2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

DeepRING: Learning Roto-translation Invariant Representation for LiDAR based Place Recognition

Sha Lu[1], Xuecheng Xu[1], Li Tang[2], Rong Xiong[1] and Yue Wang[1†]

***Abstract*— LiDAR based place recognition is popular for loop closure detection and re-localization. In recent years, deep learning brings improvements to place recognition by learnable feature extraction. However, these methods degenerate when the robot re-visits previous places with a large perspective difference. To address the challenge, we propose DeepRING to learn the roto-translation invariant representation from LiDAR scan, so that robot visiting the same place with a different perspective can have similar representations. There are two keys in DeepRING: the feature is extracted from sinogram, and the feature is aggregated by magnitude spectrum. The two steps keep the final representation with both discrimination and roto-translation invariance. Moreover, we state place recognition as a one-shot learning problem with each place being a class, leveraging relation learning to build representation similarity. Substantial experiments are carried out on public datasets, validating the effectiveness of each proposed component, and showing that DeepRING outperforms the comparative methods, especially in dataset level generalization.**

I. INTRODUCTION

Place recognition plays a significant role in autonomous driving applications. It retrieves the closest place from the past trajectory of robot for loop closure detection in SLAM systems to reduce the accumulated error. Because of the robustness to ever-changing environmental conditions, LiDAR sensors have been widely used for place recognition in recent years. As the field of view of LiDAR is wide, the LiDAR scans can have significant overlap when a robot revisits a previous place with a different perspective. However, there still remains a challenge for place recognition when a large perspective difference presents between the current scan and the scan taken at the previous place.

A popular way to deal with the challenge first occurs in handcrafted methods. By explicitly designing the rotation invariant representation, scans taken by robot spinning in spot can keep the same. Therefore, the perspective difference visiting the same place can be suppressed, place can be determined by finding the most similar previous scan to the current one. Several methods have been proposed for achieving this property, including histogram, polar gram (PG), and principal component analysis [1, 2, 3, 4, 5, 6]. More recently, translation invariance is also derived for the scan representation to address large perspective difference [7]. However, the handcrafted feature extraction limits the discrimination of this line of works, most of whom employ simple features e.g. occupancy.

To improve the features, a deep network is employed for feature learning from data [8, 9]. However, their networks do not inherently yield invariant representations, thus calling for data augmentation with artificially added rotation and translation to improve the robustness of scan representation again perspective change. Motivated by the representation design in handcrafted methods, in [10, 11, 12, 13, 14], the histogram, PG, and range image are inserted into neural networks to explicitly build rotation invariant representations. Together with deep learning, discrimination can also be improved. Nevertheless, the rotation invariance loses when there is an obvious translation difference between the scans. It remains unclear how to keep deep features invariant when both rotation and translation differences present.

To address the problem, we propose a neural network architecture, named DeepRING, to learn *roto-translation invariant* representation for a LiDAR scan by endowing the deep feature extraction with the RING architecture [7]. As shown in Fig. 1, we show that by formulating the LiDAR scan as sinogram (SG), taking it as input to the convolution network, and calculating the magnitude spectrum of the output, the resultant representation is roto-translation invariant, and can be learned from data for better discrimination in an end-to-end manner, bridging the advantages of two lines of works. Furthermore, we state the place recognition as a one-shot learning problem. Specifically, each place is regarded as a class, with the scan taken at that place as a shot, building a multi-class support set, while the current scan and the place form a query set. Such statement leverages the relation learning to replace the popular Siamese network and Euclidean triplet/quadruplet loss frequently used in place recognition. The experimental results validate the effectiveness of the proposed modules, showing that DeepRING outperforms the comparative methods, especially in data level generalization. In summary, the main contributions of our method comprise:

- An end-to-end learning framework, DeepRING, to endow scan feature extraction with roto-translation invariant property, which tackles the problem of large perspective difference.
- Statement of place recognition as one-shot learning, to leverage relation learning for building better similarity. The efficient implementation saves the computation.
- Validation of the proposed method on two large-scale benchmark datasets, showing superior performance of DeepRING, especially for generalization, verifying the

*This work was supported in part by the National Key RD Program of China under Grant 2021ZD0114500 and the Natural Science Foundation of Zhejiang Province under grant number LGG21F030012.

[1]State Key Laboratory of Industrial Control and Technology, and the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, 310058, China.
[2]Alibaba Group, Hangzhou, 310052, China.
[†]Corresponding author wangyue@iipc.zju.edu.cn.

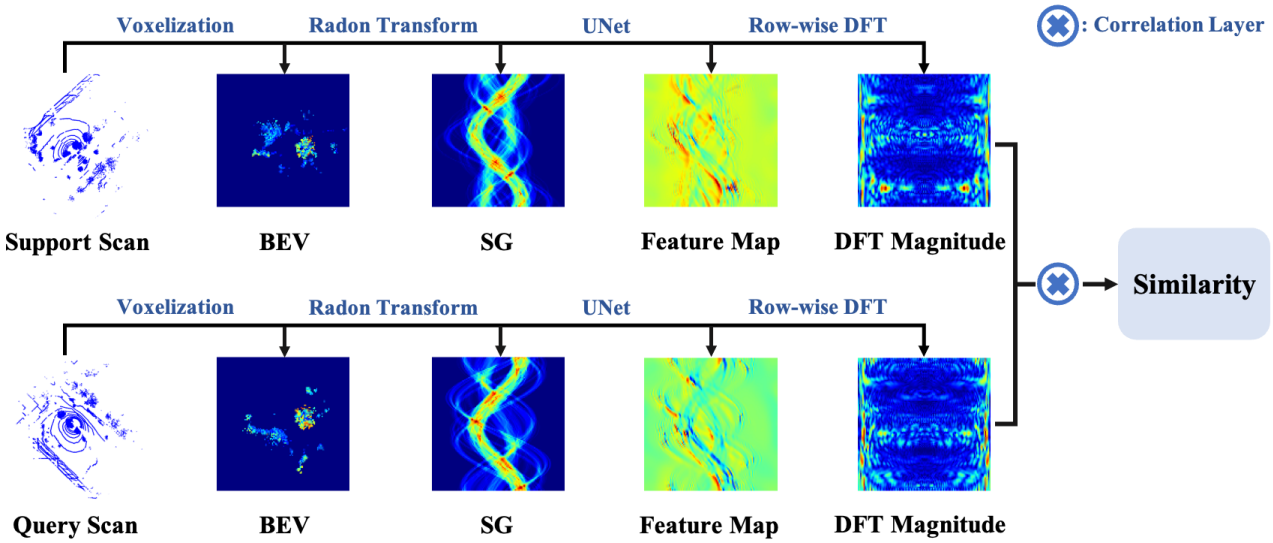


Fig. 1: Overall framework of the proposed method DeepRING.

effectiveness of invariance design and one-shot learning.

II. RELATED WORKS

In this section, we embark on related works review in terms of handcrafted methods and learning-based methods. In addition, we also introduce one-shot learning in brief.

A. Handcrafted Methods

Fast Histogram [2] leverages the range of 3D points and encodes it into histogram as the global descriptor. M2DP [15] projects the LiDAR scan into multiple 2D planes and generate the global descriptor that is robust to rotation change via PCA (Principal Component Analysis). Scan Context series [3, 5] utilize an egocentric spatial descriptor encoded by the maximum height of points. Likewise, LiDAR-Iris [4] extracts the LiDAR-Iris binary image and transforms it into frequency domain to achieve rotation invariance. Inspired by [16], RING [7] proposes sinogram to represent point clouds for both place recognition and pose estimation. In these methods, the similarity computation generally calls for an exhaustive search. Moreover, the feature extraction in these methods limits the discrimination of the representation.

B. Learning-based Methods

PointNetVLAD [8] extracts features from the raw 3D point cloud with PointNet [17] and aggregates them to global descriptor with NetVLAD [18]. LPD-Net [9] proposes a graph-based neighborhood module to aggregate the extracted adaptive local features, which enhances the place description of the global descriptor, especially for large-scale environments. Locus [19] aggregates the topological relationship with temporal information of point clouds to improve the place description ability. Except for global descriptors, some works [20, 21, 22] learn both global and local features of LiDAR scans to address a 6DoF localization problem. Without explicit invariant representation, these methods achieve robustness against perspective difference by data augmentation. LocNet [10] encodes the range histogram as fingerprint and learns a semi-learning neural network to achieve rotation invariance. OverlapNet [12] exploits multiple cues from scans and predicts the overlap along with the relative yaw angle between two scans using a Siamese network. DiSCO [11] converts the point cloud in the cylindrical coordinate, extracts features using an encoder-decoder network and transforms these features to frequency domain to reach rotation invariance. RINet [14] further exploits the stage of inserting feature extraction to keep the rotation invariance. However, the rotation invariance in these methods is sensitive to translation difference. A larger translation may degenerate their performances.

C. One-shot Learning

With the progress of deep learning in the past decades, learning based methods have presented excellent performance. A typical scenario is that the class in the test phase are all seen in the training phase, and in each class there are lots of samples. However, this is not true in tasks like face recognition, or place recognition, where each class only has few samples, and the class in test phase are all unseen. To deal with this problem, one-shot learning is proposed, and a typical method is distance metric learning based via "learning to compare with distance metrics" [23]. For instance, Matching Network [24] employs cosine similarity as the distance metric. Prototypical Network [25] utilizes Euclidean distance in the embedding space to compare different classes. Relation Network [26] designs a CNN to learn a distance metric to compare the relation of images. Inspired by the works in this direction, we state the place recognition problem as one-shot learning to leverage fruitful relation learning methods to build the similarity between scans.

III. METHODOLOGY

We propose a one-shot learning framework based on sinogram (SG) representation, named DeepRING, to construct roto-translation invariance for robust place recognition.

A. Rotation Equivariant Representation

Sinogram: In this subsection, we convert a LiDAR scan to a sinogram representation which is visualized in Fig. 1. Given a raw 3D scan, we first remove the uninformative points of ground plane and quantize the preprocessed point cloud into a finite number of pillars. Taking advantage of the occupancy information in each pillar, we encode the 3D point cloud into a bird-eye view (BEV) image, denoted as $f(x, y)$. After that, we apply Radon transform to the BEV image, yielding a resultant sinogram (SG) denoted as $R_f(\theta, \tau)$. The mathematical formula of Radon transform along a couple of parallel lines is

$$\begin{aligned}\mathcal{R}_f(\theta,\tau) &= \int_{L:\ x\cos\theta+y\sin\theta=\tau} f(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(\tau - x\cos\theta - y\sin\theta)\mathrm{d}x\mathrm{d}y\end{aligned} \tag{1}$$

where L represents the integrated line parameterized as $x\cos\theta + y\sin\theta = \tau$, $\theta \in [0, 2\pi)$ is the angle between L and the y axis, and $\tau \in (-\infty, \infty)$ is the perpendicular distance from the origin to L.

Radon sinogram transforms a rotation in the BEV image space (corresponding to the 3D Cartesian space) to an equivariant circular shift along θ axis in the radon image space, which is also rotation equivariant. Specifically, applying the rotation by an angle α to the raw 3D scan, the resultant sinogram shifts along θ axis by the distance α, which can be written as

$$\mathcal{R}_f(\theta,\tau) \xrightarrow{\alpha} \mathcal{R}_f(\theta+\alpha,\tau) \tag{2}$$

Aside from rotation transformation, a translation by a vector $\vec{d} = (\Delta x, \Delta y)^T$ reflects a linear shift in the variable τ of sinogram, equal to the projected length of vector $\vec{d}$ onto line $xcos\theta + ysin\theta = \tau$, namely, $\Delta\tau = (\cos\theta, \sin\theta)\vec{d} = \Delta xcos\theta + \Delta ysin\theta$. The corresponding relationship is

$$\mathcal{R}_f(\theta,\tau) \xrightarrow{\vec{d}} \mathcal{R}_f(\theta,\tau-\Delta\tau) \tag{3}$$

Other Representations: In terms of intermediate representations, polar transform and spherical projection are widely used to convert a point cloud from the 3D Cartesian space to the 2D image space. Similar to SG, a rotation in the 3D space is projected onto a cyclic shift along the rotation-related dimension in the image space, arriving at rotation equivariance. However, a translation causes a nonlinear shift along the rotation-related dimension of PG and range image, resulting in a changing scale of these two representations. The nonlinear shift along the axis of PG is mathematically written as

$$\begin{aligned}&\mathcal{P}_f(r,\theta) \xrightarrow{\vec{d}} \mathcal{P}_f(r',\theta') \\ &r' = \sqrt{(r\cos\theta-\Delta x)^2 + (r\sin\theta-\Delta y)^2} \\ &\theta' = \arctan\frac{r\sin\theta-\Delta y}{r\cos\theta-\Delta x}\end{aligned} \tag{4}$$

where $\mathcal{P}_f(r, \theta)$ is the result after polar transform, i.e. PG. By comparing Eq. 4 with Eq. 3, we can see that the translation variance severely corrupts the rotation equivariance property of PG. Same for range image, the translation influences the scale of the range image, which is also harmful to the rotation equivariance.

Under this circumstance, it is difficult to eliminate translation variance, which simultaneously affects the strict rotation equivariance of these representations in practice.

B. Feature Extraction Network

After the intermediate rotation equivariant representation construction, we utilize a feature extraction network to generate a more discriminative place representation. In order to maintain the rotation equivariance, relative operations are required, which corresponds to translation equivariant operations in our case. Standard convolutions are equivariant to translations so that they are successfully applied in many image tasks. Nevertheless, the circular shift resulted from rotation is circularly bounded in $[0, 2\pi)$. Consequently, conventional linear convolutions can not guarantee strict rotation equivariance because they fail to deal with data in the border of the image. To address this problem, we choose circular convolutions which extract consistent features along the upper and lower boundaries of the image, formulated as

$$\begin{aligned}E_f[i,j] &= (R_f \circledast K)[i,j] \\ &= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1}(R_f[(i-m)\%M, (j-n)\%N]K[m,n])\end{aligned} \tag{5}$$

where $\circledast$ denotes the circular convolution operation, $\%$ denotes the mod operation, E_f is the output feature map, $R_f \in \mathbb{R}^{M\times N}$ is the input sinogram, and $K \in \mathbb{R}^{M\times N}$ is the convolution kernel. Note that E_f is translation equivariant to R_f resulting from the translation equivariant property of circular convolutions. Combining Eq. 3 and Eq. 5, we can arrive at

$$E_f^d(\theta_i,\tau) = E_f(\theta_i,\tau-\Delta\tau) \tag{6}$$

where E_f^d represents the output feature map after translation on the input scan by $\vec{d}$.

After the last circular convolution layer of the feature extraction network, the output E_f is passed through row-wise DFT to eliminate translation changes. Referring to the translation invariant property of DFT, the translation invariance is arrived by taking the magnitude of the frequency spectrum. Denote the resultant magnitude spectrum as M_f, then we have

$$\begin{aligned}M_f^d(\theta_i,\omega) &= |\mathcal{F}(E_f^d(\theta_i,\tau))| \\ &= |\mathcal{F}(E_f(\theta_i,\tau-\Delta\tau))| \\ &= |\mathcal{F}(E_f(\theta_i,\tau))||e^{-j2\pi\omega\Delta\tau}| \\ &= |\mathcal{F}(E_f(\theta_i,\tau))| = M_f(\theta_i,\omega)\end{aligned} \tag{7}$$

where $|\cdot|$ is the magnitude operation, $\mathcal{F}(\cdot)$ is the 1D DFT operation, M_f^d is the DFT magnitude of E_f^d, ω is the discrete sampled frequency and j is the imaginary unit. With translation invariance, the impact of translation disturbance on rotation equivariance is further alleviated. In this paper, we utilize UNet [27] equipped with circular convolutions and DFT as the feature extraction network, as depicted in Fig. 1.

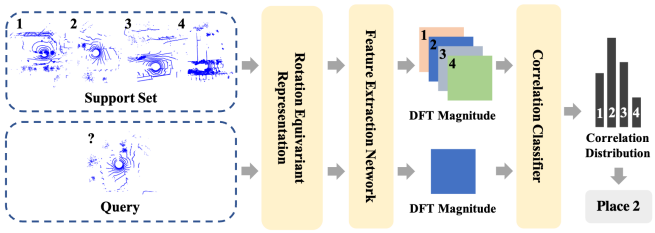
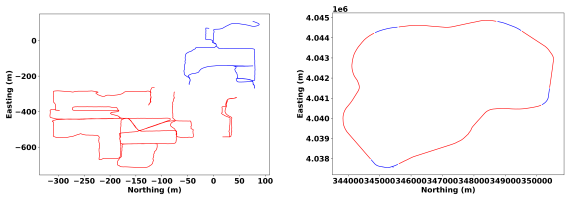


Fig. 2: One-shot learning algorithm for place recognition.



(a) Disjoint NCLT Dataset (b) Disjoint MulRan Dataset

Fig. 3: Disjoint sessions for training (red) and test (blue).

C. Statement as One-shot Learning

Place recognition aims to recognize the same place from all places in the database, which can be seen as a multi-way one-shot learning problem. Each place can be regarded as a class, and the scan taken from this place is a shot. Therefore, we tailor the one-shot learning algorithm for place recognition, as illustrated in Fig. 2. In the training stage, the support set and query set are randomly chosen from the map database. In the test stage, given a query scan of an unknown place, the neural network compares it to a set of reference scans from unknown places to determine which place it belongs to. The similarity functions in one-shot learning pipelines vary with different feature representations. Cosine and Euclidean distances are frequently used to calculate similarity for common features. However, these distance metrics do not maintain the internal relations between the feature maps learned from equivariant SG representations.

To better leverage the relations between SG representations, we propose a correlation-based distance metric to compare the similarity between places. The circular cross-correlation layer is employed after the feature extraction network to compute the correlation between support and query scans. Based on rotation equivariant and translation invariant feature maps, the correlation distance metric is symmetric and roto-translation invariant, defined as

$$\mathcal{C}(M_{f_1}, M_{f_2}) = \max_{\alpha} \sum_{\theta} \sum_{\omega} M_{f_1}(\theta+\alpha, \omega) M_{f_2}(\theta, \omega) \quad (8)$$

where $\mathcal{C}(M_{f_1}, M_{f_2})$ denotes the resultant correlation value between M_{f_1} and M_{f_2}, and α corresponds to the shift at the best alignment. For each query, we can obtain a correlation vector $\mathfrak{C} \triangleq \{\mathcal{C}(M_{f_Q}, M_{f_S}^{(i)})\}$ consisting of the correlation values between the query and each shot in the support set.

After that, we normalize the correlation values in $\mathfrak{C}$ with a softmax layer, given to the classifier to predict the class.

$$\tilde{\mathfrak{C}} = Softmax(W\mathfrak{C} + b) \quad (W, b \in \mathbb{R}) \quad (9)$$

where $\tilde{\mathfrak{C}} \in [0, 1]$ indicates the final correlation vector after softmax normalization, which is shown in Fig. 2.

Instead of using BCE loss for a binary classifier like [14, 28], we utilize cross-entropy loss considering place recognition as a multi-class classification problem.

D. Implementation Details

In the training phase, we train the model for 20 epochs, each of which incorporates 60 episodes. We use Adam [29] optimizer with the initial learning rate of 10^{-3} and decay of 10^{-4}. To schedule the learning rate, we use MultiStepLR with gamma of 0.1 and milestones at 5 and 12 epochs. Following the episode construction proposed in [24], we randomly select 24 classes from the training data and sample 7 examples including 1 shot and 6 queries within these classes to form support set S and query set Q respectively in each episode. After each episode, the model parameters are updated and the process above is repeated multiple times with different support and query examples. To distinguish different classes in the place recognition problem, we consider that scans whose positions are within $10m$ from each other belong to the same class and scans whose positions are beyond $20m$ from each other belong to different classes.

IV. DATASETS

We evaluate our method on disjoint regions of the sessions from two benchmark datasets NCLT [30] and MulRan [31] in Fig. 3. We evaluate the method with sparse data: the sampling distance of database and query trajectory is $20m$ and $5m$.

A. NCLT Dataset

The NCLT Dataset contains data of 27 mapping sessions over 15 months collected by a Segway robot. It includes a variety of environmental changes covering seasonal, temporal, and structural changes. Velodyne HDL-32E 3D LiDAR is used for point cloud acquisition. Among these sessions, we choose "2012-02-04" and "2012-03-17" sessions for model training and test, where "2012-02-04" serves as database session and "2012-03-17" serves as query session.

B. MulRan Dataset

The MulRan Dataset is a multi-modal range dataset for large-scale place recognition evaluation. It acquires scan data using Ouster OS1-64 LiDAR under various environments in South Korea. In the experiments, we select Sejong01 as database session and Sejong02 trajectories for training and test, where Sejong01 serves as database session and Sejong02 serves as query session.

C. Evaluation Metric

To evaluate place recognition performance of the proposed method comprehensively, we utilize four standard metrics: *Recall@1* reports the percentage of correctly recognized place using top 1 candidate; *F1 Score* takes the harmonic mean of precision and recall, combining them into a single metric; *Precision-Recall Curve* represents the performance of a classifier across a number of thresholds; *AUC* computes the area under the *Precision-Recall Curve*.

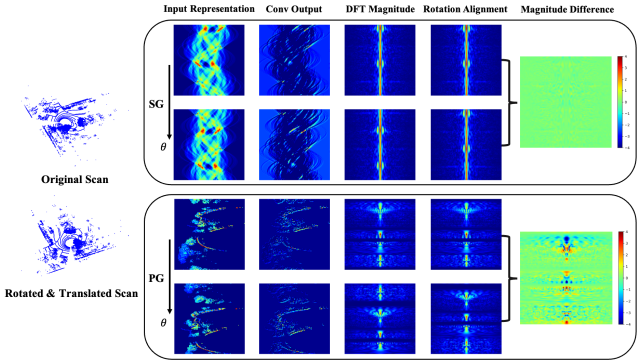
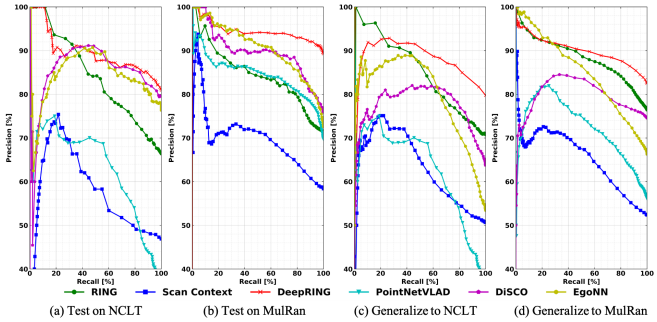


Fig. 4: Case study for roto-translation invariance achievement based on SG and PG, where the vertical arrow represents the rotation related θ axis of SG and PG.



(a) Test on NCLT (b) Test on MulRan (c) Generalize to NCLT (d) Generalize to MulRan

Fig. 5: Precision-Recall Curve of comparative methods.

V. EXPERIMENTAL RESULTS

In this section, based on the datasets and metrics above, we conduct a case study to visualize the roto-translation invariant representation, a comparative study to evaluate the performance and generalization, and several ablation studies to validate each proposed component.

A. Case Study

We first design a case study to validate that SG is a better representation than PG for roto-translation invariance as mentioned in Sec. III-A. We apply a random transformation comprising rotation and translation to generate a pair of point clouds, and then convert them to SG and PG representations respectively. Then we employ a circular convolution layer followed by row-wise DFT as a simplified implementation of feature extraction network on these two representations. Finally, we manually align the vertical axis of DFT magnitudes from SG and PG to eliminate the effect of rotation. In theory, if the representation is rotation equivariant and translation invariant, the difference between two scan representations should be zero. As shown in Fig. 4, SG based representation leads to obviously smaller difference, validating its invariance.

B. Comparison with State-of-the-Art

We compare our method with the state-of-the-art handcrafted and learning-based methods: Scan Context [3], RING [7], PointNetVLAD [8], DiSCO [11] and EgoNN [21]. For all approaches, we leverage the same data preprocessing for fair comparison. We utilize the same resolution of 120×120 for quantization in Scan Context, RING and our method to compare the performance of different representations. For EgoNN, we use the pre-trained model released publicly for evaluation on MulRan dataset and retrain the model on NCLT dataset for evaluation on NCLT dataset. We retrain PointNetVLAD and DiSCO on both datasets. Additionally, all parameters and details keep the same as the original papers.

Evaluation Performance: Tab. I and Fig. 5 depict the place recognition performance of all methods on the test trajectory of NCLT and MulRan datasets. In terms of two handcrafted methods, Scan Context reaches rotation invariance by brute-force matching and RING is roto-translation invariant taking advantage of the translation invariant property of DFT, which explains the advantage over PointNetVLAD. But limited by the discrimination of handcrafted features, the performances of these two methods are inferior to DiSCO that combines rotation invariance and feature learning. EgoNN shows approximate performance with DiSCO owing to its more complex feature extraction, as well as approximated rotation invariant representation. With both roto-translation invariant representation design and the learnable feature extraction, DeepRING outperforms the other learning based methods and handcrafted methods on the test datasets.

Generalization Performance: To evaluate the cross-domain generalization, we train the deep learning models on MulRan (NCLT) dataset and evaluate them on NCLT (MulRan) dataset, with the results depicted in Tab. II and Fig. 5. In a learning-free manner, Scan Context and RING show stable performance on both two datasets. In contrast, the learning based methods have degraded performance in the generalized domain, especially for PointNetVLAD due to the less inductive bias on invariance. Compared with PointNetVLAD, EgoNN refers to PG representation, accounting for better generalization to unseen places. DiSCO exceeds EgoNN by the explicit rotation invariant representation. Benefited from the explicit roto-translation invariant representation and one-shot learning, DeepRING shows the best generalization in new domain.

C. Ablation Study

To investigate the specific contribution of representation, aggregation and loss modules, we design multiple ablation studies on NCLT dataset and show the results in Tab. III.

Representation: We compare the performance of two rotation equivariant representations, PG and SG, in terms of place recognition. As we can see, SG based model outperforms PG based model, which is consistent with the case study in Sec. V-A. The underlying reason is that rotation and translation transformations are decoupled in SG representation, rotation is not sensitive to translation difference. In contrast, translation variance deteriorates the rotation equivariance property as analyzed in Sec. III-A.

Aggregation: After feature extraction, feature aggregation is employed to build scan representation. In the proposed framework, we employ row-wise DFT to aggregate the features into a global translation invariant representation. In

TABLE I: Place Recognition Performance Comparison

Dataset	Approach	Recall@1	F1 Score	AUC
NCLT	Scan Context	0.469	0.638	0.570
	RING	0.664	0.798	0.839
	PointNetVLAD	0.508	0.539	0.627
	DiSCO	0.793	0.887	0.849
	EgoNN	0.763	0.866	0.842
	DeepRING (ours)	**0.859**	**0.894**	**0.893**
MulRan	Scan Context	0.584	0.737	0.698
	RING	0.713	0.833	0.843
	PointNetVLAD	0.697	0.821	0.842
	DiSCO	0.743	0.863	0.897
	EgoNN	0.726	0.841	0.906
	DeepRING (ours)	**0.893**	**0.943**	**0.943**

* The compared methods are evaluated on the test trajectory.

addition, operations like global max pooling and average pooling can also arrive at translation invariance, which is widely used to aggregate features in many works. Therefore, we compare DFT against these two aggregation approaches according to the results in Tab. III. Global average pooling with *Recall@1* of 0.639 outperforms global max pooling with *Recall@1* of 0.534, because of the non-linearity caused by the max operation. Since the global pooling operation squeezes the translation dimension while DFT following by magnitude operation does not change the size of feature maps, we increase the kernel number of the last convolution layer to yield a multi-channel feature map so that the resultant representation after pooling keeps the same size as that after DFT. Compared to the improved global average pooling with *Recall@1* of 0.743, the DFT based method with *Recall@1* of 0.859 still improves the performance substantially. Therefore, we consider that DFT is learning-free, thus making it a better aggregation method to build data independent translation invariance while keeping discrimination.

Loss: The mainstream metric learning based methods take triplet loss or quadruplet loss as the loss function. For comparison, we replace the cross-entropy loss of our method with triplet loss and evaluate the final place recognition performance. The triplet loss based model results in inferior performance, with *Recall@1* decreasing from 0.859 to 0.644. The triplet loss is usually accompanied by Euclidean distance which does not support the rotation equivariance of DFT magnitude, but the cross-entropy loss of the proposed method actually finds the relative rotation and similarity at the same time. For a more fair comparison, we further employ an extra column-wise DFT after DFT magnitude to eliminate the effect of rotation. As a result, *Recall@1* increases to 0.747, which is still surpassed by our method. It demonstrates that the cross-entropy loss based on correlation distance is more suitable for roto-translation invariant feature learning.

Class Number for one-shot learning: With respect to one-shot learning algorithm, we conduct experiments on NCLT dataset to examine the effect of classes/ways number on place recognition performance. As shown in Tab. IV, the number of classes/ways used in the training stage does not obviously facilitate the final performance. We attribute this to the roto-translation invariance design of our method, which shrinks the intra-class variation. Therefore, the way number shows little influence on the place recognition performance and we take 24-way 1-shot learning algorithm for evaluation in all experiments.

TABLE II: Cross-domain Generalization Evaluation

Dataset	Approach	Recall@1	F1 Score	AUC
NCLT	Scan Context	0.507	0.673	0.629
	RING	0.708	0.829	0.847
	PointNetVLAD	0.369	0.539	0.626
	DiSCO	0.604	0.779	0.773
	EgoNN	0.562	0.696	0.799
	DeepRING (ours)	**0.797**	**0.887**	**0.881**
MulRan	Scan Context	0.524	0.688	0.653
	RING	0.764	0.866	0.883
	PointNetVLAD	0.561	0.719	0.722
	DiSCO	0.745	0.854	0.796
	EgoNN	0.663	0.797	0.850
	DeepRING (ours)	**0.826**	**0.904**	**0.901**

* The compared methods are evaluated on the whole trajectory. For generalization evaluation, the learning-based methods are trained on one dataset and generalized to the other dataset.

TABLE III: Ablation Study on Different Components

Ablation	Input	Aggregation	Loss	Recall@1
Representation	PG	DFT	CE	0.834
Aggregation	SG	GMP	CE	0.534
	SG	GAP	CE	0.639
	SG	Multi-GAP	CE	0.743
Loss	SG	DFT	Triplet	0.644
	SG	2DFT	Triplet	0.747
Ours	SG	DFT	CE	0.859

* PG: Polar Gram, SG: Sinogram, DFT: Discrete Fourier Transform, CE: Cross-Entropy, GMP: Global Max Pooling, GAP: Global Average Pooling, Multi-GAP: Multi-channel Global Average Pooling.

TABLE IV: Ablation Study on Class Number for Training

No. Class	8	16	24	32
Recall@1	0.834	0.855	**0.859**	0.855

VI. CONCLUSIONS

In this paper, we propose a novel framework to learn roto-translation invariant representation for LiDAR based place recognition. Leveraging Radon transform, we convert a LiDAR scan to a rotation equivariant representation SG. Based on circular convolutions and DFT, we exploit rotation equivariant and translation invariant feature maps. Considering place recognition as a one-shot learning problem, we learn the relation between inputs based on correlation distance metric for similarity comparison, which is roto-translation invariant. Thanks to the one-shot learning framework with the explicit design of roto-translation invariant representation, our method achieves outstanding performance and recognizes unseen places easily.

REFERENCES

[1] Radu Bogdan Rusu et al. "Fast 3d recognition and pose using the viewpoint feature histogram". In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2010, pp. 2155–2162.

[2] Timo Röhling, Jennifer Mack, and Dirk Schulz. "A fast histogram-based similarity measure for detecting loop closures in 3-d lidar data". In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2015, pp. 736–741.

[3] Giseop Kim and Ayoung Kim. "Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2018, pp. 4802–4809.

[4] Ying Wang et al. "Lidar iris for loop-closure detection". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 5769–5775.

[5] Giseop Kim, Sunwook Choi, and Ayoung Kim. "Scan context++: Structural place recognition robust to rotation and lateral variations in urban environments". In: *IEEE Transactions on Robotics* (2021).

[6] Lin Li et al. "SSC: Semantic scan context for large-scale place recognition". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021, pp. 2092–2099.

[7] Sha Lu et al. "One RING to Rule Them All: Radon Sinogram for Place Recognition, Orientation and Translation Estimation". In: *arXiv preprint arXiv:2204.07992* (2022).

[8] Mikaela Angelina Uy and Gim Hee Lee. "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4470–4479.

[9] Zhe Liu et al. "LPD-Net: 3D Point Cloud Learning for Large-Scale Place Recognition and Environment Analysis". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[10] Huan Yin et al. "Locnet: Global localization in 3d point clouds for mobile vehicles". In: *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2018, pp. 728–733.

[11] Xuecheng Xu et al. "Disco: Differentiable scan context with orientation". In: *IEEE Robotics and Automation Letters* 6.2 (2021), pp. 2791–2798.

[12] Xieyuanli Chen et al. "OverlapNet: Loop closing for LiDAR-based SLAM". In: *arXiv preprint arXiv:2105.11344* (2021).

[13] Junyi Ma et al. "OverlapTransformer: An Efficient and Rotation-Invariant Transformer Network for LiDAR-Based Place Recognition". In: *arXiv preprint arXiv:2203.03397* (2022).

[14] Lin Li et al. "RINet: Efficient 3D Lidar-Based Place Recognition Using Rotation Invariant Neural Network". In: *IEEE Robotics and Automation Letters* 7.2 (2022), pp. 4321–4328.

[15] Li He, Xiaolong Wang, and Hong Zhang. "M2DP: A novel 3D point cloud descriptor and its application in loop closure detection". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2016, pp. 231–237.

[16] Xiaqing Ding et al. "Translation Invariant Global Estimation of Heading Angle Using Sinogram of LiDAR Point Cloud". In: *arXiv preprint arXiv:2203.00924* (2022).

[17] Charles R Qi et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 652–660.

[18] Relja Arandjelovic et al. "NetVLAD: CNN architecture for weakly supervised place recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5297–5307.

[19] Kavisha Vidanapathirana et al. "Locus: Lidar-based place recognition using spatiotemporal higher-order pooling". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 5075–5081.

[20] Juan Du, Rui Wang, and Daniel Cremers. "Dh3d: Deep hierarchical 3d descriptors for robust large-scale 6dof relocalization". In: *European Conference on Computer Vision*. Springer. 2020, pp. 744–762.

[21] Jacek Komorowski, Monika Wysoczanska, and Tomasz Trzcinski. "EgoNN: Egocentric Neural Network for Point Cloud Based 6DoF Relocalization at the City Scale". In: *IEEE Robotics and Automation Letters* 7.2 (2021), pp. 722–729.

[22] Daniele Cattaneo, Matteo Vaghi, and Abhinav Valada. "Lcdnet: Deep loop closure detection and point cloud registration for lidar slam". In: *IEEE Transactions on Robotics* (2022).

[23] Wei-Yu Chen et al. "A closer look at few-shot classification". In: *arXiv preprint arXiv:1904.04232* (2019).

[24] Oriol Vinyals et al. "Matching networks for one shot learning". In: *Advances in neural information processing systems* 29 (2016).

[25] Jake Snell, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning". In: *Advances in neural information processing systems* 30 (2017).

[26] Flood Sung et al. "Learning to compare: Relation network for few-shot learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1199–1208.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[28] Gregory Koch et al. “Siamese neural networks for one-shot image recognition”. In: 2015.

[29] Diederik P Kingma. “&Ba J.(2014). Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2015).

[30] Nicholas Carlevaris-Bianco, Arash K Ushani, and Ryan M Eustice. “University of Michigan North Campus long-term vision and lidar dataset”. In: *The International Journal of Robotics Research* 35.9 (2016).

[31] Giseop Kim et al. “Mulran: Multimodal range dataset for urban place recognition”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 6246–6253.