2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

Outram: One-shot Global Localization via Triangulated Scene Graph and Global Outlier Pruning

Pengyu Yin[1], Haozhi Cao[1], Thien-Minh Nguyen[1], Shenghai Yuan[1], Shuyang Zhang[2], Kangcheng Liu[1], and Lihua Xie[1], *Fellow, IEEE*

***Abstract*— One-shot LiDAR localization refers to the ability to estimate the robot pose from one single point cloud, which yields significant advantages in initialization and relocalization processes. In the point cloud domain, the topic has been extensively studied as a global descriptor retrieval (i.e., loop closure detection) and pose refinement (i.e., point cloud registration) problem both in isolation or combined. However, few have explicitly considered the relationship between candidate retrieval and correspondence generation in pose estimation, leaving them brittle to substructure ambiguities. To this end, we propose a hierarchical one-shot localization algorithm called Outram that leverages substructures of 3D scene graphs for locally consistent correspondence searching and global substructure-wise outlier pruning. Such a hierarchical process couples the feature retrieval and the correspondence extraction to resolve the substructure ambiguities by conducting a local-to-global consistency refinement. We demonstrate the capability of Outram in a variety of scenarios in multiple large-scale outdoor datasets. Our implementation is open-sourced: `https://github.com/Pamphlett/Outram`.**

I. Introduction

LiDAR-based localization problems can be stated in the following general form. We are given a point cloud $\mathcal{P}$ produced by a LiDAR, and a reference point cloud Q, which can be either another frame of LiDAR scan [1], an accumulated submap [2], or even the entire mapping space [3]. Given a point correspondence $i \in \mathcal{I}$, $\mathbf{p}_i \in \mathcal{P}$ and $\mathbf{q}_i \in \mathcal{Q}$ can be associated and represented in the residual function $r_i := \|\mathbf{T}\mathbf{p}_i - \mathbf{q}_i\| \to [0, \infty)$ where $\mathbf{T}$ is random rigid transformation.

While the estimation problem can be relatively easy in special cases (e.g., the size of point clouds is constrained or an approximation $\mathbf{T}_{initial}$ is known a priori), it can be hard in general [4], [5] due to limited descriptiveness of local features and computational complexity. These general, prior-free cases are what we encountered in relocalization or global localization problems. To address this problem, prior work [6], [7], [8], [9], [10] usually break it down into a retrieval phase and a pose estimation phase, where several candidates are generated first and verified later for final pose estimation.

[1]Authors are with the Centre for Advanced Robotics Technology Innovation (CARTIN), School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. pengyu001, haozhi002, thienminh.nguyen, shyuan, kangcheng.liu, elhxie@ntu.edu.sg

[2]Authors are with the Department of Electronic and Computer Engineering, the Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, China. szhangcy@connect.ust.hk

This research is supported by the National Research Foundation, Singapore under its Medium Sized Center for Advanced Robotics Technology Innovation.

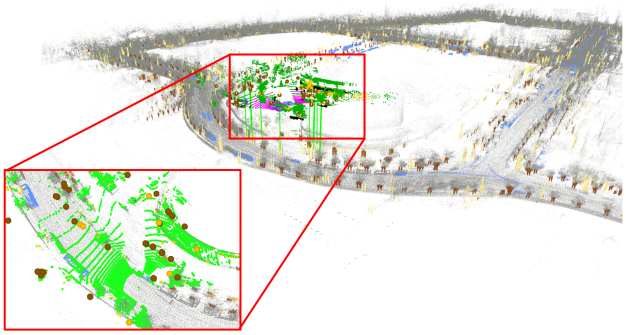


Fig. 1. Illustration of the proposed global localization algorithm on MulRan DCC dataset [11]. We generate 3D scene graphs and leverage their substructures for locally consistent correspondence generation. With these raw correspondences, we exploit a graph-theoretic outlier pruning process for globally consistent inlier extraction (green lines). The point cloud transformed by the estimated pose is shown in green in the enlarged area. The position of the semantic segmented point cloud and the grey map point cloud is presented for visualization purposes only.

These decoupled approaches, however, suffer from local ambiguities where the substructures in different areas are similar, leading to false candidate retrieval.

Rather than attempting to find the most appearance-similar keyframe through one single retrieval, we propose an algorithm called Outram for one-shot, accurate, and efficient global localization directly against a reference map. Different from existing works, we rely on local substructures of a 3D scene graph to generate locally consistent correspondences directly with the map, and further find the inlier correspondence that globally unifies substructures-wise consistency. This leads to the proposed local-to-global, hierarchical global localization algorithm that has the following contributions:

- We propose a novel representation encoding substructure of 3D scene graphs for efficient locally consistent correspondence generation.
- Together with a subsequent graph-theoretic pruning module, we propose an accurate, efficient, and one-shot global localization pipeline for large-scale outdoor environments.
- Extensive experiments are conducted on publicly available datasets on a global localization setup, showing superior robustness compared with current state-of-the-arts. We further open-sourced our implementation to benefit the community.

II. Related Work

In the literature, LiDAR-based relocalization or global localization methods can be broadly categorized into two

branches by whether a movement of the robot is needed, namely one-shot localization and iterative localization. Iterative localization methods are usually formulated in a Monte Carlo localization manner [12], [13], where the movement of the robot will provide more environment observations and thus update the weight of each particle till convergence. With the accumulative submap and generated dense segments, Segmap [14] proposed a data-driven retrieval mechanism for global localization.

On the contrary, one-shot localization is referred to as algorithms that solve the global localization problem in a full prior-free manner. We further divide the literature on this into two groups: loop closure detection-based methods and registration-based methods. The next subsection provides more details on each of the two groups.

A. Loop closure detection-based global localization

Loop closure detection (LCD) methods identify previously visited places by encoding current LiDAR measurements into global descriptors and comparing them to a database constructed from historical frames. The construction process can be divided into global and local ones. Very similar to its visual-based counterpart, several local methods [15], [16] detect 3D key points and aggregate them into a global representation, and the retrieval process is arranged in a bag-of-words manner. Alternatively, global methods directly encode the whole LiDAR scan. The encoding pattern can be either handcrafted or deep-learned. Scan Context [6] encodes the geometric information of a point cloud into a bird-eye-view global descriptor. Following the same data structure, several variants [17], [10] have been proposed to enhance the descriptiveness by extending the original geometric-only representation by either the inherent intensity information [17] or high-level semantics [10]. Several methods leverage deep neural networks to generate global descriptors directly [18], [19]. Uy et al. [18] combined PointNet [20] and NetVLAD [18] to generate compact global representations. Chen et al. [19] proposed an LCD network that estimates the overlap ratio between point clouds. More recently, techniques exploiting the graph structure of local features have been proposed [21], [9], [22], [7]. Methods involving point cloud semantics [21], [9], [22] usually encode the scene by instances and the spatial layout in between. LCD is accomplished by conducting similarity checks between these semantic instance graphs. Yuan et al. [7] proposed to aggregate local point features to form a triangle-based descriptor (the simplest graph). Leveraging the side length of each triangle as the hash table key, loop closure candidates can be found by a voting scheme. Further, the transformation in between is calculated and verified by planes in the scene.

It is natural to extend LCD methods to global relocalization due to the similar retrieval mechanism [23]. Nevertheless, local ambiguities and scene changes can make the descriptor-based algorithms brittle. Moreover, while geometric verification is widely employed after the retrieval process, it can be both time-consuming and inaccurate [24]. Additionally, even if the geometrically proximate candidate is retrieved from the database, LCD methods usually rely on local point features for the subsequent pose estimation, which could potentially suffer from feature degeneracy [4], [5], resulting in impossible pose estimation.

B. Registration-based methods

Instead of finding one single keyframe in the pre-built LCD database, registration-based methods seek to solve the global localization problem in a point cloud registration manner. There exist two concerns [25] when leveraging point cloud registration in solving global localization problems: correspondence generation and outlier pruning. In correspondence extraction, it is not computationally feasible to directly extract correspondences on the point level. Subsequently, several methods leverage high-level representation in replace of geometric points. Ankenbauer et al. proposed [26] semantic object maps for global localization by formulating a global registration problem, where all-to-all correspondence is built between semantic objects within local and global maps. In a very similar sense, all-to-all correspondence is also employed in [27] for global localization, while semantic clusters to be registered are from different modalities.

In the outlier correspondence pruning stage, the aforementioned two methods both send prebuilt correspondences into a graph-theoretic inlier selection module, where the inlier set is modeled as the maximum clique (MC) of the consistency graph [28]. Additionally, RANSAC-based methods [8] are also ubiquitous in the outlier pruning stage, while being proven to be brittle to high amounts of outliers [28], [29].

In comparison with these methods, we proposed a novel substructure representation, instead of semantic clusters only, of a semantic segmented point cloud for more informative correspondence extraction. We empirically demonstrate in IV how our proposition is more computationally tractable and accurate than the previous state-of-the-art.

III. Methodology

In this section, we first formulate the global localization problem that is considered herein. Next, we present our proposed one-shot global localization algorithm Outram with two sub-modules, where we first leverage local substructures in a 3D scene graph for correspondence generation and further prune the correspondence globally with substructure-wise consistency check.

A. Registration-based One-shot Global Localization

The point cloud registration problem is formulated as acquiring the pose transformation $\mathbf{T}$ of a single query LiDAR point cloud $\mathcal{P} = \left\{\mathbf{p}_i \in \mathbb{R}^3\right\}_{i=1}^{n}$ against a prebuilt point cloud map $\mathcal{M} = \left\{\mathbf{m}_j \in \mathbb{R}^3\right\}_{j=1}^{m}$ accumulated by a series of scans in the world frame collected during a time span $[1, t]$. The pose transformation in between is defined as:

$$\mathbf{T} \triangleq [\mathbf{R}, \mathbf{t}] \in \mathrm{SO}(3) \times \mathbb{R}^3, \tag{1}$$

where $\mathbf{R}$ represents the rotation and $\mathbf{t}$ is the translation. With the unknown ground truth transformation, corresponding

points in the query scan and the reference map can be associated as:

$$\mathbf{m}_j = \mathbf{R}\mathbf{p}_i + \mathbf{t} + \mathbf{o}_{ij}, \tag{2}$$

where the measurement error $\mathbf{o}_{ij}$ is either Gaussian or random depending on whether the correspondence is an inlier or not. Finding the pose transformation typically includes three steps: find an initial data association $\mathcal{I} \in [n] \times [m] := \{1, \dots, n\} \times \{1, \dots, m\}$, prune the initial correspondences set $\mathcal{I}$ for the inlier set $\mathcal{I}^\star$, and estimate the pose transformation with the inlier set $\mathcal{I}^\star$ [30]. In the following sections, we will detail how we leverage scene graphs for informative correspondence generation and efficient outlier pruning. We also empirically demonstrated in IV how our proposed method is superior to current existing works [26] leveraging all-to-all correspondence for global localization problems in terms of scalability.

B. Triangulated 3D Scene Graph

Since point feature level correspondence generation is not computationally feasible in global localization problems, we leverage 3D scene graphs [31] for correspondence extraction at a higher instance level. Different from previous works [26], [32] that build all-to-all correspondence or leverage instance-only descriptor, we present a new representation, triangulated 3D scene graphs, for informative and efficient correspondence generation.

Given the query point cloud $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^n$, we employ the state-of-the-art point cloud semantic segmentation network [33] to match a point $\mathbf{p}_i$ with a semantic label $l \in \mathcal{L}$. The network thus acts as a mapping $\lambda(\mathbf{p}_i) \colon \mathbb{R}^3 \to \mathcal{L} \subset \mathbb{N}$. Hence, we can define the following semantic $\mathcal{S}$ point cloud as:

$$\mathcal{S} = \{s_i | s_i = (\mathbf{p}_i, \lambda(\mathbf{p}_i)), \forall \mathbf{p}_i \in \mathcal{P}\}. \tag{3}$$

Subsequently, we leverage the projection-based clustering method [34] to generate instances from the semantic point cloud with the same label l:

$$\begin{aligned} \mathbf{C}^l = \{C_k \subset \mathcal{P} | k = 1, \dots N; \\ l = \lambda(\mathbf{p}_i) = \lambda(\mathbf{p}_j),\ \forall \mathbf{p}_i, \mathbf{p}_j \in C_k\}. \end{aligned} \tag{4}$$

We further enhance these semantic clusters by approximating each of them as Gaussian distributions:

$$\boldsymbol{\mu}_k = \frac{1}{|C_k|} \sum \mathbf{p}_i, \boldsymbol{\Sigma}_k = \frac{1}{|C_k|} \sum (\mathbf{p}_i - \boldsymbol{\mu}_k)^\top (\mathbf{p}_i - \boldsymbol{\mu}_k). \tag{5}$$

The query scan and reference semantic map can be represented by a set of semantic Gaussian distributions $\mathbb{C}_\mathcal{A} = \left\{\mathcal{A}_i^l \sim \mathcal{N}(\mathbf{a}_i, \boldsymbol{\Sigma}_{\mathcal{A}_i})\right\}$ and $\mathbb{C}_\mathcal{B} = \left\{\mathcal{B}_j^l \sim \mathcal{N}\left(\mathbf{b}_j, \boldsymbol{\Sigma}_{\mathcal{B}_j}\right)\right\}$ respectively. These semantic Gaussian distributions will then act as primitives for establishing correspondences and later pose estimation. This representation is beneficial for correspondence generation as the covariance depicts the shape information of each semantic instance, which serves as another metric for similarity check. Such semantic lifting also structures the query scan as a two-layer scene graph [31] where we have the semantic segmented points as the metric-semantic layer and instances at the upper layer. Each edge in the upper layer encodes the spatial relationship as well as the semantic topological information between instances.

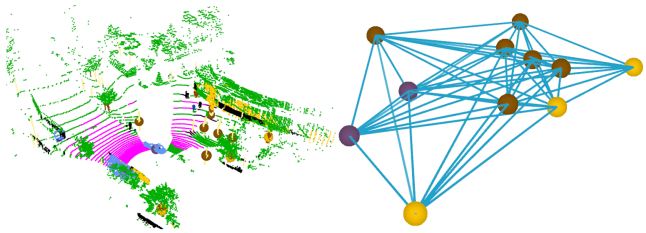


Fig. 2. One LiDAR scan from MulRan DCC dataset and its corresponding triangulated 3D scene graph. Colored spheres are the centroids of each semantic cluster with cars being purple, tree trunks being brown, and poles being yellow.

C. Correspondence Generation via Local Substructures

We then leverage substructures of the scene graph for correspondence generation. Inspired by STD [7], we triangulate each scene graph to form a series of triangles as in the minimal representation for local similarity measurement and subsequent correspondence generation. To be more specific, each anchor semantic cluster $\mathcal{A}_i \sim \mathcal{N}(\mathbf{a}_i, \boldsymbol{\Sigma}_{\mathcal{A}_i})$ is associated with K nearest clusters $\{\mathcal{A}_j\}_{j=1}^K$. Afterward, we exhaustively select two of the neighbors, together with the anchor cluster, i.e., $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$, to form one triangle representation of current scene graph. By an abuse of notation, we denote it as $\Delta(\mathcal{A}_{1,2,3})$ which comprises of the following attributes:

- $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$: centroids of the semantic clusters;
- $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, $\boldsymbol{\Sigma}_3$: corresponding covariance matrices;
- d_{12}, d_{23}, d_{31}: three side lengths, $d_{12} \leq d_{23} \leq d_{31}$;
- l_1, l_2, l_3: three semantic labels associated with each vertex of the triangle.

Similar to STD [7], we build a hash table using only the sorted side length d_{12}, d_{23}, and d_{31} as the key value due to its simplicity and permutation invariance. Other attributes are left for verification purposes. In the searching process, we have the triangulated scene graph in the query scan and reference map:

$$\begin{aligned} \Delta Query &= \left\{\Delta\left(\mathcal{A}_{1,2,3}^n\right)\right\}_{n=1}^N, \\ \Delta Map &= \left\{\Delta\left(\mathcal{B}_{1,2,3}^m\right)\right\}_{m=1}^M, \end{aligned} \tag{6}$$

where n and m are indexes for triangle descriptors in the query and map scene graph respectively. We drop the subscript and denote $\Delta\left(\mathcal{A}_{1,2,3}^n\right)$ as $\Delta\mathcal{A}^n$ for clarity. As shown in Fig. (3), query each of the triangles (e.g., $\Delta\mathcal{A}^1$) against the hash table constructed by the reference semantic scene graph will produce multiple responses $\{\Delta\mathcal{B}^q\}_{q=1}^Q$ as similar substructures could exist throughout the whole mapping region. We further leverage the semantic labels l, as well as the covariance matrix $\boldsymbol{\Sigma}$, associated with each vertex for another round of similarity-check for semantic and shape resemblance. For semantic labels, we simply employ the equality condition. For the covariance matrices, Wasserstein distance is applied for similarity measurement.

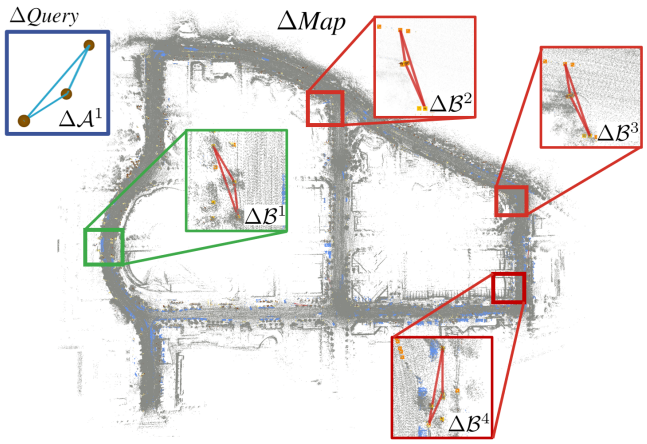


Fig. 3. Illustration of the substructure ambiguities and the proposed correspondence generation process. One triangle representation of the query scan (with three vertices labeled as tree trunk in brown) is shown in blue and multiple responses from different regions of the reference map are shown in green (true location) and red (false location) due to substructure ambiguities. Correspondence generation between all these substructures ensures local consistency while also retaining the possibility of exploiting scene-wise global consistency.

After querying all triangles in the query scene graph, a set of raw cluster-wise correspondence $\mathcal{I}_{\text{raw}}$ can be naturally built as the sorted side length offers direct mapping between semantic clusters.

Although the descriptor-based retrieval process presented above is very similar to the one in any ordinary LCD, we highlight here that we neither solve for the pose nor produce multiple candidates. Instead, we leverage these locally similar substructures, i.e., the triangulated scene graph, to build the coarse correspondence. This is different from both local feature-aggregation-based LCD methods [15], [16] and global feature-based LCD methods [6], [8] where local features are only stacked for retrieval purpose in the former, and post-retrieval correspondence generation for the latter. We implicitly formulate the candidate selection process in the correspondence generation stage which guarantees the local similarity while also retaining the possibility of exploring scene-wise global similarity in the next stage.

D. Global Graph-theoretic Outlier Pruning

With the prebuilt correspondence set $\mathcal{I}_{\text{raw}}$ that associates subareas of the current scene with locally similar ones in the reference map, we seek to find an area that maximizes the number of mutually consistent correspondences as well as maintains the consistency between these local structures:

$$\begin{aligned} &\max_{\mathcal{I}\subset\mathcal{I}_{\text{raw}}} \quad |\mathcal{I}| \\ &\text{s.t. } \mathcal{D}\left(\mathcal{I}_i,\mathcal{I}_j\right) \leq \epsilon,\ \forall \mathcal{I}_i,\mathcal{I}_j \in \mathcal{I}, \end{aligned} \tag{7}$$

where $\mathcal{D}$ is a metric consistency check indicating whether two correspondences are mutually consistent with each other and ϵ is the threshold. Namely, for two correspondences $\mathcal{I}_i$ and $\mathcal{I}_j$, with their corresponding semantic clusters $\mathcal{A}_i, \mathcal{B}_i$ and $\mathcal{A}_j, \mathcal{B}_j$, a consistency check is defined as

$$\mathcal{D}\left(\mathcal{I}_i,\mathcal{I}_j\right) \triangleq \text{dist}\left(\mathcal{A}_{ij},\mathcal{B}_{ij}\right), \tag{8}$$

with $\mathcal{A}_{ij} := \mathcal{A}_i - \mathcal{A}_j$ and $\mathcal{B}_{ij} := \mathcal{B}_i - \mathcal{B}_j$ distribution differences between semantic clusters. It is worth noting that the similarity check $\mathcal{D}$ can vary. From the simplest Euclidean distance-based [28], [26] to distribution distance-based [5].

Problem 7 can be solved by formulating the correspondence set to a consistency graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each vertex in $\mathcal{V}$ represent one correspondence and each edge in $\mathcal{E}$ represents two correspondences are consistent with each other evaluated by similarity check $\mathcal{D}$. Afterward, finding the inlier set is equivalent to searching for the maximum clique of the consistency graph. We invite interested readers to refer to [28] for more details. Maximum clique is a classic combinatorial problem in graph theory and is NP-complete. We leverage the PMC library [35] to solve it. We will also show how current state-of-the-art registration-based global localization algorithms [26] are brittle to the hardness of the maximum clique problem when the problem scales up even with a powerful parallel solver.

From a holistic view, such graph-theoretic outlier pruning strategy embeds global consistency on top of the coarse correspondence set generated by locally consistent triangulated scene graphs. Such a local-to-global scheme resists constructing the hard association between the query scan and one single scan in the keyframe database, which is usually conducted by voting or similarity check. Instead, it first exploits local structures of one scene to generate correspondences associating similar substructures. Afterward, relationships between these local fragments are considered in search of a place in the reference that is globally consistent between the local substructures. The process presented here is very similar to feature re-ranking methods [24] for LCD where the computation of pose does not happen immediately after candidate retrieval but goes through another re-ranking process for frame-wise consistency verification. While our proposed method works on a more informative lower level, the substructures of scene graphs, thus have a better possibility of reaching global consistency.

E. Pose Estimation

With the estimated inlier set $\mathcal{I}$, we formulate the objective function in Eq. (2) into the following truncated least squares (TLS) form to resist potential outliers further [36]

$$\hat{\mathbf{R}}, \hat{\mathbf{t}} = \underset{\mathbf{R}\in SO(3), \mathbf{t}\in\mathbb{R}^3}{\arg\min} \sum_{ij\in\hat{\mathcal{I}}} \min\left(\left\|\mathbf{p}_i - \mathbf{R}\mathbf{q}_j - \mathbf{t}\right\|_2, c_{ij}\right), \tag{9}$$

with c_{ij} the truncation parameter. Eq. (9) is then solved by leveraging Black-Rangarajan duality [37] and graduated non-convexity (GNC) [36].

IV. Experimental Results

In this section, we compare our proposed method with several state-of-the-art one-shot global localization methods. All mentioned algorithms are implemented in C++ and tested on a PC with Intel i9-13900 and 32Gb RAM.

Experimental Setup. We evaluate our proposed method, Outram, on six different sequences of two publicly available datasets: MulRan [11] and MCD [38]. To mimic a real global

TABLE I
DETAILS OF EVALUATION DATASETS

Mapping/Loc. Sequence	Length	Scan Number	Time Diff.
Mapping:			
MulRan DCC 03	5.7 km	7479	-
MulRan KAIST 02	6.3 km	8941	-
MCD NTU 01	3.8 km	6023	-
Localization:			
MulRan DCC 01	4.9 km	5542	20 days
MulRan DCC 02	5.2 km	7561	1 month
MulRan KAIST 01	6.3 km	8226	2 months
MulRan KAIST 03	6.4 km	8629	10 days
MCD NTU 02	0.64 km	2288	2 hours
MCD NTU 13	1.23 km	2337	2 days

localization or relocalization scenario, different from a loop closure detection setting, we intentionally involve temporal diversity between the mapping or descriptor generation session and the localization session from days to months. For each mapping sequence, we concatenate semantically annotated scans [33] by the ground truth pose to generate the semantic segmented reference map for registration-based methods. Three representative semantic classes are used for all semantic-related methods: pole, tree trunk, and car. For LCD-based global localization methods, frames in the mapping sequences are encoded to form a database for retrieval using scans in localization sequences. Statistics of the benchmark datasets are presented in Table I. The criteria for choosing the mapping sequence is the sequence that has the most coverage of the target area. We also highlight the time differences between the mapping and localization sessions ranging from several days to months, making our setup suitable for benchmarking global localization algorithms.

Baselines. We involved a variety of state-of-the-art loop closure detection methods, STD [7], Scan Context [6] and GosMatch [21], as well as recently developed registration-based global localization algorithms [26] to benchmark the performance of each of the method. STD also leverages substructures of a scene for loop closure detection in a voting manner and GosMatch leverages semantic clusters for global descriptor generation. As they have certain sub-modules that are the same as our proposed one, we involve them to confirm that our proposition, the local-to-global method, can exploit more structural information of a LiDAR scan and have a better chance to be localized in a one-shot manner. As most of the methods have an open-sourced implementation [7], [21], [6], we directly use them for comparison. As to the method proposed by Ankenbauer et al. [26], since it also leverages semantic objects for global localization, we share the same semantic clusters for a fair comparison.

Metric. We employ the ordinary relative pose error (RPE) to evaluate the accuracy of estimated pose $\hat{\mathbf{T}}$ with respect to the ground truth $\mathbf{T}$:

$$e_{trans} = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2},$$

$$e_{rot} = \arccos\left(trace(\Delta \mathbf{T})/2 - 1\right),$$

with $\Delta \mathbf{T} = \hat{\mathbf{T}} \cdot \mathbf{T}^{-1}$ the transformation difference, and Δx, Δy, Δz the positional entries of $\Delta \mathbf{T}$. We regard global localization results with $e_{trans} < 5$ meters and $e_{rot} < 10$ degrees are valid, which is generally the convergence region for local registration methods [30] for subsequent refinement.

Results. We present the experimental results in four aspects, including successful global localization rate, error distribution analysis, runtime analysis, and storage analysis.

A. Successful Rate of Global Localization

We present the results of LCD-based global localization methods on the upper side of Table II while the registration-based ones are on the lower side. Our proposed algorithm, Outram, outperforms all other methods by a margin.

We observed that without the proposed triangulated scene graph for correspondence generation, the method proposed by Ankenbauer et al. [26] can hardly scale to bigger size problems as it generates correspondences in an all-to-all fashion. For smaller-size data sets (e.g., the reference semantic map of NTU MCD only includes 1192 clusters), the method performs relatively well as in this case, the straightforward all-to-all correspondence guarantees the full inclusion of the inlier correspondences while being also computationally amenable. However, when the reference map scales to a bigger size (e.g., 5136 and 5103 semantic clusters in MulRan DCC and KAIST respectively), the original method quickly becomes computationally intractable, where we observe the algorithm drained all 32Gb RAM of the test platform, making the program to crash. In such scenarios, we modified the method to a constrained version where we limited the size of semantic clusters in the query scan by random downsampling (i.e., Ankenbauer et al. (Cons.) in Table II). However, the constrained algorithm performs poorly due to the failure of inlier inclusion. A pure geometric variant of Outram is also involved (Outram Pure Geo.) in the ablation study of semantic labels. In the implementation of the method, we disable the semantic labels of each semantic cluster and produce a set of triangle descriptors with pure geometric information, similar to the representation proposed in STD [7]. Conversely, these pure geometric triangle descriptors are used to build up correspondence followed by a graph-theoretic outlier pruning process, like what is proposed in this paper. These comparisons demonstrate the effectiveness of the proposed triangulated scene graph in terms of more informative correspondence extraction (compared with existing registration-based methods) and the superior performance of the whole registration-based pipeline (compared with LCD-based methods), which verifies our claims in I.

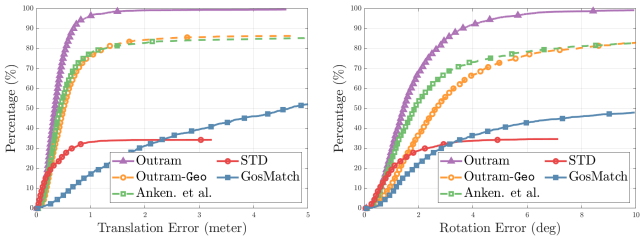
B. Error Distribution Analysis

The average translation error (ATE) and average rotation error (ARE) in Table II are calculated using the successfully localized ones only. On average, point-based loop closure detection-based methods [7] have the most accurate localization result. The phenomenon can also be verified in Fig. (4) the cumulative distribution function (ECDF). We observe that in the lower left corner, the point-based global localization method, STD, surpasses all other methods in

TABLE II

GLOBAL LOCALIZATION PERFORMANCE COMPARISON

	Dataset	Successful Global Localization Rate [%] ↑						ATE [Meter] ↓	ARE [Deg] ↓	Time [ms] ↓
		MulRan DCC		MulRan KAIST		MCD NTU				
	Localization Seq.	01	02	01	03	02	13			
LCD	STD [7]	17.57	18.06	49.61	38.96	66.64	34.27	**0.23**	**0.54**	**7.18**
LCD	GOSMatch [21]	48.61	50.17	35.93	51.98	55.01	42.53	1.92	2.06	12.08
Regis.	Ankenbauer et al. (Original) [26]	-	-	-	-	77.92	82.54	0.47	2.03	1708.3
Regis.	Ankenbauer et al. (Cons.) [26]	0.072	0.032	0.025	0.012	-	-	2.81	3.17	345.3
Regis.	Outram Pure Geo. (**Ours**)	50.65	66.93	41.90	42.17	82.11	82.76	0.69	2.70	323.8
Regis.	Outram (**Ours**)	**82.53**	**90.48**	**84.41**	**85.64**	**99.65**	**95.42**	0.40	1.83	306.8



(a) ECDF of translation error. (b) ECDF of rotation error.

Fig. 4. The empirical cumulative distribution function (ECDF) of translation and rotation error on sequence 13 of MCD NTU dataset. The x-axis represents different translation and rotation errors, and the corresponding y-axis is the probability of one specific method producing an estimation with a smaller error.

terms of translation error. This is because STD detects point-level stable features for pose estimation, while all other methods work on the object level and rely on centroids of semantic clusters for pose estimation. Centroids could be different due to viewpoint variation or occlusion. However, this does not hinder the robustness of our proposed method to outperform others, which is the most important evaluation metric for global localization.

C. Runtime

In the last column of Table II, the average runtime of each method is presented. We noticed LCD-based methods are more computationally efficient compared with registration-based methods due to the simple retrieval-based design. All registration-based methods leverage maximum clique for outlier pruning, which could be time-consuming when the graph size is big.

To understand each of the sub-modules of our proposed method better, we analyzed the time breakdown of each module and plotted it in Fig. (5). Three main components are considered, namely, the time to generate a triangulated scene graph, the time to search for correspondence establishment, and the time to solve the subsequent maximum clique problem. Please note logarithmic scale is used in the y-axis for better visualization. We note that solving for the maximum clique (i.e., finding the globally consistent inlier correspondences out of the prebuilt one) requires 190 ms on average and is the most computationally expensive process. The time required for the process is jointly determined by the size of the 3D scene graph and the reference map. Although our system cannot run in real-time, the runtime of Outram stays manageable for the global localization task.

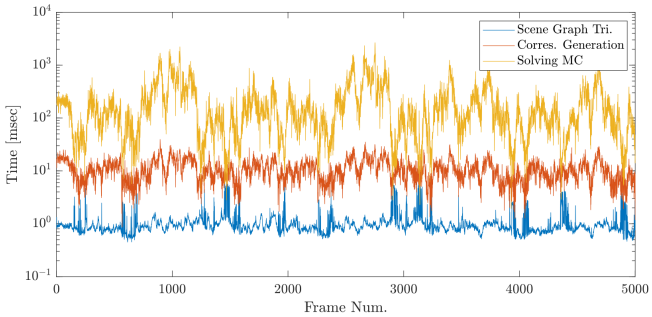


Fig. 5. Runtime breakdown of Outram on MulRan DCC sequence 01.

TABLE III

STORAGE CONSUMPTION OF EACH METHOD ON MULRAN KAIST

Method	STD	GosMatch	Regis.-based
Storage Usage ↓	62.4 MB	4.8 MB	**513 kB**

D. Storage Efficiency

We further analyze the storage consumption of each method on MulRAN KAIST sequence 02 containing 8941 frames in Table III. For average global descriptor-based LCD methods [21], it requires several MB to store the vectorized descriptors for each frame to build up the database. STD [7] requires storing planes detected in every frame for geometric verification, thus more space is consumed. While Outram only requires to store centroids. The result reveals the potential of leveraging Outram for relocalization in bigger size environments.

V. CONCLUSIONS

In this paper, we proposed Outram, a one-shot LiDAR global localization algorithm leveraging triangulated 3D scene graphs and graph-theoretic outlier pruning. Substructures of scene graphs are leveraged for locally consistent correspondence generation, and the subsequent outlier pruning process ensures the global consistency between the substructures and finds the inlier correspondences. We demonstrated the effectiveness of our method on various datasets where Outram surpasses several state-of-the-art LCD-based global localization methods, albeit at the cost of real-time performance. In the future, we plan to work on a proper indicator of the global localization quality and a theoretical guarantee for localization results.

REFERENCES

[1] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time." in *Robotics: Science and systems*, vol. 2, no. 9. Berkeley, CA, 2014, pp. 1–9.

[2] K. Chen, B. T. Lopez, A.-a. Agha-mohammadi, and A. Mehta, "Direct lidar odometry: Fast localization with dense point clouds," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2000–2007, 2022.

[3] T.-M. Nguyen, D. Duberg, P. Jensfelt, S. Yuan, and L. Xie, "Slict: Multi-input multi-scale surfel-based lidar-inertial continuous-time odometry and mapping," *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 2102–2109, 2023.

[4] H. Lim, S. Yeon, S. Ryu, Y. Lee, Y. Kim, J. Yun, E. Jung, D. Lee, and H. Myung, "A single correspondence is enough: Robust global registration to avoid degeneracy in urban environments," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8010–8017.

[5] P. Yin, S. Yuan, H. Cao, X. Ji, S. Zhang, and L. Xie, "Segregator: Global point cloud registration with semantic and geometric cues," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2848–2854.

[6] G. Kim, S. Choi, and A. Kim, "Scan context++: Structural place recognition robust to rotation and lateral variations in urban environments," *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1856–1874, 2021.

[7] C. Yuan, J. Lin, Z. Zou, X. Hong, and F. Zhang, "Std: Stable triangle descriptor for 3d place recognition," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1897–1903.

[8] D. Cattaneo, M. Vaghi, and A. Valada, "Lcdnet: Deep loop closure detection and point cloud registration for lidar slam," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2074–2093, 2022.

[9] X. Kong, X. Yang, G. Zhai, X. Zhao, X. Zeng, M. Wang, Y. Liu, W. Li, and F. Wen, "Semantic graph based place recognition for 3d point clouds," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8216–8223.

[10] L. Li, X. Kong, X. Zhao, T. Huang, W. Li, F. Wen, H. Zhang, and Y. Liu, "Ssc: Semantic scan context for large-scale place recognition," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 2092–2099.

[11] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "Mulran: Multimodal range dataset for urban place recognition," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6246–6253.

[12] H. Kuang, X. Chen, T. Guadagnino, N. Zimmerman, J. Behley, and C. Stachniss, "Ir-mcl: Implicit representation-based online global localization," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1627–1634, 2023.

[13] L. Wiesmann, T. Guadagnino, I. Vizzo, N. Zimmerman, Y. Pan, H. Kuang, J. Behley, and C. Stachniss, "Locndf: Neural distance field mapping for robot localization," *IEEE Robotics and Automation Letters*, 2023.

[14] 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.

[15] Y. Zhong, "Intrinsic shape signatures: A shape descriptor for 3d object recognition," in *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*. IEEE, 2009, pp. 689–696.

[16] B. Steder, M. Ruhnke, S. Grzonka, and W. Burgard, "Place recognition in 3d scans using a combination of bag of words and point feature based relative pose estimation," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2011, pp. 1249–1255.

[17] H. Wang, C. Wang, and L. Xie, "Intensity scan context: Coding intensity and geometry relations for loop closure detection," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2095–2101.

[18] M. A. Uy and G. H. Lee, "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4470–4479.

[19] X. Chen, T. Läbe, A. Milioto, T. Röhling, J. Behley, and C. Stachniss, "Overlapnet: A siamese network for computing lidar scan similarity with applications to loop closing and localization," *Autonomous Robots*, pp. 1–21, 2022.

[20] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[21] Y. Zhu, Y. Ma, L. Chen, C. Liu, M. Ye, and L. Li, "Gosmatch: Graph-of-semantics matching for detecting loop closures in 3d lidar data," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5151–5157.

[22] G. Pramatarov, D. De Martini, M. Gadd, and P. Newman, "Boxgraph: Semantic place recognition and pose estimation from 3d lidar," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 7004–7011.

[23] D. Xu, J. Liu, Y. Liang, X. Lv, and J. Hyyppä, "A lidar-based single-shot global localization solution using a cross-section shape context descriptor," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 189, pp. 272–288, 2022.

[24] K. Vidanapathirana, P. Moghadam, S. Sridharan, and C. Fookes, "Spectral geometric verification: Re-ranking point cloud retrieval for metric localization," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2494–2501, 2023.

[25] F. Pomerleau, F. Colas, R. Siegwart, *et al.*, "A review of point cloud registration algorithms for mobile robotics," *Foundations and Trends® in Robotics*, vol. 4, no. 1, pp. 1–104, 2015.

[26] J. Ankenbauer, P. C. Lusk, and J. P. How, "Global localization in unstructured environments using semantic object maps built from various viewpoints," *arXiv preprint arXiv:2303.04658*, 2023.

[27] S. Matsuzaki, K. Koide, S. Oishi, M. Yokozuka, and A. Banno, "Single-shot global localization via graph-theoretic correspondence matching," *arXiv preprint arXiv:2306.03641*, 2023.

[28] 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.

[29] 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.

[30] F. Pomerleau, F. Colas, R. Siegwart, and S. Magnenat, "Comparing icp variants on real-world data sets," *Autonomous Robots*, vol. 34, no. 3, pp. 133–148, 2013.

[31] A. Rosinol, A. Violette, M. Abate, N. Hughes, Y. Chang, J. Shi, A. Gupta, and L. Carlone, "Kimera: From slam to spatial perception with 3d dynamic scene graphs," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1510–1546, 2021.

[32] L. Zhang, T. Digumarti, G. Tinchev, and M. Fallon, "Instaloc: One-shot global lidar localisation in indoor environments through instance learning," *arXiv preprint arXiv:2305.09552*, 2023.

[33] H. Cao, Y. Xu, J. Yang, P. Yin, S. Yuan, and L. Xie, "Multi-modal continual test-time adaptation for 3d semantic segmentation," *arXiv preprint arXiv:2303.10457*, 2023.

[34] P. Zhou, X. Guo, X. Pei, and C. Chen, "T-loam: truncated least squares lidar-only odometry and mapping in real time," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[35] R. A. Rossi, D. F. Gleich, and A. H. Gebremedhin, "Parallel maximum clique algorithms with applications to network analysis," *SIAM Journal on Scientific Computing*, vol. 37, no. 5, pp. C589–C616, 2015.

[36] H. Yang, P. Antonante, V. Tzoumas, and L. Carlone, "Graduated non-convexity for robust spatial perception: From non-minimal solvers to global outlier rejection," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1127–1134, 2020.

[37] M. J. Black and A. Rangarajan, "On the unification of line processes, outlier rejection, and robust statistics with applications in early vision," *International journal of computer vision*, vol. 19, no. 1, pp. 57–91, 1996.

[38] N. Thien-Minh, Y. Shenghai, N. Thien Hoang, Y. Pengyu, C. Haozhi, X. Lihua, W. Maciej, J. Patric, T. Marko, Z. Justin, and B. Noel, "MCD: Diverse large-scale multi-campus dataset for robot perception," in *Conference on Computer Vision and Pattern Recognition 2024*, 2024. [Online]. Available: https://openreview.net/forum?id=7u2auPZrwV