Proceedings of 2025 IEEE
International Conference on Robotics and Biomimetics
December 3 - 7, 2025, Chengdu, China

Multi-scale Triplet Descriptors for Global LiDAR Localization with Maximum Clique-based Enhancement

Shuyue Lin, Jinsong Li, and Yuxiang Sun

Abstract— **This paper presents an accurate and robust global localization system by matching a single LiDAR scan against a global map. To enhance global pose estimation accuracy in environments with sparse semantic information, we first introduce a triplet descriptor based on multi-scale edge structures. By segmenting edge lengths with multiple thresholds, the method constructs triangular structures at different scales, enabling the extraction of hierarchical vertex descriptors that better enhance discriminability and maximize the use of limited information. To support the proposed descriptor structure, we further design a dynamic maximal clique enhancement strategy that enhances inlier selection accuracy in sparse semantic scenes while avoiding redundant information in semantically rich environments, maintaining computational efficiency. Experimental results on public datasets demonstrate that our proposed method outperforms existing state-of-the-art approaches in terms of both descriptor discriminability and pose estimation accuracy.**

I. Introduction

Global vehicle localization without relying on GPS plays an important role in autonomous driving [1]–[6]. It enables vehicles to determine their poses in environments where GPS is unavailable. LiDAR sensors provide high-precision point-cloud data that are robust to changes in lighting conditions, and can directly capture spatial structure information, making them more reliable than visual cameras in challenging environments. So, LiDAR-based global localization methods have received widespread attention [7].

LiDAR-based localization frameworks can be divided into methods using low-level and high-level features [8]. Low-level features [9] primarily leverage geometric structures in raw sensor data, or local texture information in projected 2D images (e.g., range images), focusing on capturing fine-grained details. High-level features, on the other hand, emphasize global structures and semantic information within the sensor data, resulting in more abstract but semantically rich representations. Compared to low-level features, high-level semantic features exhibit stronger robustness and can effectively handle challenges such as occlusions and lighting variations, demonstrating better generalization capability in large-scale and diverse scenarios [10]. Low-level features depend on geometric structures and can still provide rich information even in degraded data conditions. However, when a scene consists of predominantly single or few object categories, the descriptive capability of high-level semantic descriptors built on semantic instances is limited. Although high-level features can enhance discriminability in large-scale scenes, effectively utilizing the limited information in sparse semantic environments remains still a major challenge.

In addition, scan-to-map global localization based on descriptors often suffers from a large number of outliers caused by many similar scans in the map, leading to matching failures. To mitigate the impact of outliers caused by similar scans, accurate outlier removal is essential for pose estimation. During descriptor computation, kernel-based outlier removal methods are typically employed to address the uncertainty in point cloud geometric structures. However, these M-estimation frameworks [11] usually rely on single thresholds or strategies, which are insufficient to handle complex outlier distributions in large-scale and diverse environments, resulting in decreased pose estimation accuracy. By contrast, outlier removal methods based on maximum consensus [12], such as Random Sample Consensus (RANSAC) [13] and maximum clique algorithms [14], estimate an optimized inlier model, thereby ensuring better generalization across various scenarios. It should be noted that for different descriptor structures, inlier optimization strategies must be specifically designed to guarantee the efficiency and accuracy of pose optimizers.

To improve the discriminability of descriptors in scenarios with sparse semantic information and to enhance the accuracy of pose estimation, we propose a global localization system that implements a multi-scale edge-structured triplet descriptor along with a dynamic maximal clique enhancement strategy. Our code is open-sourced[1]. The contributions of this work are summarized as follows:

1) We propose a multi-scale edge-based triplet descriptor. Specifically, the method forms multi-scale triangles around each semantic cluster by varying the neighborhood range, and derives descriptors accordingly.
2) We propose a dynamic maximum clique enhancement strategy, which adaptively incorporates higher-scale descriptor vertices when the low-scale triplet matches are insufficient. This process continues until a predefined inlier threshold is reached, which improves pose estimation accuracy.
3) We integrate the proposed strategies into the TripleLoc framework [15]. The system effectiveness is validated on public datasets. The results demonstrate improvements in both localization accuracy and scene adaptability.

[1]https://github.com/lab-sun/M2Loc

This work was supported in part by the Hong Kong Research Grants Council under Grant 15222523, and in part by City University of Hong Kong under Grant 9231601. *(Corresponding author: Yuxiang Sun.)*

The authors are with the Department of Mechanical Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong (e-mail: shuyue.lin@cityu.edu.hk; jinsong.li@my.cityu.edu.hk; yx.sun@cityu.edu.hk, sun.yuxiang@outlook.com).

2025 IEEE International Conference on Robotics and Biomimetics (ROBIO) | 979-8-3315-5747-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/ROBIO66223.2025.11376127

The rest of this paper is organized as follows: Section II reviews related work. Section III presents the details of our system. Section IV shows the experimental results. The last section concludes our work.

II. Related Work

A. Low-Level vs. High-Level Features in Place Recognition

In LiDAR-based place recognition, the spatial sparsity of point clouds and the lack of texture limit the descriptive capability of traditional low-level features. To enhance feature representation, Shan et al. [16] uses LiDAR intensity to generate high-resolution images and extracts ORB features for fast retrieval. Meanwhile, Di Giammarino et al. [17] proposed a visual place recognition system that projects intensity data into panoramic images through cylindrical projection. Kim et al. [18] designed structural descriptors to ensure stable recognition under feature degradation. However, in large-scale and dynamic environments, low-level features remain vulnerable to interference, which leads to decreased recognition accuracy.

To address this limitation, research has increasingly focused on high-level features to improve generalization [19]. SeqOT [20] and AdaFusion [21] respectively extract global temporal features using deep networks, although they require high computational resources and offer limited interpretability. Semantic-based high-level features have also attracted attention. For example, constructing descriptors by object-level semantic topology in [22], while Yin et al. [23] performed matching and localization based on hierarchical relationships in 3D scene graphs. These methods demonstrate higher robustness in complex or long-term changing environments. Nevertheless, in scenarios with sparse semantics or few category distribution, the limited number of semantic instances reduces the discriminability of descriptors. How to effectively utilize such limited semantic information to enhance recognition remains an open challenge.

B. M-estimation vs. Consensus Maximization Inlier Selection Methods

Global localization essentially integrates place recognition and pose estimation, where the latter is typically formulated as a nonlinear optimization problem. Due to the inherent uncertainty in descriptor extraction and matching, the information matrix is often used to quantify the uncertainty and guide the rejection of outliers during optimization [24]. The widely used graph optimization framework General Graph Optimization (G2O) [25] adopts kernel-based loss functions to penalize inconsistent matches and improve robustness. These techniques can be categorized as M-estimations, in which a fixed threshold is applied across all scenarios. However, their robustness is limited because a single rule often fails to handle diverse outlier distributions.

To address this limitation, consensus maximization methods have been widely adopted. Shan et al. [16] integrated visual feature matching with a perspective-n-point (PnP) and RANSAC framework to achieve robust pose verification. A consensus set maximization algorithm is introduced in [26] that maintains reliability even under low inlier ratios. TripleLoc [15] employs a maximum clique solver to identify the most stable inlier subset. Since the inlier distribution is highly dependent on the design of the descriptors, it is essential to develop tailored and efficient outlier rejection mechanisms that align with the characteristics of specific descriptors.

For LiDAR-based global localization, enhancing the generalization ability in large-scale environments requires effectively place recognition and pose estimation in sparse scenes. To this end, we investigate a multi-scale edge-structure-based triplet descriptor along with a corresponding inlier selection strategy.

III. The Proposed Method

A. The Overall Architecture

The proposed system consists of a multi-scale edge-based triplet descriptor and a dynamic maximal clique enhancement scheme. The overall framework is illustrated in Fig. 1. Before performing global localization, we first apply an unrefined Cylinder segmentation network [27] to segment a point cloud into semantic instances. These instances are then clustered into cluster centroids. Each semantic instance is represented using the index and spatial position of its corresponding cluster centroid, which facilitates the construction of a semantic graph and serves as the foundation for triangle structure generation. The prior map consists of an instance-level map and a road surface normal map. The road surface normal (RSN) map [15] is constructed following the approach proposed in TripleLoc, while the instance-level map and the descriptors of the query scan are generated using the proposed multi-scale edge-based triplet descriptor. Specifically, for each cluster centroid, neighboring points are searched within a predefined edge-length threshold. These neighbors are grouped into different scales based on their edge lengths, and triangles at each scale are formed to construct the proposed triplet descriptor.

After the prior map is constructed, vertex matching of descriptors is performed across all scale levels, and the multi-scale matching pairs are fed into the pose estimation module. This module initially searches for the maximal clique from the matching pairs at the smallest scale. If the number of points in the clique is insufficient, matching pairs from larger edge-length scales are gradually incorporated to enhance the maximal clique until the required number of points is met.

B. The Multi-scale Edge-based Triplet Descriptor

To enhance the discriminative capability of the triplet-based descriptors in TripleLoc [15], a length ratio histogram is introduced in addition to the angle histogram [28]. Considering the limited field of view of the query scan, which often fails to observe all instances in the reference map, especially those located in peripheral regions, TripleLoc selects triangle edges shorter than 20 meters for descriptor construction. This strategy aims to mitigate structural inconsistencies caused by differences in sensor coverage. However, empirical observations indicate that an overly conservative edge length threshold leads to a large number of isolated points. Isolated points refer to semantic cluster points that do not participate in the

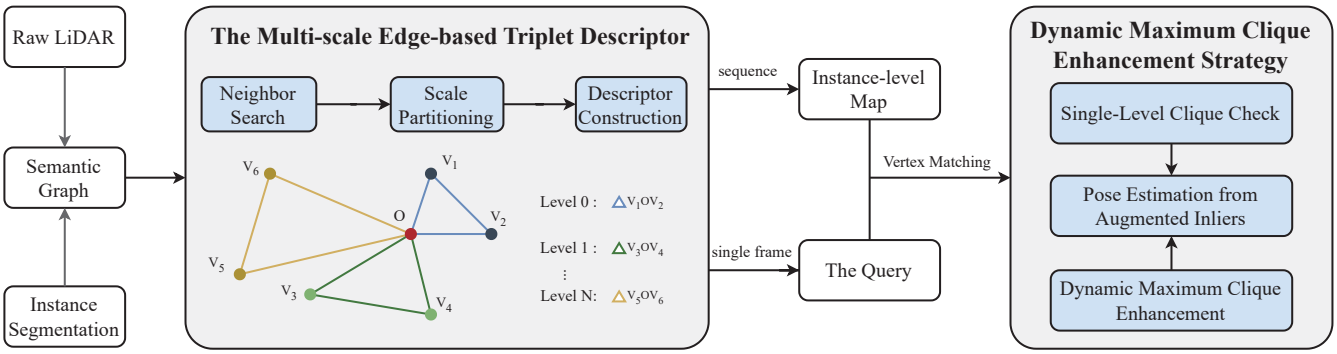

Fig. 1: The overall structure of our proposed network. The system mainly consists of a multi-scale edge-based triplet descriptor and a dynamic maximal clique enhancement strategy. It constructs descriptors at multiple scales by grouping neighboring points based on edge length thresholds. For pose estimation, it dynamically expands the search range of matching points to find larger maximal cliques.

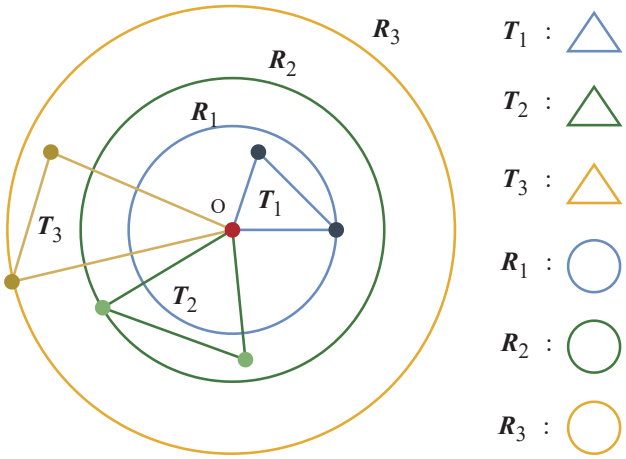

Fig. 2: Illustration of multi-scale edge-based triplet construction. The red point **O** denotes the cluster center of a semantic instance. Blue, green, and brown points represent neighboring cluster points within three edge length intervals: $[0, R_1]$, $(R_1, R_2]$, and $(R_2, R_3]$, respectively. Triangles formed by point **O** and its neighbors within each range are shown in corresponding colors to represent different scales.

construction of any descriptor. This issue is particularly prominent in sparse semantic environments, where a limited variety of instances combined with sparse spatial distribution often results in insufficient descriptive information due to the strict edge length constraint, thereby reducing the discriminative ability of the descriptors.

To balance construction efficiency with the avoidance of boundary instances, and to better capture structural relations among cluster points, a multi-scale edge-based triplet descriptor is proposed. The construction of the proposed descriptor is shown in Fig. 2. Instead of using a fixed edge length threshold, triangle construction is performed across multiple scales, generating combinations with varying edge lengths. Since triangles with longer edges often include structures formed by shorter edges, a segmented thresholding scheme is adopted. Specifically, a set of thresholds $L_0, L_1, \ldots, L_n$ is defined such that $0 < L_0 < L_1 < \ldots < L_n$. Based on these thresholds, N groups of triplet descriptors are constructed, each corresponding to one of the length intervals $[0, L1], (L1, L2], \ldots, (L_{n-1}, L_n]$. This strategy enables the modeling of scene structure at different scales and allows for the selective use of descriptors according to task-specific requirements in place recognition. We define the set of triplet-based descriptors constructed from multi-scale edge structures in the current frame as:

$$\mathcal{D} = \bigcup_{k=1}^{n} \mathcal{D}_{(L_{k-1}, L_k]}, \tag{1}$$

where the set $\mathcal{D}_{(L_{k-1},L_k]} = \{\mathbf{d}_{ijk}^{(L_{k-1},L_k]}\}$ denotes the triplet descriptors constructed within the edge length interval $(L_{k-1}, L_k]$. Each $\mathbf{d}_{ijk}^{(L_{k-1},L_k]}$ encodes the local geometric information of the corresponding triplet vertex, including the angles and the length ratios of the edges.

Ultimately, the multi-scale edge-based triplet descriptor improves discriminability while maintaining adaptability to sparse semantic scenes and robustness in handling boundary instances, achieving more robust global descriptive capability. For a more detailed understanding of the construction of angle and edge histograms in triplet-based descriptors, it can be found in works [15] and [28], where the underlying principles are thoroughly discussed.

C. The dynamic clique enhancement strategy

In global localization tasks, the scan-to-map matching is typically used to obtain a set of 3D matching points for optimization, and a nonlinear least squares problem is constructed to solve the global pose of the query. To ensure the accuracy of nonlinear optimization, outliers need to be removed from the initial matching point set. Considering that the triplet descriptor is constructed based on graph structures from semantic cluster points and that maximum clique search can be directly performed on the graph structure, this approach simplifies implementation complexity and ensures that the selected point set has high geometric consistency. At least three pairs of valid matching points are required to estimate

TABLE I: Comparison of different metrics across query sequences and methods. The best results are highlighted in bold.

Metrics	Method	DCC05	DCC06	KAIST05	KAIST06	Round02	Round03	Town01	Town02	River05	River06	Bridge02	Bridge03
P (%)	TripletLoc	97.75	95.30	78.55	70.54	75.89	81.95	55.26	58.52	79.89	75.84	36.92	41.68
	Multi-2-En	97.13	96.40	89.28	83.04	84.95	92.09	68.09	63.02	84.87	84.24	48.12	53.93
	Multi-3-En	**98.09**	**99.02**	**93.91**	**91.96**	**86.78**	**93.85**	**76.29**	**67.04**	**88.80**	**89.83**	**50.94**	**57.89**
RTE (m)	TripletLoc	1.18 ± 0.95	0.84 ± 0.62	0.99 ± 0.83	0.75 ± 0.65	0.89 ± 0.65	0.80 ± 0.64	1.37 ± 1.28	1.35 ± 1.04	1.36 ± 0.77	1.32 ± 0.73	1.67 ± 1.07	1.68 ± 1.01
	Multi-2-En	1.02 ± 0.88	0.70 ± 0.40	0.74 ± 0.59	0.57 ± 0.49	0.79 ± 0.56	0.68 ± 0.51	1.09 ± 1.03	1.06 ± 0.79	1.15 ± 0.61	1.07 ± 0.51	1.44 ± 0.91	1.43 ± 0.87
	Multi-3-En	**0.97 ± 0.87**	**0.66 ± 0.35**	**0.69 ± 0.54**	**0.52 ± 0.41**	**0.77 ± 0.53**	**0.65 ± 0.47**	**0.94 ± 0.91**	**0.98 ± 0.76**	**1.09 ± 0.55**	**0.99 ± 0.37**	**1.39 ± 0.88**	**1.35 ± 0.78**
RRE (°)	TripletLoc	2.29 ± 1.62	2.24 ± 1.63	2.80 ± 2.09	2.40 ± 2.16	2.83 ± 1.80	2.26 ± 1.63	2.91 ± 2.02	2.86 ± 1.91	3.16 ± 1.89	3.06 ± 1.88	3.57 ± 2.14	3.32 ± 2.11
	Multi-2-En	1.73 ± 1.30	1.72 ± 1.19	2.08 ± 1.73	1.70 ± 1.51	2.57 ± 1.62	1.85 ± 1.41	2.50 ± 1.94	2.16 ± 1.59	2.37 ± 1.51	2.15 ± 1.43	2.81 ± 1.83	2.74 ± 1.86
	Multi-3-En	**1.53 ± 1.10**	**1.53 ± 1.10**	**1.93 ± 1.68**	**1.39 ± 1.30**	**2.49 ± 1.57**	**1.73 ± 1.39**	**2.24 ± 1.77**	**1.96 ± 1.46**	**2.09 ± 1.37**	**1.90 ± 1.26**	**2.57 ± 1.75**	**2.46 ± 1.74**
MaxClique	TripletLoc	17.00 ± 8.54	19.00 ± 9.69	11.00 ± 8.83	14.00 ± 10.00	13.00 ± 9.48	16 ± 11.74	5.00 ± 3.16	6.00 ± 3.46	10.00 ± 6.00	12.00 ± 7.68	7.00 ± 5.56	8.00 ± 6.08
	Multi-2-En	23.00 ± 11.00	26.00 ± 13.00	15.00 ± 12.12	20.00 ± 12.88	18.00 ± 11.35	22 ± 13.78	8.00 ± 4.47	8.00 ± 5.29	15.00 ± 7.34	17.00 ± 9.60	9.00 ± 6.70	10.00 ± 7.61
	Multi-3-En	**26.00 ± 11.83**	**30.00 ± 14.28**	**16.00 ± 12.84**	**22.00 ± 14.07**	**19.00 ± 12.16**	**24 ± 14.45**	**8.00 ± 5.09**	**9.00 ± 5.83**	**16.00 ± 7.93**	**18.00 ± 10.34**	**10.00 ± 6.92**	**11.00 ± 7.87**
DescTime (ms)	TripletLoc	0.30 ± 0.26	0.34 ± 0.24	0.19 ± 0.09	0.27 ± 0.24	0.21 ± 0.22	0.24 ± 0.25	0.15 ± 0.11	0.14 ± 0.14	0.19 ± 0.19	0.22 ± 0.10	0.22 ± 0.20	0.19 ± 0.21
	Multi-2-En	0.43 ± 0.24	0.73 ± 0.55	0.26 ± 0.23	0.51 ± 0.42	0.32 ± 0.20	0.41 ± 0.28	0.14 ± 0.07	0.16 ± 0.14	0.28 ± 0.17	0.38 ± 0.25	0.25 ± 0.23	0.26 ± 0.22
	Multi-3-En	0.56 ± 0.34	1.02 ± 0.83	0.35 ± 0.38	0.72 ± 0.71	0.44 ± 0.36	0.58 ± 0.35	0.19 ± 0.22	0.20 ± 0.15	0.37 ± 0.23	0.52 ± 0.33	0.27 ± 0.27	0.38 ± 0.26
MatchTime (ms)	TripletLoc	8.27 ± 2.50	10.66 ± 3.56	3.29 ± 1.80	5.00 ± 2.71	8.54 ± 3.91	10.08 ± 3.82	1.12 ± 0.72	1.04 ± 0.68	2.80 ± 1.23	3.67 ± 1.62	8.79 ± 6.33	8.41 ± 6.68
	Multi-2-En	15.10 ± 4.47	21.36 ± 6.80	6.32 ± 3.42	10.35 ± 5.36	17.40 ± 7.53	20.26 ± 7.39	1.96 ± 1.15	2.15 ± 1.40	5.78 ± 2.35	7.35 ± 3.05	16.82 ± 11.06	18.62 ± 12.69
	Multi-3-En	22.15 ± 6.70	31.29 ± 10.16	9.21 ± 5.20	15.35 ± 7.88	25.61 ± 11.51	29.77 ± 10.81	2.85 ± 1.78	3.10 ± 2.09	8.45 ± 3.50	10.84 ± 4.52	23.32 ± 16.57	28.37 ± 18.12

the relative pose from the matching points. However, even after removing some outliers by the maximum clique method, the retained inliers may still have uncertainties, which can cause estimation failures during subsequent nonlinear optimization. Therefore, significantly more than three pairs of matching points are usually required in practical estimation to improve solution accuracy.

To address this, we propose using a multi-scale edge-based triplet descriptor that effectively supplements the number of inliers by integrating vertices from higher-scale descriptors. It should be noted that in scenes with sufficient semantic information, introducing multi-scale descriptors does not significantly improve estimation accuracy and may instead increase redundant computation. Therefore, we introduce a dynamic maximum clique enhancement strategy to ensure adequate observation in sparse scenes while avoiding unnecessary computational burden in dense semantic scenes.

Specifically, we first check whether the matching point set generated by the single-scale triplet descriptor meets the minimum maximum clique size requirement. If not, vertices from higher-scale triplet descriptors are progressively added to expand the candidate point set until the minimum inlier number required for estimation is satisfied. This strategy effectively balances robustness and computational efficiency and demonstrates good adaptability across different semantic scenarios. We obtain a multi-scale set of matching points based on the correspondence between the query frame and the global map. Let the set of semantic points in the map be denoted as follows:

$$\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{N}, \quad \mathbf{p}_i \in \mathbb{R}^3, \tag{2}$$

and the corresponding matching observations in the query frame are:

$$\mathcal{Q} = \{\mathbf{q}_i\}_{i=1}^{N}, \quad \mathbf{q}_i \in \mathbb{R}^3. \tag{3}$$

We aim to estimate the rotation matrix $R \in \mathrm{SO}(3)$ and the translation vector $\mathbf{t} \in \mathbb{R}^3$ such that the transformed map point set aligns as closely as possible with the query frame point set:

$$R^*, \mathbf{t}^* = \arg \min_{R \in \mathrm{SO}(3), \mathbf{t} \in \mathbb{R}^3} \sum_{i \in \mathcal{I}} \|R\mathbf{p}_i + \mathbf{t} - \mathbf{q}_i\|^2, \tag{4}$$

where $\mathcal{I} \subseteq \{1, 2, \ldots, N\}$ denotes the set of inlier indices used for the estimation. To remove outliers in the matches, we construct a matching consistency graph based on geometric consistency $G = (V, E)$. Each vertex $v_i \in V$ corresponds to a matched point pair $(\mathbf{p}_i, \mathbf{q}_i)$.

$$C = \arg \max_{C' \subseteq \{1,\ldots,N\}} |C'|, \quad \forall i, j \in C',\ e_{ij} \in E. \tag{5}$$

Considering that in sparse semantic scenes the maximum clique size constructed at a single scale may be insufficient to support stable pose estimation, we introduce a multi-scale descriptor structure. Let l denote the maximum clique index set extracted at the C_l scale, then the final matching point pair index set used for estimation is constructed by the following strategy:

$$C_{\text{final}} = C_0 \cup \left(\bigcup_{l=1}^{L} C_l \right), \quad \text{s.t. } |C_{\text{final}}| \geq T. \tag{6}$$

where, T is the minimum threshold for the number of matching points, which is usually set to no less than 10 to ensure estimation stability. Finally, nonlinear least squares optimization is performed on the set of point pairs $\{(\mathbf{p}_i, \mathbf{q}_i) \mid i \in C_{\text{final}}\}$ to estimate the pose $(R^*, \mathbf{t}^*)$ of the current query frame relative to the global map.

Considering that the number of cluster points in semantic scenes is significantly reduced compared to the raw point cloud, and that the number of reliable vertices in a single-scale triplet descriptor further decreases after inlier filtering, the pose estimation may ultimately fail. Through extensive experiments, we confirm that to ensure stable convergence of the estimator, the maximum clique must contain at least ten vertices. However, in sparse semantic environments, the single-layer descriptor often has many isolated points and insufficient matching points, which makes pose estimation less accurate. Our approach alleviates the above issues.

IV. Experiments

To validate the superiority of the proposed system in global localization tasks, we conduct comparative experiments on public datasets against the method TripleLoc. In addition, LiDAR point cloud semantic segmentation is performed in advance to support semantic instance clustering during validation. Detailed experimental settings and final result comparisons are presented in the following sections. It is

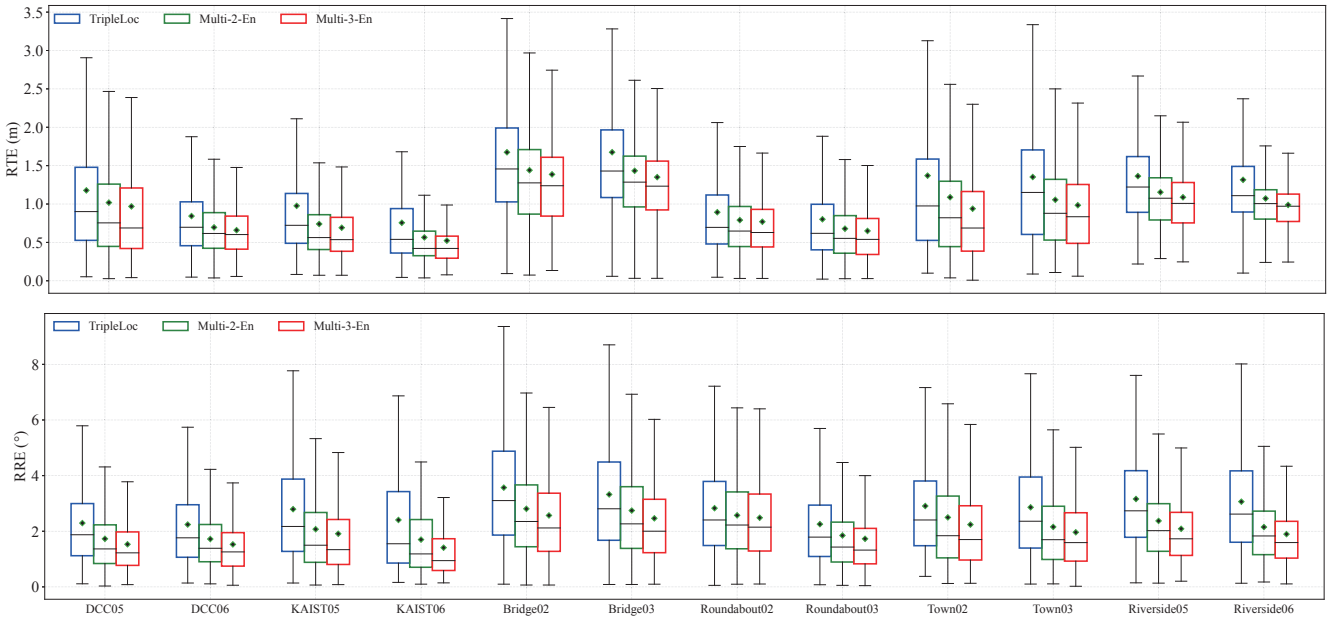

Fig. 3: Visualization of RTE and RRE box plots for each sequence. In all sequences, Multi-3-En shows improvement compared to Multi-2-En and TripleLoc.

worth noting that all the methods are executed on Ubuntu 20.04 with an Intel Core i7-14700F processor without GPU parallelization.

A. Dataset

The HeLiPR dataset [29] is designed to promote research on place recognition across heterogeneous LiDAR sensors, particularly in complex environments with spatial and temporal variations. The dataset collects data from four different types of LiDAR sensors and covers various representative scenes, including urban roads, narrow alleys, and bridge areas. The bridge scenes contain many regions with similar appearances but subtle differences in detail, which increases the difficulty of place recognition. HeLiPR supports both intra-session recognition within a single data collection and inter-session recognition across different time periods, demonstrating its potential for long-term place recognition. The dataset provides a rich testing environment and high-quality annotation benchmarks for research on heterogeneous LiDAR-based place recognition. For simplicity in experiments, this study uses only the data collected by the Ouster OS2-128 sensor.

B. LiDAR Semantic Segmentation

Considering that Cylinder3D [27] leverages the spatial characteristics of LiDAR point clouds more effectively through its innovative cylindrical voxelization and efficient convolutional design, it achieves improved accuracy and robustness in semantic segmentation. Compared with SPVNAS, Cylinder3D demonstrates clear advantages in both precision and efficiency, making it particularly suitable for handling sparse and large-scale point cloud tasks. Unlike TripleLoc, we ultimately adopt the unrefined Cylinder3D to perform instance segmentation of roads, trunks, poles, and traffic signs.

C. Comparative Results

1) Efficiency: As shown in the Tab. I, regarding descriptor computation and matching time, the computational overhead exhibits a certain increasing trend as the method progresses from TripleLoc to Multi-2-En and then to Multi-3-En. Specifically, the descriptor computation time increases by approximately 0.12 ms on average with each additional neighborhood level, indicating that the expansion of the neighborhood range has a limited impact on descriptor generation speed. The increase in matching time is more pronounced, with an average rise of about 5.98 ms and 5.57 ms for each upgrade respectively, reflecting that the expansion of the neighborhood range requires processing more vertices during matching, thereby increasing the computational burden.

2) Accuracy: In terms of success rate (P), the average increases from 67.7% in TripleLoc to 77.0% in Multi-2-En and further to 81.5% in Multi-3-En, resulting in a total improvement of approximately 13.8%. For example, KAIST05 improves significantly from 78.55% to 93.91%, Town01 increases from 55.26% to 76.29%, and Bridge02 rises from 36.92% to 50.94%, indicating that a broader neighborhood range effectively enhances local matching stability and global estimation robustness. The average relative translational error (RTE) decreases from 1.16 meters in TripleLoc to 0.94 meters in Multi-2-En, representing a reduction of about 19%, and further to 0.87 meters in Multi-3-En, with an additional 7.4% decrease. For instance, the RTE of KAIST06 drops from 0.75 meters to 0.52 meters, and Town01 decreases from 1.37 meters to 0.94 meters, reflecting the effectiveness of the dynamic maximum clique enhancement strategy. The relative rotational error (RRE) also consistently decreases, with the average falling from 2.75 degrees in TripleLoc to 2.25

degrees in Multi-2-En, a reduction of approximately 18.2%, and further to 2.06 degrees in Multi-3-En, an additional 8.4% decrease. Notably, KAIST06 decreases from 2.40 degrees to 1.39 degrees, and Bridge03 decreases from 3.32 degrees to 2.46 degrees. Meanwhile, the box plots of RTE and RRE for all sequences are visualized in Fig. 3, respectively.

The growth of the maximum clique on each sequence also reflects that the triplet descriptor of multi-scale edge structure can provide more useful information, and the dynamic clique enhancement strategy fully guarantees the number of correct inliers, improving the accuracy of pose estimation.

Overall, the multi-scale edge structure of the triplet descriptor and the dynamic maximum clique enhancement strategy provide consistent average performance improvements and achieve substantial gains in several challenging sequences. Although they introduce some additional computational cost, they offer greater enhancements in robustness and pose estimation accuracy for most application scenarios.

V. Conclusions and Future Work

We propose a LiDAR-based global localization method that implements a multi-scale edge-structured triplet descriptor together with a dynamic maximal clique enhancement strategy, which improves descriptor discriminability and enhances pose estimation accuracy, especially in sparse environments. However, we observe that the proposed descriptor still lacks sufficient discriminability in scenarios with semantic repetition and similar geometric structures. In such cases, object-level semantic descriptors fail to capture the subtle distinctions introduced by geometric textures. The future work will focus on exploring more comprehensive and efficient utilization of environmental texture information to improve global localization performance in scenes with repetitive semantics and geometric structures.

References

[1] Y. Zhang, P. Shi, and J. Li, "Lidar-based place recognition for autonomous driving: A survey," *ACM Computing Surveys*, vol. 57, no. 4, pp. 1–36, 2024.

[2] W. Ma, S. Huang, and Y. Sun, "Skyloc: Cross-modal global localization with a sky-looking fish-eye camera and openstreetmap," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 5, pp. 5832–5842, 2025.

[3] Z. Du, S. Ji, and K. Khoshelham, "3-d lidar-based place recognition techniques: A review of the past ten years," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–24, 2024.

[4] H. Xu, H. Liu, S. Meng, and Y. Sun, "A novel place recognition network using visual sequences and lidar point clouds for autonomous vehicles," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2023, pp. 2862–2867.

[5] W. Ma, H. Yin, L. Yao, Y. Sun, and Z. Su, "Evaluation of range sensing-based place recognition for long-term urban localization," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 5, pp. 4905–4916, 2024.

[6] H. Xu, H. Liu, S. Huang, and Y. Sun, "C2l-pr: Cross-modal camera-to-lidar place recognition via modality alignment and orientation voting," *IEEE Transactions on Intelligent Vehicles*, vol. 10, no. 2, pp. 1128–1144, 2025.

[7] P. Shi, Y. Zhang, and J. Li, "Lidar-based place recognition for autonomous driving: A survey," *CoRR*, 2023.

[8] 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*, vol. 132, no. 8, pp. 3139–3171, 2024.

[9] 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)*, 2023, pp. 1897–1903.

[10] K. Singh and J. J. Leonard, "Open-set semantic uncertainty aware metric-semantic graph matching," *arXiv preprint arXiv:2409.11555*, 2024.

[11] R. K. Ross, P. N. Zivich, J. S. Stringer, and S. R. Cole, "M-estimation for common epidemiological measures: introduction and applied examples," *International Journal of Epidemiology*, vol. 53, no. 2, p. dyae030, 2024.

[12] 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, 2021.

[13] J. M. Martínez-Otzeta, I. Rodríguez-Moreno, I. Mendialdua, and B. Sierra, "Ransac for robotic applications: A survey," *Sensors*, vol. 23, no. 1, p. 327, 2022.

[14] Q. Dai, R.-H. Li, M. Liao, H. Chen, and G. Wang, "Fast maximal clique enumeration on uncertain graphs: A pivot-based approach," in *Proceedings of the 2022 international conference on management of data*, 2022, pp. 2034–2047.

[15] W. Ma, H. Yin, P. J. Y. Wong, D. Wang, Y. Sun, and Z. Su, "Tripletloc: One-shot global localization using semantic triplet in urban environments," *IEEE Robotics and Automation Letters*, vol. 10, no. 2, pp. 1569–1576, 2025.

[16] T. Shan, B. Englot, F. Duarte, C. Ratti, and D. Rus, "Robust place recognition using an imaging lidar," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 5469–5475.

[17] L. Di Giammarino, I. Aloise, C. Stachniss, and G. Grisetti, "Visual place recognition using lidar intensity information," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 4382–4389.

[18] 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, 2022.

[19] I. D. Miller, A. Cowley, R. Konkimalla, S. S. Shivakumar, T. Nguyen, T. Smith, C. J. Taylor, and V. Kumar, "Any way you look at it: Semantic crossview localization and mapping with lidar," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2397–2404, 2021.

[20] J. Ma, X. Chen, J. Xu, and G. Xiong, "Seqot: A spatial–temporal transformer network for place recognition using sequential lidar data," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 8, pp. 8225–8234, 2023.

[21] H. Lai, P. Yin, and S. Scherer, "Adafusion: Visual-lidar fusion with adaptive weights for place recognition," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 12 038–12 045, 2022.

[22] 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, 2023.

[23] P. Yin, H. Cao, T.-M. Nguyen, S. Yuan, S. Zhang, K. Liu, and L. Xie, "Outram: One-shot global localization via triangulated scene graph and global outlier pruning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 13 717–13 723.

[24] S. S. Lee, D. H. Lee, D. K. Lee, and C. K. Ahn, "Improved nonlinear finite-memory estimation approach for mobile robot localization," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 5, pp. 3330–3338, 2022.

[25] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "G2o: A general framework for graph optimization," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 3607–3613.

[26] L. Luo, S.-Y. Cao, Z. Sheng, and H.-L. Shen, "Lidar-based global localization using histogram of orientations of principal normals," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 771–782, 2022.

[27] X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, W. Li, H. Li, and D. Lin, "Cylindrical and asymmetrical 3d convolution networks for lidar segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9939–9948.

[28] W. Ma, S. Huang, and Y. Sun, "Triplet-graph: Global metric localization based on semantic triplet graph for autonomous vehicles," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3155–3162, 2024.

[29] M. Jung, W. Yang, D. Lee, H. Gil, G. Kim, and A. Kim, "Helipr: Heterogeneous lidar dataset for inter-lidar place recognition under spatiotemporal variations," *The International Journal of Robotics Research*, vol. 43, no. 12, pp. 1867–1883, 2024.