ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Engineering Journals
Available on IEEE Xplore®

Research Article
https://doi.org/10.1631/ENG.ITEE.2025.0170

Relation-aware place recognition network for large-scale point clouds

Wen HAO[1,2]✉, Wenjing ZHANG[1], Yan LV[1], Li GUO[3]

[1]*Department of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China*
[2]*Shaanxi Key Laboratory for Network Computing and Security Technology, Xi'an 710048, China*
[3]*Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M15 6BX, UK*

Abstract: Most existing point cloud-based place recognition methods emphasize feature extraction from individual points or local regions, while largely neglecting the relational information embedded within local neighborhoods. As a result, they often fail to capture discriminative relational patterns, leading to reduced recognition accuracy in scenes containing geometrically similar structures. In this paper, we propose a novel relation-aware network (RA-Net) for place recognition. RA-Net jointly exploits local relational cues and global contextual information to learn discriminative scene representations for large-scale point cloud-based place recognition. First, a spatial relation feature extraction (SRFE) module is proposed to exploit relational information embedded within local neighborhoods. By learning relation-aware weights and adaptively aggregating neighborhood information, the proposed module captures discriminative relational patterns by jointly considering feature discrepancies and spatial offsets. Furthermore, a global feature extraction (GFE) module is introduced to aggregate global feature statistics and integrate them with pointwise representations, enabling local features to be enhanced with global contextual cues. Experimental results on four benchmark datasets demonstrate that RA-Net can generate more discriminative global descriptors and achieve promising performance. It exhibits strong generalization capabilities for unseen scenes.

Key words: Relation-aware; Point cloud; Place recognition; Deep learning

1 Introduction

Place recognition is a fundamental problem in robotics and autonomous navigation, aiming to retrieve the most similar place corresponding to a given query. Traditional visual-based place recognition methods (Zuo et al., 2025; Tang et al., 2026) rely on images, which can be sensitive to changes in illumination, viewpoint, and seasonal variations. In contrast, point cloud-based place recognition, leveraging 3D geometric structures, offers a more robust alternative that is invariant to lighting conditions and provides richer spatial information. With the increasing availability of light detection and ranging (LiDAR) and other 3D sensing technologies, point cloud-based place recognition has gained significant attention.

Early point cloud-based place recognition methods (Röhling et al., 2015; He et al., 2016; Rizzini, 2017; Kim and Kim, 2018) mainly depend on handcrafted features. These methods generate distinctive descriptors through statistical analysis, but they are highly sensitive to scene changes and exhibit limited generalization capability. Numerous researchers have applied convolutional neural networks to place recognition due to the rapid development of deep learning technology, yielding promising results. PointNetVLAD (Uy and Lee, 2018) is a pioneering deep learning network for point cloud-based place recognition that is built on PointNet (Qi et al., 2017). It leverages PointNet to extract local features of points and subsequently combines them into global descriptors using the NetVLAD (Arandjelovic et al., 2016) layer. After the pioneering work of PointNetVLAD, numerous efficient works such as pointwise multi-layer perceptron (MLP)-based methods (Zhang WX and Xiao, 2019; Cattaneo et al., 2022), graph-based methods (Liu Z et al., 2019; Sun et al., 2020), and Transformer-based methods (Ma et al., 2022, 2024; Li et al., 2024) have been proposed for place recognition tasks. Existing

✉ Wen HAO, haowensxsf@163.com
Wen HAO, https://orcid.org/0000-0002-0487-2784
Wenjing ZHANG, https://orcid.org/0009-0002-2201-5556
Yan LV, https://orcid.org/0009-0002-5568-617X
Li GUO, https://orcid.org/0000-0003-1272-8480
CLC number: TP391

Received: Dec. 8, 2025; Revision accepted: July 17, 2026;
Crosschecked: Aug. 28, 2026

© The Authors 2026. Published by Zhejiang University Press Co., Ltd. This is an open access article distributed under the terms of the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0/)

feature aggregation strategies typically treat all neighboring points equally, which limits their ability to capture informative contextual relationships within local neighborhoods. In large-scale urban environments, many scene elements, such as buildings, trees, and lamp posts, exhibit similar geometric characteristics across different locations. Consequently, relying solely on geometric feature representations may produce ambiguous scene representations. In contrast, the relative spatial arrangements and contextual interactions among neighboring structures often provide more distinctive cues for place recognition, as places with similar local appearances may still differ significantly in their spatial configurations.

Motivated by this observation, we propose a novel relation-aware network (RA-Net) for large-scale point cloud-based place recognition, which jointly exploits local relational information and global features to learn more discriminative scene representations. The main contributions of this paper are summarized as follows:

1. To enhance the discriminative capability of global descriptors, we propose a spatial relation feature extraction (SRFE) module that learns relation-aware contextual information within local neighborhoods. Instead of treating neighboring points equally during feature aggregation, the SRFE module adaptively exploits contextual cues by jointly considering feature discrepancies and relative spatial information among neighboring points. By capturing informative local relational features, the proposed module enables RA-Net to generate more robust and discriminative scene representations for large-scale place recognition.

2. A global feature extraction (GFE) module is proposed to enhance scene representation by incorporating global contextual information into pointwise features. Specifically, max pooling is employed to capture the most salient feature responses across the scene, while feature interaction between the global context representation and point features is used to enrich the feature embedding. This process leads to a more discriminative global descriptor for place recognition.

3. The effectiveness of the proposed approach is rigorously evaluated on four benchmark datasets, demonstrating superior performance compared to state-of-the-art (SOTA) methods. RA-Net achieves top-1 average recall rates of 92.71%, 95.48%, 87.90%, and 86.84% on the four benchmark datasets. The promising performance across unseen datasets highlights the model's strong generalization capability.

2 Related works

Numerous efficient methods for place recognition based on point clouds have been introduced in recent years (Zhang YJ et al., 2024). According to the form of data processing, these existing learning-based methods could be roughly divided into voxel-based methods and point-based methods.

1. Voxel-based methods

Voxel-based methods first divide the 3D points into a regular voxel representation. Three-dimensional convolution is then used to learn features in a hierarchical manner. Voxel-based representation learning (VBRL) (Siva et al., 2020) voxelizes the point clouds and extracts multiple types of features from voxels. It learns the importance of voxels and feature modalities by incorporating structured sparsity, inducing norms as regularizations into its optimization framework. MinkLoc3D (Komorowski, 2021) adapts a feature pyramid network (FPN) architecture to extract local features based on the sparse voxelized point cloud representation. Then, the local features are aggregated using a generalized-mean pooling layer. Later, EgoNN (Komorowski et al., 2022) uses the MinkLoc3D backbone and uses sparse 3D convolutions for extracting the global descriptors and local descriptors. MinkLoc++ (Komorowski et al., 2021) enhances the MinkLoc3D framework by integrating camera images and the sparse voxelization of 3D point clouds. A channel attention mechanism is used to improve the recognition performance. MinkLoc3Dv2 (Komorowski, 2022) improves the performance of MinkLoc3D by incorporating additional convolutional, transposed convolution blocks and an expanded number of channels. The series of MinkLoc3D networks consistently applies data augmentation techniques to boost performance. Three-dimensional point-voxel Transformer network (PVT3D) (Xia et al., 2022) combines a sparse voxel branch for low-resolution information extraction and a pointwise branch for capturing detailed local information. SVT-Net (Fan et al., 2022) leverages two different Transformers to effectively capture local features and contextual relationships between points based on the voxelization of point clouds. CASSPR (Xia et al., 2023) fuses point-based and voxel-based methods based on the cross-attention mechanism. The voxel branch extracts the geometric features by using a simplification of PointNet, and the point branch applies convolution on sparse voxels in spherical coordinates. LoGG3D-Net (Vidanapathirana et al., 2022) uses convolutional layers to learn local neighborhood details across different receptive field sizes in the voxel-based branch. Meanwhile, it employs MLP layers to retain high-resolution information in the point-based branch. The fusion of two branches ensures the comprehensiveness of extracted features. PTC-Net (Chen et al., 2023) uses sparse convolution (SPConv) to extract local voxel features and Transformers to capture global contextual features of points. A feature transformation unit (FTU) is used to convert features from a voxel format to a point-based format.

The performance of voxel-based methods is inherently influenced by voxel size. Voxel-based methods also ignore the impact of points with different categories on feature extraction.

2. Point-based methods

Point-based methods operate directly on the 3D points, preserving the original geometric structure of point clouds. These approaches always use PointNet and its variants to handle unordered point sets.

After PointNetVLAD, attention mechanisms are introduced to place recognition tasks. Point contextual attention network (PCAN) (Zhang WX and Xiao, 2019) explores an attention mechanism on the NetVLAD layer for feature aggregation, focusing on identifying and enhancing local features that are most important. Self-attention and orientation encoding network (SOE-Net) (Xia et al., 2021) proposes a point orientation encoding (PointOE) module to extract local features of points from eight orientations. Then an attention mechanism is introduced in the descriptor aggregation part, aiming at capturing long-range dependencies among local descriptors. However, encoding features based on a limited number

of orientations cannot fully explore the relationships between points. Point Transformer network with pyramid learnable tokens (PTNet-PLT) (Wen et al., 2023) uses two different attention-based modules to extract local features and aggregate the regional features based on the learnable tokens, separately. The number of tokens is reduced progressively at each layer while features are aggregated over larger regions increasingly. The final global descriptor is formed by concatenating regional tokens across all layers. HOTFormerLoc (Griffiths et al., 2025) leverages a hierarchical octree Transformer to learn multiscale scene representations for LiDAR place recognition. By introducing octree-based attention, cylindrical attention windows, and relay-token-based global–local interactions, the method effectively captures local structural information and long-range contextual dependencies. Recognizing the efficiency of graph networks in capturing underlying geometric and shape properties, efficient point cloud learning network (EPC-Net) (Hui et al., 2022) proposes a ProxyConv module for the aggregation of local geometric features in point clouds. It then uses a grouped vector of locally aggregated descriptors (VLAD) network to produce global descriptors. Lightweight parameter-shared network (LPS-Net) (Liu CX et al., 2024) proposes bidirectional perception units (BPUs) based on edge convolution and grouped self-attention modules to extract multiscale contextual features. Additionally, parameter-shared NetVLADs are used for descriptor aggregation. Spatial attention and geometric encoding network (SAGE-Net) (Hao et al., 2024) employs a graph neural network to explore the contextual information between features. At the same time, it leverages low-level geometric features for better discriminative ability. Hierarchical bidirected graph convolution network (HiBi-GCN) (Shu and Kwon, 2024) defines a hierarchical kernel based on a hierarchical graph structure using unsupervised clustering. It pools these hierarchical graphs in a fine-to-coarse manner by connecting nodes with pooling edges and then merges the pooled graphs back in a coarse-to-fine manner using fusing edges.

Relation-aware scene representation has emerged as an active research direction in place recognition and scene understanding. Instead of relying solely on geometric appearance, relation-aware methods represent scenes as graphs, where nodes correspond to semantic entities and edges encode their spatial or semantic relationships. Kong X et al. (2020) constructed semantic scene graphs from point clouds by preserving semantic attributes and topological relationships among scene elements. The resulting place recognition task is formulated as graph matching, where a graph similarity network is used to generate place recognition results. To incorporate high-level semantic and structural information into localization, SceneGraphLoc (Miao et al., 2025) represents environments as multi-modal 3D scene graphs, where nodes correspond to object instances and edges encode inter-object relationships. The method learns object-centric embeddings that facilitate cross-modal matching between query images and scene graph representations. Three-dimensional dynamic scene graphs (Rosinol et al., 2020) employ hierarchical graph structures to jointly model places, objects, and dynamic agents for spatial perception and reasoning. Li et al. (2024) introduced a scene graph transformation network that combines dynamic object filtering and a graph Transformer for place recognition. By encoding the point cloud scene with a graph structure, it employs the Transformer-attention mechanism to extract spatial and semantic features. TopSPR-Net (Kong D et al., 2024) explores topology-aware place recognition by representing a scene as a collection of segmented point cloud regions. It leverages Transformer-based interactions among segment-level features to capture topological dependencies and spatial relationships between different scene components.

Although these approaches demonstrate the effectiveness of relational modeling, most of them rely on explicit graph construction and graph matching procedures. These procedures often require reliable object extraction and semantic representations.

Different from these methods, RA-Net does not explicitly construct a graph to represent objects or scene components. Instead, spatial dependencies are modeled directly in the point feature space through relation affinity learning based on feature representations and spatial offsets among neighboring points. The learned relations are subsequently used to guide local feature aggregation, allowing the network to capture distinctive structural patterns directly from point clouds.

3 Methods

Fig. 1 shows the overall network architecture of RA-Net. First, the coordinates (x, y, z) of points are fed into three layers of MLPs to obtain the 128-dimensional features. The 128-dimensional features are then split into two 64-dimensional features $\boldsymbol{F}_{\mathrm{u}}$ and $\boldsymbol{F}_{\mathrm{d}}$ for extracting the spatial relation features and global features, respectively.

The feature $\boldsymbol{F}_{\mathrm{u}}$ and the 3D coordinates of points are fed into the SRFE module. First, the relation-aware weight estimation (RWE) unit learns relation-aware weights within local neighborhoods by jointly considering feature representations and relative spatial information. Subsequently, the relation feature extraction (RFE) unit exploits feature differences and relative geometric information among neighboring points under the guidance of the learned relation-aware weight. By emphasizing relational patterns, the RFE unit captures informative spatial relation patterns that are not fully represented by local feature similarity alone.

At the same time, the feature $\boldsymbol{F}_{\mathrm{u}}$ is fed into the GFE module, which aggregates local features using max pooling. The max-pooling operation helps capture the most prominent features $\boldsymbol{F}_{\mathrm{m}}$. The feature $\boldsymbol{F}_{\mathrm{d}}$ is multiplied by the prominent features $\boldsymbol{F}_{\mathrm{m}}$. The result is added to the reshaped $\boldsymbol{F}_{\mathrm{m}}$ to obtain the enhanced feature $\boldsymbol{F}_{\mathrm{a}}$. The original features $\boldsymbol{F}_{\mathrm{d}}$ and $\boldsymbol{F}_{\mathrm{a}}$ are then concatenated and fed into an MLP layer. The GFE module effectively enhances key features to improve the uniqueness of the global descriptor.

Finally, a NetVLAD module aggregates the features to produce a global descriptor from the scene. RA-Net leverages the benefits of spatial relation features and globally important features for ensuring distinctiveness and robustness of the global descriptor.

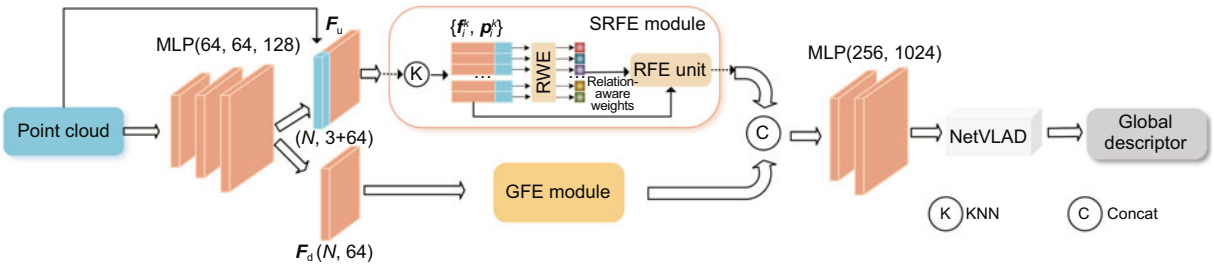

Fig. 1 Overview of the RA-Net architecture. N denotes the number of points, and k denotes the number of nearest neighbors. KNN: k-nearest neighbor

3.1 Spatial relation feature extraction

Recent point cloud-based place recognition methods primarily focus on learning discriminative point features and aggregating them into global scene descriptors. However, they often overlook the spatial relationships among neighboring points, which contain valuable structural cues for characterizing scene layouts and improving place discrimination.

Inspired by heterogeneous context modeling (Lu et al., 2021), an SRFE module is introduced to capture relational patterns within local neighborhoods in large-scale scenes. Unlike previous approaches that emphasize local similarity, the proposed method learns relation-aware weights between neighboring points by jointly considering their feature representations and spatial offsets.

Specifically, local relational features are extracted from the geometric displacements and feature differences between neighboring points. The learned relation-aware weights are subsequently used to aggregate these local relational features, enabling the network to capture informative contextual relationships beyond conventional feature similarity.

3.1.1 Relation-aware weight estimation unit

Suppose that $\boldsymbol{f}_i$ is the i^{th} feature in $\boldsymbol{F}_{\text{u}}$, and its k-nearest neighboring features are denoted as $\boldsymbol{f}_j$. For the corresponding point $\boldsymbol{p}_i$, its k-nearest neighbors in the 3D space are denoted as $\boldsymbol{p}_j$, where $\boldsymbol{p}_j \in N(i)$ and $N(i)$ represents the local neighborhood of $\boldsymbol{p}_i$.

The RWE unit learns a relation-aware weight β_{ij} between point $\boldsymbol{p}_i$ and its neighboring point $\boldsymbol{p}_j$ based on feature representations and relative spatial information:

$$\beta_{ij} = \text{Softmax}\left(\boldsymbol{W}_{\text{b}} \cdot \phi\left([\boldsymbol{f}_i \parallel \boldsymbol{f}_j \parallel \Delta\boldsymbol{p}_{ij}]\right)\right)_1, \ \Delta\boldsymbol{p}_{ij} = \boldsymbol{p}_j - \boldsymbol{p}_i, \tag{1}$$

where $\phi(\cdot)$ denotes an MLP for extracting local relational features, and "$\parallel$" denotes the feature concatenation. $\boldsymbol{W}_{\text{b}}$ is a learnable projection matrix, and $\text{Softmax}(\cdot)_1$ denotes the first value of the 2D softmax output.

β_{ij} is adopted as a relational weight. It measures the contribution of a neighboring point to the contextual representation of the local neighborhood.

A larger β_{ij} indicates that the neighboring point exhibits stronger relational distinctiveness with respect to the center point and provides contextual information that is less redundant with the dominant local neighborhood. In contrast, smaller values suggest that the neighboring point shares more consistent local contextual characteristics with the center point. Such adaptive weighting enables the model to emphasize informative relational cues while suppressing redundant ones within local neighborhoods. These relational cues are particularly valuable for place recognition, where the uniqueness of a location is determined not only by locally consistent structures but also by the spatial arrangement of surrounding elements.

3.1.2 Relation feature extraction unit

Instead of relying on semantic supervision, the RFE unit identifies structurally informative neighbors based on the learned relational cues, enabling the extraction of discriminative local relational features in large-scale point cloud scenes. Fig. 2 shows the network of the RFE unit.

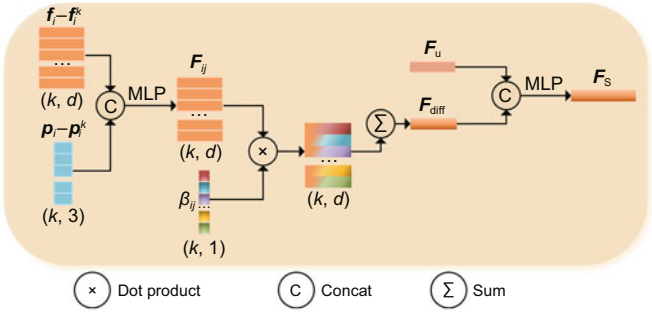

Fig. 2 Network of the RFE unit. d denotes the dimension of the features

For each point, relation descriptors are constructed by encoding feature differences and geometric displacements with respect to its neighboring points:

$$\boldsymbol{F}_{ij} = \phi\left([(\boldsymbol{f}_j - \boldsymbol{f}_i) \parallel (\boldsymbol{p}_j - \boldsymbol{p}_i)]\right), \tag{2}$$

where $\phi(\cdot)$ is an MLP that encodes spatial and feature discrepancies into a latent relational feature space.

β_{ij} is used as an adaptive weighting factor to aggregate relation descriptors:

$$\boldsymbol{F}_{\text{diff}} = \frac{\sum_{j\in N(i)} \beta_{ij} \boldsymbol{F}_{ij}}{\sum_{j\in N(i)} \beta_{ij}}. \tag{3}$$

In this manner, the RFE unit emphasizes relational patterns among neighboring points. These relational patterns provide valuable cues for place recognition in large-scale urban environments.

Finally, the original feature $\boldsymbol{F}_{\text{u}}$ and the aggregated relation feature $\boldsymbol{F}_{\text{diff}}$ are concatenated and passed through an

MLP to produce the final relation feature:

$$\boldsymbol{F}_{\mathrm{S}} = \phi\left([\boldsymbol{F}_{\mathrm{u}} \parallel \boldsymbol{F}_{\mathrm{diff}}]\right). \tag{4}$$

The resulting feature $\boldsymbol{F}_{\mathrm{S}}$ is obtained by combining the original point feature with learned relational cues from neighboring points, enabling the network to capture discriminative relational patterns within local neighborhoods and generate more distinctive scene representations.

3.2 Global feature extraction

The SRFE module learns discriminative relation-aware features by capturing relational information from feature discrepancies and spatial offsets between neighboring points. While these features effectively encode local structural relationships, they are generated from local neighborhoods only. To complement the local representations, a GFE module is proposed to summarize the overall feature distribution of the input point cloud. As we know, max pooling helps aggregate features from local neighborhoods, retaining the most distinctive ones. Therefore, we design a GFE module that employs max pooling to extract key features. We continuously enhance these significant features while aggregating them with the original features. The goal is to obtain global features that possess strong uniqueness and robustness.

Fig. 3 shows the network of the GFE module. For a given local feature $\boldsymbol{F}_{\mathrm{d}}$, the max-pooling operation is used to extract the important features. The reasons we use max pooling over more sophisticated pooling mechanisms are as follows: (1) max pooling is significantly cheaper in terms of computation and memory compared to attention mechanisms, which often require additional parameters and matrix operations; (2) our goal is to identify prominent features rather than model fine-grained relationships. Therefore, max pooling is beneficial because it retains the most significant features while discarding less important ones:

$$\boldsymbol{F}_{\mathrm{m}} = \max_\mathrm{pooling}(\boldsymbol{F}_{\mathrm{d}}). \tag{5}$$

Then, the important features are multiplied by the feature $\boldsymbol{F}_{\mathrm{d}}$. The reshaped feature $\boldsymbol{F}_{\mathrm{m}}^{\mathrm{T}}$ is added to update the feature, resulting in an aggregated feature $\boldsymbol{F}_{\mathrm{a}}$:

$$\boldsymbol{F}_{\mathrm{a}} = \boldsymbol{F}_{\mathrm{m}}^{\mathrm{T}} + \boldsymbol{F}_{\mathrm{m}} \times \boldsymbol{F}_{\mathrm{d}}, \tag{6}$$

where "×" denotes the dot product.

Next, the local feature $\boldsymbol{F}_{\mathrm{d}}$ and the feature $\boldsymbol{F}_{\mathrm{a}}$ are concatenated and fed into an MLP. The final global feature $\boldsymbol{F}_{\mathrm{G}}$ is obtained:

$$\boldsymbol{F}_{\mathrm{G}} = \phi\left([\boldsymbol{F}_{\mathrm{a}} \parallel \boldsymbol{F}_{\mathrm{d}}]\right). \tag{7}$$

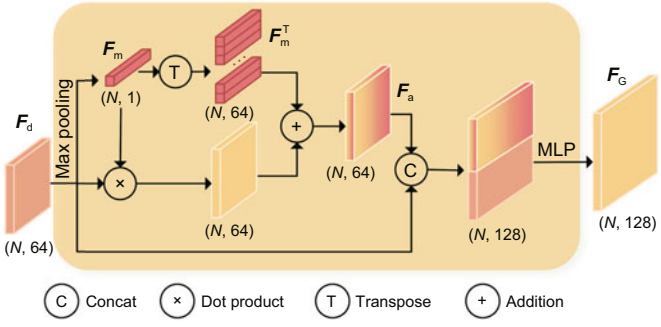

Fig. 3 Network of the GFE module

Finally, the spatial relation features $\boldsymbol{F}_{\mathrm{S}}$ and the final global feature $\boldsymbol{F}_{\mathrm{G}}$ are obtained and fed into two MLP layers.

4 Experiments

4.1 Experimental datasets and evaluation

Similar to PointNetVLAD, the same evaluation metric average recall at top N is used to evaluate the accuracy of place recognition. Average Recall@1% (AR@1%) measures the mean recall rate for correct matches found within the top 1% of retrieved candidate submaps from the test dataset. Meanwhile, Average Recall@1 (AR@1) indicates the average recall rate of true values for the first candidate point cloud retrieved by the network. As introduced in PointNetVLAD, we also adopt the Oxford RobotCar dataset (hereinafter referred to as the Oxford dataset) (Maddern et al., 2017) and three in-house datasets (Uy and Lee, 2018) to evaluate the RA-Net. The Oxford dataset includes 44 sequences along the same route and is split into 21 711 submaps for training and 3030 submaps for testing. For three in-house datasets, they are collected in different environments: a university sector (U.S.), a residential area (R.A.), and a business district (B.D.). Each region is recorded across five sequences captured at various times. In our experiments, all three in-house datasets are used for testing, aiming to validate the generalization of the proposed RA-Net.

After removing the non-informative ground planes, each submap is downsampled to 4096 points for evaluation. At the same time, the coordinate points are normalized to the range of $[-1, 1]$. A successful localization is determined if the distance between a query submap and its retrieved match is within 25 m.

We use the hard positive hard negative (HPHN) quadruplet loss (Xia et al., 2021) to train the network. All experiments of RA-Net are tested on a single NVIDIA GeForce RTX 2080Ti graphics processing unit (GPU) with 11 GB memory. Each input point cloud contains 4096 points, and the final output global descriptor is 256-dimensional.

4.2 Results and discussion

4.2.1 Result comparison

To verify the effectiveness of our method, RA-Net is compared with SOTA methods: PointNetVLAD (Uy and Lee, 2018), PCAN (Zhang WX and Xiao, 2019), LPD-Net (Liu Z et al., 2019), SVT-Net (Fan et al., 2022), SOE-Net (Xia et al., 2021), EPC-Net (Hui et al., 2022), HiBi-Net (Shu and Kwon, 2024), DH3D (Du et al., 2020), SAGE-Net (Hao et al., 2024), HiTPR-F8 (Hou et al., 2022), and LPS-Net (Liu CX et al., 2024). To ensure a fair comparison of recognition performance, LPS-Net is re-evaluated on the same hardware platform. The results of LPS-Net in Tables 1 and 2 are obtained by running it on the same computer as RA-Net. Tables 1 and 2 summarize the quantitative comparison of RA-Net with other SOTA methods on the four datasets.

In terms of AR@1% (Table 1), RA-Net reaches 97.43%, 98.58%, 94.62%, and 91.38% on the Oxford, U.S., R.A., and B.D. datasets, respectively. RA-Net achieves 1.07 percentage points (PPs), 1.04 PPs, 1.62 PPs, and 0.66 PPs higher

AR@1% scores than SAGE-Net on the four datasets. Compared with LPS-Net, RA-Net achieves performance gains of 0.42 PPs, 0.45 PPs, 1.27 PPs, and 0.97 PPs on the Oxford, U.S., R.A., and B.D. datasets. These improvements demonstrate the effectiveness of RA-Net in capturing discriminative local relational patterns and incorporating global contextual information. Such information is particularly useful in complex urban environments, where similar geometric structures may exhibit different spatial arrangements. Although RA-Net is trained exclusively on the Oxford training set and does not use any data from the three in-house datasets, it also performs well on all three in-house datasets. It can be concluded that our RA-Net has good generalization capabilities for unseen scenes.

Table 1 AR@1% for different models trained on the Oxford dataset

Method	AR@1% (%)			
	Oxford	U.S.	R.A.	B.D.
PointNetVLAD (Uy and Lee, 2018)	81.01	77.82	69.75	65.30
PCAN (Zhang WX and Xiao, 2019)	83.90	79.13	71.15	66.90
DH3D (Du et al., 2020)	84.26	81.50	72.10	68.39
LPD-Net (Liu Z et al., 2019)	94.53	95.93	90.72	87.50
EPC-Net (Hui et al., 2022)	95.20	95.42	88.10	84.24
SOE-Net (Xia et al., 2021)	96.40	93.17	91.47	88.45
SVT-Net (Fan et al., 2022)	97.80	96.50	92.70	90.70
HiTPR-F8 (Hou et al., 2022)	94.64	94.01	89.11	88.31
SAGE-Net (Hao et al., 2024)	96.36	97.54	93.00	90.72
LPS-Net (Liu CX et al., 2024)	97.01	98.13	93.35	90.41
RA-Net	97.43	98.58	94.62	91.38

Table 2 AR@1 for different models trained on the Oxford dataset

Method	AR@1 (%)			
	Oxford	U.S.	R.A.	B.D.
PointNetVLAD (Uy and Lee, 2018)	62.76	63.01	56.18	57.21
PCAN (Zhang WX and Xiao, 2019)	69.05	62.49	56.99	58.14
DH3D (Du et al., 2020)	73.28	64.44	57.92	60.08
LPD-Net (Liu Z et al., 2019)	85.62	85.68	83.39	80.89
EPC-Net (Hui et al., 2022)	86.86	86.33	79.18	77.79
SOE-Net (Xia et al., 2021)	89.37	82.46	82.92	83.33
SVT-Net (Fan et al., 2022)	93.70	90.10	84.30	85.50
HiTPR-F8 (Hou et al., 2022)	87.77	86.00	81.32	81.80
SAGE-Net (Hao et al., 2024)	90.04	93.03	86.08	86.08
HiBi-Net (Shu and Kwon, 2024)	87.46	87.81	85.76	83.03
LPS-Net (Liu CX et al., 2024)	91.56	92.14	85.49	84.70
RA-Net	92.71	95.48	87.90	86.84

Similarly, the AR@1 results of RA-Net and other advanced methods are shown in Table 2. The metric AR@1 indicates the matching rate of the top point cloud retrieved using different models, which provides a more intuitive reflection of the performance of place recognition. AR@1 of RA-Net on the U.S. dataset reaches 95.48%, which is significantly better than those of other SOTA methods. Compared to HiBi-Net, RA-Net can obtain gains with a margin of 2.14–7.67 PPs on AR@1. LPD-Net relies on handcrafted features, which are expected to perform well on datasets with regular structures. However, RA-Net still outperforms LPD-Net across three in-house datasets. Compared with SAGE-Net, RA-Net achieves substantial performance improvements of 2.67 PPs, 2.45 PPs, 1.82 PPs, and 0.76 PPs on the Oxford, U.S., R.A., and B.D. datasets, respectively. Experimental results demonstrate that RA-Net can generate more discriminative global descriptors for the retrieval task and has significant generalization capabilities.

Furthermore, we show the recall curves of PointNetVLAD, LPD-Net, EPC-Net, SOE-Net, LPS-Net, and RA-Net for the top-25 retrieval results on four benchmark datasets in Fig. 4. Obviously, RA-Net achieves superior accuracy compared to the other four SOTA methods. It is worth noting that RA-Net achieves the highest AR@1 among all compared methods, which enhances its applicability in real-world situations.

4.2.2 Result visualization

Additionally, we provide visualizations of the query submap and the top-3 retrieval results for the benchmark datasets. A query point cloud is randomly sampled from a complete run, and a different run is designated as the test map. The most similar submap within the test map is then retrieved based on the query point cloud. We select several representative queries of the Oxford dataset with different structures in Fig. 5 to verify the retrieval capability of RA-Net.

Fig. 5 shows the retrieval results of RA-Net on the Oxford dataset. The query point clouds in Figs. 5a and 5b both contain buildings and trees, while the buildings in Fig. 5b are severely incomplete. RA-Net successfully retrieves the top-1 submap that matches the query point cloud. The inherent instability in tree structures caused by the gaps between leaves can result in significant variations in the captured 3D structure, even with minimal sensor movement. However, our RA-Net is also able to find the correct match for the scene with the unstable trees and incomplete buildings. Fig. 5c contains low shrubs, a regularly shaped building wall, and curbs. The minimal geometric variation of planar walls inherently makes it challenging to extract distinctive features. Nevertheless, RA-Net effectively captures subtle differences and produces highly discriminative descriptors.

Fig. 6 shows the retrieval results of RA-Net on the three in-house datasets. Figs. 6a–6c show the top-3 retrieval results of RA-Net on the U.S., R.A., B.D. datasets, respectively. The Oxford dataset provides large-scale urban driving scenarios with diverse environmental conditions, while the three in-house datasets are collected in more structured and smaller-scale environments such as university campuses and nearby districts. They represent challenging environments characterized by numerous geometrically similar structures and repetitive

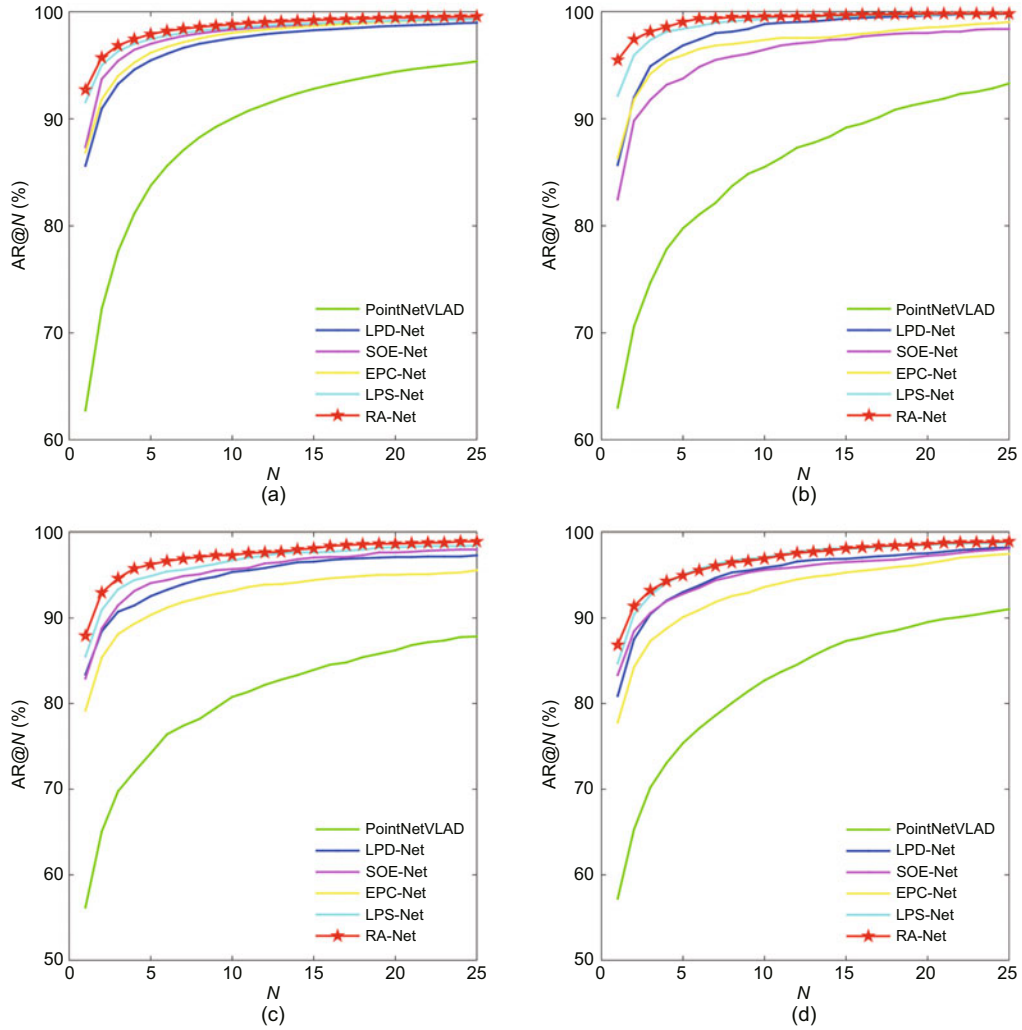

Fig. 4 AR@*N* of PointNetVLAD, LPD-Net, SOE-Net, EPC-Net, LPS-Net, and RA-Net on the Oxford (a), U.S. (b), R.A. (c), and B.D. (d) datasets

architectural patterns, making place discrimination particularly difficult. As we can see, RA-Net is also able to find the correct match for the campus-based environment.

Fig. 7 displays the top-1 retrieval results of various methods on the Oxford dataset. Despite the query point clouds in Fig. 7a containing structurally diverse trees and incomplete building walls, RA-Net successfully retrieves accurate results. The query point clouds in Fig. 7b primarily consist of regularly shaped buildings. Notably, a tree is located in front of the building on the right. RA-Net retrieves the correct scene, whereas EPC-Net and LPD-Net only capture the buildings, neglecting the trees present in the query one. This capability can be attributed to its dual consideration of relational features and comprehensive global features. By jointly exploiting these two types of features, RA-Net can effectively distinguish between structurally similar environments. In Fig. 7c, the query submap contains a significant number of tree point clouds and low-lying planes, leading EPC-Net and LPD-Net to find the buildings that do not match the query one. The retrieval results highlight that compared to EPC-Net and LPD-Net, RA-Net effectively extracts salient features from different scenes, ensuring higher retrieval accuracy.

4.2.3 Ablation studies

The ablation studies are conducted to evaluate the effectiveness of different components in RA-Net, including the SRFE module, the GFE module, and the pooling strategy. These experiments are performed across all benchmark datasets, with the results summarized in Table 3.

Table 3 Ablation studies of three components on four datasets

Method	AR@1% (%)			
	Oxford	U.S.	R.A.	B.D.
RA-Net[-SRFE]	84.29	84.90	79.64	74.00
RA-Net[-GFE]	96.74	95.93	92.40	87.90
RA-Net[-MP]	97.23	97.48	93.49	88.53
RA-Net	97.43	98.58	94.62	91.38

1. SRFE module

We only use the 64-dimensional feature $\boldsymbol{F}_{\mathrm{u}}$ as the output of the top branch in RA-Net to verify the importance of the SRFE module. AR@1% of RA-Net[-SRFE] is 13.14 PPs, 13.68 PPs, 14.98 PPs, and 17.38 PPs lower than that of RA-Net on the Oxford, U.S., R.A., and B.D. datasets, respectively. The

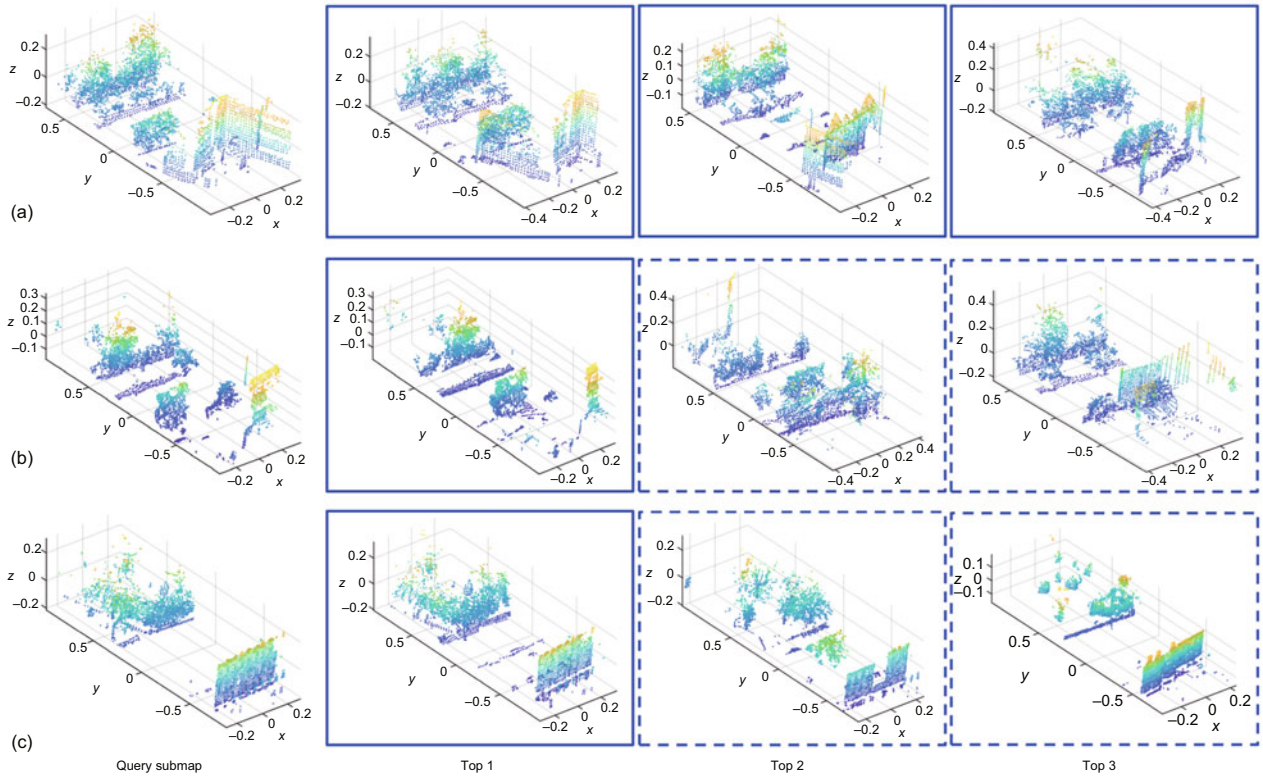

Fig. 5 Visualizations of retrieval results of RA-Net on the Oxford dataset. In these visualizations, the first column displays three different query point clouds, while the following three columns show the top-3 retrieved submaps. The query point clouds contain buildings and trees (a), highly incomplete buildings (b), and low shrubs, a regularly shaped building wall, and curbs (c). The solid-line quadrilaterals denote positive matches, and the dotted-line quadrilaterals denote incorrect matches

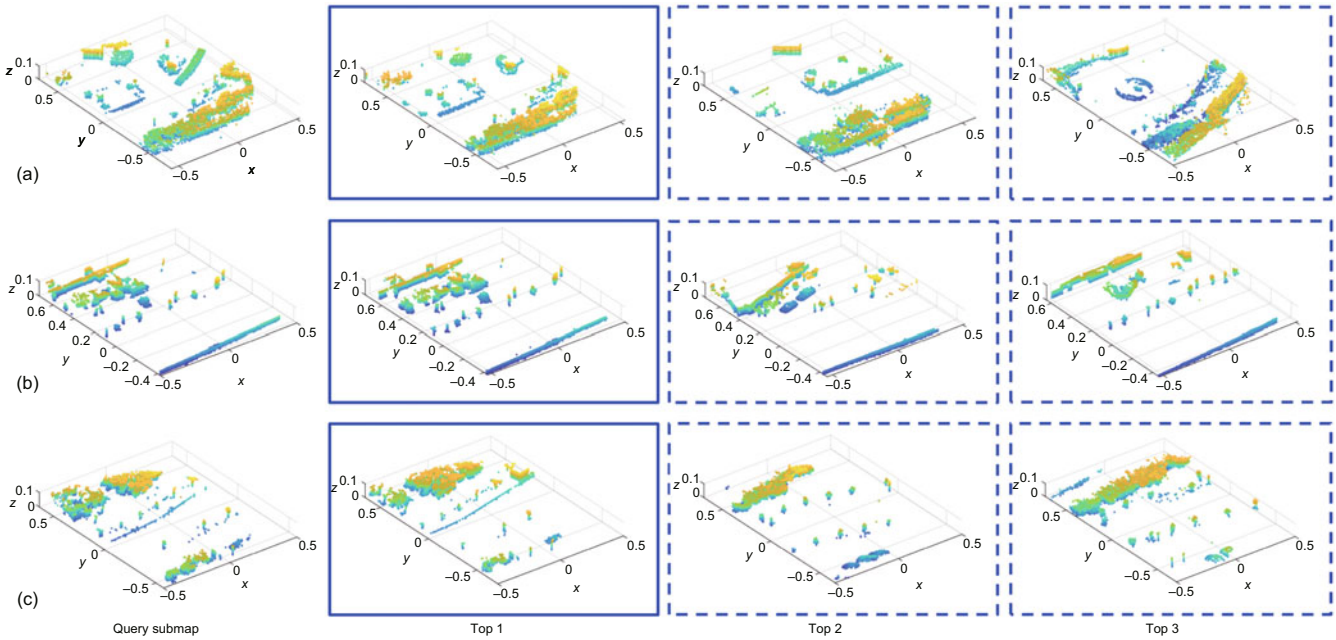

Fig. 6 Visualizations of retrieval results of RA-Net on the U.S. (a), R.A. (b), and B.D. (c) datasets. The solid-line quadrilaterals denote positive matches, and the dotted-line quadrilaterals denote incorrect matches

experimental results indicate that capturing spatial relationships is essential for robust place recognition. Without explicitly capturing relational patterns among local regions, RA-Net struggles to distinguish scenes with similar objects but different spatial configurations, leading to a noticeable degradation in recognition performance. Thanks to the SRFE module, our RA-Net achieves competitive performance on large-scale place recognition compared to SOTA methods.

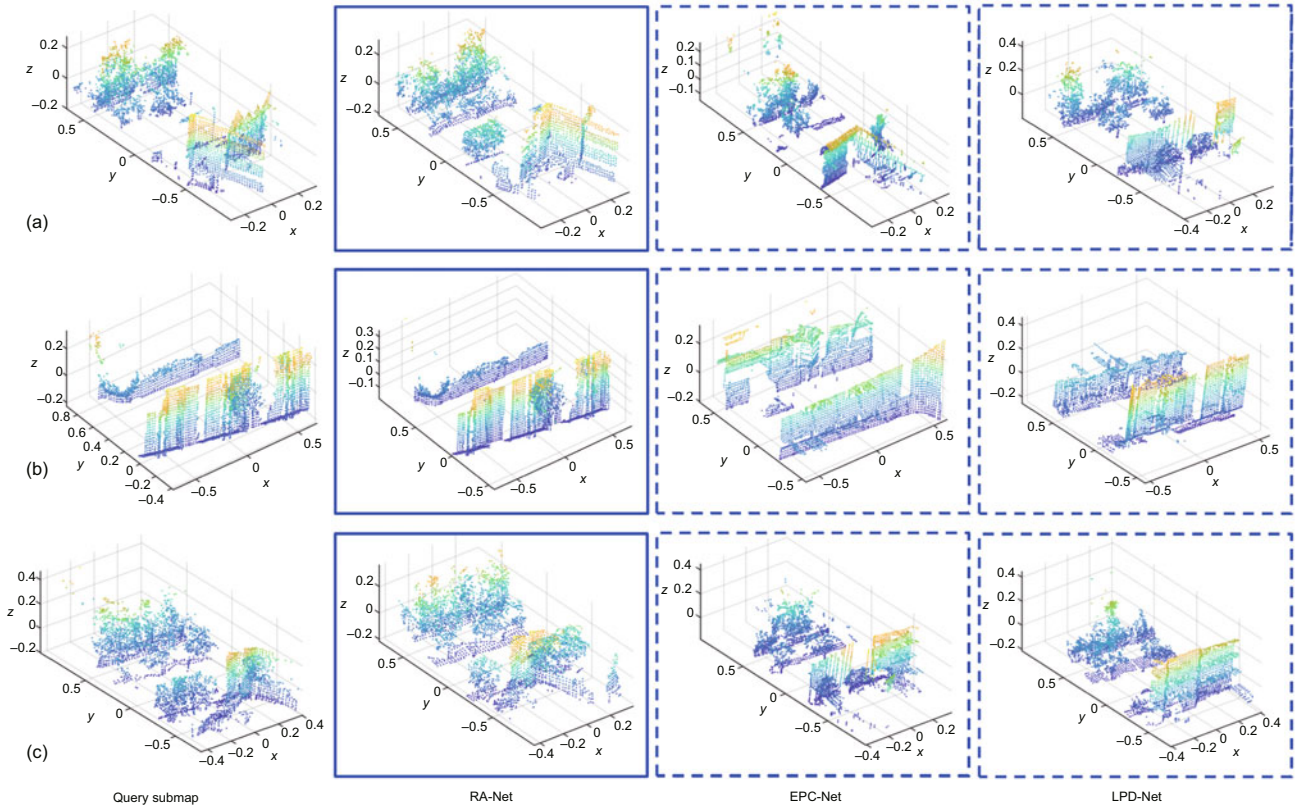

Fig. 7 Comparison of different methods on the Oxford dataset. The first column presents different query submaps, and the subsequent three columns show the retrieval results for our RA-Net, EPC-Net, and LPD-Net. The query point clouds contain structurally diverse trees and incomplete building walls (a), regularly shaped buildings (b), and trees and planes (c). The solid-line quadrilaterals represent successful retrievals, while dotted-line quadrilaterals indicate failures

2. GFE module

We conduct experiments to verify the effectiveness of the GFE module. The 64-dimensional feature $\boldsymbol{F}_{\mathrm{d}}$ is used as the output of the bottom branch in RA-Net. In Table 3, we can see that after removing the GFE module, AR@1% of RA-Net[-GFE] is 0.69 PPs, 2.65 PPs, 2.22 PPs, and 3.48 PPs lower than that of RA-Net on the Oxford, U.S., R.A., and B.D. datasets, respectively. It indicates that the GFE module has a significant effect on three in-house datasets. The enhancement of key features of points can greatly increase the recognition performance and has good generalization ability to unseen datasets.

3. Pooling strategy

To evaluate the influence of the pooling strategy, we replace the max pooling in the global feature aggregation module with mean pooling while keeping all other settings unchanged. Experimental results (RA-Net[-MP]) show that max pooling consistently achieves better performance than mean pooling.

This improvement can be attributed to the fact that place recognition mainly relies on a subset of highly discriminative structures rather than uniformly distributed scene information. Max pooling emphasizes the strongest feature activations and preserves the most distinctive scene cues, which are critical for distinguishing places with similar appearances. In contrast, mean pooling averages all features, potentially weakening salient scene characteristics and introducing redundant contextual information.

4.2.4 Computational cost analysis

Table 4 presents a comparison of the computational cost of PointNetVLAD, LPD-Net, SOE-Net, EPC-Net, and RA-Net in terms of the number of parameters, giga floating-point of operations (GFLOPs), runtime per frame, and GPU memory consumption. Runtime per frame denotes the average descriptor extraction time for a single point cloud containing 4096 points.

As shown in Table 4, RA-Net has 17.35×10^6 parameters, fewer than PointNetVLAD, LPD-Net, and SOE-Net. In terms of computational complexity, RA-Net requires 4.55 GFLOPs, which is substantially lower than those of LPD-Net and SOE-Net, while only slightly higher than those of PointNetVLAD and EPC-Net.

Despite incorporating the proposed relation-aware feature extraction mechanism, RA-Net achieves the fastest inference speed among the methods with reported runtime statistics, requiring only 9.58 ms per frame. Furthermore, the GPU memory consumption of RA-Net is 5.52 GB, which is comparable to that of PointNetVLAD and significantly lower than that of SOE-Net. These results demonstrate that the proposed relation-aware modeling introduces only a modest computational overhead while maintaining efficient runtime performance and memory usage.

Compared with EPC-Net, RA-Net introduces additional parameters and computational cost. The number of parameters increases from 4.70×10^6 to 17.35×10^6, while the

Table 4 Computational cost comparison of different methods

Method	Number of parameters ($\times 10^6$)	GFLOPs	Runtime per frame (ms)	GPU memory (GB)
PointNetVLAD (Uy and Lee, 2018)	19.78	4.21	11.32	5.28
LPD-Net (Liu Z et al., 2019)	19.81	7.89	18.91	10.66
SOE-Net (Xia et al., 2021)	19.39	26.31		16.88
EPC-Net (Hui et al., 2022)	4.70	3.25		4.83
RA-Net	17.35	4.55	9.58	5.52

GFLOPs increase from 3.25 to 4.55. Nevertheless, the increase in GPU memory consumption is relatively moderate, rising from 4.83 GB to 5.52 GB. More importantly, this additional computational overhead yields substantial improvements in recognition performance. RA-Net achieves AR@1 gains of 5.85 PPs, 9.15 PPs, 8.72 PPs, and 9.05 PPs on the Oxford, U.S., R.A., and B.D. datasets, respectively. Notably, the larger improvements on the more challenging residential and business district datasets indicate the effectiveness of the proposed relation-aware representation in capturing discriminative spatial contextual information.

Different from graph-based relational representations, the proposed RA-Net adopts a lightweight relation-aware feature extraction mechanism rather than explicit graph construction and graph message propagation. Graph-based methods typically require constructing scene graphs and performing neighborhood aggregation over graph edges, resulting in computational complexity that grows with the number of nodes and relationships. In contrast, the proposed RFE module models local spatial relationships directly in the feature space through relation affinity learning, avoiding the additional overhead associated with graph generation and graph convolution operations. Consequently, RA-Net achieves favorable efficiency and scalability while preserving the benefits of relation-aware representation learning. Nevertheless, explicit graph-based approaches may provide stronger modeling capability for long-range object interactions and highly structured scene layouts, which remains a promising direction for future research.

5 Conclusions

In this paper, we propose a novel point cloud-based place recognition network RA-Net. By learning relation-aware local representations and incorporating global contextual information, the proposed method generates more discriminative scene descriptors for robust large-scale place recognition. Extensive experiments on benchmark datasets demonstrate that the proposed RA-Net achieves superior performance compared to SOTA methods. Future work will focus on improving the robustness of RA-Net under challenging conditions such as severe point cloud sparsity, occlusions, and dynamic scene interference. Additionally, we aim to explore the integration of multi-modal sensory data and investigate more efficient feature aggregation strategies to further improve scalability and real-time applicability.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61602373), the Shaanxi Provincial Department of Education Youth Innovation Team Project (No. 24JP116), and the China Scholarship Council Program.

Author contributions

Wen HAO designed the study and supervised the research. Wenjing ZHANG conducted the experiments. Yan LV assisted with the data processing and validation. Wen HAO and Wenjing ZHANG drafted the paper. Li GUO provided critical revisions to the paper. Wen HAO revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration on the use of generative AI tools

During the preparation of this work, the authors used ChatGPT solely to improve language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Arandjelovic R, Gronat P, Torii A, et al., 2016. NetVLAD: CNN architecture for weakly supervised place recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5297-5307. https://doi.org/10.1109/CVPR.2016.572

Cattaneo D, Vaghi M, Valada A, 2022. LCDNet: deep loop closure detection and point cloud registration for LiDAR SLAM. *IEEE Trans Robot*, 38(4):2074-2093. https://doi.org/10.1109/TRO.2022.3150683

Chen LN, Wang H, Kong H, et al., 2023. Point-wise transformer with sparse convolution network for place recognition. *IEEE Robot Autom Lett*, 8(6):3414-3421. https://doi.org/10.1109/LRA.2023.3267693

Du J, Wang R, Cremers D, 2020. DH3D: deep hierarchical 3D descriptors for robust large-scale 6DoF relocalization. 16th European Conf on Computer Vision, p.744-762. https://doi.org/10.1007/978-3-030-58548-8_43

Fan ZX, Song ZB, Liu HY, et al., 2022. SVT-Net: super light-weight sparse voxel transformer for large scale place recognition. Proc 36th AAAI Conf on Artificial Intelligence, p.551-560. https://doi.org/10.1609/aaai.v36i1.19934

Griffiths E, Haghighat M, Denman S, et al., 2025. HOTFormerLoc: hierarchical octree transformer for versatile LiDAR place recognition across ground and aerial views. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6648-6658. https://doi.org/10.1109/cvpr52734.2025.00623

Hao W, Zhang WJ, Jin HY, 2024. SAGE-Net: employing spatial attention and geometric encoding for point cloud based place recognition. *IEEE Robot Autom Lett*, 9(6):4958-4965. https://doi.org/10.1109/LRA.2024.3387112

He L, Wang XL, Zhang H, 2016. M2DP: a novel 3D point cloud descriptor and its application in loop closure detection. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.231-237. https://doi.org/10.1109/iros.2016.7759060

Hou ZX, Yan Y, Xu CZ, et al., 2022. HiTPR: hierarchical transformer for place recognition in point cloud. Int Conf on Robotics and Automation, p.2612-2618. https://doi.org/10.1109/icra46639.2022.9811737

Hui L, Cheng MM, Xie J, et al., 2022. Efficient 3D point cloud feature learning for large-scale place recognition. *IEEE Trans Image Process*, 31:1258-1270. https://doi.org/10.1109/TIP.2021.3136714

Kim G, Kim A, 2018. Scan context: egocentric spatial descriptor for place recognition within 3D point cloud map. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.4802-4809. https://doi.org/10.1109/iros.2018.8593953

Komorowski J, 2021. MinkLoc3D: point cloud based large-scale place recognition. Proc IEEE Winter Conf on Applications of Computer Vision, p.1789-1798. https://doi.org/10.1109/wacv48630.2021.00183

Komorowski J, 2022. Improving point cloud based place recognition with ranking-based loss and large batch training. 26th Int Conf on Pattern Recognition, p.3699-3705. https://doi.org/10.1109/icpr56361.2022.9956458

Komorowski J, Wysoczańska M, Trzcinski T, 2021. MinkLoc++: LiDAR and monocular image fusion for place recognition. Int Joint Conf on Neural Networks, p.1-8. https://doi.org/10.1109/ijcnn52387.2021.9533373

Komorowski J, Wysoczańska M, Trzcinski T, 2022. EgoNN: egocentric neural network for point cloud based 6DoF relocalization at the city scale. *IEEE Robot Autom Lett*, 7(2):722-729. https://doi.org/10.1109/LRA.2021.3133593

Kong D, Li X, Ni PZ, et al., 2024. TopSPR-Net: topology aware segment-level point cloud learning descriptors for three-dimensional place recognition in large-scale environments. *IEEE Trans Ind Electron*, 71(10):13406-13416. https://doi.org/10.1109/TIE.2024.3363747

Kong X, Yang XM, Zhai GY, et al., 2020. Semantic graph based place recognition for 3D point clouds. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.8216-8223. https://doi.org/10.1109/iros45743.2020.9341060

Li QP, Zhuang Y, Huai JZ, et al., 2024. An efficient point cloud place recognition approach based on transformer in dynamic environment. *ISPRS J Photogramm Remote Sens*, 207:14-26. https://doi.org/10.1016/j.isprsjprs.2023.11.013

Liu CX, Chen GY, Song R, 2024. LPS-Net: lightweight parameter-shared network for point cloud-based place recognition. IEEE Int Conf on Robotics and Automation, p.448-454. https://doi.org/10.1109/icra57147.2024.10610758

Liu Z, Zhou SB, Suo CZ, et al., 2019. LPD-Net: 3D point cloud learning for large-scale place recognition and environment analysis. Proc IEEE/CVF Int Conf on Computer Vision, p.2831-2840. https://doi.org/10.1109/ICCV.2019.00292

Lu T, Wang LM, Wu GS, 2021. CGA-Net: category guided aggregation for point cloud semantic segmentation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11688-11697. https://doi.org/10.1109/cvpr46437.2021.01152

Ma JY, Zhang J, Xu JT, et al., 2022. OverlapTransformer: an efficient and yaw-angle-invariant transformer network for LiDAR-based place recognition. *IEEE Robot Autom Lett*, 7(3):6958-6965. https://doi.org/10.1109/LRA.2022.3178797

Ma JY, Xiong GM, Xu JY, et al., 2024. CVTNet: a cross-view transformer network for LiDAR-based place recognition in autonomous driving environments. *IEEE Trans Ind Inform*, 20(3):4039-4048. https://doi.org/10.1109/TII.2023.3313635

Maddern W, Pascoe G, Linegar C, et al., 2017. 1 year, 1000 km: the Oxford RobotCar dataset. *Int J Robot Res*, 36(1):3-15. https://doi.org/10.1177/0278364916679498

Miao Y, Engelmann F, Vysotska O, et al., 2025. SceneGraphLoc: cross-modal coarse visual localization on 3D scene graphs. 18th European Conf on Computer Vision, p.127-150. https://doi.org/10.1007/978-3-031-73242-3_8

Qi CR, Su H, Mo K, et al., 2017. PointNet: deep learning on point sets for 3D classification and segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.77-85. https://doi.org/10.1109/CVPR.2017.16

Rizzini DL, 2017. Place recognition of 3D landmarks based on geometric relations. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.648-654. https://doi.org/10.1109/iros.2017.8202220

Röhling T, Mack J, Schulz D, 2015. A fast histogram-based similarity measure for detecting loop closures in 3-D LIDAR data. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.736-741. https://doi.org/10.1109/iros.2015.7353454

Rosinol A, Gupta A, Abate M, et al., 2020. 3D dynamic scene graphs: actionable spatial perception with places, objects, and humans. https://arxiv.org/abs/2002.06289

Shu DW, Kwon J, 2024. Hierarchical bidirected graph convolutions for large-scale 3-D point cloud place recognition. *IEEE Trans Neur Netw Learn Syst*, 35(7):9651-9662. https://doi.org/10.1109/tnnls.2023.3236313

Siva S, Nahman Z, Zhang H, 2020. Voxel-based representation learning for place recognition based on 3D point clouds. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.8351-8357. https://doi.org/10.1109/IROS45743.2020.9340992

Sun Q, Liu HY, He J, et al., 2020. DAGC: employing dual attention and graph convolution for point cloud based place recognition. Proc Int Conf on Multimedia Retrieval, p.224-232. https://doi.org/10.1145/3372278.3390693

Tang WJ, Yang CG, An ZL, et al., 2026. EfficientVPR: toward efficient visual place recognition via scene-aware prompt tuning and adaptive feature enhancement. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.33738-33748.

Uy MA, Lee GH, 2018. PointNetVLAD: deep point cloud based retrieval for large-scale place recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4470-4479. https://doi.org/10.1109/CVPR.2018.00470

Vidanapathirana K, Ramezani M, Moghadam P, et al., 2022. LoGG3D-Net: locally guided global descriptor learning for 3D place recognition. Int Conf on Robotics and Automation, p.2215-2221. https://doi.org/10.1109/ICRA46639.2022.9811753

Wen CC, Huang H, Liu YS, 2023. Pyramid learnable tokens for 3D LiDAR place recognition. IEEE Int Conf on Robotics and Automation, p.4143-4149. https://doi.org/10.1109/icra48891.2023.10161523

Xia Y, Xu YS, Li S, et al., 2021. SOE-Net: a self-attention and orientation encoding network for point cloud based place recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11343-11352. https://doi.org/10.1109/cvpr46437.2021.01119

Xia Y, Gladkova M, Wang R, et al., 2022. PVT3D: point voxel transformers for place recognition from sparse LiDAR scans. https://arxiv.org/pdf/2211.12542v1

Xia Y, Gladkova M, Wang R, et al., 2023. CASSPR: cross attention single scan place recognition. Proc IEEE/CVF Int Conf on Computer Vision, p.8427-8438. https://doi.org/10.1109/iccv51070.2023.00777

Zhang WX, Xiao CX, 2019. PCAN: 3D attention map learning using contextual information for point cloud based retrieval. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.12428-12437. https://doi.org/10.1109/CVPR.2019.01272

Zhang YJ, Shi PC, Li JY, 2024. LiDAR-based place recognition for autonomous driving: a survey. *ACM Comput Surv*, 57(4):106. https://doi.org/10.1145/3707446

Zuo WL, Liu LG, Li YZ, et al., 2025. PRGS: patch-to-region graph search for visual place recognition. *Pattern Recogn*, 166:111673. https://doi.org/10.1016/j.patcog.2025.111673