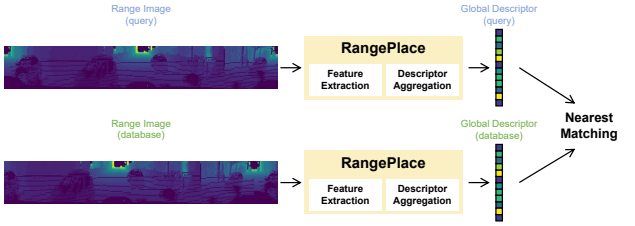
RangePlace: A Hierarchical Range Image Transformer for LiDAR-based Place Recognition

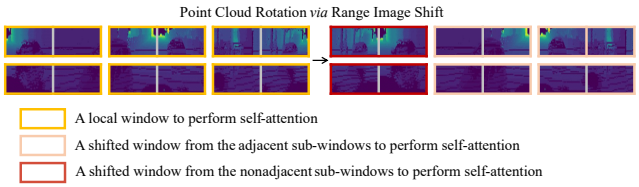
Ji Li, Qingxiao Liu, Boyang Wang, *Member, IEEE*, Haiou Liu, and Yuxuan Han

***Abstract*—LiDAR-based place recognition contributes significantly to identifying loop closure candidates for long-term simultaneous localization and mapping (SLAM) systems and obtaining global position for intelligent vehicles. Although the existing Transformer-based methods have achieved impressive results, they lack the capability of harnessing Transformers to establish long-range dependencies and suffer from high computational costs. To address these limitations, this study proposes a hierarchical range image transformer network named RangePlace to generate discriminative global descriptors for place recognition. By leveraging shifted window-based attention and depth-wise convolution, the proposed network can effectively and efficiently extract both global and local features. This not only enhances the network's capability to learn long-range information but also reduces the computational costs. In addition, a Pyramid Feature Mix module is incorporated to aggregate multi-scale feature maps to unified global descriptor, thus significantly enhancing place recognition performance. The proposed method is verified by experiments on the KITTI Odometry dataset, NCLT dataset, Ford Campus dataset and NuScenes dataset. The results demonstrate that the proposed method can achieve state-of-the-art performance in terms of recall rates. Moreover, it has good robustness to viewpoint yaw-rotation and strong generalization ability in various environments. Furthermore, it can be applied to real-time localization systems. The open-sourced codes used in this study are publicly available at https://github.com/JiLiBIT/RangePlace.**

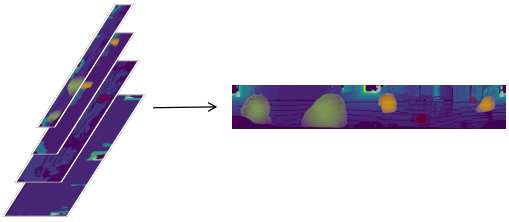
***Index Terms*—global descriptor, place recognition, range image, Transformer**



(a) Learning-based place recognition



(b) Shifted window approach in feature extraction



(c) Feature pyramid design in descriptor aggregation

Fig. 1. (a) The proposed RangePlace exploits range images as input data to extract features and generate a global descriptor; (b) the self-attention is calculated only within each window instead in the whole range image; (c) feature pyramid design enables the network to leverage the capability of self-attention across multiple scales and capture more details for retrieval purposes.

I. Introduction

PLACE recognition is an essential tasks in the field of robotics [1] and intelligent vehicles [2], [3], and has had an extensive application in simultaneous localization and mapping (SLAM) [4], satellite-absent navigation [5], [6], and three-dimensional (3D) reconstruction [7]. Compared with visual place recognition (VPR) [8], [9], LiDAR-based place recognition (LPR) [10], [11], [12], [13] has attracted more attention in applications that include vehicles working in outdoor large-scale environments. This is due to its robustness against day-and-night light variations, as well as diverse weather and seasonal conditions.

Recently, there have been numerous studied on [14], [15], [16] exploiting Transformer structure to enhance the effectiveness and robustness of place recognition. Although these studies have achieved impressive performance with larger receptive field of attention, they have a limitation in unleashing the full potential of the Transformer structure. This limitation can lead to an unexpected decrease in performance and an unacceptable computational cost when more self-attention layers are deployed into Transformer framework.

To address the aforementioned challenges, this study propose a hierarchical range image transformer network for LiDAR-based place recognition task. The proposed framework is designed to enhance the learning from range images derived from LiDAR point clouds, which allows for achieving high performance while maintaining a low computational cost. As depicted in Fig. 1b, attention is only performed in partitioned windows. Namely, when the window is shifted, self-attention

The research is funded by the National Natural Science Foundation of China under Grants No. 52172378; and the National Natural Science Foundation of China under Grants No. 52302489. (Corresponding author: Boyang Wang)

J. Li, Q. Liu, B. Wang, H. Liu and Y. Han are with the School of Mechanical Engineering, Beijing Institute of Technology, Beijing, 100081, China. (e-mail: jilibit@bit.edu.cn; 3120185271@bit.edu.cn; boyang_wang@bit.edu.cn; liuhaiou@bit.edu.cn; 3120220300@bit.edu.cn)

computation occur in windows covering both adjacent and nonadjacent sub-windows. This approach extends connections established by attention beyond the boundaries of the previous window, thus effectively establishes relation across the entire range image. Accordingly, the proposed structure can establish long-range dependencies overall range image while maintaining linear computational complexity with respect to the input size of range image. In addition, a depth-wise convolution layer is incorporated before two fully-connected layers to replace the conventional feed-forward network. This modification can improve local context awareness, which is crucial for handling the problem of low spatial resolution of input range images, and simultaneously reduce computational complexity. Furthermore, the proposed design uses feature pyramid structure to facilitate multi-scale feature maps, as illustrated in Fig. 1c. Similar to natural images, in range images, an object's size varies with its distance to the ego vehicle. To exploit this property, maps on multiple scales are employed to aggregate information of varying sizes. This approach can leverage the capability of self-attention across multiple scales and make the model learn more nuanced details. Finally, it exploits the Feature Mix module and context gating mechanism to generate global descriptor for range image retrieval.

Since range images have spatial-cues and single-channel, popular augmentation methods are not well-suited to them. masked auto encoder (MAE) [17] and range-image-based augmentation are introduced to pre-training and training phases, respectively. Applying these well-devised approaches makes unsupervised learning and data augmentation more applicable to range images.

The main contributions of this paper are as follows:

• A hierarchical range image transformer network, named RangePlace, is introduced to address LiDAR-based place recognition tasks. The proposed method demonstrates state-of-the-art performance on the KITTI Odometry dataset, NLCT dataset, Ford Campus dataset and NuScenes dataset, which demonstrates its high efficacy in place recognition;

• A shifted-window attention module is introduced and coupled with an depth-wise convolution feed-forward network to improve the ability of learning local and global features more effectively and efficiently;

• A Pyramid Feature Mix module is introduced to multi-scale feature maps. This module is designed explicitly for LiDAR-based place recognition and demonstrates a substantial performance improvement compared to alternative aggregation methods;

• The proposed method has robust generalization capabilities and remains resilient to viewpoint yaw-rotation due to its inherent rotation invariance. Furthermore, it can be applied in real-time localization systems, solidifying its practical applicability.

II. Related Works

This section focuses mainly on recent developments in the field of LiDAR-based place recognition. In this task, the main objective is to determine a particular location based on the query LiDAR data, by leveraging the geotags associated with the most pertinent LiDAR data retrieved from a reference database. Traditional handcrafted methods [18], [19], [20], [21] typically leverage statistical characteristics of LiDAR data to characterize the driving scenes, suffering from low accuracy in complex environments and long computational time. Recently, the learning-based approaches have gained increasing attention in the field of place recognition. As presented in Fig. 1a, the typical workflow of a learning-based place recognition system involves feature extraction, descriptor aggregation, and nearest matching. Feature extraction and descriptor aggregation are performed to extract geometric features of the input data and aggregate them into a compact descriptor, respectively. The effectiveness of these two techniques defines the performance of place recognition systems.

A. Feature Extraction Techniques

Mikaela *et al.* [1] used PointNet [22] to facilitate 3D place discriminative feature learning from point clouds. However, this approach omitted relationships between local features, leading to low localization accuracy under different viewpoints. Then, [11], [14], [23] exploited the modified PointNet to capture the local geometric structures, by adding orientational embeddings and graph-based network in coordinate and transformed feature spaces. Further, [24], [25] used sparse voxelized point clouds to reduce high computational costs of 3D convolutional neural networks. More recently, BEVPlace [26] exploited an additional group convolution network to extract rotation invariant features from bird's eye view images. However, these methods exhibited limitations in capturing long-range dependencies. Nevertheless, it's a crucial aspect for LiDAR-based tasks, where point clouds are often modeled as sequences with long range.

With the development of deep learning, Transformers [27] have shown remarkable performance in place recognition. Zhou *et al.* [15] incorporated the attention mechanism on transformed Normal Distributions Transform (NDT) cells to enhance the representation ability. Building on LPD-Net [23], PPT-Net introduced by Hui *et al.* [28] leveraged a pyramid point transformer to improve local feature extraction. However, it cannot learn spatial orientation differences and long-range information, which restricted its robustness to viewpoint yaw-rotation and generalization ability for unseen environments. Ma *et al.* [16] projected point clouds into range images and introduced attention mechanism to the output feature of ConvNets-based encoder. However, their work was limited in the single-Transformer framework since their experiments have shown that using a large number of Transformer blocks can lead to poor performance. In summary, the existing methods encounter difficulties in harnessing the potential of transformer architectures to achieve improved performance while simultaneously maintaining low computational costs. To address the above-mentioned challenges, our method focuses on feature extraction from range images, and utilizes a hierarchical shifted-window-based transformer network. In this way, it can effectively and efficiently enhance the ability to capture essential long-range dependencies in LiDAR-based place recognition.

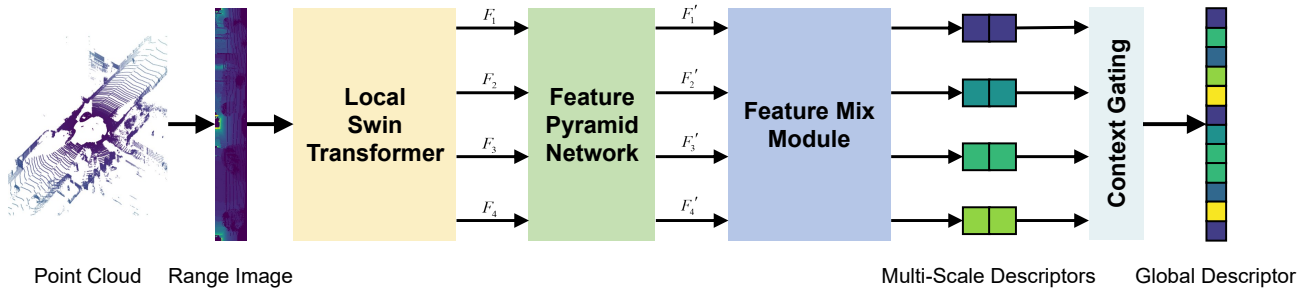


Fig. 2. The pipeline of our proposed RangePlace. Given the range images projected from point clouds, we first utilize Local Swin Transformer to capture the geometric information F_l of the range images at varying resolutions, emphasizing long-range dependencies to improve the distinction of local and global features. Following that, we build a Feature Pyramid Network to facilitate feature maps $F_l{}'$ at multiple scales. Then, we feed these feature maps to the Feature Mix module, where we generate multi-scale descriptors. In the end, we exploit a context gating mechanism to produce unified global descriptor for range image retrieval.

B. Descriptor Aggregation Techniques

Since PointNetVLAD [1] first aggregated features into a global descriptor using the deep vector of locally aggregated descriptors (VLAD) [29], most state-of-the-art methods have been designed based on VLAD. MAC [30] and Generalized Mean (GeM) [31] shifted the focus to regions of interest rather than local features, but they underperformed VLAD series in place recognition tasks. However, recent advances in isotropic architecture challenged the notion that VLAD is critical for global descriptor aggregation. For instance, [32] proposed MixVPR, a variant building on MLP-Mixer [33], which has demonstrated a remarkable performance on multiple VPR benchmarks. In light of these developments, rather than persisting with the use of VLAD, this study propose a approach for aggregating global descriptors, which involves a Pyramid Feature Mix module and a context gating mechanism. This modification provides a more effective strategy for descriptor aggregation in LiDAR-based place recognition.

III. Methods

The overarching framework of our proposed method is depicted in Fig. 2. Our proposed RangePlace comprises four main modules: a Local Swin Transformer, a Feature Pyramid Network (FPN), a Feature Mix Module, and a Context Gating Mechanism. The process begins with the raw range image projected from the raw point cloud, being input to the network. First, we utilize Local Swin Transformer to establish interactions with other pixels, both close information and far relation across the whole range image. Following that, we build a Feature Pyramid Network to facilitate multi-scale feature maps. Then, we feed these feature maps to the Feature Mix Module, where we generate multi-scale global descriptors. Finally, we employ a Context Gating Mechanism to produce a distinguishing global descriptor. For further insights into the architecture design, please refer to the details presented in Section III. Due to the devised network architecture, our proposed descriptor demonstrates rotation invariance, as reported in Section IV-B.

A. Range Image Projection

A range image can be regarded as an intermediate depiction of a singular 3D scan from a spinning LiDAR sensor. With the recent surge of multi-view learning, the depth map obtained by Monocular [34] can also be considered a form of range image. Compared with various LiDAR representations for place recognition, range image and bird's eye view (BEV) image based methods have emerged as more tractable alternatives. These methods do not require performing computationally intensive neighborhood search [35] or 3D convolution operations [36], makes them highly efficient during both training and inference phases. Furthermore, range image can retain information in vertical direction adequately, which represents an advantage compared to the BEV images.

The projection transformation to produce a range image is established relying on the LiDAR sensor parameters. To form a range image $\mathcal{R}$, a point cloud $\mathcal{P}$ undergoes projection denoted by: $\mathcal{P}:\mathbb{R}^3\rightarrow\mathbb{R}^2$, where each pixel represents a single point. A given point $p_n=(x_n,y_n,z_n)$ in $\mathcal{P}$ is converted into image coordinates (u_n,v_n) through the following transformation [16], [37]:

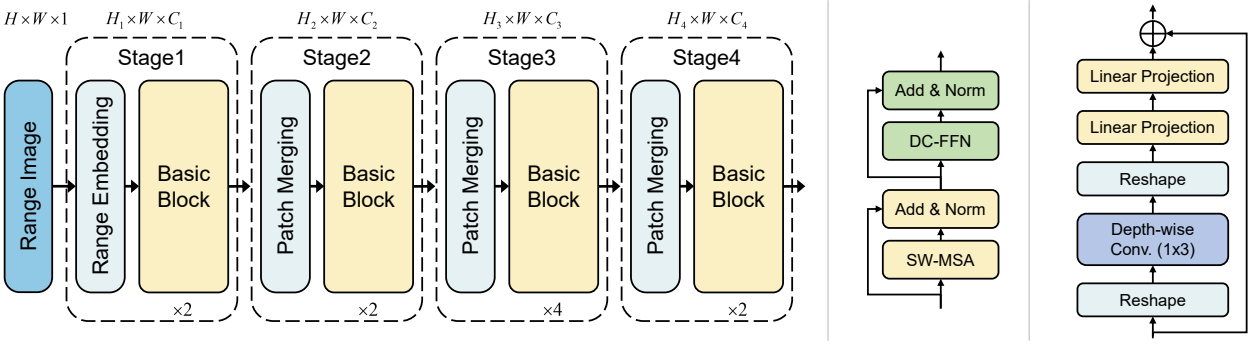
$$\begin{pmatrix}u_n\\v_n\end{pmatrix}=\begin{pmatrix}\frac{1}{2}\left[1-\arctan\left(y_n,x_n\right)\pi^{-1}\right]w\\\left[1-\left(\arcsin\left(z_n,r_n^{-1}\right)+\phi_{up}\right)\xi^{-1}\right]h\end{pmatrix}\tag{1}$$

where (u_n,v_n) represents the grid coordinate of point p_n in the range image $\mathcal{R}(u,v)$; $r_n=\sqrt{x_n^2+y_n^2+z_n^2}$ is the distance between the point and LiDAR; $\xi=|\phi_{\text{up}}|+|\phi_{\text{down}}|$ denotes the vertical field-of-views (FOVs) of the LiDAR, and ϕ_{up} and ϕ_{down} are the inclination angles in the upward and downward directions, respectively.

Since the azimuthal angle $\phi_n=arctan\left(y_n,x_n\right)$ of each point is directly related to its horizontal position u_n in the range image, any rotation R_θ with azimuthal angle θ of the point cloud results in a corresponding and predictable horizontal column shift C_s with s, via rotation-equivariant, which can be expressed as follows:

$$\begin{aligned}\begin{pmatrix}u_n{}'\\v_n{}'\end{pmatrix}&=\begin{pmatrix}\frac{1}{2}\left[1-(\phi_n+\theta)\pi^{-1}\right]w\\\left[1-\left(\arcsin\left(z_n,r_n^{-1}\right)+\phi_{up}\right)\xi^{-1}\right]h\end{pmatrix}\\&=\begin{pmatrix}u_n+s\\v_n\end{pmatrix},s=\frac{1}{2}\left[1-\theta\pi^{-1}\right]w\\\mathcal{R}C_s&=\Pi\left(R_\theta\mathcal{P}\right)\end{aligned}\tag{2}$$

Based on this characteristic, performing shifted-window-based and full receptive field operation can obtain rotation-



(a) Local Swin Transformer (b) Basic block (c) DC-FFN

Fig. 3. Illustration of Local Swin Transformer coupled with depth-wise convolution feed-forward network.

invariant descriptors easily, which is indicated in the experiments conducted in this study, as presented in Section IV-B.

B. Local Swin Transformer

We adopt a hierarchical transformer network, which comprises of overlapped range embedding, shifted window-attention based transformer and overlapped patch merging, as the backbone of proposed framework. This is motivated by the aim to reduce computational complexity, achieve larger receptive fields, and establish longer range relationship for place recognition.

Given a batch of projected range images $R(u, v)$, the process is initiated with the Range Embedding module. This module is composed of three overlapped convolutional layers, collectively mapping each pixel in the grid to a higher-dimensional embedding $F_0 \in \mathbb{R}^{C_1, H_1, W}$. Inspired by the success of Swin Transformer [38] and its follow-ups, we implement a hierarchical structure to facilitate feature maps at multiple scales. An embedded map F_0 undergoes processing through basic blocks for four stages, yielding $\{F_1, F_2, F_3, F_4\}$, respectively. To establish hierarchical structure, the reduction in the number of tokens is achieved through the incorporation of an overlapped patch merging layer as the network progresses deeper. It should be noted that our embedding and merging processes exclusively compress feature map in the vertical dimension while preserving the width dimension. This approach is adopted to alleviate discretization errors associated with rotation equivariance.

As presented in Fig. 3b, the basic block comprises a shifted window-based multi-head self-attention (SW-MSA) module, followed by a depth-wise convolution feed-forward network (DC-FFN), with GELU activation function in between. To enhance training stability, a LayerNorm (LN) layer is implemented after every MSA and DC-FFN module. The relative position bias is included in SW-MSA, which has often been used in the Transformer architecture. Further, to capture long-range features, feature maps from basic blocks are subsequently fed to windowing transformation using a window size of $[1, M]$. Next, the succeeding block adopts a shifted windowing partition that displacing the windows of preceding layer by $\left[0, \frac{M}{2}\right]$ pixels. With the shifted window partitioning approach, consecutive basic blocks are computed as follows:

$$\begin{aligned} \hat{z}^l &= \mathrm{LN}\left(\text{W-MSA}\left(\hat{z}^{l-1}\right)\right) + z^{l-1} \\ z^l &= \mathrm{LN}\left(\text{DC-FFN}\left(\hat{z}^l\right)\right) + \hat{z}^l \\ \hat{z}^{l+1} &= \mathrm{LN}\left(\text{SW-MSA}\left(z^l\right)\right) + z^l \\ z^{l+1} &= \mathrm{LN}\left(\text{DC-FFN}\left(\hat{z}^{l+1}\right)\right) + \hat{z}^{l+1} \end{aligned} \tag{3}$$

where $\hat{z}_l$ and z_l stand for the output features of the (S)W-MSA module and the DC-FFN module for block l, respectively.

After window shift operation, a partitioned window may consist of sub-windows that are originally nonadjacent. However, these sub-windows naturally have a relationship in the width direction, as they are derived from the leftmost and rightmost points of the point cloud. Therefore, the application of a mask mechanism is avoided, allowing self-attention computation to extend beyond these sub-windows.

Further, depth-wise convolution feed-forward network (DC-FFN) is integrated to capture local context effectively, serving as a replacement for the conventional feed-forward network (FFN). As presented in [39], [40], achieved by incorporating a depth-wise convolution operator, it can enhance the performance and capabilities compared to the traditional FFN approaches. Specifically, the mentioned process involves the following steps:

(a) The position of 1×3 depth-wise convolution layer is moved up to reduce computational costs by decreasing the dimension;

(b) The feature map is reshape and processed through a linear projection layer to increase the feature dimension;

(c) The resulting feature is flatten into tokens, and the channels is reduced through another linear projection layer to align with the input feature;

(d) The activation function used after each linear or convolution layer is GELU. Finally, a skip connection fuses the input back into the resulting projection.

C. Feature Pyramid Network

The feature pyramid network [41] serves as a standard approach for constructing an in-network pyramid. Since an object's size in a range image changes with its distance to the ego varies, the FPN is motivated to combine higher-resolution features from earlier stages with more-detail features from later stages; this is achieved through top-down and lateral

connections. To be specific, let's define $F_4{}' = F_4$, formulating the pyramid architecture as follows:

$$F_l{}' = \text{MLP}\left(F_l \oplus I\left(F_{l+1}{}'\right)\right) \tag{4}$$

where $l \in \{0, 1, 2, 3\}$, and $F_l{}' \in \mathbb{R}^{C \times H_l \times W}$ represents the generated pyramid feature map, $I(\cdot)$ indicates bilinear interpolation, and $\oplus$ denotes channel-wise addition.

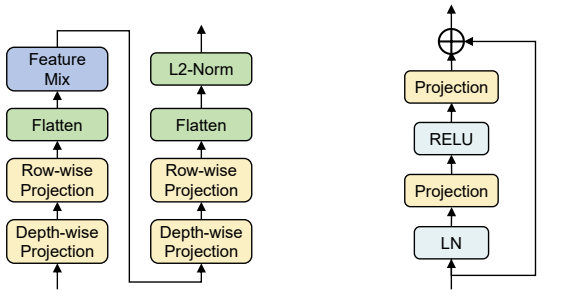
These operations collectively enable us to build a feature pyramid network with multiple scales, thus enhancing the versatility and effectiveness of feature extraction.

D. Feature Mix Module

The existing techniques, such as GeM [31], VLAD [42] typically treat feature map F as a collection of spatial descriptors. These techniques aim to learn global dense representations by integrating these feature maps. As a result, each descriptor is associated with a receptive field in the input feature map. In our pursuit of obtaining the global descriptor, inspired by [32], we introduce the Feature Mix module to LiDAR-based place recognition tasks. The intuition behind Feature Mix is to leverage the inherent capacity of fully-connected layers. The MLPs utilize their full receptive field to aggregate features autonomously and holistically. This allows each neuron insight into the entirety of the input feature maps, moving beyond a focus solely on local features.

The 3D feature map $F_l{}'$ can be regarded as a set of 2D feature maps with dimensions $H \times W$, which can be expressed as follows:

$$F_l{}' = \{X_i\}, i = \{1, \ldots, C\} \tag{5}$$



(a) the Feature Mix module (b) the Feature Mix operation

Fig. 4. The flowchart of the feature mix module in Fig. 2

.

The Feature Mix modules are employed on feature maps at multiple scales, aggregating them to efficient global descriptors for retrieval purposes. Fig. 4a illustrates the main procedures involved in Feature Mix module:

(a) Dimension Projection: First, two successive fully-connected layers are employed, first to increase the dimension in depth-wise, then to reduce in row-wise, leading to projected feature maps denoted as $F_P \in \mathbb{R}^{C_P \times 1 \times W}$.

(b) Feature Flatten: Each feature map X_i is transformed into a token-shaped representation, leading to tokenized feature maps denoted as $F_T \in \mathbb{R}^{C_P \times W}$.

(c) Feature Mix: The Feature Mix operation consists a cascade of two fully-connected layers constituted MLP blocks with an identical structure, as presented in Fig. 4b. It receives flattened features as input and integrates global relationships into each $X_i \in F_T$ as formulation below:

$$X_i \leftarrow W_2(\sigma(W_1 X_i)) + X_i, \quad i = \{1, \ldots, C_P\} \tag{6}$$

where W_1 and W_2 stand for the weights of fully-connected layers, σ denotes a ReLU activation function. And a skip connection [43] is applied after each block.

Owing to its isotropic architecture, the Feature Mix operation can generate an output $Z \in \mathbb{R}^{C_P \times W}$ having the same shape as the input $F_P \in \mathbb{R}^{C_P \times W}$. Therefore, this operation can be repeated until reach a sequence of L consecutive blocks, which can be expressed as follows:

$$Z = FM_L(FM_{L-1}(\ldots FM_1(F_T))) \tag{7}$$

where dimension of Z equals to the dimension of initially projected feature maps F_P.

(d) Dimension Projection: Finally, two other successive fully-connected layers are employed to reduce the dimension first in depth-wise, then in row-wise. This can be regarded as a learnable pooling operation that controls the output size of the global descriptor. Specifically, depth-wise projection and row-wise projection map Z from $\mathbb{R}^{C \times W}$ to $\mathbb{R}^{D \times W}$, and from $\mathbb{R}^{D \times W}$ to $\mathbb{R}^{D \times R}$, subsequently. The final output $O \in \mathbb{R}^{D \times R}$ is flattened and $L2$-normalized, as performed in [16].

E. Context Gating

Concatenating the descriptors at various scales into a unified high-dimensional descriptor will result in increased inference and matching time. To address this, a fully-connected layer is utilized to compress a high-dimension vector into a more manageable size, particularly 256 in our implementation. Subsequently, the global context is encoded through context gating mechanism [42] and transformed into one new global descriptor, which captures essential contextual information from the multi-scale descriptors.

TABLE I
DATABASE AND QUERY FRAMES OF THE DATASETS.

Sequence	KITTI			Ford
	00	05	06	02
Database	0-3000	0-1000	0-600	0-2800
Query	3200-4540	1200-2761	800-1100	3000-6103

IV. Experiments

A. Experimental Settings

Datasets. We carry out place recognition experiments on four large-scale public datasets, i.e. the KITTI Odometry dataset [44], NCLT dataset [45], Ford Campus dataset [46] and NuScenes dataset [47] to evaluate our method and other baseline methods. **KITTI Odometry dataset** involves LiDAR sensor data collected from a automobile that repeatedly traverses across different regions at distinct time points, and provides ground truth poses for 11 sequences (00–10). For dataset construction, we allocate sequences 00, 05, and 06 for

TABLE II
RECALL RATES ON THE KITTI ODOMETRY DATASET.

Method	00		05		06		Mean	
	AR@1	AR@1%	AR@1	AR@1%	AR@1	AR@1%	AR@1	AR@1%
Scan Context++ [21]	89.7	96.2	77.2	84.3	87.0	89.3	84.6	89.9
PointNetVLAD [1]	90.6	97.5	77.0	84.6	77.8	84.4	81.8	88.8
OverlapTransformer [16]	96.3	98.8	91.5	96.4	95.6	98.5	94.5	97.9
MinkLoc3D-v2 [25]	95.9	98.4	86.6	94.9	80.4	87.0	87.6	93.4
SOE-Net [14]	95.5	98.1	83.7	89.5	69.6	81.1	82.9	89.6
PPT-Net [28]	96.0	98.8	87.0	95.3	85.2	90.4	89.4	94.8
RangePlace[1] (ours)	99.1	99.6	96.1	97.8	99.3	**100.0**	98.2	99.1
RangePlace (ours)	**99.6**	**100.0**	**96.6**	**98.4**	**100.0**	**100.0**	**98.7**	**99.5**

[1] Is trained without proposed range image-based data augmentation approaches.

evaluation, since these sequences encompass sizable retraced areas. The remaining sequences were used for training and validation. **Ford Campus dataset** contains point cloud data of the research campus of Ford and the downtown area of Dearborn, Michigan. In this study, we examine the generalization capability of different methods on its sequence 02. We split aforementioned sequences into database and query frames for place recognition experiments, as summarized in Table I. **NCLT dataset** involves LiDAR sensor data collected by Velodyne HDL-32E sensor across different seasons. Following the setup of [48], we allocate sequence 2012-01-08 for training, and sequences 2012-02-05, 2012-06-15, 2013-02-23 for evaluation. **Nuscenes dataset** includes data also collected by a Velodyne HDL-32E sensor. We exploit it to evaluate the generalization ability of the methods for data acquired by different sensors. Following [49], the data from last 15 days collected in Boston are used as query, and the remaining data in Boston are used as a database.

Evaluation approaches. In our experiments on the NCLT dataset, following [48, 50] we consider a retrieval as true positive if the geometric distance between the centroids of a match and the query point cloud is less than 1.5 m. On the KITTI Odometry dataset, Ford Campus dataset and NuScenes dataset, following [16, 26, 49], we consider a retrieval as true positive if that distance is less than 5 m, and a negative if the distance is larger than 7.5 m.

We train our framework using Truncated Smooth-AP (TSAP) [25], which is an adapted Smooth-AP loss. This loss function is designed to maximize the ranking of all positives corresponding to each query, and improve computational efficiency by ensuring that only a limited number of the top-k candidates are considered. Let's define q as a query range image, P as a set comprising its k nearest positives in the descriptor space, N as the count of positives per query element, and Ω as a set encompassing all positives and negatives. The TSAP loss is calculated as the average across all batch elements B_m:

$$L_{TSAP} = \frac{1}{m}\sum_{q=1}^{m}\left(1-\frac{1}{|N|}\sum_{i\in P}\frac{1+\sum_{j\in P, j\neq i} g\left(d\left(q,i\right)-d\left(q,i\right);\tau\right)}{1+\sum_{j\in \Omega, j\neq i} g\left(d\left(q,i\right)-d\left(q,i\right);\tau\right)}\right) \tag{8}$$

where g is a differentiable approximation of the target function, given by $g(x;\tau) = (1+\exp(-x/\tau))^{-1}$; and $d(q,i)$ represents the Euclidean distance between the descriptor of the query range image q and the i-th range image; τ regulates the sharpness of the approximation, and the numerator signifies a soft ranking of a positive.

To assess the effectiveness of place recognition methods, we employ the evaluation metric Recall@N, a methodology utilized in [1]. This metric quantifies correctly matched queries in percentage. We present our results using specific average recall metrics: average recall@1 (AR@1), average recall@20 (AR@20) and average recall@1% (AR@1%). These metrics can provide insights into the recognition accuracy and performance of the place recognition methods.

Implementation details. The overall architecture of proposed RangePlace is depicted in Fig 2. We work with range images having a size of $64 \times 896 \times 1$. In the Local Swin Transformer, our implementation of Range Embedding module is employing convolutional block with kernel 5×1, stride 1×1 initially for larger receptive field, and kernel 3×1, stride 2×1 for the others. The range embedding dimension C is set to 64, and the window size M is chosen as $[1, 14]$, $[1, 28]$, $[1, 56]$, $[1, 112]$, respectively. The latent ratio r of DC-FFN is set to 4. Overlapped patch merging layers are achieved by convolution with kernel 2×1, stride 2×1 for the first stage and stride 1×1 for the other stages. The number of basic blocks is configured as $\{2, 2, 4, 2\}$. Using dropout is intentionally refrained to preserve rotation equivariant properties. Furthermore, the spatial resolution of the feature maps obtained in the various stages are $[6, 896]$, $[3, 896]$, $[2, 896]$, $[1, 896]$, with the channel dimension of $64, 128, 128, 256$, respectively. The output dimension of FPN is set to 128. The cascade depth L of Feature Mix is set to 4, and output dimension R, D are set

TABLE III
RECALL RATES ON THE NCLT DATASET.

Method	2012-02-05		2012-06-15		2013-02-23		Mean	
	AR@1	AR@20	AR@1	AR@20	AR@1	AR@20	AR@1	AR@20
PointNetVLAD [1]	75.4	88.3	60.2	75.9	46.9	71.9	60.8	79.0
OverlapTransformer [16]	87.1	93.0	63.9	78.1	57.3	76.7	69.4	82.6
MinkLoc3D-v2 [25]	83.2	92.6	64.7	79.4	51.6	76.1	66.5	82.7
SOE-Net [14]	82.1	91.7	61.5	76.3	50.7	75.1	65.1	81.4
PPT-Net [28]	86.9	92.5	63.4	77.4	53.8	77.2	68.0	82.4
RangePlace (ours)	**91.5**	**95.9**	**77.8**	**89.0**	**76.5**	**88.7**	**81.9**	**91.2**

to $4, 64$, respectively.

We pre-train our Local Swin Transformer using a masked auto-encoder on sequences of KITTI dataset, excluding data which are reserved as test sets for the place recognition task. The encoder and decoder utilize the Local Swin Transformer as the backbone, as described in Section III-B. The decoder structure closely resembles the Unet structure, without the lateral connection layer between the encoder and the decoder. We introduce random mask operation on each window of size $[8, 14]$. The loss function used for pre-training calculates the mean square error (MSE) between original and reconstructed range images.

During training phase, we propose range image-based data augmentation approaches. For example, RangeShift randomly slides the scan along the azimuth direction within the span of $\left[\frac{-W}{2}, \frac{W}{2}\right]$, RangeTransfer randomly increases the uncertainty of range, and RangeMirror randomly mirrors in the width direction, within probability set to 0.2. The Local Swin Transformer backbone is initialized from pre-trained parameters and forzen during training. We employ an AdamW optimizer [51] for 100 epochs with a cosine-decay learning-rate scheduler. A mini-batch size of 32, an initial learning rate of 1e-4, an minimum learning rate of 1e-6, and a weight decay of 1e-4 are used. Sigmoid temperature τ is set to 0.01, and positives per query k is set to 4.

B. Place Recognition Results

Quantitative results. We evaluate our proposed RangePlace with a series of progressive methods: the handcraft-based method Scan Context++ [21], and learning-based methods, including OverlapTransformer [16], PointNetVLAD [1], PPT-Net [28], SOE-Net [14] and MinkLoc3D-v2 [25]. To ensure a fair comparison, we run the official open-sourced implementation codes for retraining and evaluation. A distinguishing feature that sets our method apart from PointNetVLAD-based approaches [1, 14, 25, 28] is that we do not require an initial step of organizing LiDAR scans into down-sampled submaps and eliminating non-informative ground planes. We run the official codes of [1] to prepare the datasets for retraining and evaluating PointNetVLAD-based methods.

We provide the average recall@1 (AR@1) and average recall@1% (AR@1%) of different methods in Table II, which are evaluated on the KITTI Odometry dataset. As shown in this table, our proposed RangePlace achieves the state-of-the-art performance. For instance, we observe 3.2%, 5.0%, and 4.4% improvements on AR@1 over OverlapTransformer [16].

TABLE IV
RECALL RATES AT TOP-1 ON THE RANDOMLY ROTATED KITTI ODOMETRY DATASET.

Method	00	05	06	Mean
Scan Context++ [21]	89.7	77.2	87.0	86.0
PointNetVLAD [1]	85.0	68.2	51.5	68.2
OverlapTransformer [16]	96.3	91.5	95.6	94.5
MinkLoc3D-v2 [25]	89.4	82.8	48.1	73.4
SOE-Net [14]	91.5	82.3	65.6	79.8
PPT-Net [28]	90.0	73.8	50.0	71.3
RangePlace[1] (ours)	99.1	96.1	99.3	98.2
RangePlace (ours)	**99.6**	**96.6**	**100.0**	**98.7**

[1] Is trained without proposed range image-based data augmentation approaches.

We further provide the average recall@1 (AR@1) and average recall@20 (AR@20) on NCLT dataset in Table III. On this dataset, none of the methods performed as well as KITTI Odometry dataset. Still, our proposed RangePlace works a lot better than the other methods.

Our method can learn the detailed information in the vertical direction through overlapped range embedding and patch merging. Compared to OverlapTransformer, in proposed method, long-range relationship in the width direction is simultaneously established via shifted window-attention on the neighboring windows. This dual focus enhances the discrimination of local and global features. In addition, we employ a Pyramid Feature Mix module to generate global descriptors from multi-scale feature maps. Compared to SOE-Net, the proposed strategy enables the utilization of self-attention capabilities across different scales, empowering the model to capture more nuanced details for retrieval purposes.

Rotation invariance study. We evaluate the robustness of our RangePlace against viewpoint yaw-rotation. Specifically, during the rotation-invariance test phase, we introduce ran-

domly rotations to the input point clouds along Z-axis, covering the range $[0°, 360°)$ to simulate yaw-rotation. In Table IV, we present the recall rate at Top-1 retrievals on KITTI Odometry dataset under the rotation operation. The results in Table IV emphasize that our RangePlace surpasses all other methods regarding the recall rate performance. In addition, both with and without data augmentation, our method is not affected by randomly rotation. In contrast, the performances of MinkLoc3D-v2, SOE-Net and PPT-Net decrease noticeably under the rotation of point clouds. These results validate that range image representation is meaningful in handling viewpoint yaw-rotation, and our network is rotation-invariant designed.

TABLE V
RECALL RATES ON FORD CAMPUS DATASET AND NUSCENES DATASET.

Method	Ford		NuScenes	
	AR@1	AR@1%	AR@1	AR@1%
PointNetVLAD [1]	81.2	89.6	38.5	69.9
OverlapTransformer [16]	89.6	95.4	47.2	75.1
MinkLoc3D-v2 [25]	88.2	94.2	44.6	74.4
SOE-Net [14]	86.2	90.8	43.7	73.3
PPT-Net [28]	90.0	93.2	41.8	74.2
RangePlace (ours)	**93.4**	**97.8**	**56.2**	**80.4**

Generalization ability study. We carry out the experiments on the Ford Campus dataset and NuScenes dataset using the weights which are only trained on the KITTI Odometry dataset. It should be noted that the data used for training and the data in the Ford Campus dataset were collected using a Velodyne-64, while the NuScenes dataset was collected with Velodyne-32. In Table V, we provide the recall rates for each method on these datasets. It is evident that our method exhibits strong generalization and outperforms other advanced methods by a significant margin. The performances on Ford Campus dataset and NuScenes dataset serve as additional evidence of the generalization capability of our RangePlace in entirely unseen environments, and between different types of LiDAR sensors, respectively.

TABLE VI
RECALL RATES AT TOP-1 ON KITTI ODOMETRY DATASET AND NCLT DATASET FOR ABLATION ON AGGREGATION TECHNIQUES.

Sequence	KITTI		NCLT		Mean
	00	05	2012-06-15	2013-02-23	
no FPN	98.2	94.9	74.8	72.2	85.0
FPN, last-map	98.7	95.3	75.3	75.8	86.3
FPN, 4-stage	**99.6**	**96.6**	**77.8**	**76.5**	**87.6**

C. Ablation Study

Feature pyramid design. In addition, we study the impact of feature pyramid design on the KITTI Odometry dataset and NCLT dataset. We study a baseline without feature pyramid (no FPN) which already outperforms the existing methods, as seen in Table VI. We realize the FPN architecture and context gating module as in Fig 2. It can be found that our hierarchical structure (FPN, 4-stage) achieve the best performance, increasing AR@1 by up to 1.4%, 1.7%, 4.0% and 4.3%, respectively. We note that only using the last map from feature pyramid network to generate descriptor (FPN, last-map) also enhances retrieval performance, which means the global descriptor is multiple-scale in this case.

TABLE VII
RECALL RATES AT TOP-1 ON KITTI ODOMETRY DATASET AND NCLT DATASET FOR ABLATION ON FEATURE PYRAMID DESIGN.

Sequence	KITTI		NCLT		Mean
	00	05	2012-06-15	2013-02-23	
AVG	96.5	91.9	68.8	65.2	80.6
GeM [31]	98.8	96.2	73.7	72.8	85.4
VLAD [29]	98.4	93.3	75.6	75.1	85.6
Feature Mix	**99.6**	**96.6**	**77.8**	**76.5**	**87.6**

Feature mix with other aggregations. To validate the efficacy of the Feature Mix module, we analyze its performance on KITTI Odometry dataset and NCLT dataset, and compared it to those of the existing aggregation methods in the field of LiDAR-based place recognition. We compare against Average Pooling (AVG), GeM [31] and VLAD [29], which recently demonstrated SoTA performance. For a fair comparison, we train other aggregations with the same Local Swin Transformer and feature pyramid design employed by the Feature Mix module. Additionally, the clusters number of VLAD is set to 64. The results are presented in Table IX. It is evident that Feature Mix convincingly outperforms all other techniques. For instance, Feature Mix attains a new high AR@1 of 99.6%, 96.6%, 77.8% and 76.5%, which is 1.2%, 3.3%, 2.2% and 1.4% increase over VLAD, on sequences 00, 05 of KITTI dataset and sequences 2012-06-15, 2013-02-23 of NCLT dataset, respectively.

TABLE VIII
RECALL RATES AT TOP-1 ON KITTI ODOMETRY DATASET AND NCLT DATASET FOR ABLATION ON DC-FFN.

Sequence	KITTI		NCLT		Mean
	00	05	2012-06-15	2013-02-23	
FFN	98.8	95.7	75.1	75.3	86.2
DC-FFN	**99.6**	**96.6**	**77.8**	**76.5**	**87.6**

Depth-wise Convolution Feed Forward Network. To validate the efficacy of the DC-FFN, we analyze place recognition performance on KITTI Odometry dataset and NCLT dataset against conventional FFN and DC-FFN. The results are presented in Table VIII. It is evident that the effect of place recognition network is enhanced, through DC-FFN improving the local modeling ability.

D. Running Time

We evaluate the running time of our method and others on a desktop with an Intel i5-13490F CPU and an NVIDIA RTX4070 GPU. Our method has a relatively low descriptor generation time. It is ranked second only after OverlapTransformer with last-map, and third with 4-stage. Note that when pre-processing in Python, our method only takes about 5 ms to project point clouds onto the range image, while PointNetVLAD-based methods take about 30 ms to eliminate ground planes and perform downsampling. This indicates that our method is light-weight enough to be applied in real-time localization systems, considering that most LiDAR sensors function at a frequency of 10 Hz.

TABLE IX
RUNNING TIME.

Method	Descriptor Generation [ms]	Matching [ms]
PointNetVLAD [1]	18.8	2.65
OverlapTransformer [16]	**1.44**	**0.44**
MinkLoc3D-v2 [25]	10.5	2.13
SOE-Net [14]	20.7	1.49
PPT-Net [28]	16.9	3.05
RangePlace (last-map)	5.41	0.71
RangePlace (4-stage)	12.6	0.71

V. CONCLUSION

This work proposes a hierarchical range image transformer for LiDAR-based large-scale place recognition named RangePlace. The proposed design incorporates the Local Swin Transformer backbone, using the shift window-attention and depth-wise convolution to extract both global and local features effectively and efficiently. The Pyramid Feature Mix module and context gating mechanism, which is an effective strategy for descriptor aggregation in LiDAR-based place recognition, are employed. This innovative strategy allows for leveraging self-attention capabilities across various scales, capturing intricate details for retrieval. The proposed method is verified by experiments on different datasets and compared to state-of-the-art methods. The results demonstrate that the proposed method can achieve state-of-the-art performance in terms of the recall rate and has strong generalization ability. In addition, the representation of range images and the rotation-invariant network design contribute to the robustness of the proposed method against viewpoint yaw-rotation. Moreover, the proposed method achieves real-time and practical application.

Future work could explore the learning of global descriptors from multi-view images and monocular depth maps. Further, the learning of transferable pose estimation models with supervision from global descriptors could be studied.

REFERENCES

[1] M. A. Uy and G. H. Lee, "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4470–4479.

[2] A.-D. Doan, Y. Latif, T.-J. Chin, Y. Liu, T.-T. Do, and I. Reid, "Scalable place recognition under appearance change for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9319–9328.

[3] P. Shi, Y. Zhang, and J. Li, "Lidar-based place recognition for autonomous driving: A survey," *arXiv preprint arXiv:2306.10561*, 2023.

[4] C. Cao, H. Zhu, Z. Ren, H. Choset, and J. Zhang, "Representation granularity enables time-efficient autonomous exploration in large, complex worlds," *Science Robotics*, vol. 8, no. 80, p. eadf0970, 2023.

[5] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4878–4888.

[6] P. Shi, Z. Zhu, S. Sun, Z. Rong, X. Zhao, and M. Tan, "Covariance estimation for pose graph optimization in visual-inertial navigation systems," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 6, pp. 3657–3667, 2023.

[7] J. Deng, Q. Wu, X. Chen, S. Xia, Z. Sun, G. Liu, W. Yu, and L. Pei, "Nerf-loam: Neural implicit representation for large-scale incremental lidar odometry and mapping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8218–8227.

[8] D. Xiao, S. Li, and Z. Xuanyuan, "Semantic loop closure detection for intelligent vehicles using panoramas," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 10, pp. 4395–4405, 2023.

[9] G. Peng, Y. Yue, J. Zhang, Z. Wu, X. Tang, and D. Wang, "Semantic reinforced attention learning for visual place recognition," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 415–13 422.

[10] L. Chen, H. Kong, H. Wang, W. Yang, J. Lou, F. Xu, and M. Ren, "Hvp-net: A hybrid voxel- and point-wise network for place recognition," *IEEE Transactions on Intelligent Vehicles*, pp. 1–12, 2023.

[11] S. Zhao, P. Yin, G. Yi, and S. Scherer, "Spherevlad++: Attention-based and signal-enhanced viewpoint invariant descriptor," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 256–263, 2023.

[12] P. Shi, J. Li, and Y. Zhang, "A fast lidar place recognition and localization method by fusing local and global search," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 202, pp. 637–651, 2023.

[13] P. Shi, Y. Xiao, W. Chen, J. Li, and Y. Zhang, "A new horizon: Employing map clustering similarity for lidar-based place recognition," *IEEE Transactions on Intelligent Vehicles*, 2024.

[14] Y. Xia, Y. Xu, S. Li, R. Wang, J. Du, D. Cremers, and U. Stilla, "Soe-net: A self-attention and orientation encoding network for point cloud based place recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 343–11 352.

[15] Z. Zhou, C. Zhao, D. Adolfsson, S. Su, Y. Gao, T. Duckett, and L. Sun, "Ndt-transformer: Large-scale 3d point cloud localisation using the normal distribution transform representation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5654–5660.

[16] J. Ma, J. Zhang, J. Xu, R. Ai, W. Gu, and X. Chen, "Overlaptransformer: An efficient and yaw-angle-invariant transformer network for lidar-based place recognition," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6958–6965, 2022.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[18] T. Röhling, J. Mack, and D. 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)*, 2015, pp. 736–741.

[19] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3D point cloud map," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Madrid, Oct. 2018.

[20] Y. Wang, Z. Sun, C.-Z. Xu, S. E. Sarma, J. Yang, and H. Kong, "Lidar iris for loop-closure detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5769–5775.

[21] G. Kim, S. Choi, and A. Kim, "Scan context++: Structural place recognition robust to rotation and lateral variations in urban environments," *IEEE Transactions on Robotics*, 2021.

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

[23] Z. Liu, S. Zhou, C. Suo, P. Yin, W. Chen, H. Wang, H. Li, and Y.-H. Liu, "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*, 2019, pp. 2831–2840.

[24] J. Komorowski, "Minkloc3d: Point cloud based large-scale place recognition," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1789–1798.

[25] Jacek, "Improving point cloud based place recognition with ranking-based loss and large batch training," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 3699–3705.

[26] L. Luo, S. Zheng, Y. Li, Y. Fan, B. Yu, S.-Y. Cao, J. Li, and H.-L. Shen, "Bevplace: Learning lidar-based place recognition using bird's eye view images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8700–8709.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[28] L. Hui, H. Yang, M. Cheng, J. Xie, and J. Yang, "Pyramid point cloud transformer for large-scale place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6098–6107.

[29] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, 2018.

[30] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1269–1277.

[31] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[32] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "Mixvpr: Feature mixing for visual place recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2998–3007.

[33] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "Mlp-mixer: An all-mlp architecture for vision," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 24 261–24 272.

[34] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, 2022.

[35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[36] Z. Liu, H. Tang, Y. Lin, and S. Han, "Point-voxel cnn for efficient 3d deep learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[37] X. Chen, T. Läbe, A. Milioto, T. Röhling, O. Vysotska, A. Haag, J. Behley, and C. Stachniss, "OverlapNet: Loop Closing for LiDAR-based SLAM," in *Proceedings of Robotics: Science and Systems (RSS)*, 2020.

[38] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[39] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 17 683–17 693.

[40] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[41] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[42] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," *arXiv preprint arXiv:1706.06905*, 2017.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 630–645.

[44] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[45] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of michigan north campus long-term vision and lidar dataset," *The International Journal of Robotics Research*, vol. 35, no. 9, pp. 1023–1035, 2016.

[46] G. Pandey, J. R. McBride, and R. M. Eustice, "Ford campus vision and lidar data set," *The International Journal of Robotics Research*, vol. 30, no. 13, pp. 1543–1552, 2011.

[47] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

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

[49] K. Cai, B. Wang, and C. X. Lu, "Autoplace: Robust place recognition with single-chip automotive radar," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2222–2228.

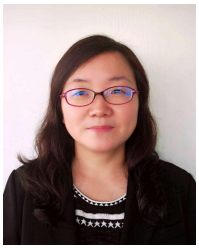
[50] L. Li, X. Kong, X. Zhao, T. Huang, W. Li, F. Wen, H. Zhang, and Y. Liu, "Rinet: Efficient 3d lidar-based place recognition using rotation invariant neural network," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4321–4328, 2022.

[51] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

Ji Li received the B.S. degree from the Beijing Institute of Technology, Beijing, China, in 2022. He is currently pursuing the M.S. degree in mechanical engineering with the Beijing Institute of Technology, Beijing, China. His research interests include robust and precise vehicle state estimation, multi-sensor integration perception and high precision localization of the intelligent vehicles.

Qingxiao Liu received the B.S. degree from the Beijing Institute of Technology, Beijing, China, in 2018. He is currently pursuing the Ph.D. degree in mechanical engineering with the Beijing Institute of Technology, Beijing, China. His research interests include intelligent vehicle environment perception and understanding, decision making, path/motion planning, and control.

Boyang Wang (Member, IEEE) received the B.S. degree from the Beijing Institute of Technology , Beijing, China, in 2015, and the Ph.D. degree from the Beijing Institute of Technology , Beijing, China, in 2020. He is currently an associate Researcher, School of Mechanical Engineering. His research interests include intelligent vehicle path/motion planning, and control.



Haiou Liu received the B.S. and Ph.D. degrees from the Beijing Institute of Technology, Beijing, China, in 1998 and 2003, respectively. She is currently a Professor with School of Mechanical Engineering, Beijing Institute of Technology. Her teaching interests focus on vehicle control classes at both undergraduate and graduate levels. Her current research interests include design and control of automated manual transmission and hybrid powertrain.

Yuxuan Han received the B.S. degree from the Sichuan University, Chengdu, China, in 2021. He is currently pursuing the M.S. degree in mechanical engineering with the Beijing Institute of Technology, Beijing, China. His research interests include intelligent vehicles, trajectory prediction, multi sensor perception.