Chen, "Goreloc: Graph-based object-level relocalization for visual slam," *IEEE Robotics and Automation Letters*, 2024.
[40] 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 international conference on intelligent robots and systems (IROS)*. IEEE, 2023, pp. 1358–1365.
[41] 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.
[42] 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," *IEEE Transactions on Robotics*, vol. 38, no. 4, 2022.
[43] P. Schmuck, T. Ziegler, M. Karrer, J. Perraudin, and M. Chli, "Covins: Visual-inertial slam for centralized collaboration," in *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 2021, pp. 171–176.
[44] P.-Y. Lajoie and G. Beltrame, "Swarm-slam: Sparse decentralized collaborative simultaneous localization and mapping framework for multi-robot systems," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 475–482, 2023.
[45] 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," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1150–1157, 2021.
[46] 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 large-scale underground environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9175–9182, 2022.
[47] H. Do, S. Hong, and J. Kim, "Robust loop closure method for multi-robot map fusion by integration of consistency and data similarity," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5701–5708, 2020.
[48] 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," *The International Journal of Robotics Research*, vol. 36, no. 12, pp. 1286–1311, 2017.
[49] Y. Chang, N. Hughes, A. Ray, and L. Carlone, "Hydra-multi: Collaborative online construction of 3d scene graphs with multi-robot teams," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 10 995–11 002.
[50] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Transactions on pattern analysis and machine intelligence*, no. 5, pp. 698–700, 1987.
[51] 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.
[52] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1482–1489.
[53] F. S. Roberts, "Chapter 18 limitations on conclusions using scales of measurement," in *Operations Research and The Public Sector*, ser. Handbooks in Operations Research and Management Science. Elsevier, 1994, vol. 6, pp. 621–671. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927050705800994
[54] J. Aczél and F. S. Roberts, "On the possible merging functions," *Mathematical Social Sciences*, vol. 17, no. 3, pp. 205–243, 1989.
[55] J. Aczél, "Determining merged relative scores," *Journal of Mathematical Analysis and Applications*, vol. 150, no. 1, pp. 20–40, 1990.
[56] 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.
[57] Y. Shi, N. Wang, and X. Guo, "Yolov: Making still image object detectors great at video object detection," *arXiv preprint arXiv:2208.09686*, 2022.
[58] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
[59] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5021–5028.
[60] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," *arXiv preprint arXiv:1801.09847*, 2018.
[61] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.
[62] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.
[63] M. Zaffar, S. Garg, M. Milford, J. Kooij, D. Flynn, K. McDonald-Maier, and S. Ehsan, "Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2136–2174, 2021.
[64] P.-E. Sarlin, M. Dusmanu, J. L. Schönberger, P. Speciale, L. Gruber, V. Larsson, O. Miksik, and M. Pollefeys, "Lamar: Benchmarking localization and mapping for augmented reality," in *European Conference on Computer Vision*. Springer, 2022, pp. 686–704.
[65] M. Grupp, "evo: Python package for the evaluation of odometry and slam." https://github.com/MichaelGrupp/evo, 2017.
[66] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: an open-source library for real-time metric-semantic localization and mapping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1689–1696.
[67] M. Abate, Y. Chang, N. Hughes, and L. Carlone, "Kimera2: Robust and accurate metric-semantic slam in the real world," in *International Symposium on Experimental Robotics*. Springer, 2023, pp. 81–95.