



A fast and robust 3D LiDAR-based loop closure detection method using BEV images

Chizhou Zhang^{1,2} · Mingsong Chen² · Guanqiang Wang³ · Yongcheng Lin⁴ · Zehao Li²

Received: 27 August 2025 / Accepted: 28 November 2025 / Published online: 12 February 2026
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd. 2025

Abstract

In this paper, a novel method for loop closure detection (LCD) is presented, which utilizes multi-layer occupied bird's eye view (BEV) images. Initially, the original point clouds undergo preprocessing and are then projected into multi-layer occupied BEV images. Subsequently, to expedite the extraction of features from the BEV images, we employ a lightweight U-shaped network incorporating the Atrous Spatial Convolution Pooling Pyramid (ASPP) module. Lastly, a similarity measurement is devised, leveraging the rotation-invariance characteristic of the descriptor. The primary innovations center around the design of the lightweight U-shaped network, successfully achieved through the use of depth-wise separable convolution (DW-conv) and larger separable convolution kernels during the down-sampling process. Additionally, the incorporation of the ASPP module before the up-sampling process enriches the feature information derived from point clouds. Experimental comparisons conducted on the NCLT and MulRan datasets highlight significant improvements in both precision and efficiency achieved by the proposed method. Notably, the inference time is reduced by more than 20% when compared to the advanced DiSCO method. Furthermore, our LCD method has been seamlessly integrated into a comprehensive LiDAR-based simultaneous localization and mapping (SLAM) system, providing validation of its effectiveness.

Keywords Loop closure detection · Bird's eye view · Lightweight · Atrous spatial convolution pooling pyramid

1 Introduction

In the context of SLAM technology, LCD is critical for establishing exact positioning and mapping (Agunbiade 2025). The reason is that it helps to reduce the accumulated localization drift by finding the proper loop candidate frames. In the past few decades, the LCD method, which is based on

image processing technology, has been widely investigated. In the first research stage, the artificially designed features are designed to characterize image information. Based on lots of investigation of handcrafted features, it was found that this method has a limited ability to represent images. Convolutional neural networks (CNNs) have the ability to extract more detailed characteristics from images, potentially enhancing the visual LCD method's resilience. Then, LCD based on the CNN method has been intensively investigated in recent years. (Wang et al. 2018) introduced a novel place recognition method based on an omnidirectional camera, which utilizes an omnidirectional convolutional neural network (O-CNN) to deal with drastic changes in camera pose. (Orhan and Baştanlar 2021) adopted a novel slip-window searching technology based on panoramic images obtained by a CNN method to improve the accuracy and speed of retrieval simultaneously. However, the visual LCD method is greatly affected by weather, viewing angle, and light intensity, often resulting in the occurrence of incorrect loop candidate frames. Hence, due to the better robustness and reliability of the extraction of point cloud features, the

✉ Mingsong Chen
chenms18@csu.edu.cn

✉ Guanqiang Wang
wgq001101@gmail.com

¹ School of Artificial Intelligence, Hubei University of Automotive Technology, Shiyian 442002, China

² Light Alloy Research Institute of Central South University, Central South University, Changsha 410083, China

³ School of Mechanical and Electrical Engineering, Guilin University of Electronic Technology, Guilin 541004, China

⁴ School of Mechanical and Electrical Engineering, Central South University, Changsha 410083, China

point cloud LCD method based on 3D LIDAR raises the attention of many scholars (Zhang et al. 2024).

For an efficient LCD system based on 3D LiDAR, it is principal to solve these problems resulted from the disorder, sparsity and non-uniformity of point clouds. On the one hand, a robust 3D data representation needs to be obtained that enables the network to learn point cloud features for LCD. Commonly, there are four representation methods, i.e., unordered point sets (Zhou et al. 2021), voxel-based (Li et al. 2024), bird's-eye view (Lang et al. 2019), and image-like (Żywanowski et al. 2020). Li et al. (He et al. 2016) obtained an efficient and accurate descriptor by dealing with 2D features after the projection of the original 3D point cloud. On the other hand, the high-precision point cloud registration technology is sensitive to the initial rotation transformation. For instance, the matching based on the advanced point cloud registration (Zhao et al. 2025) will fail when the initial rotation angle of the point cloud is large. Therefore, it is necessary to estimate the yaw angle between loop candidate frames and history ones, and abandon the loop candidate frames if the yaw angle is too large. In addition, another issue that has to be resolved is the enormous number of model parameters and lengthy inference time for the deep-learning approach of place recognition. This is because the long-term real-time LCD requires a fast detection speed and a stable operating environment resulting from tiny model parameters.

Based on the analysis presented above, existing LCD methods are still confronted with several issues: (1) Descriptors constructed from single features extracted from point clouds lack strong discriminative capabilities for place recognition; (2) Accurate estimation of the yaw angle is missing in some methods; (3) Common similarity calculation methods are sensitive to the rotation of point clouds. To address these challenges, we propose a novel fast and robust LCD method based on 3D LIDAR place recognition. This method extends and applies the DiSCO recognition method (Xu et al. 2021) to enhance the performance of 3D scanning. Construction of a learned-based descriptor based on multi-occupancy-based BEV images and a DiSCO-based similarity measurement method are proposed.

To summarize, the primary contributions of this work are as follows:

(1) A lightweight place recognition network based on LiDAR point cloud called Fast-DiSCO is proposed to extract rotation-invariant features and estimate the yaw angle. The preprocessed BEV images are sent to an improved U-shaped backbone for a better and quicker feature fusion. On the one hand, the down-sampling process is lightweight by using the DW-conv and larger separable convolution kernels. On the other hand, the ASPP module is added before the

up-sampling process to generate more useful feature information on point clouds.

(2) A Fast-DiSCO-based similarity measurement method is created to calculate the similarity score by comparing the cosine distance between our designed matched features.

(3) The experimental results on NCLT and MulRan datasets demonstrate that our method outperforms existing state-of-the-art point cloud place recognition methods. Moreover, the physical experiments demonstrate that integrating our proposed into a LiDAR-based SLAM system improves the quality of the map and trajectory.

The rest of this paper is organized as follows: Sect. 2 describes related work about LiDAR-based LCD methods; Sect. 3 introduces the details implementation for the proposed LCD method; Sect. 4 presents experiment results based on public datasets and the SLAM application; Sect. 5 shows the conclusion of this work and an outlook for future work.

2 Related work

Feature-based Method. For the traditional point cloud LCD methods, the statistical histogram-based or signature-based methods are widely used to analyze the hand-made local and global features of point clouds. For the histogram-based method, the PFH (Point Feature Histograms) is the most classical one (Rusu et al. 2008). The PFH encodes the pose information related to the query points and surrounding points to achieve a good matching of point clouds. Moreover, the FPFH (Fast Point Feature Histograms) (Rusu et al. 2009) is proposed to minimize the temporal complexity of PFH features while keeping the original local point cloud features. By projecting point clouds onto many 2D planes and creating spatial density distributions for each plane, M2DP (He et al. 2016) first creates intermediate signatures. A manually created descriptor called Scan Context (Kim and Kim 2018) is effectively calculated from a BEV scan representation in polar coordinates. The human iris characteristic serves as inspiration for the LiDAR Iris method (Wang et al. 2020a), which uses it to efficiently identify loop closures. However, since these handcrafted methods focus on acquiring local features (Ren et al. 2025), the generalization ability of traditional descriptor extraction methods is poor, particularly in the case of reverse loops.

Segmentation-based Method. The method based on segmentation encodes the point cloud into a set of distinguishing features to reduce error matching. A common point cloud LCD approach, SegMatch (Dubé et al. 2017) is based on 3D segment matching. It uses Random Forest in conjunction with a RANSAC-based geometric verification phase to accomplish point cloud matching and characterize

the semantic components of a scene. Unlike SegMatch, SegMap (Dube et al. 2020) method better realizes LCD by integrating semantic information into point cloud segmentation. SemSegMap (Cramariuc et al. 2021) is a segment-based method that employs colour and semantic data in the environment to increase location accuracy and resilience. STV-SC (Tian et al. 2022) includes a three-step loop detection procedure in addition to a range image-based 3D point segmentation technique. A one-stage segmentation and temporal verification procedure and a two-stage candidate loop search process comprise the three-stage technique.

Learning-based Method. Their core ideas are to extract robust point cloud descriptors from preprocessed point clouds through deep network learning (Du et al. 2020; Milad et al. 2023; Uy et al. 2018) for LCD and place recognition. On the one hand, some methods directly extract the feature descriptors from the raw point clouds. For example, (Yan et al. 2021) introduced a self-attention and orientation encoding network (SOE-Net) that fully explores the relationship between points and incorporates long-range context into point-wise local descriptors. On this basis, (Yan et al. 2023) developed a place recognition method to fuse point-based and voxel-based approaches using cross attention transformers. On the other hand, a deep learning approach that does not process directly from the raw point cloud is to represent the scan of the point cloud as a bird-eye view representation. Xu et al. (Xu et al. 2021) proposed

an oriented differentiable scanning context, which solves the rotational invariance of the point cloud, but the network performance needs to be further improved. Different from the above, OverlapNet (Chen et al. 2021) projected the raw point cloud onto a range image and then calculated the overlap of the image to provide an estimated value of the yaw angle between pairs of point cloud frames. With the development of natural language in re-localization, Text2loc (Xia et al. 2024) proposed a text-submap global place recognition approach which is lighter faster and more accurate than previous methods.

3 Proposed method

3.1 System overview

Figure 1 depicts the framework of our proposed method, comprising two primary modules: data preprocessing, feature extraction and similarity calculation. Here is an overview of the process: (1) The initial step involves projecting the raw point clouds into multi-occupancy-based BEV images after undergoing data processing. (2) The descriptor is then extracted using the Fast-DiSCO block. (3) Simultaneously, the yaw angle is estimated through Fourier transform. (4) Next, potential loop candidates are identified based on the descriptors. (5) Finally, the system computes

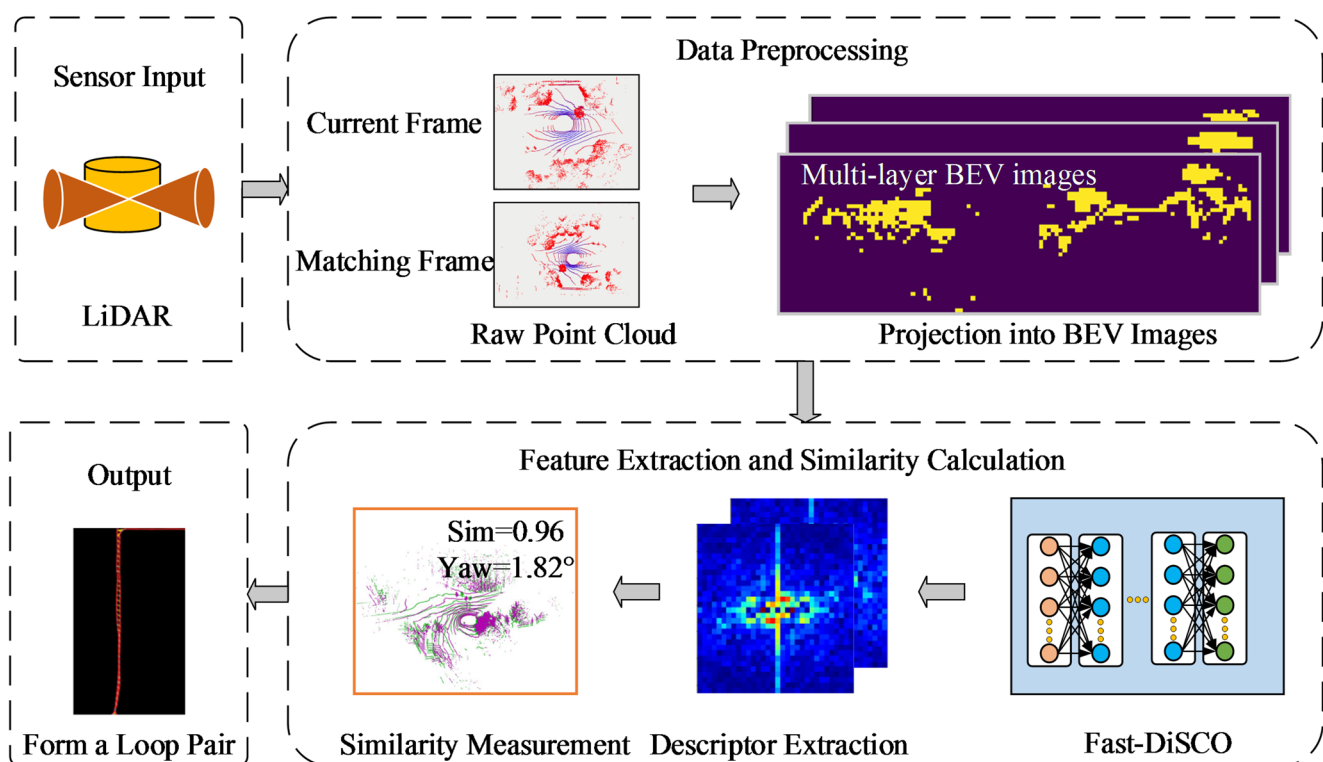


Fig. 1 Overview of our proposed method

the Manhattan distance between matched rotation-invariant descriptors to determine whether the matched point cloud frames form a loop-closing pair. Regarding the similarity measurement, the loop closure similarity can be calculated by measuring the similarity between the current frame and any frame in the database scans.

3.2 Construction of our proposed descriptor

The input 3D LiDAR point cloud $P_i(x_i, y_i, z_i)$ is quantized as a multi-layer BEV image (set the BEV representation as a function B), and the pixel values of the image are determined by the voxel values of the same height level of the point cloud voxelization (the empty voxels are considered to be unoccupied). To achieve the rotation invariance, a simple feature extraction module F was designed:

$$F(B(P, 0)) = F(B(P, \psi)) \quad (1)$$

Regarding the initial conditions, we set $\psi = 0$ to represent the BEV of the P without rotation. For ψ , it denotes the relative orientation.

Because for the revisited scene, there must be differences between point cloud frames. When database scan P_s is identified, the relative yaw angle estimation of the query scan P_q and P_s can be expressed as an identity mapping $\mathcal{H}$ and a global solver estimator φ :

$$\psi = \varphi(\mathcal{H}(B(P_s, 0)), \mathcal{H}(B(P_q, \psi))) \quad (2)$$

An intuitive approach is that letting $\mathcal{H}$ be an identity mapping and φ an exhaustive solver enables finding the globally optimal φ by enumerating all potential rotations. For revisited scenes, point clouds differ slightly, causing $\mathcal{H}$ to break down. Moreover, enumerating all possibilities is time-consuming for the exhaustive solver. Thus, a better solution is to seek an efficient global solver φ along with a matched mapping $\mathcal{H}$.

Finally, the Fourier transform F_f is applied to transform the multilayer BEV image from the original domain to the frequency domain representation. Since the amplitude components in the frequency domain are translation invariant, the final signature is also rotation invariant. Based on the above principles, the method implements an operation to construct the signature S as the output of F in the equation:

$$S = F(\cdot) = |F_f \circ F_c \circ F_p| \quad (3)$$

where the symbol $\circ$ represents the convolution operation.

To estimate the relative yaw angle at the same time, an optimal solver for two images is achieved by an exhaustive searcher:

$$\begin{aligned} \theta &= \arg \min_{\delta} \|G_q - G_s(\delta)\| \\ G_s(\theta) &= F_c \circ F_p(B(P_s, \theta)) \end{aligned} \quad (4)$$

Two frequency spectrums can be multiplied point-wise to obtain the convolution of two BEV images. Therefore, the yaw angle can be further specified as:

$$\theta = \arg \max_{\delta} F_f^{-1}(F_f(G_q) \otimes F_f(G_s(\delta))) \quad (5)$$

where $\otimes$ represents the point-wise multiplication.

Different from DiSCO method, in this paper, to obtain a more robust descriptor at a faster speed, an improved U-shaped network is used as the backbone network of the point clouds feature extractor. The long connections of U-Net (Huang et al. 2020) are adjusted to nested dense short connections to realize the fusion of spatial information between different network layers from shallow to deep. Meanwhile, the modules of the same layer are densely jumped. The connection makes the semantic information fusion more closely and minimizes the loss of information during feature extraction. The backbone network is divided into three parts: encoder, skip connection and decoder, as shown in Fig. 2. The module A represents the depth-wise separable convolution operation in the down-sampling process; the module B represents the depth-wise separable atrous spatial pooling pyramid (ASPP) operation before up-sampling, and the module C represents the feature output module after Fourier transform.

In Fig. 2, the U-shaped backbone network is used for deeper feature extraction and up-sampling restoration features, and the proposed modification lies in the feature extractor, where the orange module is proposed by DiSCO method.

In a more detailed description, the input to the backbone network is denoted as x_e^1 with dimensions $40 \times 120 \times 20$, representing the BEV feature map. The output feature map from the backbone network is denoted as x_{out} , with dimensions $40 \times 120 \times 1$.

Following Fourier transformation, this yields corresponding amplitude and phase information, resulting in an output feature size of $40 \times 120 \times 2$. Subsequently, a second-order norm operation is applied, and the central portion of the resulting data is extracted to generate a point cloud feature signature with a dimensionality of 1024. To enhance feature contrast, this 1024-dimensional feature is mapped into a 32×32 feature map. As illustrated in Fig. 3, the overall pipeline for this process can be visualized. The redundancy between adjacent multi-layer BEV images is implicitly addressed through the feature extraction network. The U-shaped backbone, enhanced with depth-wise separable convolution and the ASPP module, learns to compress and fuse multi-scale features from overlapping BEV layers.

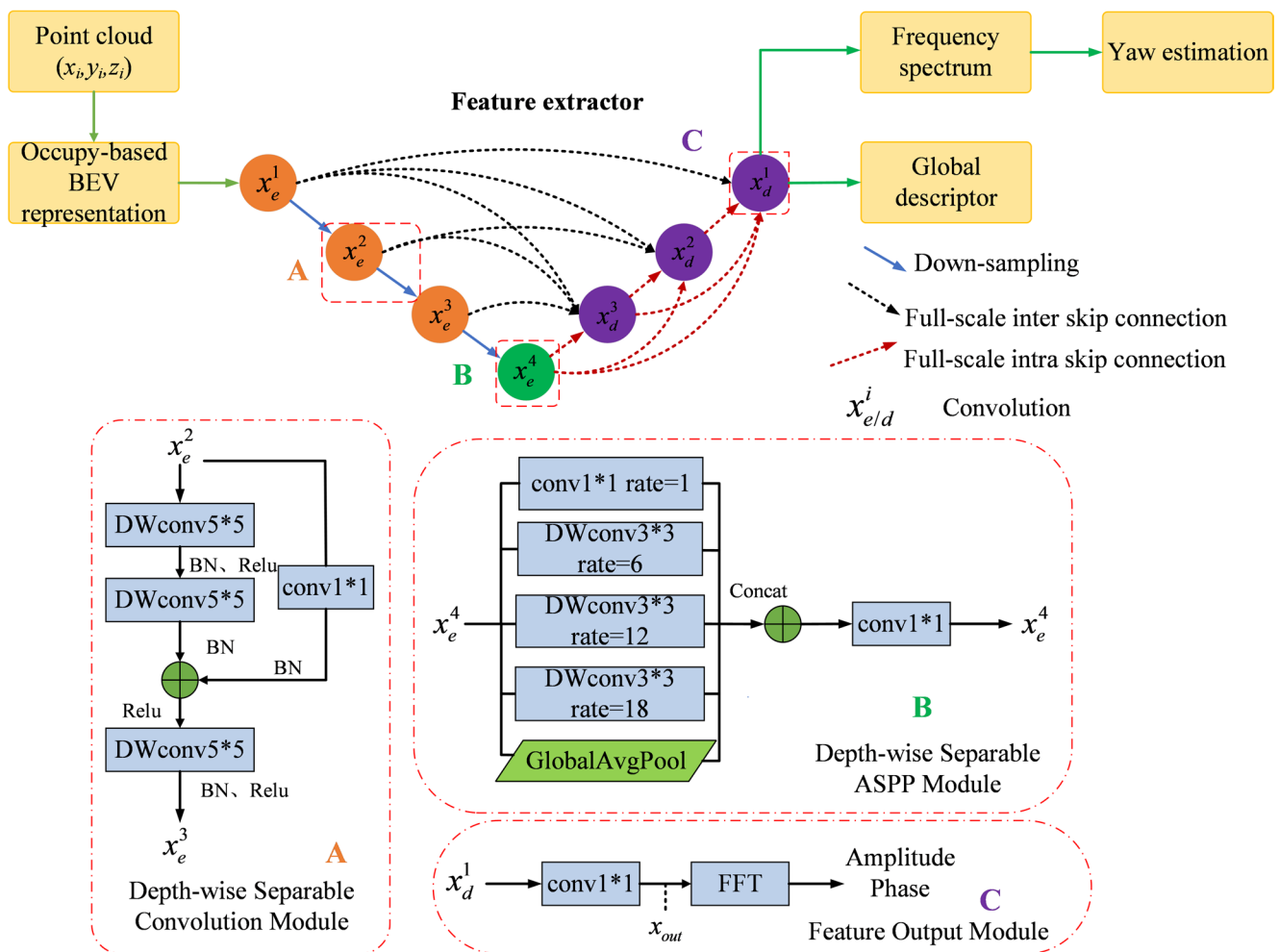


Fig. 2 Overall processing steps of the Fast-DiSCO architecture to extract descriptor and estimate yaw angle

Although not explicitly modeled as mutual information, the network effectively minimizes redundancy by emphasizing distinctive spatial features across layers during training, leading to a more compact and discriminative descriptor. Notably, it demonstrates that the feature maps of loop frame pairs within similar locations exhibit a high degree of similarity, whereas those of loop frame pairs in dissimilar locations show significant disparities.

Regarding yaw angle estimation, this can be accomplished using an estimator positioned after the feature extraction module and Fourier transformation, following the principles outlined in Formula (5). The process of yaw angle θ estimation, termed “phase correlation,” aligns with the methodology employed in the DiSCO method.

3.2.1 Depth-wise separable Convolution

In this paper, to carry out a lightweight transformation of DiSCO, a DW-conv is used in the feature extraction network to replace the standard convolution (Chollet and Xception

2017). The conventional convolution is split into depth-wise and point-wise convolutions by the DW-conv.

Take a set of feature maps of size $D_m \times D_n \times C_i$ as input, a set of feature maps of size $D_p \times D_q \times C_o$ as output, D_m and D_n are the width and height of the input feature map, D_p and D_q are the output feature maps. Width and height of C_i and C_o are the number of input image channels and the number of output image channels, respectively. The normal convolution kernel and depth-wise convolution kernel sizes are represented by $D_k \times D_k$, whereas the point-by-point convolution kernel has a size of 1×1 . The specific operation is shown in Fig. 4.

When only multiplication is considered, as to the depth-wise separable convolution and ordinary convolution, the ratio of the calculation amount is:

$$\frac{D_k \times D_k \times C_i \times D_m \times D_n + 1 \times 1 \times C_i \times C_o \times D_m \times D_n}{D_k \times D_k \times C_i \times C_o \times D_m \times D_n} = \frac{1}{C_o} + \frac{1}{D_k^2} \quad (6)$$

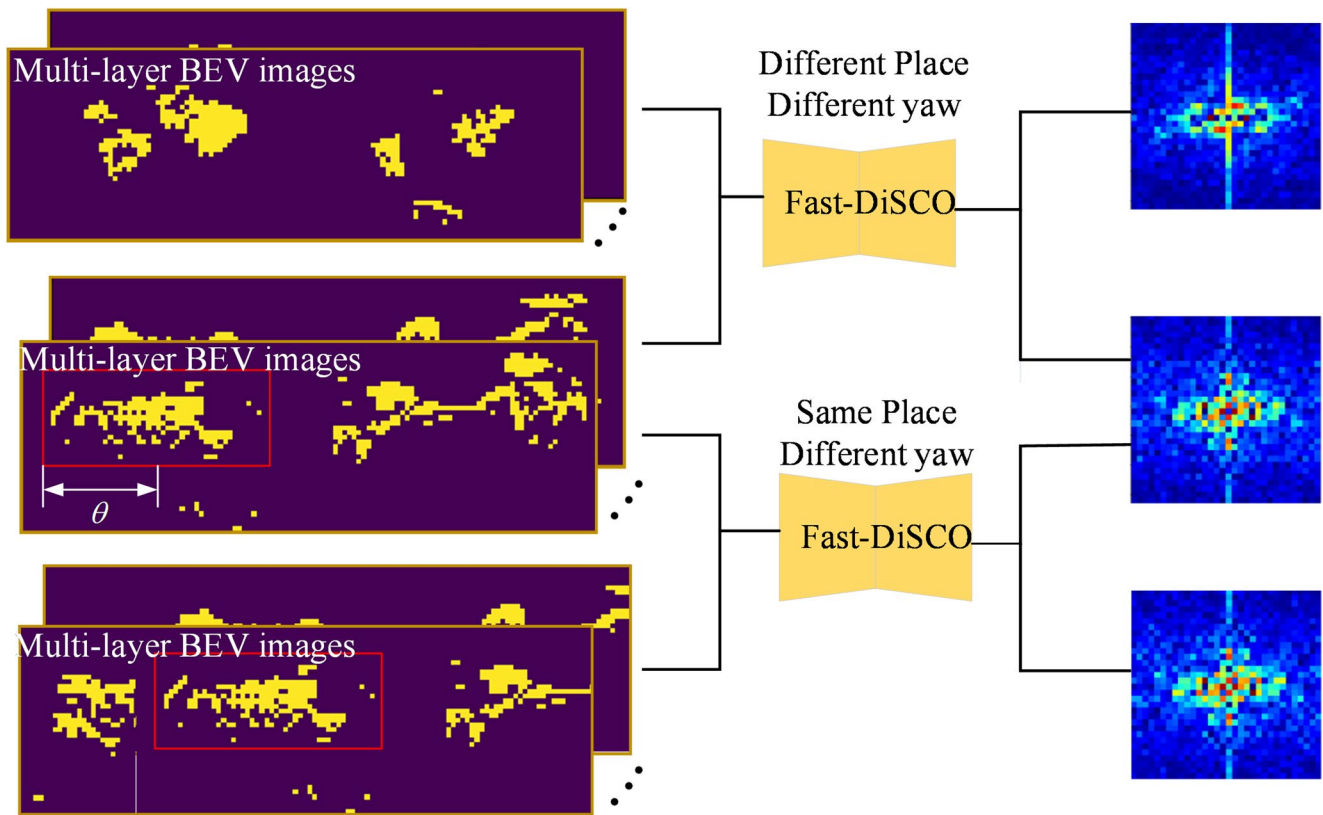


Fig. 3 Output visualization from different steps

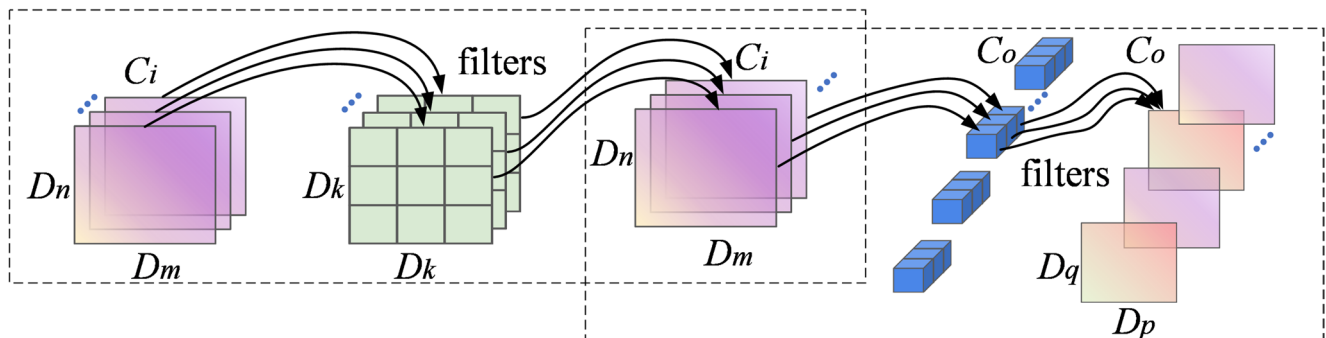


Fig. 4 Schematic diagram of depth-wise separable convolution

In addition, the ratio of its parameters is:

$$\frac{C_i \times D_k^2 + 1 \times 1 \times C_i \times C_o}{C_i \times C_o \times D_k^2} = \frac{1}{C_o} + \frac{1}{D_k^2} \quad (7)$$

It can be seen from the above analysis that the depth-wise separable convolution has the advantage of greatly reducing the amount of computation and model parameters while maintaining similar performance to standard convolution. To expand the receptive field and improve the network accuracy, the encoder in the point cloud feature extraction network will use 5×5 depth-wise separable convolution instead of 3×3 ordinary convolution. Although the

depth-wise separable convolution kernel is changed from 3×3 to 5×5 , which brings a small amount of computation, the receptive field is greatly increased. The increased receptive field is beneficial to improving the ability of the network to obtain important information. In addition, since the maximum continuous pooling operation in the encoding path will lose the feature map resolution, we use the convolutional layers to replace the pooling layers in the backbone to reduce the error rate of the network (Chen et al. 2018). The standard convolution and maximum pooling operations of the 1st ~3rd layer encoder is replaced by depth-wise separable convolution.

A better-performance network is more complicated, and there are some problems, such as the degradation of the model and even the disappearance or explosion of gradients. While the identity map is performed, its output is superimposed with the output of the stacked layers to solve the network degradation problem. The specific architecture is shown in Fig. 5. When the input and output feature dimensions are the same, the input feature and the output feature map of the residual learning unit are directly fused to obtain the final feature map; when the input feature and output feature dimensions are different, the input feature dimension is adjusted by 1×1 convolution, and then added and fused with the output feature of the residual learning unit to obtain the final feature map.

Given the excellent performance of the deep residual network in feature extraction, the network architecture of this paper uses the deep residual architecture as the backbone network to form a more efficient point cloud feature extractor.

3.2.2 Atrous Spatial Convolution pooling pyramid

The current encoder has a small standard convolutional receptive field, which leads to the problem of insufficient comprehensive capture of multi-scale feature information (Wang et al. 2020a, b). However, the atrous convolution can ensure the resolution attribute of the feature map. It realizes the effective perception of the convolution kernel with fewer parameters and a small amount of calculation. The expansion effect of the field reasonably aggregates the contextual information.

The ASPP module of multi-scale information fusion was first proposed by the Google team (Chen et al. 2017). Atrous convolution with multiple levels of atrous sampling rates is used to sample the feature map in parallel. At the same time, to reduce the parameters of the ASPP module, we use a DW-conv to optimize its architecture, and its architecture is shown in Fig. 6.

Consequently, within our network, before the up-sampling step, the feature map containing high-level information is fed into the ASPP module. This arrangement enables the ASPP module to acquire BEV image features from various receptive fields, thus extracting features of diverse scales.

3.3 Similarity measurement based on our proposed descriptor

Prior to the similarity measurement phase, the Fast-DiSCO block is responsible for computing rotation-invariant descriptors. Subsequently, suitable loop closure candidates are selected. Given that the rotation-invariant descriptors are represented as vectors, we employ the cosine distance metric to measure similarity, as delineated in Algorithm 1. The distance between these descriptors is determined through the following calculation:

$$d(P_q, P_s) = 1 - \frac{\mathbf{f}_1 \mathbf{f}_2}{\|\mathbf{f}_1\| \|\mathbf{f}_2\|} \quad (8)$$

where $\mathbf{f}_1 = [x_1, x_2, \dots, x_{1024}]$ and $\mathbf{f}_2 = [y_1, y_2, \dots, y_{1024}]$ are 1024-dimensional descriptors of P_q and P_s respectively using Fast-DiSCO block.

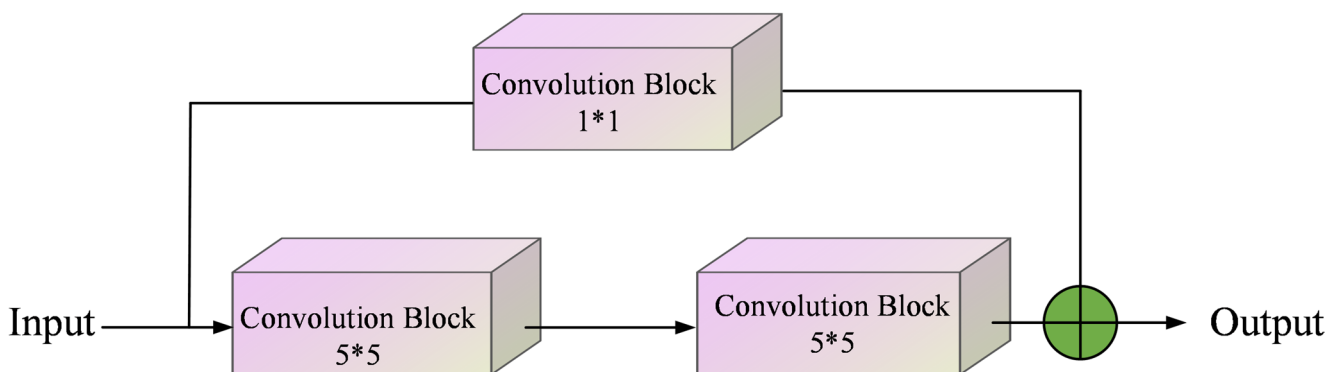


Fig. 5 Schematic diagram of residual architecture network

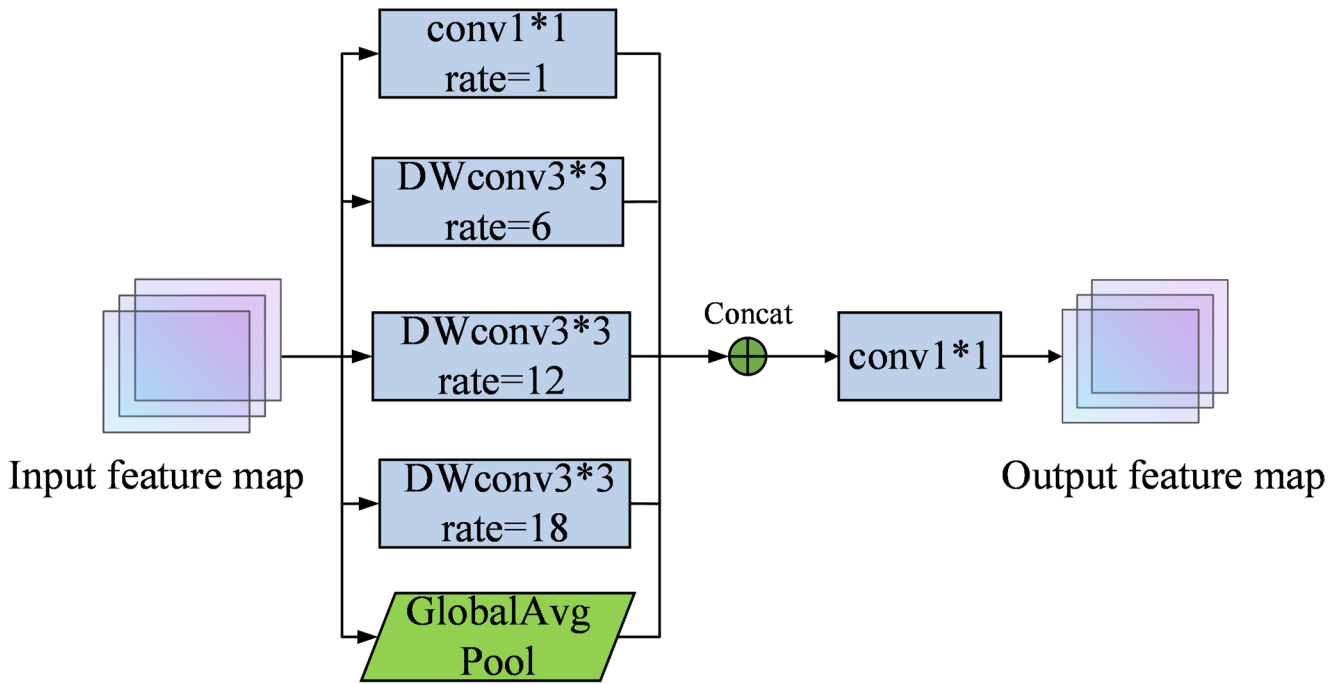


Fig. 6 Schematic diagram of ASPP module based on depth-wise separable convolution

Input: The Current Point Cloud: P_q , Database of Historical Point Cloud: $D_h = \{P_h^i\}(i = 1, 2, \dots)$

Output: The Index of Historical Point Cloud on the Database

```

1: while  $P_q$  is not Empty do
2:   Compute descriptor of  $P_q$  and  $P_h^i(i = 1, 2, \dots)$  by Fast-DiSCO Block
3:   Build the Kd-Tree of  $P_h^i(i = 1, 2, \dots)$ 
4:   if Not Find Similar Descriptors in Kd-Tree of  $P_h^i(i = 1, 2, \dots)$  then
5:     Break;
6:   else
7:     Find Ten Most Similar Descriptors in Kd-Tree of  $P_h^j(j = m, \dots, m + k)$ 
8:   end if
9:   for  $i = 1$  to  $k$  do
10:    Calculate Distance  $C_h^i(j = m, \dots, m + k)$  between  $P_q$  and  $P_h^j(j = m, \dots, m + k)$  by Using Formula (8)
11:  end for
12:  if  $\exists C_h^j(j = m, \dots, m + k) < d_{threled}$  then
13:    Return index;
14:  else
15:    Break;
16:  end if
17: end while

```

Algorithm 1 The Proposed LCD Method

4 Experimental validation

In this part, we use two public datasets to illustrate the recall and efficiency of the proposed Fast-DiSCO. Additionally, we present a thorough experimental evaluation of progressive processes and individually examine the place recognition and orientation estimation models. Furthermore,

we extend our LCD method to a complete LiDAR-SLAM system.

The effectiveness of our method is verified on self-collected datasets by our robot platform. Additionally, we evaluate the impact of our LCD method on the positioning error in the z-direction through real physical experiments. Figure 7 shows the robotic platform equipped with two sets of motors and power drivers from Agile-X Robotics, a high-density lithium battery, a Realsense D435i camera for visualizing the surrounding environment, and an embedded NVIDIA XAVIER AGX device that runs the Robot Operating System (ROS) with Ubuntu 18.04 operating system. Additionally, an LS-C16 LiDAR is used to sense the surroundings and estimate the robot's position.

4.1 Implementation and training

Our model is trained and tested on a desktop equipped with a computer with AMD Ryzen 7 4800 H CPU @ 2.90Ghz and NVIDIA GeForce GTX 1650 Ti with 4GB memory. During training, the resolution of BEV images is 40×120 , the number of image channels is 20, and each image pixel corresponds to 1 m. The deep learning framework used by this network is Pytorch, and the Adam optimizer is used for training. The algorithm parameters are set as follows. The margin of the quadruplet loss function α_1 is 0.5, while α_2 is 0.2. The network learning rate η is 0.00001. The mini-batch size is set as 16, while the number of training rounds is 25.

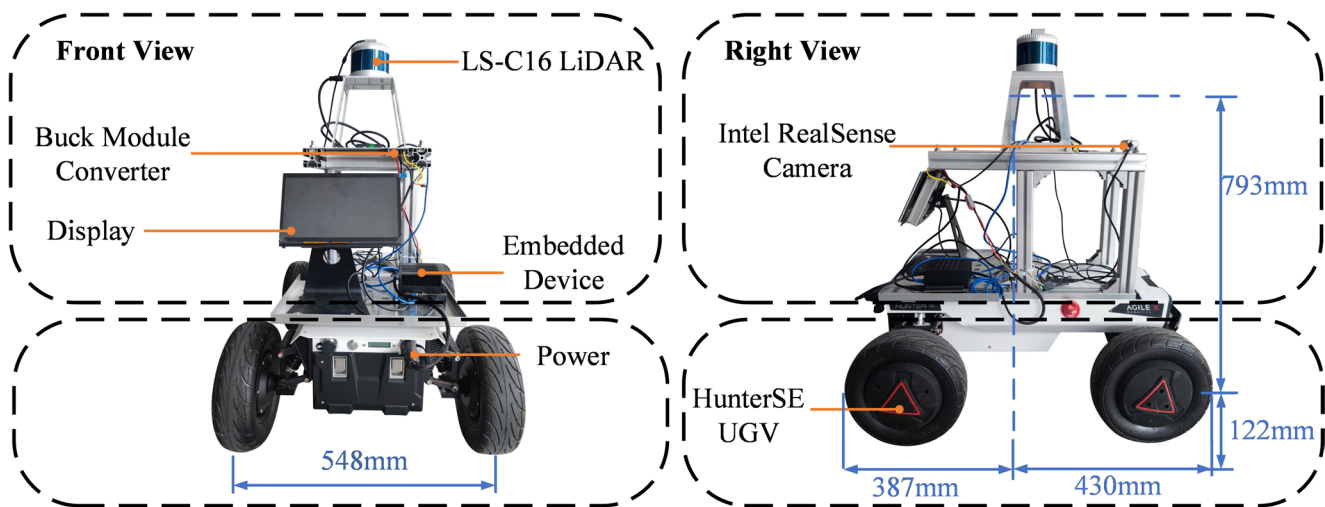


Fig. 7 Our mobile robot platform

4.2 Benchmark datasets and evaluation measures

4.2.1 Benchmark datasets

We use the NCLT (Carlevaris-Bianco et al. 2016) and MulRan (Kim et al. 2020) datasets for real-world experimental validation. Before inputting each frame of the point cloud, the information of the ground plane will be filtered out because the ground point cloud is repetitive. After that, all these point clouds are down-sampled to $[-1, 1]$ m.

The NCLT dataset is a long-term time LiDAR dataset collected by a Segway robotic platform at the University of Michigan North Campus. The inspection hardware equipment in this platform is the Velodyne HDL-32E 3D LiDAR with a collection frequency of 10hz. In the NCLT dataset, point cloud pairs separated by less than 1.5 m are regarded as positive samples, while point cloud pairs separated by more than 2 m are regarded as negative samples. We used session “2012-02-04” to train the lightweight model and used sessions “2012-03-17” “2012-06-15” “2012-08-20” for testing. Finally, 9650 point cloud frames are generated for training our network, and 2429 point cloud frames for testing. If the retrieved point cloud location is within 1.5 m in the testing stage, the query frame point cloud is recognized.

The MulRan dataset is a multimodal range dataset used for urban place recognition. This data is provided by an Ouster OS1-64 3D LIDAR with a collection frequency of 10hz. Like the NCLT dataset, we also divided positive and negative samples. At the same time, we used session “DDC03” for training the model, and sessions “DCC01” and “DCC02” for evaluating the generalization ability of the model. Finally, 785 point cloud frames are generated for training our network, and 878 point cloud frames are for

testing. If the retrieved point cloud location is within 3 m in the testing stage, the query frame point cloud is recognized.

Our established dataset consists of short-term LiDAR dataset collected by our mobile robot platform in both indoor and outdoor environments, as illustrated in Fig. 8. Among it, colors in point cloud maps encode the intensity values of the point clouds. Point cloud intensity is closely related to factors like the reflective properties of objects’ materials, the incident angle of LiDAR, and the distance between objects and the sensor. For example, objects with high reflectivity (such as metal surfaces) often appear in warmer hues (like yellow or orange), while areas with lower reflectivity (such as some non-metallic or rough surfaces) tend to show cool hues (like blue or green). The indoor environment is collected from a long corridor, and the outdoor dataset is collected from roads on the campus. Due to the low vertical resolution of low-line lidar, the SLAM method exhibits significant drift in the z-axis direction over an extended period. To validate the effectiveness of our proposed LCD method for SLAM positioning, only the positioning data along the z-axis is considered. We also use a leveling tool to measure the vertical distance and evaluate the vertical accuracy.

4.2.2 Evaluation measures

The testing dataset is evaluated by picking a query location and matching it to a collection of remaining locations known as the database. The Euclidean distance of the descriptors obtained by different methods for these sites is used for matching. The Recall@X assessment is predicated on the assumption that the recognition of a location is considered correct if, among the top X most similar locations retrieved from the database, there exists at least one location within the threshold distance from the query location. Recall@1 and Recall@1% are deemed to be the most

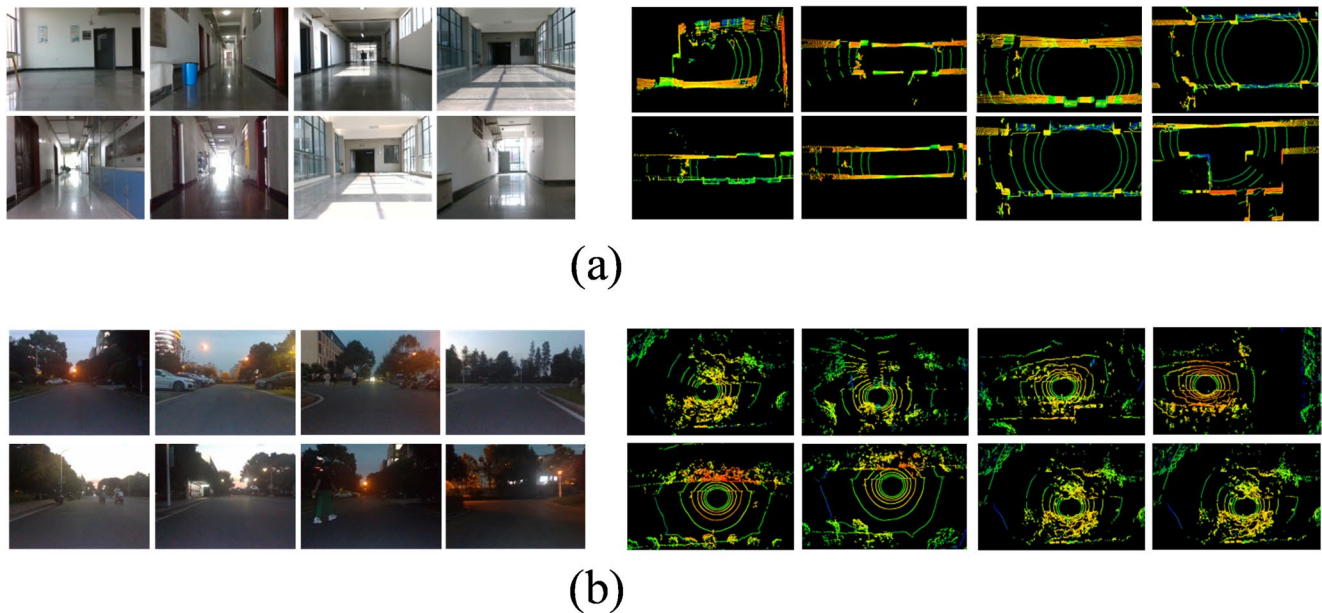


Fig. 8 Visualization diagram of self-collected datasets. (a) Indoor corridor; (b) Outdoor campus

crucial evaluation metrics in LCD. Additionally, the evaluation metrics for heading consist of the average mean errors of heading (Mean) and the standard error of heading (Std). Regarding the comparison of efficiency, the evaluation metrics include the number of model parameters and the inference time of the network.

4.3 Results

4.3.1 Experiments on public datasets

Four competitive place recognition methods, Scan Context (Kim and Kim 2018), PointNetVLAD (Uy et al. 2018), DiSCO (Xu et al. 2021) and OverlapNet (Chen et al. 2021), are selected for comparison.

Scan Context is a handcrafted point cloud descriptor. A stable point cloud global descriptor is extracted from the original point cloud by transforming it into a bird's-eye view of size $\text{Ring} \times \text{Sector}$. The ring and sector parameters respectively are set to 40×120 .

PointNetVLAD is the first deep-learning-based point cloud place recognition method. Among them, PointNet is used to extract point cloud features, and NetVLAD is used to encode features. Finally, it can obtain 1024-dimensional feature descriptors based on Tensorflow version.

DiSCO uses multi-layer BEV images (the size is $40 \times 120 \times 20$) as input, and obtain 1024-dimensional feature descriptors by a feature extractor. Robust scene recognition and more accurate yaw angle estimation can be performed simultaneously through the network.

OverlapNet can solve the place recognition problem of point clouds by learning the overlap ratio and relative yaw angle in the range image. Due to the inefficiency in the retrieval process, only the relative yaw angle part can be evaluated.

1) Place recognition analysis.

More than one place candidate is required for testing in an actual place retrieval situation to ensure robustness. Thus, the top N candidates are used to assess the overall performance of the place recognition. Corresponding to the previous state-of-the-art methods, the N value for both the NCLT and MulRan datasets is 10. Then we run our method and the comparison method respectively to get the scene recognition recall results. In the NCLT dataset, we use section “2012-02” with many loopback scenarios as the basic database, and take section “2012-06” with loopback scenarios as an example to test and obtain the corresponding PR curve, the experimental results are shown in Fig. 9; Table 1

Figure 9(a) and Fig. 9(b) show that the scene recognition ability embodied by our method has a large competitive advantage. Among them, to test the online recognition ability of our algorithm, we select enough 2000 new position pairs (positive and negative sample ratio is 1:1) in Section “2012-06-15” in the NCLT data set for verification, as online unseen scene recognition results. It can be seen from Fig. 9(c) that our method has the best recognition ability for unseen scenes compared with other methods. PointNetVLAD exhibits relatively poor detection performance due to its failure to account for rotational changes in scenes. This shortcoming is particularly evident in its inability to detect reverse loop closures and the rapid decline in curve

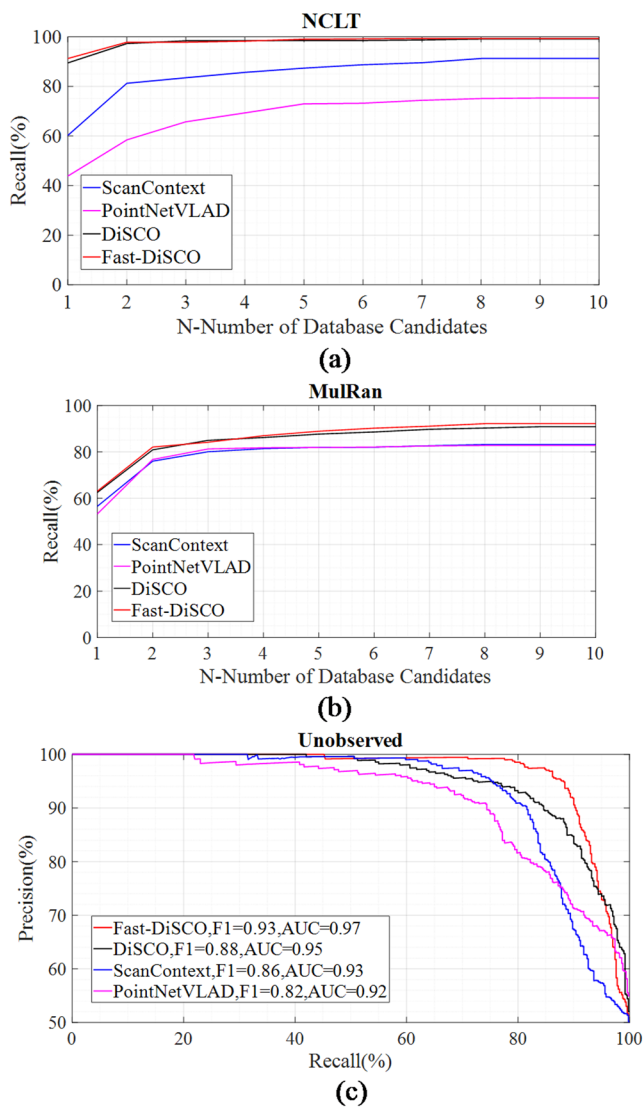


Fig. 9 Comparison of place recognition performance between Fast-DiSCO and other methods on NCLT and MulRan datasets: (a) average recall rate for location recognition on NCLT dataset; (b) average recall rate for location recognition on MulRan dataset; (c) precision-recall curves obtained by running the unobserved place recognition in Section “2012-06-15” of NCLT dataset

Table 1 Average recall rate of different methods on NCLT and MulRan datasets

Dataset	Method	Recall@1%	Recall@1
NCLT	ScanContext	91.25%	60.21%
	PointNetVLAD	75.32%	43.82%
	DiSCO	99.10%	89.47%
	Ours	99.35%	91.23%
MuRan	ScanContext	83.21%	56.48%
	PointNetVLAD	82.75%	53.20%
	DiSCO	85.94%	62.48%
	Ours	84.36%	63.02%

Table 2 Recall rate on the rotated NCLT dataset

Dataset	Method	Recall@1%	Recall@1
NCLT	ScanContext	86.05%	58.76%
	PointNetVLAD	64.21%	34.21%
	DiSCO	95.18%	88.72%
	Ours	96.12%	90.67%

slopes. ScanContext, while theoretically possessing rotation invariance, heavily relies on vertical height information, limiting its performance when changes in vertical height are not pronounced in the scene. In contrast, both DiSCO and our proposed Fast-DiSCO method exhibit relatively strong detection performance, primarily relying on multi-occupied BEV images to address the rotation transformation challenge. Differing slightly from the DiSCO method, the Precision-Recall curves of our method show a marginal decrease in precision rate with increasing recall rate.

For the sake of comparison, we test ScanContext method on the NCLT and MulRan datasets, respectively, achieving a Recall@1% of 91.25% and a Recall@1 of 60.21% on the NCLT dataset, and on the MulRan dataset the Recall @1% is 83.21% and Recall@1 is 56.48%. For the PointNetVLAD method, the Recall@1% is 75.32% and Recall@1 is 43.82% on the NCLT dataset, and the Recall@1% on the MulRan dataset is 82.75% and Recall@1 is 53.20%. By using DiSCO method, and the Recall@1% on the NCLT dataset is 99.10% and Recall@1 is 89.47%, and the Recall@1% on the MulRan dataset is 85.94% and Recall@1 is 62.48%. It can be noticed the DiSCO method is competitive because of its great performance.

Our method Fast-DiSCO shows a better performance compared to the DiSCO method. On the NCLT dataset, the Recall@1% is 99.35% and Recall@1 is 91.23%. The performance on the MulRan dataset is also quite competitive, where Recall@1% is 84.36% and Recall@1 is 63.02%. This proves the feasibility and superiority of the improved method in large-scale point cloud location recognition.

2) Rotation invariance analysis.

To substantiate the rotation invariance of our proposed descriptor, we conducted an experiment involving the random rotation of point clouds within the NCLT Section “2012-06-15” to simulate changes in viewpoint. As depicted in Table 2, our method exhibits a higher recall rate when compared to other methods. When considering the results from both Tables 1 and 2, it becomes evident that our proposed descriptor indeed preserves rotation invariance. However, it is worth noting that PointNetVLAD struggles to generalize effectively to the rotated NCLT dataset. This limitation can be attributed to the higher sensitivity of the PointNetVLAD descriptor to rotation.

3) Heading estimation analysis.

Based on the ground truth orientation of point clouds in the NCLT and the MulRan datasets, the accuracy of yaw angle estimation for the proposed method is evaluated.

The yaw angle estimation errors calculated according to the DiSCO, OverlapNet and Fast-DiSCO were listed in Table 3. The results show that the mean and standard estimated errors for DiSCO method and our method not only on the NCLT dataset but also on the MulRan dataset are much lower than the values calculated by OverlapNet method. Moreover, the estimated error for our method in the MulRan dataset is lower than 3° . It means that the optimized network can extract features better and has the best accuracy for the estimation of direction.

4) Progressive experimental studies.

To ensure the place recognition ability of the lightweight network, the main modifications can be summarized as follows:

Point1: Replacing the pooling layer in the down-sampling module with a depth-wise separable convolution;

Point2: Replacing all 3×3 depth-wise separable convolutions with 5×5 depth-wise separable convolutions;

Point3: Introducing residual architecture in the encoding process;

Point4: Adding ASPP module before up-sampling.

The improvement of the network performance with the above four modifications is verified based on the MulRan dataset which has a shorter training time to evaluate. First, the original network without the above-mentioned four changes is used as the baseline model. After that, the above-mentioned changes are applied in turn on the original network to carry out ablation experiments. During ablation experiments, the recall rate and yaw angle estimation error are used as evaluation criteria, and the experimental results are shown in Table 4. It can be seen that the model was underfitting at the beginning. With the introduction of improved points, the place recognition performance of the model gradually improved.

Furthermore, our analysis also extends to the influence of improvement point 4 on scene recognition capabilities. As demonstrated in Fig. 10, the inclusion of the ASPP module in the model allows for the generation of more valuable feature information from the point clouds. In contrast, the model lacking the ASPP module tends to lose certain

Table 3 Yaw angle estimation for different methods on NCLT and MulRan datasets

Dataset	Method	Mean(deg)	Std(deg)
NCLT	OverlapNet	13.52	26.18
	DiSCO	2.93	2.93
	Ours	1.26	1.85
MulRan	OverlapNet	4.25	15.32
	DiSCO	3.67	11.25
	Ours	2.55	2.99

critical feature information. This highlights the significance of improvement point 4 in enhancing the model's scene recognition abilities by preserving essential feature details. Additionally, the colormaps are applied to visually signify the activation strength in the feature maps, with warmer colors (e.g., red) indicating higher activation.

5) Improvement of LiDAR-based SLAM.

In this experiment, the proposed LCD approach is integrated into the FAST-LIO2 (Xu et al. 2022) and a loop closure constraint is established by using FastGICP (Koide et al. 2021) as a geometric verification method to match the current point clouds frame with historical frames when loop closure is detected. After that, the LiDAR odometry factor acquired at the front end of FAST-LIO2 and the loop closure factor generated in our LCD approach are both added into the global factor graph for mapping after back-end optimization. Figure 11 depicts the estimated trajectory on NCLT sequence "2012-06" after adding our loop closure restriction into FAST-LIO2. The detected loop closures are represented with black lines in this figure, while the trajectory is highlighted with red line. It can also indicate that the loop closures observed on the NCLT section are entirely closed in the map and have no overlap in the estimated trajectory. Moreover, the blue point cloud map shows few ghosts, which further shows that our LCD can improve the quality of the map.

To provide a visual representation of the loop candidate extraction process, we selected a point cloud frame within the green box from Fig. 11 for verification, as depicted in Fig. 12. In this illustration, the current frame is projected into multi-layer BEV images. Next, Fast-DiSCO is used to compute the descriptor of the current frame. The next step involves the rapid retrieval of the ten most similar loop candidates from the Kd-tree, based on cosine distance

Table 4 Performance evaluation of each modification

Method	Improvement point 1	Improvement point 2	Improvement point 3	Improvement point 4	Place recognition(%)		Yaw angle(deg)	
					Ave recall@1%	Ave recall@1	Mean	Std
1	×	×	×	×	79.25	55.12	3.23	8.15
2	√	×	×	×	80.12	59.38	3.12	6.32
3	√	√	×	×	82.08	61.23	2.95	5.14
4	√	√	√	×	83.95	62.78	2.63	3.29
5	√	√	√	√	84.36	63.02	2.55	2.99

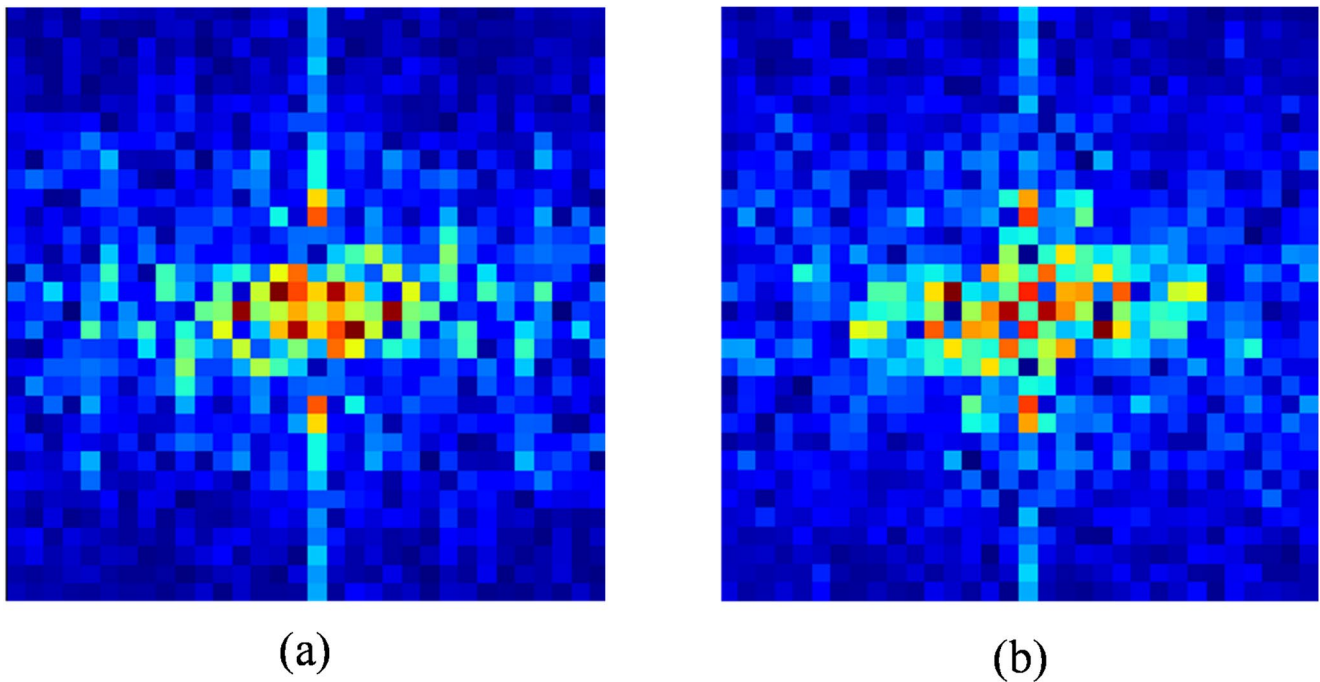


Fig. 10 Visualization of feature map w/o ASPP module. **(a)** model without ASPP; **(b)** model with ASPP



Fig. 11 The resulting trajectory and map of NCLT sequence “2012-06-15”

measurements. Following this retrieval, geometric verification is performed, resulting in the identification of a final loop closure candidate frame. Additionally, it is worth noting that our method demonstrates effective detection

capabilities even in scenarios involving reverse loops, as evident from the illustration.

6) Runtime evaluation.

Generally, a better detection performance is related to a more complex model. However, the number of model

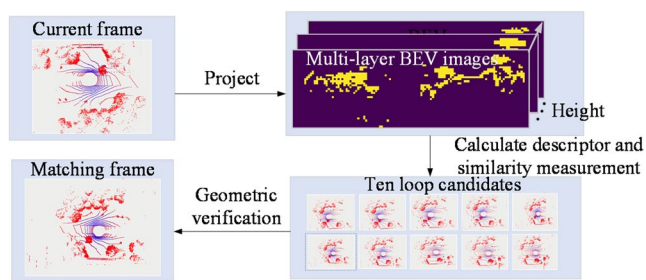


Fig. 12 Execution result of selection of loop candidate

Table 5 Comparison of model parameters

Model	Parameters(M)
DiSCO (filters: [64,128,256,512])	7.77
Ours (filters: [32,64,128,256])	1.74

parameters will increase when the model is more complex. Therefore, a balance between accuracy and complexity should be made. Table 5 shows the number of parameters for the proposed network and its original DiSCO network. It can be seen that our method compresses the number of model parameters, and effectively maintains the network performance. Compared with the DiSCO, the number of model parameters is only 22.39% of that.

Moreover, we use the two approaches described below to assess the efficiency and computing demands of our LCD method using the two points listed below: (1) Examine the temporal complexity of our LCD approach, which is $O(L \times W + K)$. The BEV image's dimensions are L and W , and its number of point cloud frames is K . (2) Calculate the average execution time of each point cloud frame on MulRan dataset. More specifically, we calculate the time required for the processing of each point cloud frame from the initial preprocess to the final selection of the most comparable candidate point cloud frames. Furthermore, we divided the overall processing time into four parts: pre-processing of the point cloud frame, signature generation, yaw estimation, and similarity measurement with Kd-Tree. Table 6 shows the average computation execution time per frame of point clouds for DiSCO and our method.

As can be seen, the time cost of the improved algorithm in the signature generation module and the yaw angle estimation module is less than that of DiSCO, the total execution time is reduced by 7.75 ms, and the speed is increased by 30.60%. Our LCD method is 49.71 ms, which has met the real-time requirement since the data sampling frequency for most LiDAR is about 10hz.

4.3.2 Physical experiments

The effectiveness of our LCD method is validated through a series of indoor and outdoor experiments, focusing on its impact on SLAM. We utilize the FAST-LIO2 method as the SLAM method and integrate our loop detection algorithm into this method. Mapping results in the indoor and outdoor environments are depicted in Fig. 13. As shown in the red area of Fig. 13(a) and Fig. 13(c), the map generated by the SLAM method without loop closure exhibits ghosting. In contrast, the map produced by the SLAM algorithm incorporating our LCD method shows no ghosting, and the trajectories of the closed loop coincide, as illustrated in Fig. 13(b) and Fig. 13(d). In the outdoor environment, a loop occurs around 320s, and the position at the initial moment is passed again. At this time, the positioning in the z-axis direction is corrected, as shown in the black area in Fig. 14. Therefore, the loopback detection method significantly reduces the cumulative positioning error in the z-axis direction.

4.4 Discussion

The core conceptual advancement of our proposed approach is its unified approach to achieving robust viewpoint and domain generalization through a rotation-invariant descriptor learned from multi-scale contextual features. While implemented via architectural refinements (DW-conv, ASPP), these choices are driven by key algorithmic insights:

Generalization to Unseen Viewpoints: The method explicitly enforces rotation-invariance not just in the descriptor's design, but throughout the learning process. By combining the Fourier transform's inherent property with a network trained on multi-layer BEV images, it learns features that are inherently disentangled from the robo's in-plane rotation. This directly enables reliable recognition in reverse loops and from significantly altered viewpoints, a common failure point for non-invariant methods.

Improved Global Context Modeling: The integration of the ASPP module is not merely a performance booster; it is critical for hierarchical feature fusion. It allows the network to simultaneously capture fine-grained local structures and broader semantic patterns (e.g., the layout of an intersection vs. a corridor) from the BEV representation. This multi-scale understanding is the algorithmic key to distinguishing perceptually similar but globally distinct locations, thereby reducing false positives.

Table 6 Comparison of the average computation execution time in the MulRan dataset

Dataset	Method	Preprocessing (ms)	Signature generation(ms)	Yaw estimation(ms)	Similarity measurement
MulRan	DiSCO	0.45	19.65	5.23	/
	Ours	0.45	15.28	1.85	32.13

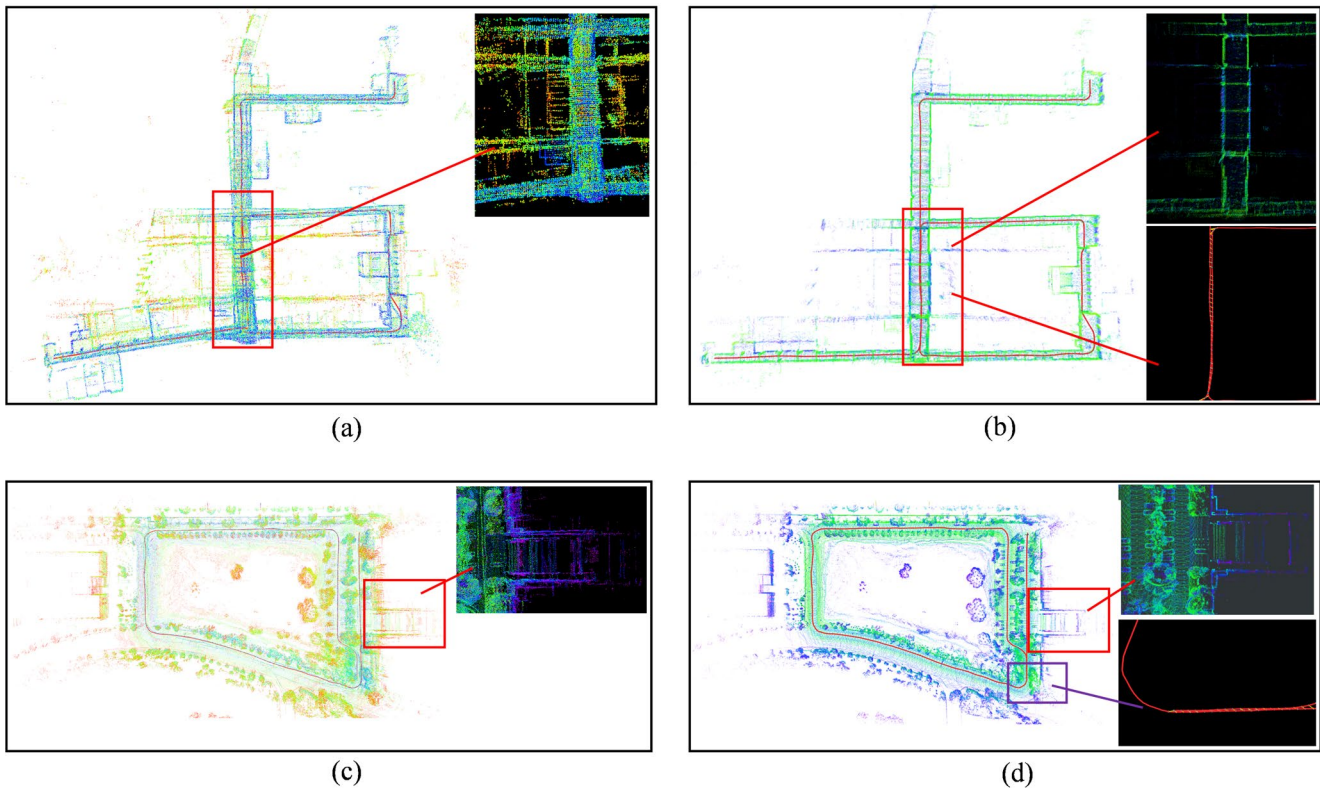


Fig. 13 mapping results in the indoor and outdoor environments. (a) mapping result without LCD in the indoor long corridor environment. (b) mapping result with our proposed LCD in the indoor long corridor

environment. (c) mapping result without LCD in the outdoor campus environment. (d) mapping result with our proposed LCD in the outdoor campus environment

Handling Cross-Domain Changes: The lightweight design, based on DW-conv, does more than just speed up computation. It acts as a strong regularizer, forcing the network to learn a more essential and robust feature set from the BEV images with far fewer parameters. This reduces overfitting to the training data's specific noise and density patterns, leading to improved performance across different environments (indoor corridor) and datasets (NCLT/MulRan) as demonstrated, which is a hallmark of better cross-domain adaptability.

While the proposed LCD method demonstrates strong performance, it has certain limitations. First, the descriptor's construction from multi-layer BEV images inherently relies on the vertical structure of the environment. Consequently, it may be sensitive to significant vertical misalignments caused by factors like strong pitch or roll errors in the LiDAR odometry or large changes in terrain elevation, which could distort the occupancy information across the BEV layers. Second, the method's effectiveness is tied to the vertical resolution and scanning pattern of the LiDAR sensor. It is optimized for and tested on specific LiDAR models, such as the 16-channel LS-C16 unit employed in our experimental evaluations. Consequently, its performance may vary when used with sensors that exhibit significantly

different vertical field-of-view or resolution characteristics. This includes, for example, solid-state LiDARs characterized by an exceptionally narrow vertical aperture.

5 Conclusion

In this paper, we propose a novel method for loop closure detection. On the one hand, a novel global descriptor of point clouds has been constructed by the Fast-DiSCO. On the other hand, a similarity measurement based on our proposed descriptor is proposed. Based on the experimental NCLT and MulRan datasets, it can be found that the proposed Fast-DiSCO can simultaneously estimate the position candidate frames and relative orientations with a fast inference speed and slight parameter quantity. Furthermore, our LCD method is integrated into a complete LiDAR-based SLAM system and its effectiveness is verified on the self-collected datasets. The main conclusions are listed as follows:

(1) The proposed Fast-DiSCO improves the Recall@1% and Recall@1 for place recognition compared to the DiSCO, ScanContext, and PointNetVLAD methods. The fast-DiSCO method shows a great improvement of Recall@1%

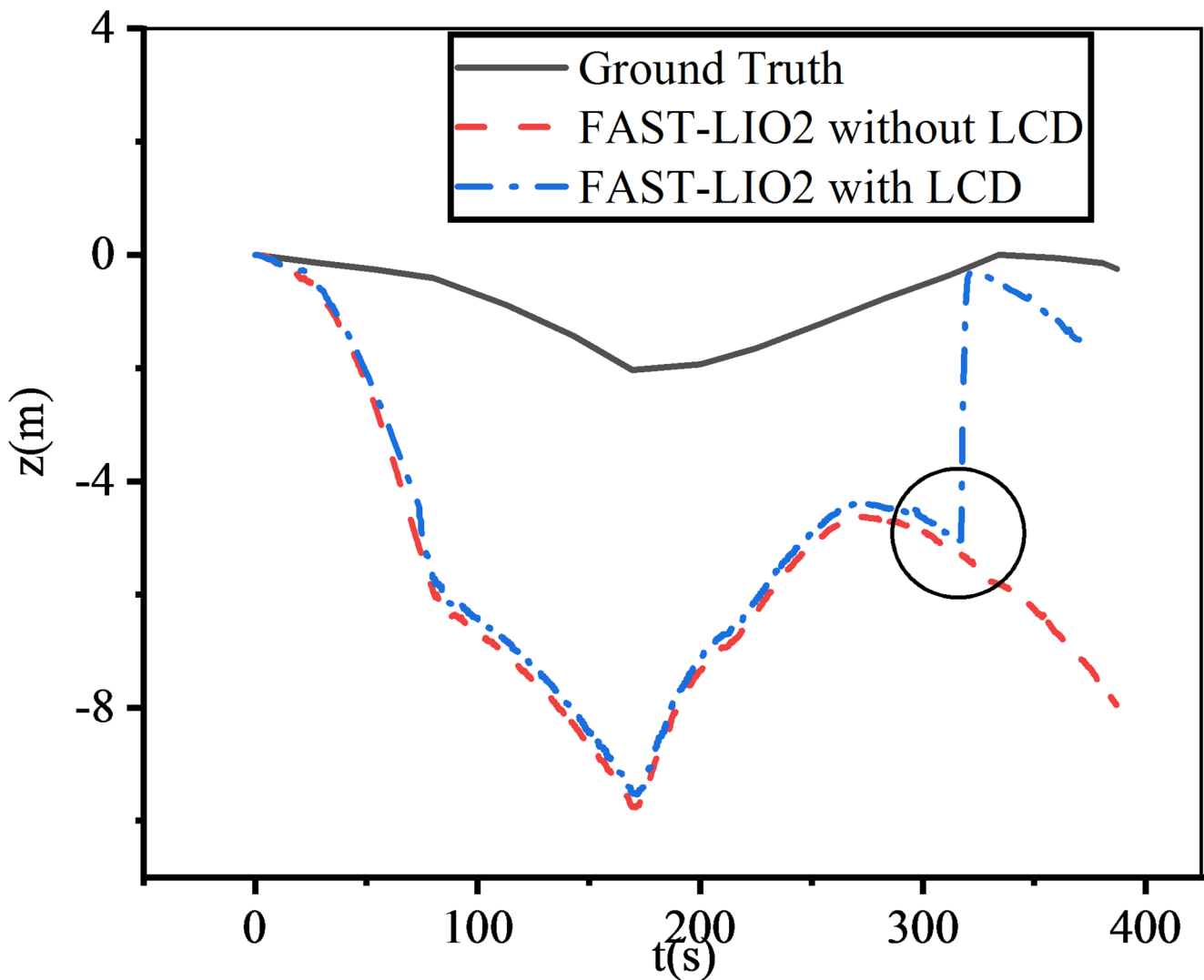


Fig. 14 Location result of z-axis in the outdoor environment

and Recall@1 not only on the NCLT dataset but also on the MulRan dataset, compared to the ScanContext and PointNetVLAD methods. Furthermore, compared to the DiSCO method, the Fast-DiSCO method also improves the Recall@1 by 1.76% and 0.54% on the NCLT dataset and the MulRan dataset, respectively.

(2) The proposed Fast-DiSCO reduces the estimated error for yaw angle compared to the DiSCO and Overlap methods. When the Fast-DiSCO method is applied, the standard errors of estimated yaw angle are further decreased by 1.08° and 8.26° on the NCLT dataset and MulRan dataset, respectively.

(3) The proposed overall LCD method can improve the performance of the LiDAR-based SLAM system, which can enhance the quality of the map and trajectory.

(4) The proposed method shows a good meeting on real-time place recognition. Compared to the DiSCO method,

the inference time and the number of model parameters are dropped by 30.60% and 77.61%, respectively. The overall time of our LCD method has a better real-time performance.

The main novelty of Fast-DiSCO lies in its lightweight U-shaped backbone, which integrates depth-wise separable convolution to reduce computational cost, and an ASPP module to enhance receptive fields and feature representation. In the future, we plan to explore the integration of 3D attention models and visual BEV information to further improve loop closure detection performance.

Acknowledgements The authors wish to thank the anonymous reviewers for their valuable comments and suggestions.

Author contributions Investigation, visualization, methodology, conceptualization, Chizhou Zhang; writing—original draft preparation, validation, data curation, Chizhou Zhang and Guanqiang Wang; project administration, supervision, methodology, Mingsong Chen; methodology, supervision, Yongcheng Lin; formal analysis, software,

Zehao Li; All authors have read and agreed to the published version of the manuscript.

Funding This work was supported by the General Program of Hunan Provincial Natural Science Foundation No. 2025JJ5027.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

- Agunbiade, O.Y.: Enhancing the MCL-SLAM algorithm to overcome the issue of illumination variation, non-static environment and kidnapping to present the NIK-SLAM. *International Journal of Intelligent Robotics and Applications* **9**(2), 531–552 (2025)
- Carlevaris-Bianco, N., Ushani, A.K., Eustice, R.M.: University of Michigan North campus long-term vision and lidar dataset. *Int. J. Rob. Res.* **35**(9), 1023–1035 (2016)
- Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2017)
- Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the 2018 European Conference on Computer Vision (ECCV)*, pp. 801–818 (2018)
- Chen, X., Labe, T., Milioto, A., Röhling, T., Vysotska, O., Haag, A., Behley, J., Stachniss, C.: OverlapNet: Loop closing for LiDAR-based SLAM. [arXiv: 2105.11344](https://arxiv.org/abs/2105.11344) (2021)
- Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: *Proceedings of the 2017 IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1251–1258 (2017)
- Cramariuc, A., Tschopp, F., Alatur, N., Benz, S., Falck, T., Brühlmeier, M., Hahn, B., Nieto, J., Siegwart, R.: SemSegMap–3D segment-based semantic localization. In: *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1183–1190 (2021)
- Du, J., Wang, R., Cremers, D.: DH3D: Deep hierarchical 3D descriptors for robust large-scale 6DoF relocalization. In: *Proceeding of the 2020 European Conference on Computer Vision (ECCV)*, pp. 744–762 (2020)
- Dube, R., Cramariuc, A., Dugas, D., Sommer, H., Dymczyk, M., Nieto, J., Siegwart, R., Cadena, C.: SegMap: Segment-based mapping and localization using data-driven descriptors. *Int. J. Rob. Res.* **39**(2–3), 339–355 (2020)
- Dubé, R., Dugas, D., Stumm, E., Nieto, J., Siegwart, R., Cadena, C.: Segmatch: Segment based place recognition in 3D point clouds. In: *Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5266–5272 (2017)
- He, L., Wang, X., Zhang, H.: M2DP: A novel 3D point cloud descriptor and its application in loop closure detection. In: *Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 231–237 (2016)
- Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y., Wu, J.: UNet 3+: A full-scale connected unet for medical image segmentation. In: *Proceedings of the 2020–2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1055–1059 (2020)
- Kim, G., Kim, A.: Scan context: Egocentric spatial descriptor for place recognition within 3D point cloud map. In: *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4802–4809 (2018)
- Kim, G., Park, Y.S., Cho, Y., Jeong, J., Kim, A., Mulran: Multimodal range dataset for urban place recognition. In: *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6246–6253 (2020)
- Koide, K., Yokozuka, M., Oishi, S., Banno, A.: Voxelized GICP for fast and accurate 3D point cloud registration. In: *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11054–11059 (2021)
- Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12697–12705 (2019)
- Li, Y.J., Gladkova, M., Xia, Y., Wang, R., Cremers, D.: Vxp: Voxel-cross-pixel large-scale image-lidar place recognition. *ArXiv Preprint arXiv: 2403.14594* (2024)
- Milad, R., Wang, L., Knights, J., Li, Z., Pounds, P., Moghadam, P.: Pose-graph attentional graph neural network for lidar place recognition. *IEEE Rob. Autom. Lett.* **9**(2), 1182–1189 (2023)
- Orhan, S., Baştanlar, Y.: Efficient search in a panoramic image database for long-term visual localization. In: *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1727–1734 (2021)
- Ren, M., Zhang, X., Liu, B., Zhu, Y.: Hierarchical loop closure detection with weighted local patch features and global descriptors. *Appl. Intell.* **55**(4), 266 (2025)
- Rusu, R.B., Marton, Z.C., Blodow, N., Beetz, M.: Persistent point feature histograms for 3D point clouds. In: *Proceedings of the 2008 10th Int Conf Intel Autonomous Syst (IAS-10)*, pp. 119–128 (2008)
- Rusu, R.B., Blodow, N., Beetz, M.: Fast point feature histograms (FPFH) for 3D registration. In: *Proceedings of the 2009 IEEE international conference on robotics and automation (ICRA)*, pp. 3212–3217 (2009)
- Tian, X., Yi, P., Zhang, F., Lei, J., Hong, Y.: STV-SC: Segmentation and Temporal Verification Enhanced Scan Context for Place Recognition in Unstructured Environment. *Sensors* **22**(22), 8604 (2022)
- Uy, M.A., Lee, G.H.: Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In: *Proceedings of the 2018 IEEE Conference on Computer Vision and Recognition (CVPR)*, pp. 4470–4479 (2018)
- Wang, T., Huang, H., Lin, J., Hu, C., Zeng, K., Sun, M.: Omnidirectional CNN for visual place recognition and navigation. In: *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2341–2348 (2018)
- Wang, Y., Sun, Z., Xu, C., Sarma, S.E., Yang, J., Kong, H.: Lidar iris for loop-closure detection. In: *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5769–5775 (2020)
- Wang, Z., Zou, N., Shen, D., Ji, S.: Non-local u-nets for biomedical image segmentation. In: *Proceedings of the 2020 Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 6315–6322 (2020)
- Xia, Y., Shi, L., Ding, Z., Henriques, J.F., Cremers, D.: Text2Loc: 3D point cloud localization from natural language. In: *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14958–14967 (2024)
- Xu, X., Yin, H., Chen, Z., Li, Y., Wang, Y., Xiong, R.: Disco: Differentiable scan context with orientation. *IEEE Robotics Autom. Lett.* **6**(2), 2791–2798 (2021)
- Xu, W., Cai, Y., He, D., Lin, J., Zhang, F.: FAST-LIO2: Fast direct lidar-inertial odometry. *IEEE T. Robot.* **38**(4), 2053–2073 (2022)

- Yan, X., Xu, Y., Li, S., Wang, R., Du, J., Cremers, D., Stilla, U.: SOE-Net: A self-attention and orientation encoding network for point cloud based place recognition. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11348–11357 (2021)
- Yan, X., Gladkova, M., Wang, R., Li, Q., Stilla, U., Henriques, J.F., Cremers, D.: CASSPR: Cross attention single scan place recognition. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8461–8472 (2023)
- Zhang, C., Chen, M., Wang, G., Lin, Y., Li, K., Wu, M., Li, Z., Wang, Q.: LIWOM-GD: enhanced LiDAR-inertial-wheel odometry and mapping by fusion with ground constraint and dynamic points elimination[J]. *IEEE Sens. J.* **24**(19), 30287–30303 (2024)
- Zhao, F., Huang, H., Hu, W.: An optimized hierarchical point cloud registration algorithm. *Multimedia Syst.* **31**(1), 14 (2025)
- Zhou, Z., Zhao, C., Adolfsson, D., Su, S., Gao, Y., Duckett, T., Sun, L.: NDT-transformer: Large-scale 3D point cloud localisation using the normal distribution transform representation. In: Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 5654–5660 (2021)
- Żywanowski, K., Banaszczyk, A., Nowicki, M.R.: Comparison of camera-based and 3D LiDAR-based place recognition across weather conditions. In: Proceedings of the 2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV), pp. 886–891 (2020)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Chizhou Zhang received a bachelor's degree in mechanical design, manufacture and automation from Xiangtan University in 2017, a master's degree in mechanical engineering from Nanchang University in 2020 and a Ph.D. degree in mechanical engineering from Central South University in 2025. Since now, he has been a lecturer in the School of Artificial Intelligence, Hubei University of Automotive Technology. His research interests include VLN, SLAM, reinforcement learning, and multi-sensor fusion technology.

Mingsong Chen was born in Xinshao, Hunan province, China. He received a B.S. degree in mechanical engineering from Xiangtan University, Hunan province, China, in 2006, and a Ph.D. degree in mechanical engineering from Central South University, Hunan province, China, in 2011. Since 2011, he has been a lecturer (2011–2014), associate professor (2014–2019), and professor in the School of Mechanical and Electrical Engineering, and Light Alloy Research Institute, Central South University, China. He is the author of two books, more than 80 articles, and more than 30 inventions. His research interests include intelligent forging, machine vision, intelligent wall-climbing robot, multi-rotor UAV and nonlinear control system.

Guanqiang Wang received a bachelor's degree in mechanical design, manufacture and automation from Xiangtan University in 2017, and a Ph.D. degree in mechanical engineering from Central South University in 2025. Since now, he has been a lecturer in the School of Mechanical and Electrical Engineering, Guilin University of Electronic Technology. His research interests include machine vision, deep learning, and control system design.

Yongcheng Lin was born in Liuyang, Hunan province, China. He received B.S degree (2000) and M.S degree (2003) in mechanical engineering from Liaoning Shihua University, Liaoning province, China, and a Ph.D. degree in mechanical engineering from Tianjing University, Tianjing, China, in 2006. Since 2006, he has been associate professor (2006–2011), and professor in the School of Mechanical and Electrical Engineering, Central South University, China. He is the author of two books, more than 200 articles, and more than 40 inventions. His research interests include intelligent forging, machine vision, intelligent wall-climbing robot, multi-rotor UAV and nonlinear control system.

Zehao Li received his Bachelor's degree in Mechanical Design, Manufacturing and Automation from Changsha University of Science and Technology, China in 2022 and a master's degree in mechanical engineering from Central South University in 2025. His current research interests include wall-climbing robot design, robot control technology, SLAM, and multi-sensor fusion technology.