RING++: Roto-Translation Invariant Gram for Global Localization on a Sparse Scan Map

Xuecheng Xu, Sha Lu, Jun Wu, Haojian Lu, *Member, IEEE*, Qiuguo Zhu, Yiyi Liao, Rong Xiong, *Senior Member, IEEE*, and Yue Wang, *Member, IEEE*

***Abstract*—Global localization plays a critical role in many robot applications. LiDAR-based global localization draws the community's focus with its robustness against illumination and seasonal changes. To further improve the localization under large viewpoint differences, we propose RING++ that has roto-translation-invariant representation for place recognition and global convergence for both rotation and translation estimation. With the theoretical guarantee, RING++ is able to address the large viewpoint difference using a lightweight map with sparse scans. In addition, we derive sufficient conditions of feature extractors for the representation preserving the roto-translation invariance, making RING++ a framework applicable to generic multichannel features. To the best of our knowledge, this is the first learning-free framework to address all the subtasks of global localization in the sparse scan map. Validations on real-world datasets show that our approach demonstrates better performance than state-of-the-art learning-free methods and competitive performance with learning-based methods. Finally, we integrate RING++ into a multirobot/session simultaneous localization and mapping system, performing its effectiveness in collaborative applications.**

***Index Terms*—Global localization, place recognition, simultaneous localization and mapping (SLAM).**

I. Introduction

GLOBAL localization aims to estimate the pose of a robot on a map using onboard sensor measurements without priors. This task is essential for many robotics applications, including loop closures and map alignments in simultaneous localization and mapping (SLAM) systems, and relocalization in navigation systems. Vision-based methods have advanced rapidly in the last decade by exploiting image cues. However, these approaches are sensitive to illumination and seasonal changes between current and mapping session [1], [2]. Recent studies have shown that LiDAR can be employed to overcome the difficulty [3], [4], [5], [6].

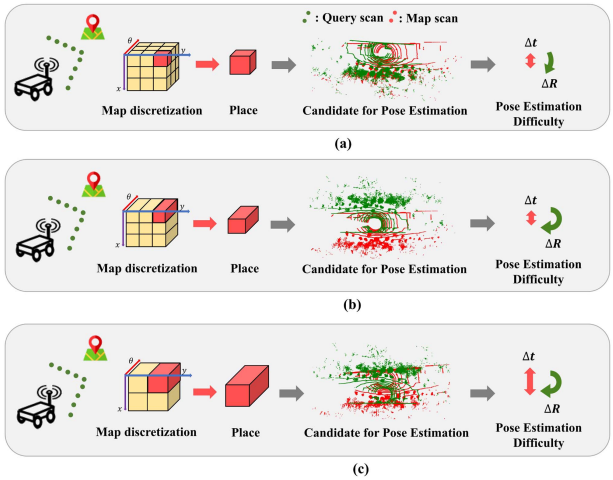


Fig. 1. Demonstration of global localization pipelines with place recognition and pose estimation. Three approaches to global localization are presented. The cost of place recognition is significantly reduced as the map discretization becomes sparser, but it also raises the challenges of roto-translation invariance for place recognition and global convergence for pose estimation. (a) Normal dense place representation. (b) Dense place representation with rotation invariance. (c) Sparse place representation with roto-translation invariance (ours).

However, even with LiDAR-based methods, reliable localization is still challenging when the current scan and mapping sessions have trajectories with large viewpoint differences.

To clarify the challenge, we define global localization as a search problem: the query is the current LiDAR scan, and the search space is the pose space covering the entire map. A typical global localization pipeline consists of two steps: 1) place recognition; and 2) pose estimation, solving the problem in a coarse-to-fine process. As shown in Fig. 1, these two steps are divided by the *place*, which encodes a representative scan and its pose. At the coarse level, the search space is discretized into many places. Accordingly, place recognition aims at finding the place to which the robot is closest. At the fine level, the search space is the continuous pose space centered at the pose of the place. Taking the pose of the place as initialization, the pose estimation process aims to obtain the accurate pose of the query. When the viewpoint difference is large, a query and a map scan taken in the same place can be different, which brings ambiguity to place recognition and calls for a large convergence basin in pose estimation.

Manuscript received 23 May 2023; revised 21 July 2023; accepted 31 July 2023. Date of publication 18 August 2023; date of current version 6 December 2023. This paper was recommended for publication by Associate Editor Ayoung Kim and Editor Sven Behnke upon evaluation of the reviewers' comments. This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0114500 and in part by the Natural Science Foundation of Zhejiang Province under Grant LGG21F030012. *(Xuecheng Xu and Sha Lu contributed equally to this work.) (Corresponding author: Yue Wang.)*

The authors are with Zhejiang University, Zhejiang 310027, China (e-mail: xuechengxu@zju.edu.cn; lusha@zju.edu.cn; wujun_csecyber@zju.edu.cn; luhaojian@zju.edu.cn; qgzhu@iipc.zju.edu.cn; lyyecho1119@gmail.com; rxiong@zju.edu.cn; wangyue@iipc.zju.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TRO.2023.3303035.

Digital Object Identifier 10.1109/TRO.2023.3303035

We focus on global localization in autonomous driving scenarios, which inherently assumes that the sensor is mounted roughly parallel to the ground plane. Thus, the viewpoint difference is three-degree-of-freedom (DoF), which includes one-DoF rotation and two-DoF translation. Early works utilize three-DoF discretization [7], [8], [9], [10], regarding scans with slight viewpoint differences as different places. This is illustrated in the first row of Fig. 1. Such formulation relaxes the place recognition to only consider small viewpoint differences and simplifies the pose estimation to be locally convergent, demonstrating good performance. However, it calls for large memory to save dense places. The second row of Fig. 1 illustrates a pioneering work, Scan Context [3], [11], which regards scans with arbitrary rotation as the same place by explicitly modeling the rotation invariance for place representation. Besides, it proposes a global convergent rotation solver in the pose estimation stage, as no rotation initialization is available from place recognition with rotation invariance. Thus, discretization is reduced to two-DoF (2-D translation). Unfortunately, the rotation invariance is sensitive to the translation difference, hence requiring dense discretization in translation space. Considering the advantage of explicitly modeling rotation invariance, we raise a question: *Is it possible to build a representation that is invariant to both rotation and translation?*

In this article, we address the large viewpoint difference in global localization using a sparse scan map. As illustrated in the third row in Fig. 1, we propose a framework RING++ that achieves *roto-translation invariance* for place recognition and *global convergence* for both rotation and translation estimation, theoretically guaranteeing the performance under large viewpoint differences. Therefore, only a sparse discretization of map pose space in two-DoF translation is sufficient. RING++ takes a bird's-eye view (BEV) of the scan and processes in two passes: the representation pass for building roto-translation-invariant representation and the solving pass for place recognition and pose estimation. In the representation pass, we exploit the properties of Radon transform (RT) and Fourier transform on BEV representation to generate roto-translation-invariant place representation and derive sufficient conditions of the feature extractor for the representation to preserve the roto-translation invariance. In the solving pass, the roto-translation-invariant representation leads the successful place recognition even when large viewpoint differences exist between the query and map scans. By employing the cross correlation, the rotation and translation are further estimated with global convergence. To the best of our knowledge, RING++ is the first learning-free framework to address all the subtasks of global localization in the sparse scan map. The experimental results validate our method in both place recognition and pose estimation and show a superior performance compared to existing methods. To summarize, this article presents the following contributions.

1) We propose a novel learning-free framework named RING++ for global localization on the sparse scan map, which can simultaneously solve place recognition and pose estimation tasks.
2) We theoretically prove the roto-translation invariance property of our place representation. Meanwhile, the pose estimation solvers are presented with global convergence.
3) We further derive sufficient conditions of feature extractors for the representation preserving the roto-translation invariance, enabling the framework to aggregate multichannel features.
4) We validate RING++ on three real-world datasets and multirobot SLAM applications in the wild. We make the code, SLAM system, and evaluation tools all publicly available.

In the preliminary conference paper [12], we proposed a method that utilizes a single-channel occupancy map to generate roto-translation-invariant place representation. In this article, we formally state and prove the roto-translation invariance of our representation to provide a theoretical guarantee. Then, we introduce the aggregation method to take multichannel features into the framework, yielding better discrimination to enhance both place recognition and pose estimation. We also derive the sufficient conditions of the feature extractor that can preserve the roto-translation invariance of our representation. Besides, we develop a global and effective correlation-based translation solver to address the outlier sensitivity in the conference version, which is nontrivial as the scans with large viewpoint differences must have less overlap. With all efforts above, the overall success rate of RING++ increases by almost 20%. We also validate the performance of RING++ with more ablation studies, substantial experiments with both learning-free and learning methods in diverse scenarios, as well as multirobot collaborative SLAM.

II. Related Works

In this section, we provide a literature review on LiDAR-based recognition and estimation tasks with two main lines embedded in previous works. To begin with, local point features are evolving to represent the point cloud in a compact and efficient manner. To overcome noise and variances, different aggregation strategies are proposed to generate robust global representations with invariance properties from local features.

A. Feature Extraction

In the early stages of scan-based recognition, researchers focus on how to generate compact representations of scans. There are many ways to achieve this goal, but the early consensus is to find compact and effective local features. Geometric relations are first explored in many approaches. Stein and Medioni [13] propose structural indexing (SI), which constructs a representation from 3-D curves and splashes. Rusu et al. [14] build point feature histograms (PFHs) by aggregating four handcrafted features which stem from normals and the distance between k-nearest point pairs. In the large-scale scene interpretation task, Weinmann et al. [15] provide eight semantic point features calculated by normalized eigenvalues of each point with its neighbors. With the wide use of local features, researchers found that fine-resolution features are easily influenced by noises. To overcome this limitation, some coarser features are presented. Spin Image [16] counts point numbers in each volume of cylinder support around a keypoint. The ensemble of shape functions (ESF) [17] method uses voxel grids to approximate the surface and encodes shape properties. SHOT [18], [19] combines signature and histogram in a local reference frame and encodes the

cosine value of the angle between the normal and the local z-axis at the feature point.

Although these features are widely used for scan-based recognition tasks, they are unstable when applied to the LiDAR-based place recognition scenario. The poor generalization performance is caused by the sensor characteristics of the newly equipped LiDAR. Points collected by LiDAR are unstructured, and the sparsity of points varies along with the sensor range. In order to tackle this problem, learning-based methods are proposed. Due to the strong descriptive ability of the 2-D convolution neural network (CNN), some researchers extend it to 3-D cases by representing point clouds in 3-D volumes [20], [21], [22]. However, these CNN-based methods usually introduce quantization errors and high computational costs. To alleviate the drawbacks brought by the CNN, PointNet [23] first learns features directly from the raw 3-D point cloud data. PointNet++ [24] further introduces hierarchical feature learning to learn local features with increasing scales. To better acquire relationships between local points, graph-based [10], [25] and kernel-based [26] networks are proposed. Other than geometry information, OverlapNet [27] and SSC [28] also take semantic clues into account. Some recent works propose fusion frameworks for incorporating image features as well [29], [30], [31]. Despite the fact that different methods provide different local features, they are all presented in a multichannel manner. With this characteristic in mind, we propose a multichannel framework that can be viewed as a feature aggregation module, allowing us to use different local features and improve discrimination with roto-translation invariance.

B. Aggregation and Invariance

With local features extracted, global descriptors are often generated by aggregating the local features. Early aggregation methods can be divided into two classes: signature and histogram.

The signature describes the 3-D support by defining a local reference frame and encoding local features into the subset of the support. Histogram describes the support by encoding counts of local features. SI [13] is one of the first methods to use signatures to capture 3-D curves. 3-D SURF [32] extends the mature 2-D SURF [33] by voxelizing the 3-D data and utilizing Haar wavelet response to determine the saliency of each voxel.

More previous works prefer the histogram since it provides a coarser representation of the point cloud that is robust to slight variance. PFH [14], [34] and viewpoint feature histogram (VFH) [35] are the early methods that explicitly introduce translation and rotation invariance. VFH finds the viewpoint directions and counts the angles between the normals in a histogram. SHOT [18] collects the histogram of normal angles in a spherical bin around a keypoint to build the descriptor. Rohling [7] first utilizes histogram-based similarity measures in robotics systems to find loop closures.

With the appearance of learning-based local features, new aggregation methods are proposed in the place recognition task. NetVLAD [36], first introduced in visual place recognition, modifies VLAD [37] with learnable weights and integrates it into a CNN. Following the idea of NetVLAD, several methods [9], [10] apply NetVLAD to supervise the feature learning. DiSCO [38] utilizes Fourier transform to generate a global descriptor in the frequency domain. Apart from these specially designed aggregation methods, MinkLoc3D and its extensions [6], [39], [40] adopt the generalized-mean (GeM) [41] pooling layer to generate discriminative global descriptors.

Inspired by early works on rotation invariance, many approaches also focus on achieving invariance against viewpoint differences. M2DP [8] presents a multiview projection on the point cloud and uses principal component analysis (PCA) to compensate viewpoint difference. LocNet [42] aggregates the distance of consecutive points in the same ring to a rotation-invariant histogram and adopts a siamese network for feature learning. Iris [4] explicitly models rotation invariance using Fourier transform on polar images. Previous works only present rotation-invariant representation for place recognition, Scan context and its extension [3], [11] also realize lateral invariance and simultaneously estimate one-DoF rotation and one-DoF lateral translation, which is effective in autonomous driving. However, in the Scan context, translation invariance is achieved through augmentation based on urban road assumptions. When there is a large difference in viewpoint, translation invariance is invalid. DiSCO [38] adopts the invariance property of the Fourier transform to simultaneously estimate rotation and achieve invariance. Since the polar transform used in DiSCO is not translation invariant, it suffers from distortion caused by large translation variance. To overcome the influence brought by the large translation, Ding et al. [43] utilized RT to estimate one-DoF rotation with robust translation invariance. In this article, we further derive a multichannel framework based on the RT, which can estimate the three-DoF pose and preserve roto-translation invariance.

III. Overview

LiDAR-based global localization aims to estimate the pose T_Q of the query scan P_Q in a scan map $\mathfrak{M} \triangleq \{P_i, T_i\}$, which is a database populated by map scans P_i with registered poses T_i. In general, global localization is solved in two stages: place recognition and pose estimation. In the place recognition stage, a map scan with a large overlap with the query scan is retrieved from the map, denoted as P_n, which indicates that P_Q is collected at the place near P_n in the map. Then, in the pose estimation stage, the metric pose T_Q is achieved by estimating the relative pose T_{nQ} between P_Q and P_n, and applying T_{nQ} to T_n.

When the robot trajectory in the current session is different from the map session, P_Q and P_n can overlap but with T_Q and T_n being different, especially in rotation, e.g., opposite direction. At the same time, $|\mathfrak{M}|$, the place density of a map, is expected to be sparse for storage and efficiency. As a result, the main challenge for place recognition is to build a representation that is similar under large viewpoint differences, i.e., variance of the relative pose T_{nQ}, which further causes the challenge in pose estimation: the estimator should be globally convergent, as no reliable initial value for T_{nQ} is available.

A. Equivariance and Invariance

To deal with the pose variance in place recognition, the invariant representation of the scan is built. However, when the

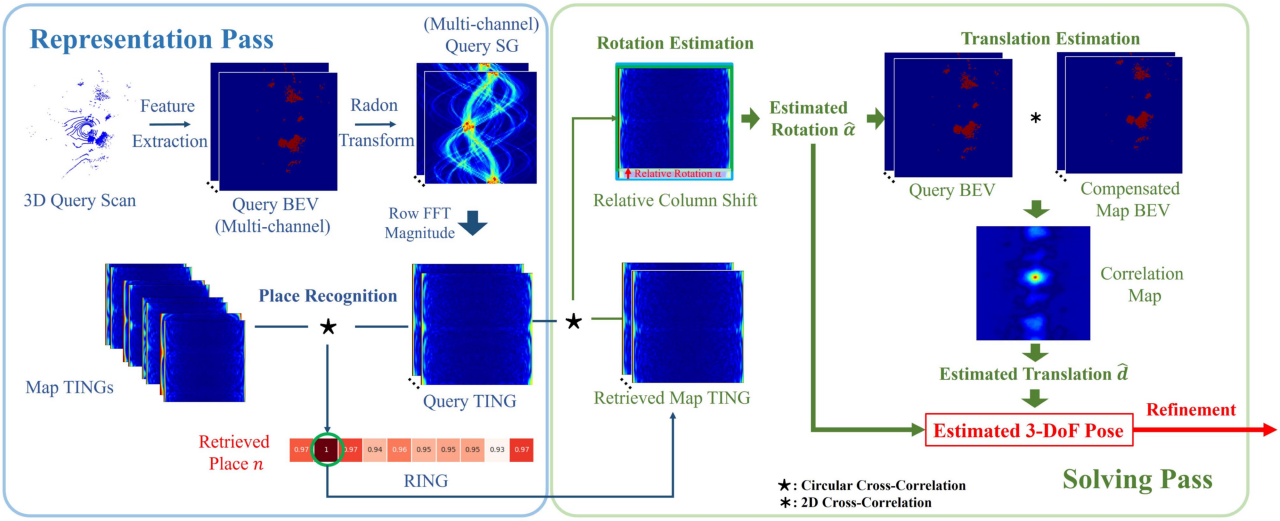


Fig. 2. Overall framework of the proposed method. RING representation is used for place recognition, TING representation is utilized for rotation estimation, and BEV representation is leveraged for translation estimation.

representation is invariant to the relative pose, the pose estimation becomes impossible to solve, calling for additional representations bridging the raw scan and the invariant representation. Following this idea, we formally introduce equivariance and invariance in the context of global localization. The representation is denoted as an operation g taking the scan P as input. P_d and P_α indicate that the scan is translated by d, or rotated by α. g_d and g_α indicate that the representation is translated by d or rotated by α. Then, we have the definitions.

Definition 1 (Translation equivariance): If the operation g is translation equivariant, it satisfies the equation

$$g_d(P) = g(P_d). \tag{1}$$

Definition 2 (Rotation equivariance): If the operation g is rotation equivariant, it satisfies the equation

$$g_\alpha(P) = g(P_\alpha). \tag{2}$$

Definition 3 (Translation invariance): If the operation g is translation invariant, it satisfies the equation

$$g(P) = g(P_d). \tag{3}$$

Definition 4 (Rotation invariance): If the operation g is rotation invariant, it satisfies the equation

$$g(P) = g(P_\alpha). \tag{4}$$

B. Framework

As shown in Fig. 2, RING++ extracts features from scans and builds representations for place recognition, rotation estimation, and translation estimation. Totally, there are two passes in RING++: A forward representation pass and a backward solving pass.

In the representation pass, the rotation-equivariant scan feature of P_Q is represented as sinogram (SG) by RT, then represented as translation-invariant rotation-equivariant gram (TING) by discrete Fourier transform (DFT), and finally represented as roto-translation-invariant gram (RING) by batch circular cross correlation.

In the solving pass, RING is utilized for place recognition with different query/mapping trajectories and low place density, by which the caused pose difference becomes invariant in RING. Then, TING is utilized for relative rotation estimation given the retrieved map scan P_n, since the translation in TING is invariant. Finally, BEV is utilized for relative translation estimation after the rotation in BEV is compensated by the estimation. Thanks to the decoupling leveraged by invariance, rotation and translation in T_{nQ} can be solved independently via globally convergent solvers.

IV. Representation and Solving

A. Representation Pass

1) Bird's-Eye View: The first step of RING++ is to extract the features from the scan. Following common pipelines, we eliminate the ground from the 3-D point cloud. Then, we voxelize the scan into a 3-D volume, in which each voxel encodes the occupancy, indicating whether there is a point inside. With the extracted features, we accumulate the height dimension of 3-D volume to generate a 2-D BEV representation $f(x, y) \in \mathbb{R}$. Specifically, if any of the voxels in the height dimension is occupied, the reduced 2-D grid in the BEV representation is set as occupied.

2) Sinogram: Given $f(x, y)$, we apply RT $\mathcal{R}$ to yield an SG. RT is a linear integral transform that maps $f(x, y)$ from the original image space (x, y) to the Radon parameter space (θ, τ), which is demonstrated in Fig. 3. Denoting the line for integral in RT as L, we have

$$L: \ x\cos\theta + y\sin\theta = \tau \tag{5}$$

where $\theta \in [0, 2\pi)$ represents the angle between L and the y-axis, and $\tau \in (-\infty, \infty)$ represents the perpendicular distance from

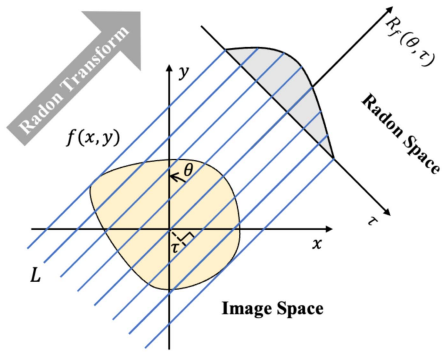


Fig. 3. Graph illustration of RT, which demonstrates a single row of the SG with a constant θ after RT.

the origin to L. By formalizing the integral results into a 2-D function with the Radon parameter as axes, we have $\mathcal{R}_f(L) = \mathcal{R}_f(\theta,\tau)$, namely SG. Specifically, the RT is calculated as

$$\begin{aligned}\mathcal{R}_f(\theta,\tau) &= \int_{x\cos\theta+y\sin\theta=\tau} f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(\tau - x\cos\theta - y\sin\theta)dxdy\end{aligned} \tag{6}$$

where $\delta(\cdot)$ is the Dirac delta function.

When the robot revisits the same place with different rotation α and translation $d \triangleq (\Delta x, \Delta y)^T$, SG can reflect such relative pose as follows:

Rotation: A rotation of $f(x,y)$ by a rotation angle α results in a circular shift along θ axis of SG:

$$\mathcal{R}_f(\theta,\tau) \xrightarrow{\alpha} \mathcal{R}_f(\theta+\alpha,\tau). \tag{7}$$

A rotation leads to a uniform shift along the θ axis of SG, which satisfies Definition 2 of rotation equivariance, arriving at the following result.

Lemma 1: SG is rotation equivariant.

Translation: Denote r_θ as a directional vector, i.e., $r_\theta \triangleq (\cos\theta, \sin\theta)^T$. A translation of $f(x,y)$ by d results in an angle-dependent shift of τ parameter on SG

$$\mathcal{R}_f(\theta,\tau) \xrightarrow{d} \mathcal{R}_f(\theta,\tau - r_\theta \cdot d). \tag{8}$$

Different from the rotation, translation leads to a nonuniform shift along the τ-axis of SG.

Pose: With both rotation α and translation d, SG can be described by combining (7) and (8), as follows:

$$\mathcal{R}_f(\theta,\tau) \xrightarrow{\alpha,d} \mathcal{R}_f(\theta+\alpha,\tau - r_{\theta+\alpha}\cdot d). \tag{9}$$

3) Translation-Invariant Rotation-Equivariant Gram: To further eliminate the coupled effect of translation on the τ-axis of SG in (8), we apply row-wise 1-D DFT to SG along the τ-axis and then calculate the magnitude of the resultant frequency spectrum. We name this representation as TING, denoted as $M_f(\theta,\omega)$, where ω is the sampled frequency in the discrete frequency spectrum.

Lemma 2: TING is rotation equivariant and translation invariant.

Proof of Lemma 2: Suppose that $M'_f(\theta,\omega)$ is the TING representation constructed from transformed BEV $f'(x,y)$ by random translation d. Referring to the shift property of the Fourier transform, we have

$$\begin{aligned}M'_f(\theta_j,\omega) &= |\mathcal{F}(R_f(\theta_j,\tau - r_{\theta_j}\cdot d)| \\ &= |\mathcal{F}(R_f(\theta_j,\tau))e^{-i2\pi w r_{\theta_j}\cdot d}| \\ &= |\mathcal{F}(R_f(\theta_j,\tau))||e^{-i2\pi w r_{\theta_j}\cdot d}| \\ &= |\mathcal{F}(R_f(\theta_j,\tau))| \\ &= M_f(\theta_j,\omega)\end{aligned} \tag{10}$$

where $|\cdot|$ is the operation of taking magnitude, $\mathcal{F}(\cdot)$ is the DFT operator, and θ_j means a row in SG and the same row in TING. Therefore, TING satisfies Definition 3 of translation invariance.

As we only apply DFT to the τ-axis of SG, the θ-axis of TING keeps the same as that of SG; thus, the rotation equivariance is reserved according to Lemma 1. ■

4) Roto-Translation-Invariant Gram: The final step of the representation pass is to eliminate the effect of rotation. According to (7), note that the rotation only leads to the cyclic shift of θ-axis of TING; we build the rotation-invariant representation in two steps. First, we employ a batch circular cross correlation between TING M_Q of query scan P_Q and TING M_i of every map scan P_i in $\mathfrak{M}$, resulting in a batch of correlation maps as

$$\begin{aligned}\mathfrak{C}_i(k_\theta,k_\omega) &= M_Q(\theta,\omega) * M_i(\theta,\omega) \\ &= \sum_{\theta_j}\sum_{\omega_m} M_Q(\theta_j+k_\theta,\omega_m+k_\omega)M_i(\theta_j,\omega_m)\end{aligned} \tag{11}$$

where $\mathfrak{C}_i(k_\theta,k_\omega)$ is the resultant 2-D correlation map, k_θ and k_ω are the axes of the correlation map, and $*$ is the 2-D cross-correlation operation between two images.

Since TING is translation invariant by Lemma 2, the ω-axis does not make any difference in 2-D cross-correlation calculation. Therefore, 2-D cross correlation between two TINGs can be reduced to 1-D cross correlation along the θ-axis, which is derived by substituting $k_\omega = 0$ into (11) as follows:

$$\begin{aligned}\mathfrak{C}_i(k_\theta) &= M_Q(\theta,\omega) \star M_i(\theta,\omega) \\ &= \sum_{\theta_j}\sum_{\omega_m} M_Q(\theta_j+k_\theta,\omega_m)M_i(\theta_j,\omega_m) \\ &= \sum_{\theta_j} M_Q(\theta_j+k_\theta)M_i(\theta_j)^T\end{aligned} \tag{12}$$

where $\mathfrak{C}_i(k_\theta)$ is the resultant 1-D correlation map, k_θ is the axis of $\mathfrak{C}_i(k_\theta)$, and $\star$ is the 1-D circular cross correlation that we derive from the 2-D one. Then, we take the max pooling on the correlation map $\mathfrak{C}_i$ of each TING pair (M_Q, M_i), comprising

the final vector representation named RING $N_Q \in \mathbb{R}^{|\mathfrak{M}|}$:

$$N_{Q,i} = \max_{k_\theta} \mathfrak{C}_i(k_\theta) \tag{13}$$

where $N_{Q,i}$ is the ith element of N_Q. Then, we arrive at the first theorem.

Theorem 1: RING is roto-translation invariant.

Proof of Theorem 1: Applying the rotation α and translation d to P_Q, we denote the resultant SG as R'_Q as (9):

$$R'_Q(\theta, \tau) = R_Q(\theta + \alpha, \tau - r_{\theta+\alpha} \cdot d). \tag{14}$$

Based on Lemma 2, the TING M'_Q of transformed SG R'_Q is

$$\begin{aligned} M'_Q(\theta_j, \omega) &= |\mathcal{F}(R'_Q(\theta_j, \tau))| \\ &= |\mathcal{F}(R_Q(\theta_j + \alpha, \tau - r_{\theta_j+\alpha} \cdot d))| \\ &= M_Q(\theta_j + \alpha, \omega). \end{aligned} \tag{15}$$

The correlation map between M'_Q and M_i is formulated as

$$\begin{aligned} \mathfrak{C}'_i(k_\theta) &= M'_Q(\theta, \omega) \star M_i(\theta, \omega) \\ &= M_Q(\theta + \alpha, \omega) \star M_i(\theta, \omega) \\ &= \mathfrak{C}_i(k_\theta + \alpha). \end{aligned} \tag{16}$$

Based on (16), the invariance of maximum value with respect to the shift leads to the equality between RING N_Q and RING N'_Q derived as

$$N_{Q,i} = \max_{k_\theta} \mathfrak{C}_i(k_\theta) = \max_{k_\theta} \mathfrak{C}_i(k_\theta + \alpha) = N'_{Q,i}. \tag{17}$$

Thus, RING is roto-translation invariant. ■

Now, we summarize the representations in forward representation pass in brief: the rotation-equivariant SG, the rotation-equivariant and translation-invariant TING, and the roto-translation-invariant RING. RING of a scan stays the same even if large viewpoint changes are present.

B. Solving Pass

1) Place Recognition: Put Theorem 1 in a real scenario. Due to the finite scan range occlusion, the invariance of RING cannot be guaranteed when translation is significant with respect to the scan range occlusion. Therefore, we have RING that is invariant given an arbitrary rotation change, but gradually degenerated with respect to larger translation changes. Following this result, we can set the range of a place by measuring the change of RING. On the other hand, the density of the place for successful global localization is able to reflect the robustness against the difference between the trajectory in the query and mapping session.

Based on the difference between RINGs of query and map scans, we can retrieve the minimum one as the place n where the query scan lies, arriving at place recognition:

$$n = \arg\min_i \|N_Q - N_i\|. \tag{18}$$

In (18), the computation complexity of one RING N is $O(|\mathfrak{M}|)$, proportional to the number of map scans. To calculate all map RINGs N_i, the total computation complexity is $O(|\mathfrak{M}|^2)$, which may not be affordable for onboard processing in a large-scale environment.

To reduce the computation, we introduce an approximation to (18) by checking the maximum element in N_Q:

$$n = \arg\max_i \tilde{N}_{Q,i} \tag{19}$$

where $\tilde{N}$ is the normalized RING calculated from normalized TING $\tilde{M}$ with zero mean and unit variance. The aim of normalization is to eliminate the effect from the finite scan range, e.g., number of valid scan points. With the help of normalization, we only need to calculate the current query RING N_Q instead of all map RINGs for place recognition. Consequently, the computation complexity is reduced from $O(|\mathfrak{M}|^2)$ to $O(|\mathfrak{M}|)$. In this way, we avoid the computation and storage of RINGs for all map scans. The equivalence between (18) and (19) is proved in Appendix A of the supplementary material in [44]. We show that the prerequisite for (19) is practical in real applications and, hence, is employed in the experiments.

2) Pose Estimation: After place recognition, we estimate the relative pose T_{nQ} between P_Q and P_n. Denote f_n and M_n as the BEV and TING of the retrieved map scan P_n, respectively. Given the local planar ground surface [45], the relative pitch, roll, and height between the query and map pose should be small. Therefore, we first estimate a reduced three-DoF relative pose globally without initial value, comprising one-DoF rotation α and two-DoF planar translation d, and then estimate the six-DoF relative pose with the resultant three-DoF relative pose as an initial value. To address the first step, we leverage TING to estimate rotation α, which is then utilized to compensate BEV for the estimation of translation d. For the second step, with the pose above as the initial value, we refine the metric pose using iterative closest point (ICP) [46].

a) Rotation estimation: Rotation estimation is achieved based on the TING. In (13), the maximum value occurs when M_Q is equal to M_n. Given correct place recognition, as TING is translation invariant according to Lemma 2, M_Q and M_n should be related by a relative rotation, and the equality is achieved when the shift k_θ equals the real relative rotation.

Therefore, we regard the shift achieving the maximum value along the θ-axis as the estimation of relative rotation between P_Q and P_n, denoted as $\hat{\alpha}$:

$$\hat{\alpha} = \arg\max_{k_\theta} \tilde{\mathfrak{C}}_n(k_\theta) \tag{20}$$

where $\tilde{\mathfrak{C}}$ is the normalized correlation map calculated from the normalized TING $\tilde{M}$. Normalization is employed to relieve the effect of finite scan range and discretization in practice. It is important that the estimator be globally convergent without dependence on any initial values. It is actually an exhaustive search, thus keeping the optimality. With the aid of fast Fourier transform (FFT) and parallel computing, this exhaustive process is fast. Refer to Appendix B of the supplementary material in [44] for derivation.

b) Translation estimation: Based on the rotation estimation, we can rotate the BEV $f(x, y)$ by $-\hat{\alpha}$ to eliminate the relative rotation between $f_Q(x, y)$ and $f_n(x, y)$, yielding a compensated BEV $f'_n(x, y)$. With the rotation variance eliminated,

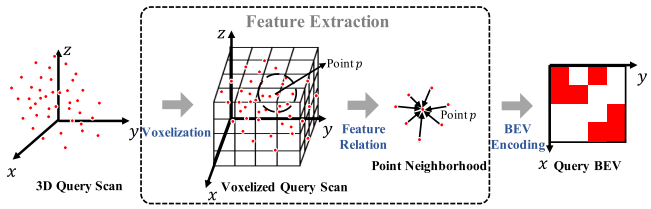


Fig. 4. Feature extraction module achieved by BEV representation in our method. In the representation pass, we first voxelize a raw 3-D scan into voxels and extract the features of every point from its nearest neighbors, yielding the final BEV by accumulating the features along the dimension of height.

2-D cross correlation can be applied to solve the translation

$$\mathfrak{C}_n(k_x, k_y) = f_Q(x, y) * f'_n(x, y) \tag{21}$$

where $\mathfrak{C}(k_x, k_y)$ is the correlation map of two BEVs. k_x and k_y are the axes of the correlation map. As for rotation, the maximum correlation in (21) should peak at the real relative translation. Denoting the estimated translation in correlation map as $\hat{d}$, we have

$$\hat{d} = \arg\max_{k_x, k_y} \mathfrak{C}_n(k_x, k_y). \tag{22}$$

Benefiting from the cross correlation, the translation estimation is globally convergent.

c) Refinement: With the rotation $\hat{\alpha}$ and translation $\hat{d}$ estimated above, an initial value is built by setting the relative pitch, roll, and height with zeros. Theoretically, the estimation accuracy is up to the resolution of the BEV and RT, thus keeping the convergence of the local refinement algorithm with high probability. In this article, ICP is employed for refinement. Finally, we arrive at the six-DoF pose $\hat{T}_{nQ}$, which is further applied to T_n as the query pose estimation $\hat{T}_Q$.

C. Perspective of Feature Aggregation

A typical pipeline [37] for representation in place recognition consists of feature extraction and feature aggregation. In feature extraction, local features of sparse keypoints [33], [34], [35], [47] or dense points [10] are extracted, which is fed to the feature aggregation module to build a global feature for the scan, say global max pooling, GeM pooling [41], or VLAD [37], [48], etc.

Our method also fits the pipeline; the representation pass illustrated in Fig. 2 consists of both feature extraction and feature aggregation. Specifically, the feature extraction module is achieved by BEV representation from a raw scan, as illustrated in Fig. 4, and the feature aggregation module is achieved by RT and FFT operations. The advantage of the proposed feature aggregation is the representation property, which is roto-translation invariant for place recognition, and equivariant for pose estimation. As a general feature aggregation $g(\cdot)$, the feature extraction part can be switched to others, which is discussed in the next section.

V. Multichannel Framework

An intuitive improvement is to replace the simple binary BEV feature with one that takes more structured information in the height dimension. Naturally, we investigate the question: what are the requirements for feature extraction that is able to preserve the equivariance/invariance of SG, TING, and RING?

A. Feature Extraction

We begin with the second theorem to present sufficient conditions for feature extraction.

Theorem 2: Let $E(\cdot)$ be a feature extractor, and $g(\cdot)$ be a roto-translation-invariant feature aggregation. Then, the representation produced by $g(E(\cdot))$ preserves both translation and rotation invariance if both of the following requirements are satisfied.

1) The translation of $E(\cdot)$ is either equivariant or invariant.
2) The rotation of $E(\cdot)$ is either equivariant or invariant.

Proof of Theorem 2: Let P be a point cloud as the initial input of $E(\cdot)$. Utilizing the feature aggregation method $g(\cdot)$ to the features exploited by the feature extractor $E(\cdot)$, we can obtain the aggregated result $g(E(P))$. Since $g(\cdot)$ is invariant to both translation and rotation, $g(E(P)) = g(E_d(P))$ by Definition 3 and $g(E(P)) = g(E_\alpha(P))$ by Definition 4. Theorem 2 includes a total of four cases, so we divide the proof into four subproofs corresponding to the four cases. We only demonstrate one case here; the other three cases can be found in Appendix C of the supplementary material in [44].

Case 1: If $E(\cdot)$ is equivariant to both translation and rotation, then $E_d(P) = E(P_d)$ by Definition 1 and $E_\alpha(P) = E(P_\alpha)$ by Definition 2.

Combining $E_d(P) = E(P_d)$ and $g(E(P)) = g(E_d(P))$, we have

$$g(E(P)) = g(E_d(P)) = g(E(P_d)). \tag{23}$$

Combining $E_\alpha(P) = E(P_\alpha)$ and $g(E(P)) = g(E_\alpha(P))$, we have

$$g(E(P)) = g(E_\alpha(P)) = g(E(P_\alpha)). \tag{24}$$

Referring to Definitions 3 and 4, we can conclude that the final representation $g(E(P))$ is translation invariant and rotation invariant. ■

As stated in Theorem 1, the proposed aggregation method $g(\cdot)$ possesses translation and rotation invariance, which satisfies the requirement of an aggregation method in Theorem 2. Thus, we build rotation- and translation-equivariant features to preserve the property of the aggregation. We select six rotation- and translation-equivariant features from previous literature [15] shown in Table I. Note that popular features such as Harris3D [47], SIFT3D [49], FPFH [34], and SHOT [18] can also fit in the framework, some of which are leveraged for ablation study in Section VII-E. Specifically, the local feature is built in two steps. First, we downsample the original scan with a given voxel size, i.e., 0.1 m, and for each point p in the scan, we extract features from its 30 nearest neighbors, as shown in Table I. Then, we accumulate the point features along the dimension of height. Specifically, for all point features belonging to the same BEV grid, we use the channelwise max pooling, resulting in a six-channel BEV, $f(x, y) \in \mathbb{R}^6$. Denote each channel of $f(x, y)$ as $f_c(x, y) \in \mathbb{R}$, where $c \in \{1, 2, 3, 4, 5, 6\}$ indicates the channel dimension. Compared with the occupancy, this feature

TABLE I
EXTRACTED FEATURES OF RING++

Feature	Formulation
Change of curvature	$\frac{\lambda_3}{\sum_{j=1}^{3}\lambda_j}$
Omnivariance	$\frac{\sqrt[3]{\prod_{3}^{j=1}\lambda_j}}{\sum_{j=1}^{3}\lambda_j}$
Eigenvalue entropy	$-\sum_{j=1}^{3}(\lambda_j \ln \lambda_j)$
2-D linearity	$\frac{\lambda_{2D,2}}{\lambda_{2D,1}}$
Maximum height difference	$Z_{\max} - Z_{\min}$
Height variance	$\sum_{k=1}^{30} \frac{(Z_k - \frac{\sum_{k=1}^{30} Z_k}{30})^2}{30}$

* $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ represents the ordered eigenvalues of the symmetric positive-definite covariance matrix of the neighborhood.

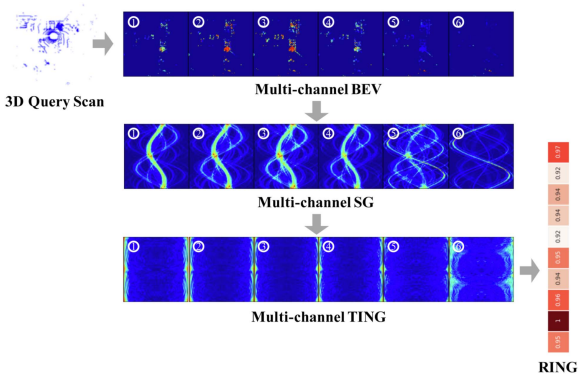


Fig. 5. Visualization of the multichannel representation pass. With six extracted local features, the corresponding BEV, SG, and TING are six-channel.

better encodes the information in the height dimension such as the building facade.

After feature extraction, we embark on feature aggregation, following the same pipeline demonstrated in Section IV to represent. One remaining problem is to aggregate the multichannel feature into one for solving.

B. Multichannel Representation Pass

For clear demonstration, we visualize the multichannel representation pass in Fig. 5. Based on the six-channel BEV $f(x, y)$, we apply RT to each BEV channel $f_c(x, y)$, yielding an SG channel $R_{f_c}(\theta, \tau)$. By concatenating all $R_{f_c}(\theta, \tau)$ along channel dimension c, we can obtain a six-channel SG $R_f(\theta, \tau)$.

After that, we employ row-wise DFT to $R_{f_c}(\theta, \tau)$ and take the magnitude spectrum of the result, denoted as $M_{f_c}(\theta, \omega)$, to achieve translation invariance. In the same manner, we concatenate all $M_{f_c}(\theta, \omega)$ along c dimension, yielding a six-channel TING $M_f(\theta, \omega)$.

Next, we implement 1-D circular cross correlation between query TING $M_Q(\theta, \omega)$ and all map TINGs $M_i(\theta, \omega)$. For each channel TING pair $(M_{Q_c}(\theta, \omega), M_{i_c}(\theta, \omega))$, we can get a 1-D correlation map $\mathfrak{C}_{i_c}(k_\theta)$:

$$\begin{aligned}\mathfrak{C}_{i_c}(k_\theta) &= M_{Q_c}(\theta, \omega) \star M_{i_c}(\theta, \omega) \\ &= \sum_{\theta_j} M_{Q_c}(\theta_j + k_\theta) M_{i_c}(\theta_j)^T. \end{aligned} \tag{25}$$

Then, we sum all $\mathfrak{C}_{i_c}(k_\theta)$ along c dimension, generating a single-channel 1-D correlation map $\mathfrak{C}_i(k_\theta)$

$$\mathfrak{C}_i(k_\theta) = \sum_{c=1}^{6} \mathfrak{C}_{i_c}(k_\theta). \tag{26}$$

Finally, we take the maximum value of $\mathfrak{C}_i(k_\theta)$ as the ith element of the final representation RING N_Q, which is the same as (13).

C. Multichannel Solving Pass

In the solving pass, we perform place recognition with roto-translation-invariant global descriptor RING N_Q first. Utilizing the same operation in (19), we retrieve the closest map scan P_n from $\mathfrak{M}$ by finding the index of the maximum element in $\tilde{N}_Q$.

Then, we carry out pose estimation including three steps: rotation estimation, translation estimation, and refinement. Taking advantage of the summed correlation map $\mathfrak{C}_n(k_\theta)$ between the query TING M_Q and the retrieved map scan TING M_n, the relative rotation can be easily estimated by (20), which is the optimal shift when M_Q best aligns with M_n. To solve the relative translation d, we begin with compensating the BEV $f_Q(x, y)$ via rotating $f_n(x, y)$ by $-\hat{\alpha}$. Under the multichannel framework, we first calculate the correlation map of each channel $\mathfrak{C}_{n_c}(k_x, k_y)$ using (21) and then sum the correlation maps along the channel dimension to obtain the globally optimal translation $\hat{d}$

$$\hat{d} = \arg\max_{(k_x, k_y)} \sum_{c=1}^{6} \mathfrak{C}_{n_c}(k_x, k_y). \tag{27}$$

The refinement for the multichannel framework is exactly the same as the single-channel one. The estimated three-DoF relative transformation between P_Q and P_n is utilized as the initial value for ICP to acquire the refined six-DoF pose $\hat{T}_Q$.

VI. DATASET AND EVALUATION CRITERIA

In this section, we describe the datasets and evaluation criteria that we employ for performance validation.

A. Dataset

We perform our RING++ method on three widely used datasets for place recognition evaluation: NCLT [50], MulRan [51], and Oxford [52] datasets. In order to validate the translation and rotation invariance of our approach, we select several trajectories with large translation and rotation changes. Furthermore, we utilize different sequences for multisession place recognition evaluation. The characteristics of these sequences are detailed in the subsections.

1) NCLT Dataset: It is a large-scale and long-term dataset collected by a Segway robot on the University of Michigan's North Campus. It contains 27 different sessions from January 8, 2012 to April 5, 2013 biweekly. Covering the same trajectory over 15 months, the dataset includes a large variety of environmental changes: dynamic objects like moving people, seasonal changes like winter and summer, and structural changes like the construction of buildings.

2) MulRan Dataset: It is a multimodal range dataset containing Radar and LiDAR data especially collected in the urban environment. It covers four different target environments: DCC, KAIST, Riverside, and Sejong City, providing both temporal and structural diversity for place recognition research.

3) Oxford Radar RobotCar Dataset: It is an extension to the *Oxford RobotCar Dataset* [53] for autonomous driving research. The data comprise 32 traversals of a central Oxford route in January 2019. A variety of weather and lighting conditions are encompassed in this dataset. A pair of Velodyne HDL-32E 3-D LiDARs is mounted on the left and right sides of the vehicle to improve 3-D scene understanding performance. For the convenience of place recognition evaluation, we concatenate point clouds collected by these two LiDARs into one single scan.

B. Evaluation Metrics

1) Revisited Threshold: The revisited threshold is for place recognition, which determines whether the query scan and retrieved map scan are a "true" loop. Two scans are considered a true positive if the distance between them is less than the revisited threshold. Without specific instructions, the revisited threshold is set to half of the map sampling distance in our experiments.

2) Recall@1: We utilize *Recall@1* [54] metric to evaluate a place recognition system in terms of the number of true loop candidates. *Recall@1* is defined as the ratio of top 1 true positives to total positives, which is formulated as

$$Recall@1 = \frac{\text{TP}_{\text{top1}}}{\text{TP}_{\text{top1}} + \text{FN}_{\text{top1}}} \tag{28}$$

where TP_{top1} denotes the number of true positives with top 1 retrieval, and FN_{top1} denotes the number of false negatives with top 1 retrieval.

3) Precision–Recall Curve: For the place recognition task, *precision* [54] is defined as the ratio of true positives to total matches, and *recall* [54] is defined as the ratio of true positives to total positives. The mathematical expressions are

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{29}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{30}$$

where TP denotes the number of true positives, FP denotes the number of false positives, and FN denotes the number of false negatives. *Precision–recall curve* [55] is plotted under various thresholds. A point in the *precision–recall curve* depicts the *precision* and *recall* values corresponding to a specific threshold.

4) F1 Score–Recall Curve: F1 score [56] is the harmonic mean of precision and recall, combining precision and recall metrics into a single metric

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{31}$$

which is a suitable metric to balance *precision* and *recall*.

5) TE and RE: We calculate average translation error *TE* and rotation error *RE* following [40], [57] for pose estimation evaluation. Since our method only yields three-DoF poses, we evaluate two-DoF translation and one-DoF rotation error in the majority of evaluations. The mathematical formulas are

$$\text{TE} = \|\hat{d} - d^*\| \tag{32}$$

$$\text{RE} = |\hat{\alpha} - \alpha^*| \tag{33}$$

where $\hat{d}$ and $\hat{\alpha}$ are the estimated two-DoF translation and one-DoF rotation, respectively. d^* and α^* are the ground-truth translation and rotation, respectively. Pose error calculation of incorrectly matched scans (i.e., not from the same place) is meaningless. Thus, we only compute the estimated pose errors of successfully matched pairs to compare the pose estimation performance.

6) Global Localization Success Rate: To quantitatively evaluate the overall global localization performance, we leverage *Success Rate* [40], [57] metric and further extend it to involve both place recognition and pose estimation. Rather than the original *success rate* in [40] and [57], which only reflects the pose estimation performance, *global localization success rate* is a compound performance of place recognition and pose estimation. The mathematical formula is

$$\text{GL success rate} = \frac{\text{TP}_{TE<2\ m\ \&\ RE<5^\circ}}{\text{TP} + \text{FP}} \tag{34}$$

where $\text{TP}_{TE<2\ m\ \&\ RE<5^\circ}$ denotes the number of true positives whose translation error is below 2 m and whose rotation error is below 5°.

7) Absolute Trajectory Error: Absolute trajectory error (*ATE*) is used to evaluate the performance of SLAM systems. It directly measures the difference between points of the true and the estimated trajectory. The ATE of the ith frame is formulated as follows:

$$\text{ATE}_i = \|\text{trans}(Q_i^{-1} S P_i)\|, \tag{35}$$

where Q_i is the ground-truth pose and P_i is the estimated pose. $S \in SE(3)$ is the transformation matrix to align the two poses, which is calculated by the least square method. $\text{trans}(\cdot)$ represents the translation part of the pose difference. In this article, we use the average ATE to evaluate the performance of SLAM systems.

C. Comparative Methods

In terms of place recognition and pose estimation, we compare our approach against state-of-the-art learning-free and learning-based methods.

1) *M2DP* [8] projects the point cloud to multiple 2-D planes and leverages the singular vector of all signatures for points in each plane as the global descriptor to detect loop closure.
2) *Fast Histogram* [7] utilizes the histogram of range distance extracted from the 3-D point cloud as the global descriptor of the point cloud for loop closure detection.
3) *Scan Context* [3] encodes the 3-D scan to a representation called Scan Context, which contains 2.5-D information, compared to histogram-based methods. It is invariant to rotation changes via a two-phase search algorithm.
4) *LiDAR Iris* [4] utilizes several LoG-Gabor filter and threshold operations on the binary signature image. It

achieves rotation invariance by adopting Fourier transform on the LiDAR-Iris representation.

5) *Intensity Scan Context* [58] explores both geometry and intensity characteristics based on the Scan Context representation.
6) *Scan Context++* [11] introduces two subdescriptors to combine topological place recognition with one-DoF semimetric localization, so as to enhance the performance of Scan Context. Depending on the coordinate selection, the original article named the two subdescriptors Polar Context (PC) and Cart Context (CC). In this article, we use the abbreviations SC++ (PC) and SC++ (CC) to distinguish these two descriptors. The SC++ (PC) provides an estimation of the yaw angle, and the SC++ (CC) estimates the lateral translation.
7) *PointNetVLAD* [9] combines PointNet and NetVLAD to extract the global descriptor for a 3-D point cloud by end-to-end training.
8) *DiSCO* [38] encodes a 3-D scan to a scan context and then uses an encoder–decoder network to extract features. It constructs a rotation-invariant place descriptor by taking the magnitude of the frequency spectrum for place recognition and designs a correlation-based rotation estimator for one-DoF pose estimation.
9) *EgoNN* [40] designs a fully convolutional architecture to extract global and local descriptors. It uses the global descriptor for coarse place recognition and the local descriptors for six-DoF pose estimation.
10) *LCDNet* [59] introduces a relative pose head based on the unbalanced optimal transport theory and can simultaneously identify previously visited places and estimate the six DoFs relative pose.

D. Our Methods

In Sections IV and V, we solve the place recognition problem utilizing RING representation constructed from TING and estimate the relative rotation based on TING. To validate the translation invariance of RING generated by TING, we directly construct RING representation based on SG representation skipping the TING formation process for place recognition and then estimate the relative rotation and translation based on SG, which serves as a variant of our method. Taking feature extraction into account, we then have four versions of our method: RING (SG), RING, RING++ (SG), and RING++. By comparing the four versions, we can figure out the effects of translation invariance design and feature extraction on place recognition and pose estimation results.

1) *RING (SG)* encodes occupancy information of a point cloud to SG and RING representations. It utilizes RING representation for place recognition, SG representation for rotation estimation, and single-channel BEV representation for translation estimation.
2) *RING* encodes occupancy information of a point cloud to SG, TING, and RING representations. It leverages RING representation for place recognition, TING representation for rotation estimation, and single-channel BEV representation for translation estimation.
3) *RING++ (SG)* extracts multiple local features of a point cloud, following a multichannel framework. It utilizes RING representation for place recognition, SG representation for rotation estimation, and multichannel BEV representation for translation estimation.
4) *RING++* extracts multiple local features of a point cloud, following a multichannel framework. It leverages RING representation for place recognition, TING representation for rotation estimation, and multichannel BEV representation for translation estimation.

E. Implementation Details

For a fair comparison, we remove the ground plane of the raw point cloud and filter it to the same range $[-70\ \mathrm{m}, 70\ \mathrm{m}]$ for all the methods. For both Scan Context and our method, we set the grid size of BEV to 120×120, so the translation resolution is $140/120 = 1.17$ m/pixel and the rotation resolution is $360/120 = 3^\circ$/pixel. The number of candidates for Scan Context++ is 10. The parameters and settings of other compared methods are kept the same as those in the original papers. PointNetVLAD, DiSCO, and LCDNet are retrained on the MulRan dataset. The output dimension of these methods is set to 1024, 1024, and 256, respectively. The local descriptor dimension of PointNetVLAD and LCDNet is 1024 and 256. In the training step, places within 10 m are regarded as positive pairs, while negative pairs are at least 20 m apart. Other parameter configurations are similar to those in the original paper. For EgoNN and Product of Cross-Attention Matrices (PCAM), we use the publicly available pretrained model for better performance. We use ICP implementation from FastGICP [60] in the experiments with pose refinement. The parameters of ICP are set according to common practice: *max correspondence distance* is 3 m and *max iteration* is 64.

VII. Experimental Evaluation

In this section, we design some experiments to verify that the proposed approach:

1) has strong translation invariance and rotation invariance that is independent of translational difference;
2) detects the loops successfully when the pose difference between query and map point clouds is large;
3) estimates a three-DoF pose as a qualified initial guess for further metric refinement (ICP alignment);
4) is computation efficient with a compact representation for real-time applications;
5) is easily pluggable into SLAM systems for loop closure detection and relocalization.

A. Illustrative Toycase Study

We present two toycases to validate the effectiveness of our approach in terms of global localization. Specifically, we design

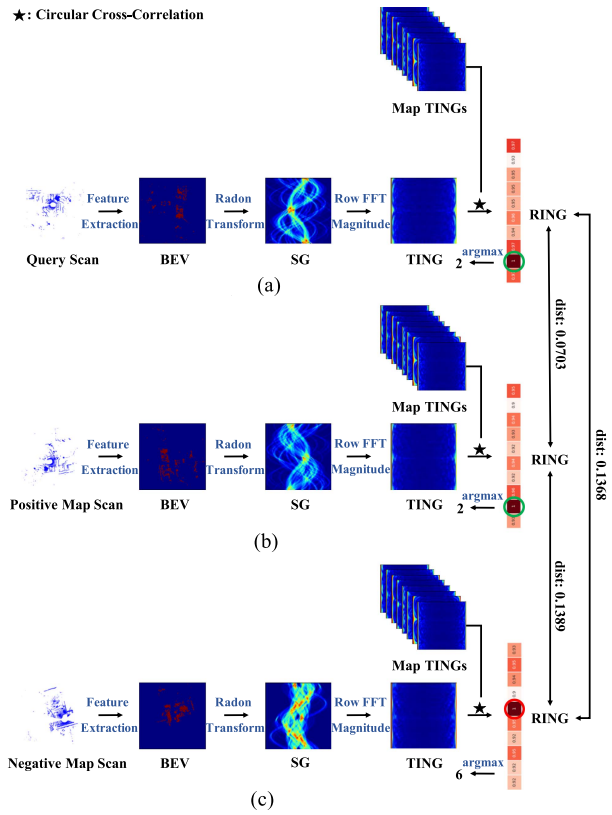


Fig. 6. Visualization of the representation pass for query scan, positive map scan, and negative map scan, to verify the feasibility of utilizing (19) to equivalently replace (18) for place recognition. (a) Query scan representation pass. The maximum value of RING representation points to the second scan in the scan map, which is the positive map scan shown in (b). (b) Positive map scan representation pass. The RING representation of the positive map scan is the closest to the RING representation of the query scan in terms of Euclidean distance. (c) Negative map scan representation pass. The Euclidean distance between query RING and negative map RING is nearly twice of that between query RING and positive map RING.

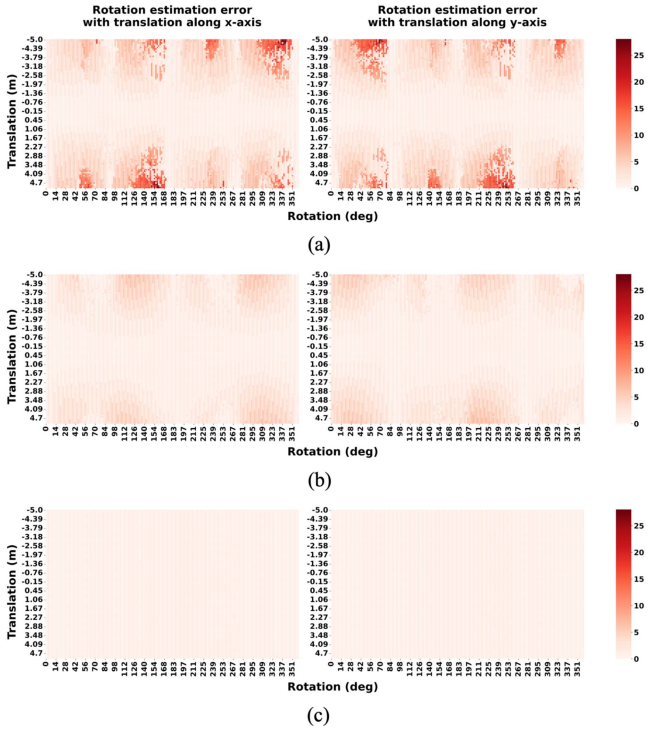


Fig. 7. Rotation estimation error using different representations. The left column shows the rotation estimation error with translation along the x-axis, and the right column shows the rotation estimation error with translation along the y-axis. (a) SC. (b) SG. (c) TING.

a toycase to verify the roto-translation invariance of our representation RING for place recognition. Moreover, we investigate the impact of translation motion on relative rotation estimation utilizing various representations (SC (Scan Context), SG, and TING), in order to highlight the advantage of our representation for pose estimation.

1) Place Recognition Illustration: We begin with a toycase study to illustrate the feasibility of our method for place recognition. In this toycase, we perform place retrieval in a scan map with ten map scans for simplification. To better illustrate the proposed method, we visualize all the intermediate representations, as shown in Fig. 6. The top block (a) of Fig. 6 shows that the query scan is located at the same place where the second map scan lies according to (19). The retrieved map scan, i.e., the second map scan, follows the representation pass, which is depicted in the bottom block (b) of Fig. 6. Furthermore, the Euclidean distance between query RING and positive map RING is the smallest, which satisfies (18), and is also consistent with the retrieved result above by (19).

2) Pose Estimation Illustration: Similar to Scan Context series [3], [11], we leverage rotation-equivariant representations SG and TING to estimate the relative rotation, referring to rotation equivariance illustrated in Lemmas 1 and 2. To distinguish our representation from others, we rotate the point cloud within $[0^\circ, 360^\circ)$ and translate the point cloud within $[-5\ \text{m}, 5\ \text{m}]$. After that, we employ the three aforementioned representations to estimate the relative rotation between the original point cloud and the transformed point cloud. The resolution of all representations for rotation estimation is the same, 3°/pixel. As we can easily see in Fig. 7, SC representation is strongly affected by translation differences. When the relative translation is beyond 1 m, the rotation estimation error rises to about 30°. In contrast, our representations SG and TING are much more robust to translation perturbation, showing smaller rotation estimation errors. In addition, TING equipped with translation invariance is almost not influenced by translation disturbance, whose rotation estimation error keeps nearly 0° for all translations within 5 m.

B. Evaluation of Place Recognition

For comprehensive place recognition evaluation, we evaluate the proposed method regarding both online loop closure detection (single-session scenarios) and long-term localization (multisession scenarios).

1) Single-Session Scenarios: Under single-session scenarios, we select "2012-02-04" sequence in NCLT dataset, "DCC01" sequence in the MulRan dataset, and "2019-01-11-13-24-51" sequence in the Oxford dataset for place recognition evaluation. To verify the advantage of our approach in the sparse scan map, we perform all the methods to detect loop

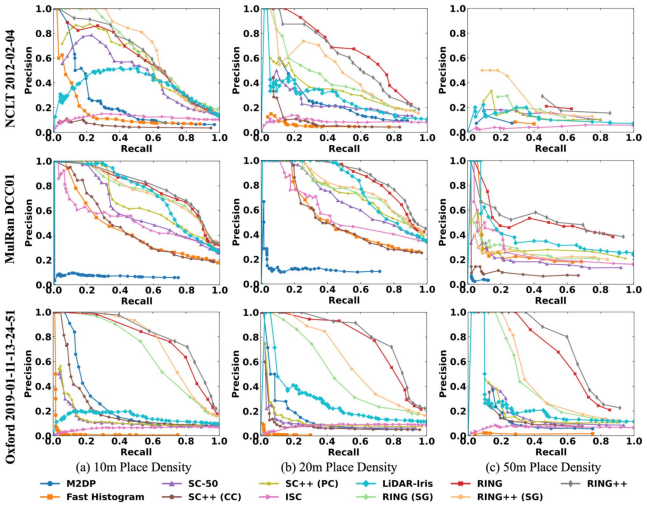


Fig. 8. Precision–recall curve for NCLT, MulRan, and Oxford datasets under single-session scenarios.

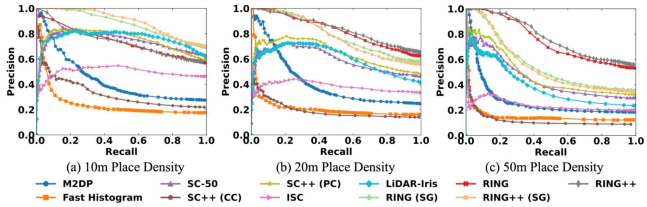


Fig. 9. Precision–recall curve for "2012-03-17" query sequence to "2012-02-04" map sequence in the NCLT dataset.

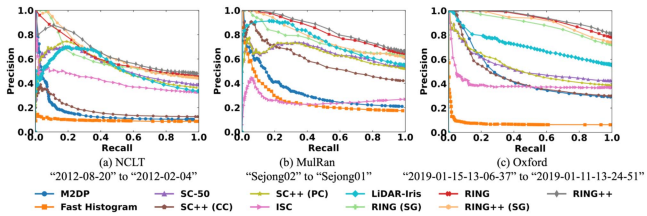


Fig. 10. Precision–recall curve for NCLT, MulRan, and Oxford datasets at 20-m place density.

closure online at different place densities (10, 20, and 50 m). *Precision–recall curve* is utilized as the evaluation metrics of single-session scenarios, as depicted in Fig. 8. As we can see, our RING-based methods are less affected by place density than others. At 10-m place density, RING-based methods show competitive performance with SC-based methods. However, with the decrease in the place density, the translation between the query scan and the retrieved map scan enlarges, so our methods outperform them by an increasing margin. Compared with RING (SG) and RING++ (SG), RING and RING++ remain high-performance even at 50-m place density, which validates the strong translation invariance of RING representation based on TING. The comparison result is consistent with the toycase illustrated in Fig. 7, validating the robustness to the translation of our methods. Compared with RING (SG) and RING, respectively, RING++ (SG) and RING++ achieve better performance, which is more obvious at lower place density. It demonstrates that the extracted local features are beneficial for place representation, thereby improving discrimination. In terms of the Fast Histogram, the histogram aggregation is invariant to the rotation difference in the dense place representation. However, without the design of translation invariance, the performance on sparse place representation is limited. In terms of M2DP, the projection strategy improve discrimination but also cannot handle the scenario in which point cloud centerings are not well aligned.

2) Multisession Scenarios: From the viewpoint of long-term autonomy, a robust place recognition system should work well when the surroundings change as time passes. To compare long-term place recognition performance, we select several intersessions covering the same region to serve as map sequence and query sequence, respectively. Under multisession scenarios, the sampling interval of query data is fixed to 5 m for all datasets. In the same manner, we compare the proposed approach against the state-of-the-art methods at different map place densities (10, 20, and 50 m) on the NCLT dataset, with the results depicted in Fig. 9. *Precision–recall curve* shows the same trend as that in Fig. 8. RING (SG), RING++ (SG), and SC-based methods show competitive performance at high place density (10 m), while RING and RING++ achieve better performance at low place density (20 and 50 m) due to rigorous roto-translation invariance design. Using six extracted local features, RING++ (SG) and RING++ perform better than RING (SG) and RING, proving the validity of feature extraction for place recognition. To validate the cross-dataset consistency, we evaluate all the methods on different datasets at 20-m place density. Besides "2012-03-17" query sequence to "2012-02-04" map sequence for the NCLT dataset, we perform long-term place recognition on "2012-08-20" to "2012-02-04" pair for the NCLT dataset, "Sejong02" to "Sejong01" pair for the MulRan dataset, and "2019-01-15-13-06-37" to "2019-01-11-13-24-51" pair for the Oxford dataset, as shown in Fig. 10. Via comparison, four versions of our approach all outperform other approaches at 20-m place density, especially for the Oxford dataset, which presents the strength of roto-translation invariance for place recognition. In Fig. 11, we visualize the revisited pairs of *top1 retrieval* using different methods on the NCLT dataset at 20-m place density. It can be easily seen that RING++ surpasses other methods by a lot, further validating the effectiveness of our approach in sparse places.

In addition to qualitative comparison, we provide the quantitative results of the handcrafted methods for long-term place recognition, as listed in Table II. Compared with other handcrafted methods, our method and its variants are capable of recognizing many more revisited places. Compared with RING (SG) and RING++ (SG), RING and RING++ make obvious improvements thanks to the translation invariance construction of RING representation based on TING.

3) Generalization: Unlike deep learning representations, our representation is training-free, so we can easily generalize to new scenes without retraining or fine-tuning. We compare RING++ with some deep learning methods across different datasets to show our superior generalization ability among deep learning

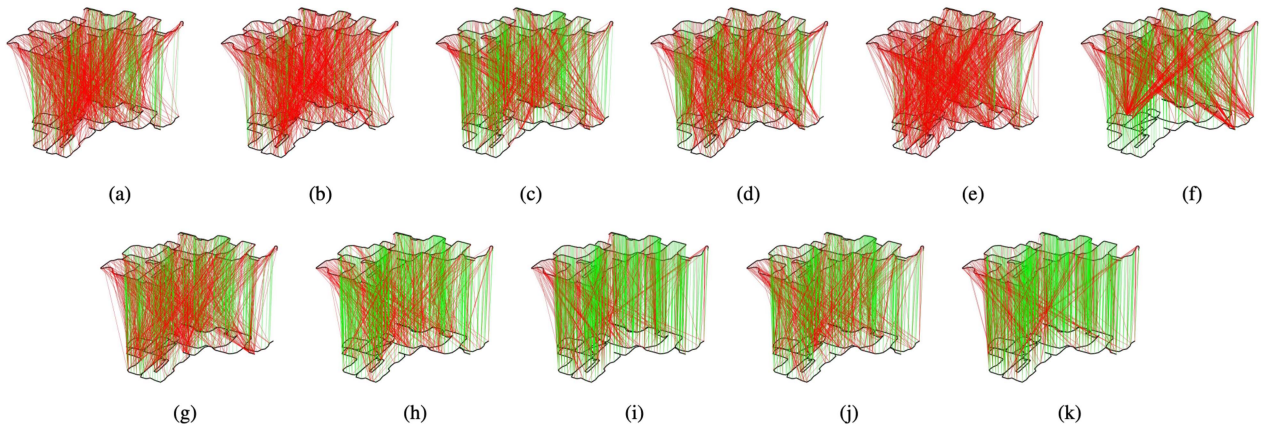


Fig. 11. Qualitative comparison of match graph for the NCLT dataset at 20-m place density, where black lines visualize trajectories, green lines visualize correctly retrieved pairs, and red lines visualize mistakenly retrieved pairs. (a) M2DP. (b) Fast Histogram. (c) SC. (d) SC++ (PC). (e) SC++ (CC). (f) ISC. (g) LiDAR-Iris. (h) RING (SG). (i) RING. (j) RING++ (SG). (k) RING++.

TABLE II
QUANTITATIVE RESULTS OF PLACE RECOGNITION EVALUATION IN MULTISESSION SCENARIOS

Approach		NCLT			MulRan*			Oxford		
		Recall@1	F1 score	AUC	Recall@1	F1 score	AUC	Recall@1	F1 score	AUC
Handcrafted	M2DP	0.2811	0.3966	0.4127	0.2149	0.3456	0.3309	0.2997	0.4496	0.4638
	Fast Histogram	0.1840	0.2785	0.2136	0.1805	0.2986	0.2614	0.0642	0.1181	0.0741
	SC-50	0.5248	0.6314	0.5934	0.5572	0.6981	0.6490	0.4375	0.5935	0.5366
	SC++ (PC)	0.3862	0.6495	0.6465	0.5384	0.6826	0.6396	0.3993	0.5561	0.5295
	SC++ (CC)	0.1355	0.2449	0.1895	0.4370	0.5922	0.5838	0.3084	0.4596	0.4540
	ISC	0.4540	0.5705	0.4735	0.2865	0.2646	0.1255	0.3654	0.5380	0.3969
	LiDAR-Iris	0.4732	0.5876	0.6055	0.5888	0.7005	0.7058	0.5653	0.7145	0.7141
	RING (SG) (Ours)	0.6582	0.7274	0.7392	0.6361	0.7642	0.7474	0.7425	0.8366	0.8787
	RING (Ours)	0.7098	0.7658	0.8246	0.6633	0.7841	0.8283	0.8013	0.8772	0.9234
	RING++ (SG) (Ours)	0.6309	0.7150	0.7776	0.6361	0.7642	0.7474	0.7612	0.8492	0.8992
	RING++ (Ours)	**0.7321**	**0.7890**	**0.8374**	0.6941	0.7983	0.8439	**0.8274**	**0.8937**	**0.9438**
Learning-based	PointNetVLAD	0.3691	0.5391	0.6260	0.6968	0.8214	0.8436	0.5247	0.7825	0.6114
	DiSCO	0.6036	0.7816	0.7722	**0.7425**	**0.8630**	0.8973	0.6562	0.7890	0.8566
	EgoNN	0.5620	0.6965	0.7988	0.7260	0.8412	**0.9060**	0.5920	0.7438	0.8471
	LCDNet	0.6199	0.7659	0.8081	0.3978	0.5692	0.6115	0.5307	0.6934	0.7797

* In the MulRan dataset, we perform place recognition evaluation on the test region used in EgoNN [40] for all methods since the compared learning-based approaches are trained on the training region of the MulRan dataset.

methods. For deep learning methods, we train the model on the MulRan dataset ("Sejong01" trajectory serves as map sequence, "Sejong02" trajectory serves as query sequence) and validate it on the NCLT dataset ("2012-02-04" trajectory serves as map sequence, "2012-03-17" trajectory serves as query sequence) and the Oxford dataset ("2019-01-11-13-24-51" trajectory serves as map sequence, "2019-01-15-13-06-37" trajectory serves as query sequence) for generalization ability evaluation, where the equidistant sampling gap between query data is 5 m and that of map data is 20 m.

As shown in Table II, RING++ has excellent performance across different datasets. DiSCO and EgoNN maintain most of their place recognition performance on generalized datasets due to their rotation-invariant structures, but the polar transform still suffers from translation variance. LCDNet performs better in NCLT and Oxford datasets with no occlusion. The performance drops drastically in the MulRan dataset, which may be caused by the occlusion.

C. Evaluation of Pose Estimation

To eliminate the effect of place recognition, we solely evaluate relative pose estimation between positive scans determined by the ground truth. At the pose estimation stage, our method yields a three-DoF pose: one-DoF rotation and two-DoF translation. Scan Context, ISC, LiDAR-Iris, and Scan Context++ only provide one-DoF yaw estimation. As M2DP and Fast Histogram do not provide pose estimation, we do not include them in the experiments. Moreover, we compare our methods with handcrafted point cloud registration: ICP [61], RANSAC with

TABLE III
POSE ESTIMATION ERROR ON THE MULRAN DATASET*

Approach	Success (%)	TE† (m)	RE† (°)
RING (SG)	53.01	0.58/1.08/4.52	0.51/1.77/5.94
RING	63.47	0.51/**0.70**/1.81	0.35/0.73/1.52
RING++ (SG)	53.01	0.60/1.17/4.54	0.62/2.54/6.24
RING++	**65.76**	**0.50/0.70/1.79**	**0.34/0.72/1.51**

† TE here represents a two-DoF translation error and RE represents a one-DoF yaw error.
* We list 50%, 75% and 95% quantiles of TE and RE in this table.

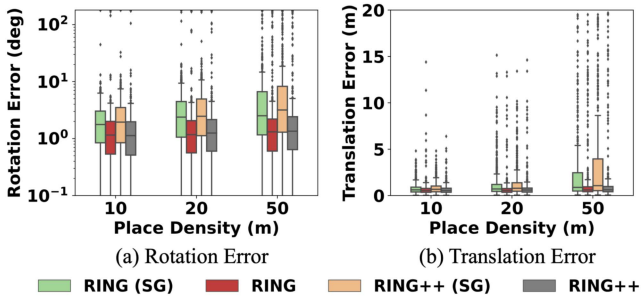


Fig. 12. Pose estimation error on the NCLT dataset.

FPFH features [34] and TEASER++ [62], and the learning-based methods: PCAM [63] and OverlapNet [27] with one-DoF yaw estimation.

We perform pose estimation on NCLT and MulRan datasets, which provide accurate ground-truth poses. As for the Oxford dataset, we observe that the ground-truth poses between different traversals of the same trajectory are slightly misaligned, so we discard it in evaluations involving pose estimation. The positive pairs are determined by the ground-truth poses and from the same trajectories used in place recognition experiments. We evaluate these positive pairs in terms of three-DoF pose (TE $<$ 2 m and RE $< 5^\circ$), two-DoF translation (TE < 2 m), and one-DoF rotation (RE $< 5^\circ$) success rate.

The pose estimation results are shown in Table IV. In our experiment setup with a 5-m query and 20-m map sampling distance, we observe that ICP achieves the best performance among previous handcrafted methods and EgoNN demonstrates superior results compared to existing learning-based approaches on the MulRan dataset. However, these methods are not robust to registration under large viewpoint changes and show a large performance drop when evaluated on the NCLT dataset. With the design of a roto-translation-invariant structure, our proposed RING++ achieves the highest success rate. Note that, rather than presenting a novel pose estimation algorithm, we focus on the reuse of the intermediate representation of place recognition for pose estimation.

D. Evaluation of Global Localization

Global localization is evaluated by the compound performance of place recognition and pose estimation. We first compare the proposed four variants. The pose estimation error on "2012-03-17" to "2012-02-04" for the NCLT dataset at different place densities is visualized by the box plot shown in Fig. 12. Together with three-DoF pose estimation on the NCLT dataset, we provide Table III, which presents *Success Rate*, *TE*, and *RE* of our approach on "Sejong02" to "Sejong01" for the MulRan dataset with a place density of 20 m. As the results show, RING and RING++ estimate more accurate rotation and translation than RING (SG) and RING++ (SG), which indicates that translation invariance benefits pose estimation. By comparing RING with RING++, we find that the extracted six local features of RING++ do not obviously facilitate pose estimation performance. In general, RING++ performs the best among all variants.

To assess the three-DoF pose estimated by our method, we employ ICP with and without the initial pose provided by the proposed method to align two matched scans after place recognition. The cases are illustrated in the supplementary material in [44]. As we can see, ICP fails without an initial guess, especially in the case of large rotation variance. Our approach offers a qualified initial guess for scan matching, close to the best alignment. With the estimated pose of RING++ (SG) and RING++, ICP converges to global minima rather than traps in local optima. Furthermore, the time cost of ICP with the initial guess is much less than without any initial poses.

After that, we compare RING++ with both handcrafted and learning-based methods in terms of global localization performance on NCLT and MulRan datasets. For each dataset, we evaluate these approaches in terms of *Success Rate*, *TE*, and *RE*. To demonstrate the effectiveness of better initial guesses, we also provide a pose estimation successful rate, which is calculated by the percentage of successfully aligned matches (TE $<$ 2 m and RE $< 5^\circ$) among all correctly retrieved pairs.

1) Comparison With Learning-Free Methods: Among the compared hand-engineering methods, M2DP and Fast Histogram are only capable of place recognition, while Scan Context, ISC, LiDAR-Iris, and Scan Context++ also produce one-DoF yaw or lateral translation estimation, leaving the remaining dimension to ICP. In contrast, our method yields a relative yaw that is a by-product in the process of place recognition and estimates a two-DoF relative translation successively in the process of pose estimation. For a fair comparison, we evaluate the pose estimation performance of all methods after refinement by ICP. As can be seen in Table V, we calculate *Success Rate* for global localization evaluation and *TE* and *RE* at 50%, 75%, and 95% for pose estimation evaluation. Compared with other methods, RING++ has better performance on both the NCLT and MulRan datasets, arriving at high *Success Rate* for global localization. Referring to the pose estimation success rate, RING++ has a better chance of achieving satisfied alignment by providing the qualified three-DoF initial guesses, which makes refinement more easily converge to the global optimal pose. Our method achieves less *TE* and *RE* at most quantiles, which verifies the efficacy of better initial guesses.

2) Comparison With Learning-Based Methods: Learning-based methods, DiSCO, EgoNN, and LCDNet are trained on the MulRan dataset and tested on both MulRan and NCLT datasets for generalization evaluation. For the MulRan dataset, EgoNN

TABLE IV
COMPARISON OF RELATIVE POSE ERRORS (ROTATION AND TRANSLATION) BETWEEN POSITIVE PAIRS DETERMINED BY GROUND TRUTH

Approach		NCLT			MulRan		
		Pose Succ. (%)	Trans Succ. (%)	Rot Succ. (%)	Pose Succ. (%)	Trans Succ. (%)	Rot Succ. (%)
Handcrafted	ICP	21.33	33.06	26.09	68.91	68.91	**100.00**
	RANSAC	4.56	6.25	41.63	68.57	73.71	82.75
	TEASER++	2.62	5.65	23.59	41.86	46.45	58.55
	SC++(PC)*	-	-	44.59	-	-	74.50
	SC++(CC)*	-	24.67	-	-	70.20	-
	ISC*	-	-	26.09	-	-	96.70
	LiDAR-Iris*	-	-	63.20	-	-	98.14
Learning-based	DiSCO*	-	-	40.32	-	-	83.25
	OverlapNet*	-	-	3.73	-	-	39.64
	LCDNet	1.01	10.99	3.33	4.59	9.32	21.97
	Egonn	4.21	6.26	65.94	87.91	87.91	**100.00**
	PCAM	8.17	37.10	8.87	76.77	77.05	97.77
Ours	**RING++**	**86.96**	**88.98**	**88.98**	**91.26**	**91.26**	96.99

* These methods only estimate the yaw angle or one-DoF translation between two scans and are, thus, not directly comparable to the others.

TABLE V
QUANTITATIVE RESULTS OF GLOBAL LOCALIZATION EVALUATION*

Approach		NCLT			MulRan		
		GL/PE Succ. (%)	TE† (m)	RE† (°)	GL/PE Succ. (%)	TE† (m)	RE† (°)
Handcrafted	SC-50 + ICP	24.97/47.58	0.30/1.02/7.67	4.14/6.50/18.11	35.82/65.45	0.45/6.54/8.17	0.52/0.95/2.01
	SC++ (PC) + ICP	25.68/46.95	0.33/1.87/8.78	4.18/6.70/19.81	36.25/66.94	0.38/6.12/8.10	0.51/0.94/1.96
	SC++ (CC) + ICP	8.09/50.98	0.36/14.30/139.55	**3.67**/7.06/177.01	18.19/62.23	0.48/138.85/140.63	0.47/0.80/1.48
	ISC + ICP	24.77/54.57	0.28/0.56/7.69	4.14/**6.33**/13.17	25.21/88.00	0.31/0.65/6.62	**0.42**/**0.72**/1.33
	LiDAR-Iris + ICP	24.47/51.71	**0.27**/**0.48**/7.21	4.25/6.60/13.99	42.26/71.78	0.41/3.44/7.99	0.45/0.78/1.62
Learning-based	DiSCO + ICP	28.31/46.91	0.35/2.73/8.73	6.09/6.52/34.69	38.36/51.66	1.23/7.45/8.25	0.71/1.02/1.84
	LCDNet + ICP	2.12/3.41	5.67/8.33/12.93	8.61/38.01/168.94	19.75/49.65	2.10/7.37/8.84	0.78/1.49/11.91
	EgoNN + ICP	6.07/10.80	3.48/8.12/15.51	12.32/27.45/134.95	62.00/85.40	0.28/0.52/13.04	0.72/1.14/1.98
Ours	**RING++ + ICP**	**42.37/57.88**	0.31/0.53/**1.36**	4.13/6.74/**12.52**	**66.09/95.22**	**0.27**/**0.42**/**1.82**	0.51/0.80/**1.19**

† As ICP refinement is applied, TE here represents a 3-DoF translation error and RE represents a 3-DoF rotation error.
* We list 50%, 75% and 95% quantiles of TE and RE in this table. The *GL Success Rate* is the compound performance of place recognition and pose estimation. The PE Success Rate is the percentage of successfully aligned matches ($TE < 2m$ & $RE < 5°$) among all correctly retrieved pairs by different methods. For MulRan dataset, we perform place recognition and pose estimation evaluation on the test dataset which is used in EgoNN [40].

shows a great performance of global localization thanks to the rotation-invariant representation and extracted local features. RING++ presents competitive performance (66.09% *GL Success Rate*) to EgoNN (62.00% *GL Success Rate*). The overall global localization is mainly limited by the place recognition performance which we discuss in Section VII-H. LCDNet's global localization performance is limited, owing mostly to pose estimation that suffers from occlusion and poor initialization. In terms of global localization error, RING++ has lower *TE* and *RE* than compared methods. For the NCLT dataset, the distribution of points is different from that of the MulRan dataset due to the different LiDAR sensors, making it hard to generalize. Therefore, the global localization performance of learning-based methods on the NCLT dataset drops a lot, especially for EgoNN and LCDNet. The underlying reason may be explained by the lack of explicit invariance design, which makes the features correlated and fails to generalize to unseen datasets. Unlike EgoNN and LCDNet, DiSCO constructs a rotation-invariant place descriptor by transforming features to the frequency domain and then taking the magnitude of the frequency spectrum, which explains the relatively smaller decline of *GL Success Rate* from the MulRan dataset generalizing to the NCLT dataset. Through comparison, RING++, as a learning-free method, arrives at the top performance on the NCLT dataset, with a high *GL Success Rate* and small *TE* and *RE*.

E. Ablation Study

To investigate the influence of resolution on place recognition and pose estimation, we present ablation studies on the grid size and corresponding resolution of RING++. We carry out experiments utilizing the same multisession pair "2012-03-17" to "2012-02-04" in the NCLT dataset, with the results displayed in Table VI. As the grid size/resolution increases, *Recall@1* of our method increases fast at first and then approaches a constant gradually. In addition, translation estimation error

TABLE VI
ABLATION STUDY ON THE RESOLUTION*

Resolution	Recall@1 (%)	TE (m)	RE (°)	Time (ms)
40×40 (3.50 m, 9°)	62.18	1.43/2.08/3.73	2.45/4.14/7.27	34.4
60×60 (2.33 m, 6°)	68.76	0.98/1.37/2.52	1.86/3.19/5.96	41.6
80×80 (1.75 m, 4.5°)	71.59	0.79/1.10/1.95	1.53/2.70/5.32	45.6
120×120 (1.17 m, 3°)	73.21	0.56/0.77/1.31	1.25/2.14/4.07	53.3
160×160 (0.88 m, 2.25°)	75.83	0.48/0.70/1.30	1.08/1.98/3.59	86.7
200×200 (0.70 m, 1.8°)	76.54	0.41/0.60/1.10	1.08/1.93/3.26	92.1
300×300 (0.47 m, 1.2°)	77.65	0.32/0.48/1.13	1.07/1.79/3.13	166.9

* We list 50%, 75% and 95% quantiles of TE and RE in this table.

TABLE VII
ABLATION STUDY ON LOCAL FEATURES*

Local Features	Recall@1 (%)	TE (m)	RE (°)
FPFH [35] w/ RING++	75.03	0.56/0.79/1.30	1.19/2.03/4.03
SHOT [19] w/ RING++	63.73	0.56/0.79/1.39	1.16/2.03/4.08

* We list 50%, 75%, and 95% quantiles of TE and RE in this table.

declines slightly, while rotation estimation benefits greatly from the increased resolution. We also provide the runtime of different resolutions. The results show that as the resolution grows, the time cost also increases.

The other ablation study is carried out to confirm RING++'s effectiveness as a feature aggregation method. Using the same multisession pair from the NCLT dataset, we test RING++ using several local features with the resolution set to 120 × 120. The results are shown in Table VII. We utilize FPFH [35] implemented in Open3D [64], which extracts 33-channel point features. Regarding SHOT [18], [19], we use its Python implementation and change the bin value to 2, which results in the final 64-channel features. Although the bin value can be increased to 11 as in the original work, we only set it to 2 because of the linear memory consumption growth during the evaluation. The results show that our RING++ can also perform well with only two signature bins. Compared to the features we selected, FPFH and SHOT show competitive performance, validating the RING++ framework as a generic feature aggregation method. To strike a balance between precision and storage, we use the six-channel features introduced in Section V-A in all other experiments.

F. Computational Cost

Considering the real-time constraint, we calculate the computation cost of our algorithm. All the methods are implemented in Python and tested on a system equipped with an Intel i7-10700 (2.9 GHz) and an Nvidia GeForce RTX 2060 SUPER with 8-GB memory. As can be seen in Fig. 13, the average computational time of RING++ is about 50 ms. The most time-consuming part is the nearest neighbor search during the local feature

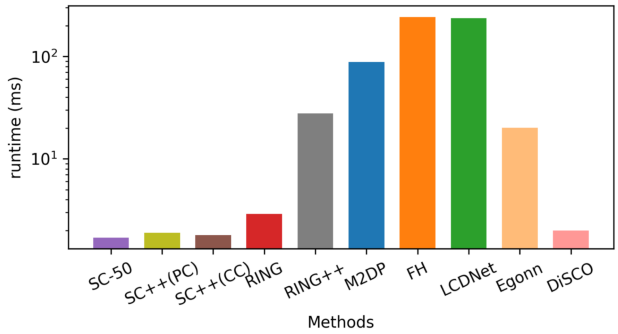


Fig. 13. Average descriptor generation time when tested on the NCLT Dataset. FH represents Fast Histogram. The result of RING++ does not include the RING generation time.

TABLE VIII
TIME COSTS (MS) FOR EACH RING++ COMPONENT

NN method	NN search	Feature extraction	TING generation	RING generation	Recall@1 (%)
Voxel	25.7	0.7	1.4	25.5	73.21
Pointcloud	68.2	1.8	1.4	25.5	75.23

calculation. In our pipeline, we first downsample and voxelize the original scans to boost this search. To reveal how much performance is gained from this step, we conduct an ablation study. The results are shown in Table VIII. It indicates that while place recognition performance only slightly decreases, the nearest neighbor search on voxelized points saves a significant amount of time. It is worth noting that the time consumption varies according to the number of channels in the local feature. It grows linearly as the feature channel increases. The RING generation time relates to the number of map scans. In our settings, the entire trajectory (∼5.5 km) of the NCLT dataset is covered with 253 map scans, and the RING generation of each query takes 25.5 ms. As for the large-scale MulRan dataset (∼23.4 km), it costs 99.7 ms to generate RING and necessitates 1124 map scans. All runtime results are tested in batches of eight in parallel on GPU.

G. Robustness

We further evaluate the robustness of our method in occluded and opposite driving direction situations. The experiment setting is the same as previous experiments where the equidistant sampling gap between query data is 5 m and that of map data is 20 m. In such challenging scenarios, we observe that the most similar scan may not be closest to the query, which makes the previous 10 m revisited criteria unreasonable. Therefore, we evaluate different algorithms using the revisited threshold ranging from 10 to 140 m.

1) Occlusion: Occlusions caused by dynamic objects make localization more difficult. In Fig. 14(a) and (b), we randomly block 90° and 150° of the scan in the NCLT dataset and evaluate the success rates of global localization, place recognition, and pose estimation. We observe that the performance of both methods significantly declines, while RING++ still outperforms

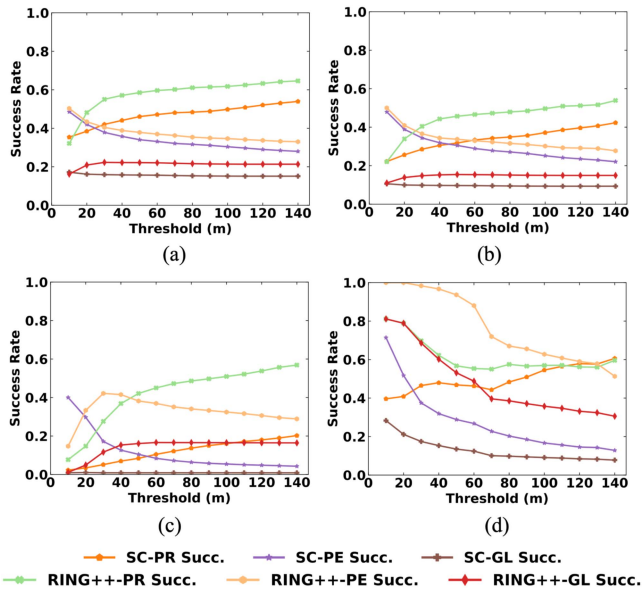


Fig. 14. Success rates under occluded and reversal scenarios. (a) NCLT (90° occlusion). (b) NCLT (150° occlusion). (c) MulRan Riverside. (d) KITTI 08.

Scan Context. In addition, we evaluate these two methods in the MulRan Riverside dataset where the scans suffer from a fixed 90° occlusion and reverse directions. The performance of both methods drops a lot, but RING++ still outperforms the other. In these difficult cases involving dynamic objects and occlusion, creating a submap from recent scans might be a possible solution to the significant performance reduction, which we would like to look into in the future.

2) Reverse Driving Direction: Another challenging scenario in urban areas is that loops are from the opposite driving direction. As mentioned above, the scenarios in MulRan Riverside include reverse direction and occlusion. The results are shown in Fig. 14(c) and (d). To eliminate the effect of occlusion in the reverse direction scenarios, we further conduct experiments on two reverse trajectory segments cropped from KITTI 08. Due to the limited number of scans in these two short reverse segments under sparse scenarios, the trend of Fig. 14(d) is different from the other three plots in Fig. 14(a)–(c). From Fig. 14, we see that utilizing 360° scans improves the performance of both approaches significantly. Despite the sparse map setup and translation difference caused by the lane change, our approach still has better performance.

H. Discussion

1) Global Localization Criteria: In real-world scenarios incorporating occlusion or reverse loops, the place recognition criteria like 10 m in the majority of our experimental settings may not be suitable for evaluation, which can be verified in Fig. 14. Taking the MulRan Riverside dataset as an example, as the revisited threshold grows, the place recognition surely provides more candidates, but the pose estimation is not able to align all these matches due to the large translation differences, so the compound criterion *GL Success Rate* finally reaches a stable level. With the help of the globally convergent solver, RING++ has more chances to align these matches, resulting in a higher *GL Success Rate* at a larger threshold. Based on this observation, we suppose that the max *GL Success Rate* evaluated with changing revisited threshold may be a better criterion for evaluating global localization.

TABLE IX
ATE AND SUCCESSFULLY ALIGNED LOOP NUMBER OF SCAN CONTEXT++ (PC) AND RING++ INTEGRATED SLAM SYSTEMS

Approach	Legged Robot Dataset		NCLT	
	ATE (m)	Succ. Loop	ATE (m)	Succ. Loop
SC++ SLAM	1.56	151	7.89	66
RING++ SLAM	**0.72**	**167**	**6.43**	**126**

2) Applications to Other Range Sensors: RING++ offers a general framework that can be easily applied to other range sensors, like radars. It should be noted that the range sensor should be mounted roughly parallel to the ground plane to generate a reasonable BEV. As for the sensors like livox LiDAR with a limited field of view, the performance is similar to the cases of strong occlusion. Moreover, we expect that submap can tackle this problem.

3) Leveraging Deep Learning Feature Extractor: With the multichannel framework derived, it is straightforward to leverage deep learning on feature extraction to provide better performance. In this research, we mainly focus on architecture derivation and offer a general framework with roto-translation invariance. Readers can refer to [65] and see how to incorporate deep learning with our proposed framework.

VIII. System Application

As a lightweight framework, we implement the proposed RING++ into a stand-alone module without any prior information. We integrate a real-time back-end manager with the pose graph optimizer implemented by GTSAM [66]. With the optimization results acquired, we rearrange the keyframes to generate a global map. These components, together with any front-end odometry, form a complete SLAM system. To validate the performance of our approach in real-world applications, we integrate FAST-LIO2 [67] into our SLAM system and compare RING++ and Scan context++ (PC) within the SLAM system.

A. Multirobot SLAM System

In real applications such as exploration and rescue, multirobot SLAM systems are expected to quickly build a precise environment map. The precision of the map is largely dependent on the correct alignment between keyframes. Because of the motion flexibility, legged robots are frequently used in such scenarios [68]. With this background, we collected data from three legged robots outfitted with an IMU-integrated Ouster64 LiDAR and a Jetson Xavier NX. The ground truth is acquired by the offline interactive SLAM [69] with automatically added and handpicked edges. In this setup, keyframes are generated at 2-m intervals. The results are shown in Table IX and Fig. 15; RING++ provides more successfully aligned loops for back-end pose optimization, which leads to lower drifts. The point cloud built by our system can be seen in Fig. 16; points generated by different robots are well aligned.

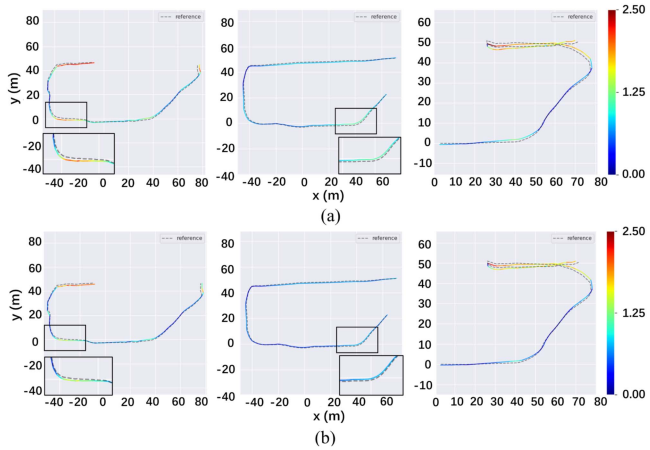


Fig. 15. Error visualization of Scan Context++ (PC) integrated SLAM system (top) compared to our approach (bottom) on legged robot dataset. (a) SC++ SLAM. (b) RING++ SLAM.

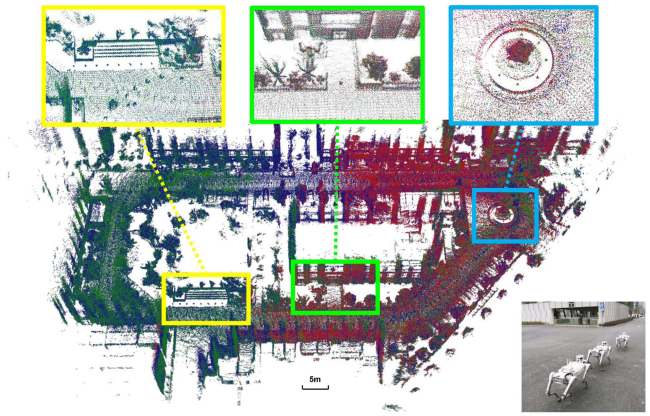


Fig. 16. Qualitative results of our RING++ SLAM on the legged robot dataset. Different colored points are generated from different legged robots.

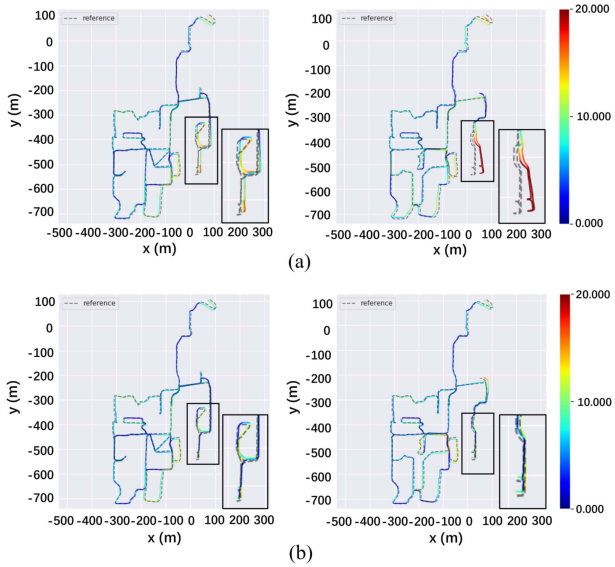


Fig. 17. Error visualization of Scan Context++ (PC) integrated SLAM system (top) compared to our approach (bottom) on two sequences of NCLT dataset. (left:"2012-05-26", right:"2012-03-17"). (a) SC++ SLAM. (b) RING++ SLAM.

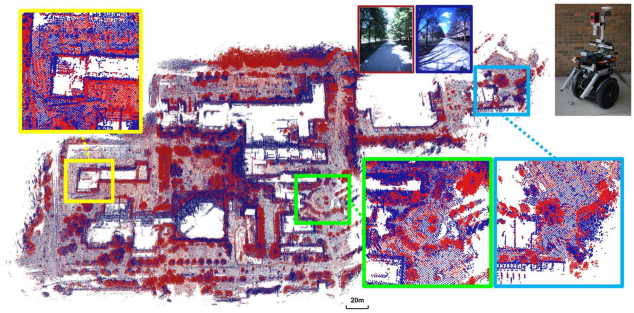


Fig. 18. Qualitative results of our RING++ SLAM on the NCLT dataset. Red points are generated from "2012-05-26," while blue points are generated from "2012-03-17."

B. *Multisession SLAM System*

Another application of RING++ is multisession SLAM. It differs from a multirobot SLAM system in the changing environment. The chosen NCLT dataset contains a wide range of environmental changes, such as dynamic objects, seasonal changes like winter and summer, and structural changes like building construction. We evaluate our system on "2012-05-26" and "2012-03-17." The data from these sessions are processed online at the same time, indicating a multirobot setup. The LiDAR points replayed online are from different sessions, forming a challenging temporal/spatial multiagent setup. To maintain a sparse representation for the very large environment, we generate keyframes at 5-m intervals. The performance can be seen in Table IX and Fig. 17; RING++ provides almost twice the number of loop closures than Scan Context++ (PC), thus having better performance. The mapping result is shown in Fig. 18, where two different colored points represent two NCLT dataset sequences. The overlapping landmarks in magnified figures indicate the low drift of our system. It should be noted that FAST-LIO2 failed to provide acceptable odometry in the final minutes of the NCLT dataset, so those parts were discarded in the evaluation.

IX. Conclusion

In this article, we presented a roto-translation-invariant framework RING++ for global localization on the sparse scan map, including representation pass and solving pass. Specifically, we extracted six local features with geometry information and aggregated the features into three representations in the representation pass: rotation-equivariant SG, translation-invariant and rotation-equivariant TING, and roto-translation-invariant RING. In the solving pass, we performed place recognition, rotation estimation, translation estimation, and pose refinement. Thanks to roto-translation-invariant representation, RING++ achieved superior performance on benchmark datasets, outperforming the state-of-the-art methods. By integrating RING++ as a stand-alone loop closure detection module into SLAM systems, we validated the effectiveness of our approach without any prior knowledge in real scenarios.

REFERENCES

[1] C. Valgren and A. J. Lilienthal, "SIFT, SURF and seasons: Long-term outdoor localization using local features," in *Proc. 3rd Eur. Conf. Mobile Robots*, 2007, pp. 253–258.

[2] S. Lowry et al., "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[3] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3D point cloud map," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 4802–4809.

[4] Y. Wang, Z. Sun, C.-Z. Xu, S. E. Sarma, J. Yang, and H. Kong, "LiDAR iris for loop-closure detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 5769–5775.

[5] K. Vidanapathirana, P. Moghadam, B. Harwood, M. Zhao, S. Sridharan, and C. Fookes, "Locus: LiDAR-based place recognition using spatiotemporal higher-order pooling," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2021, pp. 5075–5081.

[6] J. Komorowski, "MinkLoc3D: Point cloud based large-scale place recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1790–1799.

[7] T. Röhling, J. Mack, and D. Schulz, "A fast histogram-based similarity measure for detecting loop closures in 3-D LIDAR data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 736–741.

[8] L. He, X. Wang, and H. Zhang, "M2DP: A novel 3D point cloud descriptor and its application in loop closure detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 231–237.

[9] M. A. Uy and G. H. Lee, "PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4470–4479.

[10] Z. Liu et al., "LPD-Net: 3D point cloud learning for large-scale place recognition and environment analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2831–2840.

[11] G. Kim, S. Choi, and A. Kim, "Scan context++: Structural place recognition robust to rotation and lateral variations in urban environments," *IEEE Trans. Robot.*, vol. 38, no. 3, pp. 1856–1874, Jun. 2022.

[12] S. Lu, X. Xu, H. Yin, R. Xiong, and Y. Wang, "One RING to rule them all: Radon sinogram for place recognition, orientation and translation estimation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 2778–2785.

[13] F. Stein and G. Medioni, "Structural indexing: Efficient 3-D object recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 125–145, Feb. 1992.

[14] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Aligning point cloud views using persistent feature histograms," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2008, pp. 3384–3391.

[15] M. Weinmann, B. Jutzi, and C. Mallet, "Semantic 3D scene interpretation: A framework combining optimal neighborhood size selection with relevant features," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 2, no. 3, pp. 181–188, 2014.

[16] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3D scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 5, pp. 433–449, May 1999.

[17] W. Wohlkinger and M. Vincze, "Ensemble of shape functions for 3D object classification," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2011, pp. 2987–2992.

[18] F. Tombari, S. Salti, and L. D. Stefano, "A combined texture-shape descriptor for enhanced 3D feature matching," in *Proc. IEEE 18th Int. Conf. Image Process.*, 2011, pp. 809–812.

[19] F. Tombari, S. Salti, and L. d. Stefano, "SHOT: Unique signatures of histograms for surface and texture description," *Comput. Vis. Image Understanding*, vol. 125, pp. 251–264, 2014.

[20] D. Maturana and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. robots Syst.*, 2015, pp. 922–928.

[21] Z. Wu et al., "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1912–1920.

[22] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4490–4499.

[23] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[24] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5105–5114.

[25] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, 2019.

[26] Y. Shen, C. Feng, Y. Yang, and D. Tian, "Mining point cloud local structures by kernel correlation and graph pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4548–4557.

[27] X. Chen et al., "OverlapNet: Loop closing for LiDAR-based SLAM," in *Proc. Robot. Sci. Syst.*, 2020.

[28] L. Li et al., "SSC: Semantic scan context for large-scale place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 2092–2099.

[29] J. Komorowski, M. Wysoczańska, and T. Trzcinski, "MinkLoc++: LIDAR and monocular image fusion for place recognition," in *Proc. Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.

[30] Y. Pan, X. Xu, W. Li, Y. Cui, Y. Wang, and R. Xiong, "CORAL: Colored structural representation for bi-modal place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 2084–2091.

[31] H. Lai, P. Yin, and S. Scherer, "AdaFusion: Visual-LiDAR fusion with adaptive weights for place recognition," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 12038–12045, 2022.

[32] J. Knopp, M. Prasad, G. Willems, R. Timofte, and L. V. Gool, "Hough transform and 3D SURF for robust three dimensional classification," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 589–602.

[33] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[34] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 3212–3217.

[35] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3D recognition and pose using the viewpoint feature histogram," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2010, pp. 2155–2162.

[36] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[37] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3304–3311.

[38] X. Xu, H. Yin, Z. Chen, Y. Li, Y. Wang, and R. Xiong, "DiSCO: Differentiable scan context with orientation," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2791–2798, Apr. 2021.

[39] K. Żywanowski, A. Banaszczyk, M. R. Nowicki, and J. Komorowski, "MinkLoc3D-SI: 3D LiDAR place recognition with sparse convolutions, spherical coordinates, and intensity," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1079–1086, Apr. 2022.

[40] J. Komorowski, M. Wysoczanska, and T. Trzcinski, "EgoNN: Egocentric neural network for point cloud based 6DoF relocalization at the city scale," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 722–729, Apr. 2022.

[41] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, Jul. 2019.

[42] H. Yin, L. Tang, X. Ding, Y. Wang, and R. Xiong, "LocNet: Global localization in 3D point clouds for mobile vehicles," in *Proc. IEEE Intell. Veh. Symp.*, 2018, pp. 728–733.

[43] X. Ding et al., "Translation invariant global estimation of heading angle using sinogram of LiDAR point cloud," in *Proc. Int. Conf. Robot. Autom.*, 2022, pp. 2207–2214.

[44] X. Xu, "Supplementary material for RING++: Roto-translation invariant gram for global localization on a sparse scan map," 2023. [Online]. Available: https://drive.google.com/file/d/1PuYDx3__J5tk8Dg3ToKPzLGKv5iWqMUt/view?usp=share_link

[45] W. Lu, Y. Zhou, G. Wan, S. Hou, and S. Song, "L3-Net: Towards learning based LiDAR localization for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6389–6398.

[46] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Sensor Fusion IV: Control Paradigms and Data Structures*, vol. 1611. Bellingham, WA, USA: SPIE, 1992, pp. 586–606.

[47] I. Sipiran and B. Bustos, "Harris 3D: A robust extension of the Harris operator for interest point detection on 3D meshes," *Vis. Comput.*, vol. 27, no. 11, pp. 963–976, 2011.

[48] R. Arandjelovic and A. Zisserman, "All about VLAD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1578–1585.

[49] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 357–360.

[50] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of Michigan north campus long-term vision and LiDAR dataset," *Int. J. Robot. Res.*, vol. 35, no. 9, pp. 1023–1035, 2016.

[51] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "MulRan: Multimodal range dataset for urban place recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 6246–6253.
[52] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The Oxford Radar RobotCar Dataset: A radar extension to the Oxford RobotCar Dataset," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 6433–6438.
[53] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar Dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, 2017.
[54] V. Raghavan, P. Bollmann, and G. S. Jung, "A critical investigation of recall and precision as measures of retrieval system performance," *ACM Trans. Inf. Syst.*, vol. 7, no. 3, pp. 205–229, 1989.
[55] D. M. Christopher and S. Hinrich, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.
[56] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to Information Retrieval*, vol. 39. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[57] D. Cattaneo, M. Vaghi, and A. Valada, "LCDNet: Deep loop closure detection for LiDAR SLAM based on unbalanced optimal transport," 2021.
[58] H. Wang, C. Wang, and L. Xie, "Intensity scan context: Coding intensity and geometry relations for loop closure detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 2095–2101.
[59] D. Cattaneo, M. Vaghi, and A. Valada, "LCDNet: Deep loop closure detection and point cloud registration for LiDAR SLAM," *IEEE Trans. Robot.*, vol. 38, no. 4, pp. 2074–2093, Aug. 2022.
[60] K. Koide, M. Yokozuka, S. Oishi, and A. Banno, "Voxelized GICP for fast and accurate 3D point cloud registration," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2021, pp. 11054–11059.
[61] Z. Zhang, "Iterative point matching for registration of free-form curves and surfaces," *Int. J. Comput. Vis.*, vol. 13, no. 2, pp. 119–152, 1994.
[62] H. Yang, J. Shi, and L. Carlone, "TEASER: Fast and certifiable point cloud registration," *IEEE Trans. Robot.*, vol. 37, no. 2, pp. 314–333, Apr. 2021.
[63] A.-Q. Cao, G. Puy, A. Boulch, and R. Marlet, "PCAM: Product of cross-attention matrices for rigid registration of point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13229–13238.
[64] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," 2018, *arXiv:1801.09847*.
[65] S. Lu, X. Xu, L. Tang, R. Xiong, and Y. Wang, "DeepRING: Learning roto-translation invariant representation for LiDAR based place recognition," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 1904–1911.
[66] F. Dellaert, "Factor graphs and GTSAM: A hands-on introduction," Center Robot. Intell. Mach., Georgia Inst. Technol., Atlanta, GA, USA, Tech. Rep. GT-RIM-CP&R-2012-002, 2012.
[67] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "FAST-LIO2: Fast direct LiDAR-inertial odometry," *IEEE Trans. Robot.*, vol. 38, no. 4, pp. 2053–2073, Aug. 2022.
[68] C. D. Bellicoso et al., "Advances in real-world applications for legged robots," *J. Field Robot.*, vol. 35, no. 8, pp. 1311–1326, 2018.
[69] K. Koide, J. Miura, M. Yokozuka, S. Oishi, and A. Banno, "Interactive 3D graph SLAM for map correction," *IEEE Robot. Autom. Lett.*, vol. 6, no. 1, pp. 40–47, Jan. 2021.

Xuecheng Xu received the B.S. degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2019, where he is currently working toward the Ph.D. degree with the State Key Laboratory of Industrial Control Technology and Institute of Cyber-Systems and Control.

His current research interests include LiDAR simultaneous localization and mapping and multirobot systems.

Sha Lu received the B.S. degree in mechanical engineering from the CQU-UC Joint Co-op Institute, Chongqing University, Chongqing, China, in 2021. She is currently working toward the Ph.D. degree with the State Key Laboratory of Industrial Control Technology and Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China.

Her research interests include robotics and deep learning, with a focus on LiDAR simultaneous localization and mapping.

Jun Wu received the M.E. degree in control engineering in 2018 from Zhejiang University, Hangzhou, China, where she is currently working toward the Ph.D. degree with the State Key Laboratory of Industrial Control Technology and Institute of Cyber-Systems and Control.

Her research interests include robotics perception.

Haojian Lu (Member, IEEE) received the B.Eng. degree in mechatronics engineering from the Beijing Institute of Technology, Beijing, China, in 2015, and the Ph.D. degree in robotics from the City University of Hong Kong, Hong Kong, in 2019.

He is currently a Professor with the State Key Laboratory of Industrial Control and Technology and Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China. His research interests include micro/nanorobotics, bioinspired robotics, medical robotics, micro aerial vehicle, and soft robotics.

Qiuguo Zhu received the B.Eng. degree in mechanical engineering in 2008, the M.Eng. and Ph.D. degrees in control science and engineering from Zhejiang University, Hangzhou, China, in 2011 and 2020, respectively.

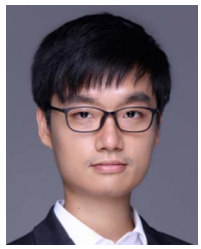
He is currently an Associate Professor with the Department of Control Science, Zhejiang University, Hangzhou, China. His research interests include the control of humanoid robots, bipedal and quadrupedal walking and running, manipulators, rehabilitation exoskeletons, and machine intelligence.

Yiyi Liao received the Ph.D. degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2018.

From 2018 to 2021, she was a Postdoctoral Researcher with the Autonomous Vision Group, University of Tubingen, Tübingen, Germany, and Max Planck Institute for Intelligent Systems, Tübingen. She is currently an Assistant Professor with Zhejiang University. Her research interests include 3-D vision and scene understanding.

Rong Xiong (Senior Member, IEEE) received the Ph.D. degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2009.

She is currently a Professor with the Department of Control Science and Engineering, Zhejiang University. Her current research interests include motion planning and simultaneous localization and mapping.



Yue Wang (Member, IEEE) received the Ph.D. degree in control science and engineering from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2016.

He is currently an Associate Professor with the Department of Control Science and Engineering, Zhejiang University. His current research interests include mobile robotics and robot perception.