electronics

MDPI

Article

SG-LPR: Semantic-Guided LiDAR-Based Place Recognition

Weizhong Jiang [1], Hanzhang Xue [1,2], Shubin Si [1,3], Chen Min [4], Liang Xiao [1,*], Yiming Nie [1,*] and Bin Dai [1]

[1] Unmanned Systems Technology Research Center, Defense Innovation Institute, Beijing 100071, China; jiangweizhong16@alumni.nudt.edu.cn (W.J.); xuehanzhang13@nudt.edu.cn (H.X.); sishubin@hrbeu.edu.cn (S.S.); daibin@alumni.nudt.edu.cn (B.D.)
[2] Test Center, National University of Defense Technology, Xi'an 710106, China
[3] College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China
[4] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China; mincheng@ict.ac.cn
* Correspondence: xiaoliang@nudt.edu.cn (L.X.); nieym@alumni.nudt.edu.cn (Y.N.)

Abstract: Place recognition plays a crucial role in tasks such as loop closure detection and re-localization in robotic navigation. As a high-level representation within scenes, semantics enables models to effectively distinguish geometrically similar places, therefore enhancing their robustness to environmental changes. Unlike most existing semantic-based LiDAR place recognition (LPR) methods that adopt a multi-stage and relatively segregated data-processing and storage pipeline, we propose a novel end-to-end LPR model guided by semantic information—SG-LPR. This model introduces a semantic segmentation auxiliary task to guide the model in autonomously capturing high-level semantic information from the scene, implicitly integrating these features into the main LPR task, thus providing a unified framework of "segmentation-while-describing" and avoiding additional intermediate data-processing and storage steps. Moreover, the semantic segmentation auxiliary task operates only during model training, therefore not adding any time overhead during the testing phase. The model also combines the advantages of Swin Transformer and U-Net to address the shortcomings of current semantic-based LPR methods in capturing global contextual information and extracting fine-grained features. Extensive experiments conducted on multiple sequences from the KITTI and NCLT datasets validate the effectiveness, robustness, and generalization ability of our proposed method. Our approach achieves notable performance improvements over state-of-the-art methods.

Keywords: LiDAR-based place recognition; semantic-guided; auxiliary task; swin transformer; U-Net

Citation: Jiang, W.; Xue, H.; Si, S.; Min, C.; Xiao, L.; Nie, Y.; Dai, B. SG-LPR: Semantic-Guided LiDAR-Based Place Recognition. *Electronics* **2024**, *13*, 4532. https://doi.org/10.3390/electronics13224532

Academic Editor: Dah-Jye Lee

Received: 11 October 2024
Revised: 11 November 2024
Accepted: 15 November 2024
Published: 18 November 2024

1. Introduction

Place Recognition (PR) technology fundamentally involves the feature-encoding of environmental information observed by sensors, followed by the retrieval or matching of identical locations within a global place feature database or map. According to this definition, PR technology is categorized under global localization for mobile robots and serves as an auxiliary tool for existing navigation methods. It is primarily applied in the closure loop detection module of Simultaneous Localization and Mapping (SLAM) systems or the re-localization module for long-term navigation tasks. This is crucial for reducing cumulative localization errors and achieving reliable localization [1,2].

Vision-based Place Recognition (VPR) has undergone early development and reached a relatively mature stage. However, for outdoor mobile robot applications, the widespread utilization of VPR is constrained by its sensitivity to variations in illumination, weather, and seasons. In contrast, LiDAR offers advantages such as long-range sensing, high accuracy, and resilience to illumination changes, making LiDAR-based Place Recognition (LPR) a growing research focus.

Unlike images that contain rich texture features, raw 3D point clouds primarily consist of geometric information, making LPR challenging [3]. Consequently, researchers have developed various deep learning models to extract features from raw 3D point clouds

and their sparse voxels or 2D projections, aiming to effectively represent places [4–8]. Recent studies have demonstrated that incorporating semantic information from scenes can significantly enhance the robustness of LPR models against challenges such as occlusion and viewpoint changes, while also enhancing their generalization ability [3,9–11]. Hence, semantic-based LPR methods are gaining increasing attention.

Most existing semantic-based LPR methods rely on semantic labels during both the training and testing phases. These labels must be extracted from the raw point clouds using an additional semantic segmentation model [10,12–14]. The extracted labels typically undergo preprocessing, constructing semantic graphs or other types of intermediate features, which are then stored locally. Subsequently, graph-based methods [10,14,15] or graph-free methods [9,11,16] encode these intermediate features to generate feature descriptors that effectively represent places. Therefore, the process from raw 3D point cloud input to the final place feature descriptor output is not end-to-end but is completed in stages involving explicit segmentation and processing of semantic labels. We refer to this process as "segmentation-then-describing", as illustrated in Figure 1a. This framework has several limitations: (1) the quality of the extracted semantic information heavily depends on the performance of the chosen semantic segmentation algorithm; (2) unnecessary errors or disturbances may be introduced during intermediate data-processing or storage stages; (3) the relatively segmented process design impedes the end-to-end training of the entire framework.

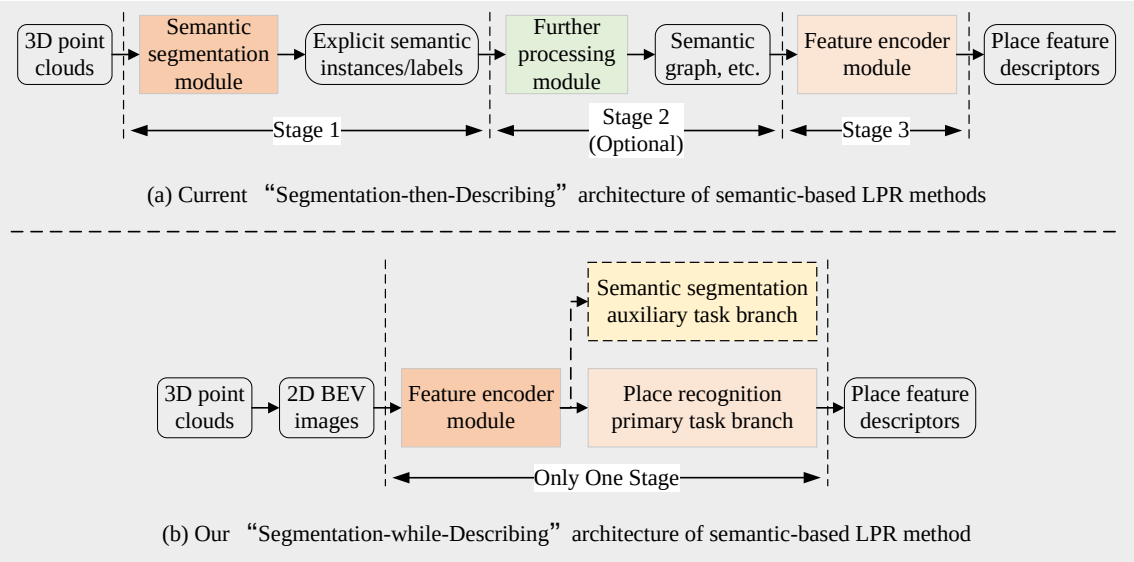


Figure 1. Comparison of system frameworks for semantic-based LPR methods. (**a**) represents the prevalent "segmentation-then-describing" framework employed by most existing semantic-based LPR methods. This framework comprises multiple distinct stages. (**b**) depicts our proposed "segmentation-while-describing" framework, which implicitly provides high-level semantic features into the primary LPR task through an auxiliary semantic segmentation task (indicated by the yellow dashed box, which is active only during model training).

Furthermore, existing semantic-based LPR methods primarily rely on label-level semantic information, semantic graphs, or other handcrafted intermediate features. This manual classification and processing inevitably lead to the loss of local information beyond the semantic categories of interest. At the same time, these methods also exhibit limitations in capturing global context and extracting fine-grained features.

To address the aforementioned issues, we propose a novel semantic-guided LPR model termed SG-LPR. Specifically, by introducing a semantic segmentation auxiliary task, we enhance the model's capacity to capture semantic information from scenes and implicitly integrate high-level semantic features into the primary place recognition task. The semantic

segmentation auxiliary task branch is trained jointly with the place recognition main task branch, with the segmentation branch functioning solely during the training phase.

It is important to note that, in the field of robotic navigation and localization, many studies have leveraged auxiliary tasks to enhance the performance of the primary task. For instance, MapLocNet [17] uses perception tasks as an auxiliary objective for pose prediction, enabling a reliable and human-like re-localization method without requiring high-precision maps. The methods most similar to ours are CGiS-Net [18], and AEGIS-Net [19], which introduce a semantic segmentation auxiliary task to provide implicit semantic information for the place recognition task. They use self-attention mechanisms to integrate this information with color and geometric features, significantly improving indoor place recognition performance. However, these methods mainly focus on indoor place recognition tasks using RGB-D camera data, and their model training process is divided into two stages: first, training the semantic encoder-decoder, followed by training the feature embedding for place recognition. In contrast, our study targets LiDAR-based place recognition for large-scale outdoor scenes, and we jointly train the semantic segmentation auxiliary task and the primary place recognition task, avoiding the need for separate training processes.

Therefore, we establish a unified framework that directly processes raw input data through a deep learning model to generate place feature descriptors, therefore circumventing the need for explicit extraction, processing of semantic information, and storage of intermediate data during testing or inference phases. This framework implements a "segmentation-while-describing" system architecture, as illustrated in Figure 1b. Furthermore, during the model design process, we integrated the strengths of the classic Swin Transformer [20] and U-Net to enhance the model's performance in capturing global contextual information and fine-grained feature extraction. To accommodate the model's structural characteristics while balancing scale and computational efficiency, we employ the 2D bird's-eye view (BEV) projection of 3D point clouds as the model input.

It is well known that the 2D BEV projection of 3D point clouds inevitably results in the loss of vertical scene information. However, the incorporation of semantic information can partially alleviate the negative impact of this information loss on model performance.

Our main contributions can be summarized as follows:

- We propose a unified semantic-guided LPR framework, characterized by a "segmentation-while-describing" structure, which eliminates the need for additional intermediate data-processing and storage steps.
- Based on this framework, we design the SG-LPR model, integrating the advantages of Swin Transformer and U-Net in capturing global contextual information and fine-grained feature extraction.
- Experimental results on the KITTI and NCLT datasets demonstrate the effectiveness of the proposed framework, with the model outperforming comparative baseline algorithms in terms of place recognition performance and generalization ability.

2. Related Work

Current LPR methods can be categorized into handcrafted feature-based methods and deep learning-based methods. In this section, we only introduce a few representative works. For a more comprehensive overview, the readers may refer to [1,2,21].

2.1. Handcrafted Feature-Based Methods

Early researchers typically relied on manually designed rules to convert 3D point clouds into 2D feature maps, histograms, or other structures, effectively extracting geometric information to construct local or global place feature descriptors. Methods such as the Scan Context series [3,22] and LiDAR Iris [23] utilize polar coordinate transformations to obtain a BEV representation of the 3D point cloud, encoding height and other relevant information to describe the scene. M2DP [24] projects the raw 3D point cloud onto six different planes, counting point density within sector grids on each plane to construct a

global feature descriptor. Methods like NDT-histogram [25] divide the 3D point cloud space into spherical grids and build statistical histograms based on various attributes of each grid. Bosse et al. [26] leverage keypoint features and geometric consistency from the 3D point cloud to construct place feature descriptors. SegMatch [27] utilizes segment features extracted from the 3D point cloud for place recognition.

Handcrafted feature-based methods are characterized by human-defined feature construction rules, which generally provide better interpretability. However, these methods are often limited to specific platforms or types of LiDAR sensors, exhibiting restricted generalization capabilities and sensitivity to viewpoint changes and occlusions [28]. With the rapid advancement of deep learning technology, deep learning-based LPR methods have demonstrated significant advantages in accuracy and efficiency, gradually becoming the mainstream approach.

2.2. Deep Learning-Based Methods

Such methods typically utilize deep neural networks to encode input data into high-dimensional feature descriptors. Depending on the data structure input to the neural network, these methods can be further categorized into those based on raw 3D point clouds or its voxels, those based on 2D projections, and those based on semantics.

Methods based on 3D points or 3D sparse voxels. Pioneering work in this category includes PointNetVLAD [7], which directly utilizes PointNet [29] to extract local features from 3D point clouds and aggregates these features into a global descriptor using NetVLAD [30]. Methods such as PCAN [31] and SOE-Net [32] introduce attention mechanisms to enhance the model's capacity to encode local features. DH3D [4] incorporates multi-level spatial context information and channel feature correlations into local features based on point convolution and attention mechanisms. DAGC [33] aggregates multi-level neighborhood features of each point using graph convolution, effectively mining the geometric structure information of the local neighborhood. PPT-Net [34] employs a pyramid point-Transformer to capture spatial relationships between local features of point clouds at different resolutions. The original 3D point cloud often contains local details important for fine-grained tasks such as segmentation and detection; however, these details may be unhelpful for LPR tasks and could even be perceived as outliers or noise, therefore burdening LPR models in understanding the scene. By voxelizing the original point cloud, it is possible to reduce irrelevant local details while retaining the overall structural information of the scene and decreasing data volume [35]. MinkLoc3D [5] designs a feature pyramid network to encode features from 3D sparse voxels. SVT-Net [35] utilizes Transformers to learn both short-range local features and long-range contextual features within 3D voxels. Methods such as LCDNet [36], LoGG3D-Net [8], and CASSPR [37] fuse features derived from both points and voxels.

Although methods based on 3D points or voxels maximize the retention of original point cloud information, they exhibit limitations in computational efficiency due to the sparsity, disorder, and large scale of point cloud structures [38].

Methods based on 2D projection. This category of methods projects the original 3D point clouds into relatively compact 2D image structures. Common projection techniques include spherical projection, BEV projection, cylindrical projection, and sinogram projection. Methods such as OverlapNet [16] and OT [39] are designed to create corresponding rotation-invariant neural network models that account for the structural characteristics of range images. DiSCO [40] and BEVPlace [6,41] focus on feature-encoding for BEV images to obtain global feature descriptors with rotation invariance. To address the issue of sparse features in single-frame point clouds and enhance the model's robustness to occlusions and viewpoint changes, Cao et al. [42] utilized cylindrical projection to transform the 3D point clouds into 2D images that capture prominent geometric structures of the scene. Furthermore, to solve the global localization problem based on sparse places, Lu et al. [43] designed a RING descriptor based on radon sinogram projection that provides a compact and unified representation of places while maintaining direction and translation invariance.

CVTNet [44] and MVSE-Net [45] integrate features from both spherical projection and BEV projection perspectives.

2D projection data typically exhibit lightweight and structurally regular characteristics, allowing the retention of certain features from the original 3D point clouds, such as local geometric structures and yaw rotation invariance, while ensuring relatively high computational efficiency. Therefore, this paper employs the 2D BEV projection of 3D point clouds as the input for the model.

Methods based on semantic. OverlapNet [16] demonstrated that incorporating semantic category information into the model's input can enhance the accuracy of LPR. SGPR [10] is a graph convolutional network that relies on semantic graph representation and graph matching. SSC [3] is a global descriptor for LPR based on semantic information. RINet [13] is a structurally rotation-invariant siamese network that uses semantic information and improves the robustness of global descriptors against viewpoint changes. Locus [14] formulated a global descriptor by aggregating multi-level features related to semantic components in a scene. SL_LPR [12] is a chained cascade network with the consistency of semantic information to eliminate the influence of dynamic objects on the LPR task.

However, most existing semantic-based LPR methods adopt a "segmentation-then-describing" framework, which consists of multiple distinct stages, as depicted in Figure 1a. To address these issues, we propose a "segmentation-while-describing" system framework, as illustrated in Figure 1b. By introducing an auxiliary semantic segmentation task, we implicitly integrate high-level semantic information into the primary LPR task, effectively eliminating the need for additional intermediate data-processing steps.

3. Preliminaries

3.1. Data Representation

In this work, we use BEV representations of 3D point clouds as inputs for our model, which includes BEV images and corresponding ground-truth semantic maps sharing the same pixel resolution. It should be noted that the ground-truth semantic maps are exclusively used to train the auxiliary semantic segmentation task and do not participate in encoding place features or the final testing phase.

Let P be a 3D point cloud, with each point denoted as $p_i(x, y, z)$. The corresponding BEV image and ground-truth semantic map of P are designated as B^I and B^S, respectively, both with size of (H, W). The pixel coordinates of p_i in B^I and B^S are calculated as follows:

$$\begin{cases} u = \dfrac{W}{2} + \lfloor \dfrac{x}{r} \rfloor, \\ v = \dfrac{H}{2} - \lfloor \dfrac{y}{r} \rfloor - 1, \end{cases} \tag{1}$$

where (u, v) denotes the pixel coordinates corresponding to point p_i in B^I and B^S, while r signifies the projection resolution, and the operation $\lfloor \cdot \rfloor$ represents the floor function. The pixel value of B^I at position (u, v) indicates the number of points projected at that location, whereas the pixel value of B^S at (u, v) corresponds to the semantic label value of the last point projected at that position.

It should be clarified that the semantic categories adopted in this study are derived from the fusion of semantic categories defined in the Semantic-KITTI dataset [46]. Specifically, we condense and merge the original 34 semantic categories into 20 categories as RINet [13].

3.2. Problem Definition

Let $D_P = \{P_i \mid i = 1, \ldots, M\}$ denote a pre-collected database of 3D point clouds defined with respect to a fixed reference frame, and its corresponding BEV database can be represented as $D_B = \{(B_i^I, B_i^S) \mid i = 1, \ldots, M\}$. Each dataset frame is geotagged with a Universal Transverse Mercator coordinate at its centroid using GPS/INS. Given a query point cloud Q_P with its corresponding BEV representation $Q_B(B_q^I, B_q^S)$, the objective of

the LPR task is to retrieve a point cloud D_P^* from D_P that is structurally similar to Q_P. To tackle this problem, we design a neural network that learns a function $f(\cdot)$ to map B_i^I to a fixed-size global feature descriptor $\mathcal{F}_{B_i^I}$. The goal is to identify a BEV image $B^{I*} \in D_B$ such that the Euclidean distance between the global descriptors $f(B^{I*})$ and $f(B_q^I)$ is minimized:

$$B^{I*} = \underset{B_i^I \in D_B}{\arg\min} \|f(B_q^I) - f(B_i^I)\|_2 = \underset{B_i^I \in D_B}{\arg\min} \|\mathcal{F}_{B_q^I} - \mathcal{F}_{B_i^I}\|_2, \tag{2}$$

where $\|\cdot\|_2$ represents the ℓ_2-norm. Ultimately, the corresponding point cloud D_P^* can be obtained according to B^{I*}.

4. Methodology

Adhering to the proposed "segmentation-while-describing" framework, we design the SG-LPR model, which integrates the strengths of both the Swin Transformer [20] and U-Net in capturing global context information and fine-grained features.

4.1. Overall Architecture

The architecture of SG-LPR is illustrated in Figure 2 and consists of three modules: the feature extraction module, the LPR task branch, and the semantic segmentation task branch. During the model training process, we employ a triplet-based training scheme by inputting an anchor point cloud along with its corresponding positive and negative samples. It is important to note that the ground-truth semantic map S_g is utilized solely for training the semantic segmentation task and does not participate in the processing of the feature extraction module. Furthermore, the semantic segmentation task branch operates only during the training phase, aiming to guide the feature extraction module in acquiring the ability to extract high-level semantic features from BEV images, which is not required during inference. Therefore, unlike other semantic-based LPR methods, SG-LPR does not require S_g or its post-processed data during the inference or testing phases.

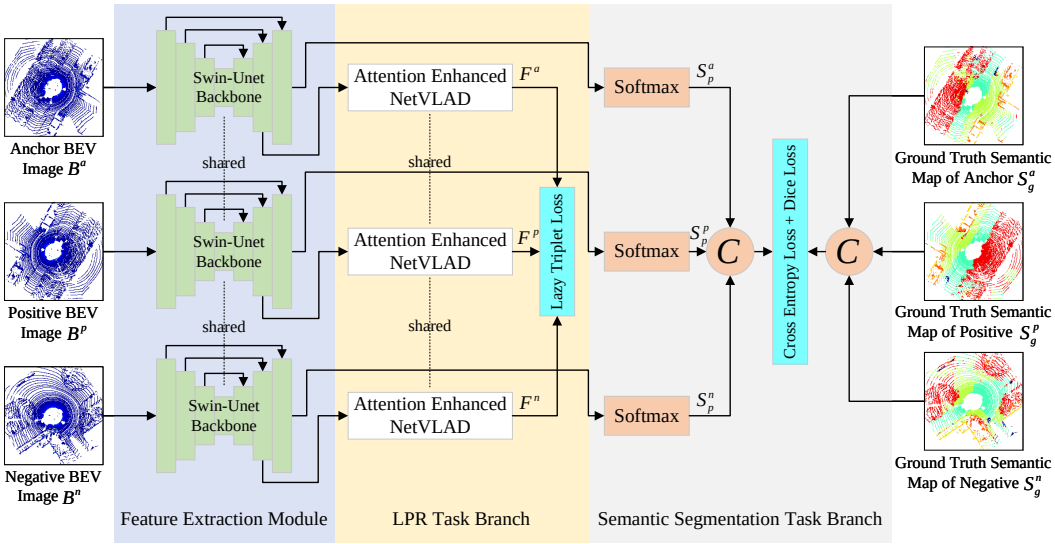


Figure 2. Overview of the proposed SG-LPR architecture. It consists of a shared feature extractor (blue area), followed by two parallel branches: one for the LPR task (yellow area) and another for the semantic segmentation task (gray area). These branches are jointly trained to implement the "Segmentation-while-describing" framework. Notably, the semantic segmentation branch is active only during training and incurs no additional computational cost during testing.

4.1.1. Feature Extraction Module

This module converts 2D BEV images into high-dimensional embeddings enriched with deep semantic features for subsequent primary LPR tasks. To improve its capacity to

capture global context and fine-grained features, we leverage the complementary strengths of Swin Transformer [20] and U-Net architectures, constructing this module's framework based on Swin-Unet [47], as shown in Figure 3a. Swin-Unet, inherently a transformer-based U-shaped encoder-decoder architecture, was originally conceived for medical image segmentation. Within Swin-Unet, the Swin Transformer block utilizes a local window self-attention mechanism and facilitates inter-window information exchange through sliding window techniques. The structure of its fundamental building block is illustrated in Figure 3b, with the detailed computation process described as follows:

$$\begin{cases} \hat{z}^l = W\text{-}MSA(LN(Z^{l-1})) + z^{l-1}, \\ z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l, \\ \hat{z}^{l+1} = SW\text{-}MSA(LN(z^l)) + z^l, \\ z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1}, \end{cases} \tag{3}$$

where $\hat{z}^l$ and z^l represent the outputs of the $(S)W\text{-}MSA$ module and the MLP module for the l^{th} block, respectively. The terms $W\text{-}MSA$ and $SW\text{-}MSA$ refer to window-based multi-head self-attention utilizing regular and shifted window partitioning configurations, respectively. The computation of self-attention is defined as [47]:

$$A(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}} + B)V, \tag{4}$$

where $Q, K, V \in R^{M^2 \times d}$ denote the query, key and value matrices. Here, M^2 and d represent the number of patches within a window and the dimensions of the query or key, respectively. The values in B are derived from the bias matrix $\hat{B} \in R^{(2M-1)\times(2M+1)}$.

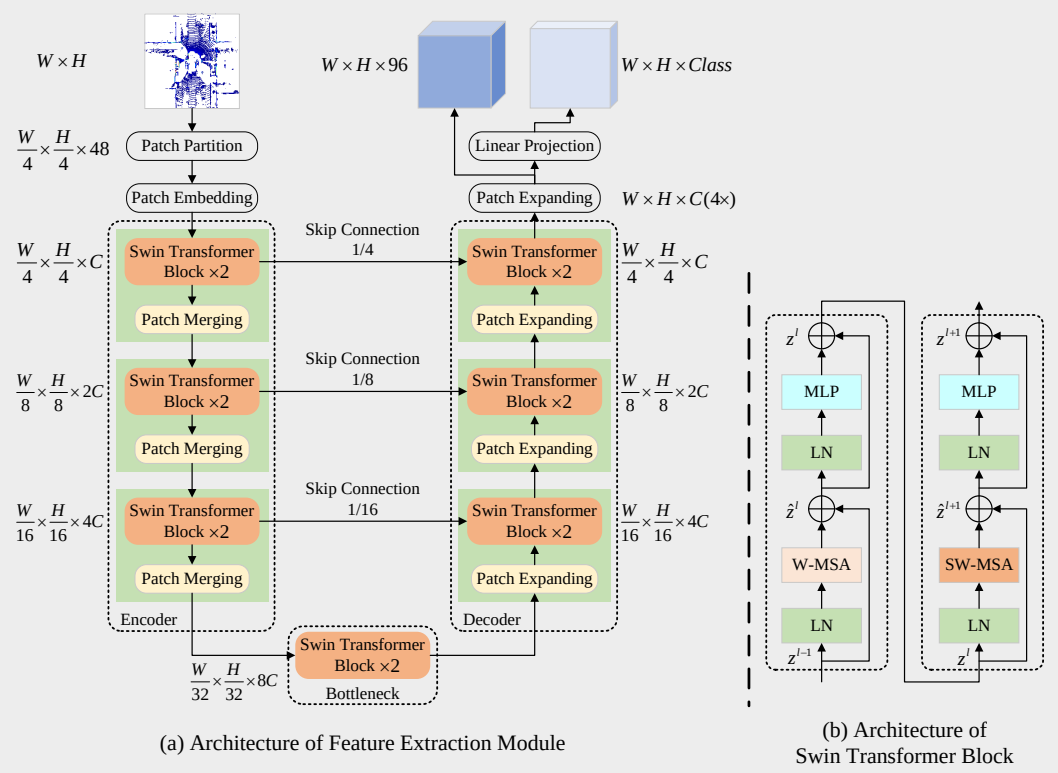


Figure 3. Architecture of the Feature Extraction Module. We construct this module based on Swin-Unet [47], with the semantic segmentation task branch guiding it to extract feature tensors that are rich in high-level semantic information from raw BEV images.

Similar to U-Net, Swin-Unet integrates skip connections to merge multi-scale features from the encoder with up-sampled features form the decoder, effectively mitigating the loss of spatial information that typically occurs during down-sampling.

Our main modification to Swin-Unet involves the addition of an output head positioned before the final linear projection layer. This newly introduced head is intended to deliver high-dimensional feature embeddings enriched with high-level semantic information for the primary LPR task module. Meanwhile, the original output head of Swin-Unet continues to function as the input for the auxiliary semantic segmentation task.

4.1.2. LPR Task Module

This module is essentially a channel-space attention-enhanced NetVLAD layer designed to process the high-dimensional feature embeddings output by the feature extraction module, generating a global feature descriptor that effectively represents the place, as illustrated in Figure 4. Inspired by CBAM [48], we first apply spatial and channel attention to weight the high-dimensional feature embeddings, preserving the key information within the input tensor while suppressing noise and irrelevant details at both the spatial and channel levels without significantly increasing the complexity of the network. Subsequently, three convolutional layers are employed to reduce the spatial resolution of the weighted feature tensor while increasing its channel dimensionality. Finally, the classic NetVLAD layer [30] aggregates the features to produce a compact low-dimensional global feature descriptor, facilitating subsequent tasks such as storage and place retrieval.

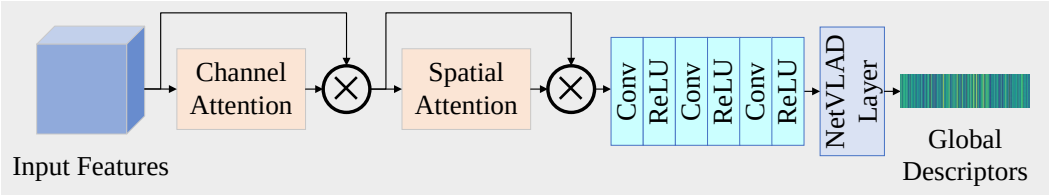


Figure 4. Architecture of the LPR task branch. We construct this module based on Swin-Unet [47], with the semantic segmentation task branch guiding it to extract feature tensors that are rich in high-level semantic information from raw BEV images.

4.1.3. Semantic Segmentation Task Module

The semantic segmentation auxiliary task branch aims to establish a mapping relationship between the 2D BEV images and the semantic information within the scene. The architecture of our feature extraction module is built based on Swin-Unet [47], a renowned model in medical image segmentation. While capable of capturing semantic information, its direct relevance to the primary LPR task is limited. To enhance the module's capability in capturing semantic information relevant to the LPR task, we introduce a semantic segmentation auxiliary task module, as depicted in the gray area in Figure 2. This task directs the module's attention to salient semantic information within the scene, implicitly providing precise semantic cues for the subsequent primary LPR task module. This design maintains consistency with our proposed "segmentation-while-describing" framework for the entire model architecture. Subsequent experiments validate the effectiveness and necessity of this auxiliary task module. Notably, ground-truth semantic maps are used exclusively to train the semantic segmentation auxiliary task during the model training phase.

4.2. Loss Functions

Our comprehensive loss function comprises a lazy triplet loss, denoted as $\mathcal{L}_{lt}$, which primarily facilitates the LPR task. Additionally, it incorporates a cross-entropy loss, represented as $\mathcal{L}_{ce}$, and a dice loss, symbolized as $\mathcal{L}_{dice}$, both supporting the semantic segmentation auxiliary task. In the following paragraphs, we provide a detailed explanation of these loss functions.

4.2.1. Lazy Triplet Loss

For the LPR task, we adopt the commonly used lazy triplet loss [7], defined as follows:

$$\mathcal{L}_{lt}(\mathcal{T}) = \max_{i}([m + \delta_p - \delta_{n_i}]_+), \tag{5}$$

where $\mathcal{T} = (\mathcal{F}_{B_q^l}, \{\mathcal{F}_{B_p^l}\}, \{\mathcal{F}_{B_n^l}\})$ denotes a training tuple, and for one query descriptor $\mathcal{F}_{B_q^l}$, we utilize k_p positive descriptors $\{\mathcal{F}_{B_p^l}\}$ and k_n negative descriptors $\{\mathcal{F}_{B_n^l}\}$, respectively. $[\cdot]_+$ signifies the hinge loss, m is a margin value, δ_p represents the distance between the global feature descriptors of the anchor sample $\mathcal{B}_a$ and its structurally similar ("positive") sample, while δ_{n_i} stands for the distance between the global feature descriptors of $\mathcal{B}_a$ and its structurally dissimilar ("negative") sample. We adhere to the training strategy outlined in [7] and consider two point clouds to be structurally similar if their geometric distance is less than ϵ meters.

4.2.2. Cross-Entropy Loss

For the semantic segmentation auxiliary task, we utilize the cross-entropy loss function to assess the disparity between predicted semantic labels and the ground-truth semantic labels, which is defined as:

$$\mathcal{L}_{ce}(Y, \hat{Y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^c \cdot log(\hat{y}_i^c), \tag{6}$$

where Y and $\hat{Y}$ represent the predicted semantic label map and its corresponding ground-truth semantic label map, respectively. N denotes the number of pixels in a sample, C signifies the number of classes, y_i^c indicates the probability of pixel i belonging to class c, and $\hat{y}_i^c$ corresponds to the ground-truth label of class for pixel i.

4.2.3. Dice Loss

BEV images generated from 3D point clouds often contain extensive non-informative background regions, while the distribution of semantic categories within meaningful foreground areas may exhibit imbalances. Categories such as dynamic or static objects tend to have relatively low proportions. To mitigate the adverse effects caused by the imbalance between foreground and background regions and the uneven distribution of semantic categories, we adopt the configurations detailed in [47] and incorporate the dice loss function into the semantic segmentation task. The formula for calculating the dice loss is presented below:

$$\mathcal{L}_{dice}(Y, \hat{Y}) = \frac{1}{C}\sum_{c=1}^{C} w_c \left(1 - \frac{2\sum_{i=1}^{N}\sum_{j=1}^{N}(y_{ij}^c \times \hat{y}_{ij}^c) + \epsilon}{\sum_{i=1}^{N}\sum_{j=1}^{N}(y_{ij}^c)^2 + \sum_{i=1}^{N}\sum_{j=1}^{N}(\hat{y}_{ij}^c)^2 + \epsilon}\right), \tag{7}$$

where C signifies the total number of classes, (M, N) is the resolution of the input BEV image, y_{ij}^c denotes the predicted class number at the specific position (i, j) within the predicted semantic label map, correspondingly, $\hat{y}_{ij}^c$ represents the ground-truth semantic label at the same position, w_c stands for the weight of the dice coefficient corresponding to the c-th class and is set to 1.0, ϵ is a smoothing term, introduced to prevent the denominator from being zero, which is set to 10^{-5} in our study.

4.2.4. Joint Loss

Our network is optimized jointly using a weighted sum of the LPR loss and the semantic segmentation loss, defined as:

$$\mathcal{L} = \mathcal{L}_{lt} + \lambda_{ce}\mathcal{L}_{ce} + \lambda_{dice}\mathcal{L}_{dice}, \tag{8}$$

where λ_{ce} and λ_{dice} are scalar hyperparameters representing the weights assigned to different loss terms, respectively.

5. Experiments and Results

5.1. Dataset and Experimental Settings

5.1.1. Dataset

KITTI [49]. This dataset includes 22 sequences of LiDAR scans from a Velodyne HDL-64E. For LPR, the first 11 sequences with precise ground-truth poses are typically used. We adopt the leave-one-out cross-validation strategy used in SGPR [10] and RINet [13] for model training on the KITTI dataset. Specifically, for each sequence in sequences 00-10, we designate one sequence as the test set while using the remaining sequences for training. In the actual experiments, we select sequences 00, 02, 05, 06, 07, and 08, which contain loop closures, as the test set. Notably, sequence 08 includes reverse loops, while the others are in the same direction.

NCLT [50]. This dataset was collected on the North Campus of the University of Michigan and encompasses both indoor and outdoor scenes. Point cloud data were acquired using a Velodyne HDL-32 LiDAR over a 15-month period. Consequently, it captures a diverse range of variations in seasons, weather conditions, lighting, viewpoints, and scene appearances. Additionally, the dataset contains a substantial number of dynamic objects, which pose greater challenges to the performance of LPR models. For evaluation purposes, We select sequences as outlined in [41], including "2012-01-15", "2012-02-04", "2012-03-17", "2012-06-15", "2012-09-28", "2012-11-16", and "2013-02-23".

5.1.2. Implementation Details

The proposed network is implemented using the PyTorch framework (v. 1.14.0) and trained from scratch on a single Nvidia A6000 GPU with 48 GB of memory. Following [10], we determine whether two point clouds represent the same place based on ground-truth poses. Specifically, point clouds are considered positive pairs if their Euclidean distance is less than 3 m; otherwise, they are negative pairs if the distance exceeds 20 m. When projecting a 3D point cloud onto the BEV plane, we use a grid size r of 0.2 m and a projection radius of 11.2 m, resulting in a BEV image size of 224×224 pixels. It should be noted that the ground-truth semantic map corresponding to the BEV image is only used to guide the training of the semantic segmentation auxiliary task and is not required during the testing. For each training tuple, we set $k_p = 1$ and $k_n = 10$. The parameters of the backbone of the feature extraction module adhere to the specifications outlined in Swin-Unet [47]. For the NetVLAD layer, we set the number of clusters to 64 and the output feature dimension to 256. During model training, we employ the Adam optimizer with an initial learning rate of 0.00001, along with an exponential scheduler for learning rate decay. The coefficients in the joint loss function $\mathcal{L}$ are configured as $\lambda_{ce} = 0.4$ and $\lambda_{dice} = 0.6$. Additionally, we apply random rotations around the z-axis on the point clouds for data augmentation during the training process.

5.1.3. Evaluation Metrics

We evaluate the performance of LPR models using the maximum F_1 score, recall rate at top-1 (Recall@1) retrieval, the precision–recall (PR) curve, and top 1 retrieval along the trajectory. The maximum F_1 score and Recall@1 are quantitative metrics, reflecting the overall performance of the model and its ability to correctly identify the target location in the first retrieved result, respectively. The PR curve and Top 1 retrieval along the trajectory are qualitative metrics, providing visual comparisons of the model's performance against other methods, as well as the discrepancy between the model's Top 1 retrieval result and the ground-truth of the loop closure. The F_1 score is defined as the harmonic mean of precision (P) and recall (R):

$$\begin{cases} P = \dfrac{TP}{TP + FP}, \\ R = \dfrac{TP}{TP + FN}, \\ F_1 = 2 \times \dfrac{P \times R}{P + R}, \end{cases} \tag{9}$$

where TP, FN, TN, and FP represent true positive, false negative, true negative, and false positive, respectively. As previously mentioned, the thresholds for classifying samples as positive or negative are based on distances of 3 m and 20 m.

5.2. Comparison with State-of-the-Art

This section presents a quantitative and qualitative comparison of the proposed method with other state-of-the-art methods on multiple raw sequences of the KITTI dataset. The compared methods include non-semantic approaches (M2DP [24], Scan Context (SC) [22], LiDAR Iris (LI) [23], PointNetVLAD (PNV) [7], OverlapTransformer (OT) [39], DiSCO [40], LoGG3D-Net [8], and BEVPlace [6]), as well as semantic-based methods (Semantic Scan Context (SSC) [3], SGPR [10], Locus [14], RINet [13], and SC_LPR [12]). The results for M2DP, SC, LI, PNV, SSC, SGPR, and RINet are taken from the RINet publication, while the results for DiSCO, LoGG3D-Net, and Locus are obtained from their respective original papers. Additionally, we replicate OT and BEVPlace under identical training parameter settings, where the input for OT is the range image data provided by the official source, with a resolution of 64×900.

5.2.1. Quantitative Results

Comparison with Non-semantic Methods. Among the compared non-semantic methods, M2DP, SC, and LI are based on handcrafted features, while the others utilize deep learning-based approaches. Table 1 shows that the proposed method outperforms other non-semantic methods in terms of both the maximum F_1 score across all sequences and the average maximum F_1 score. Specifically, in terms of the average maximum F_1 score, our method achieves an 18.5% improvement compared to the best-performing handcrafted feature-based method, SC, and a 2.1% increase compared to the best-performing deep learning-based method, BEVPlace. These results indicate that our method significantly outperforms other non-semantic methods, primarily due to the incorporation of a semantic segmentation auxiliary task. This auxiliary task enables the feature extraction module to focus more on the semantic information within scenes. Additionally, the LPR main task branch further guides the model to extract semantic features that are beneficial for the LPR task. These semantic features are implicitly encoded into the global place feature descriptor, therefore enabling a more efficient and discriminative representation of places.

Comparison with Semantic-based Methods. Among the compared semantic-based methods, SSC is classified as a handcrafted feature-based approach, while the others are deep learning-based. According to the comparative results presented in Table 1, it is evident that our method achieves a maximum F_1 score of 100% on sequences 06 and 07, and the average maximum F_1 score is improved by 1.9% compared to the second-best method. Overall, our method outperforms other approaches in terms of both maximum F_1 score and average maximum F_1 score across most sequences. This superiority can be attributed to: (1) Our model's "segmentation-while-describing" pipeline minimizes potential errors or disruptions by avoiding extra data-processing or storage steps; (2) Leveraging the strengths of Swin Transformer and U-Net, our SG-LPR excels in capturing global contextual information and fine-grained features; (3) SG-LPR utilizes high-dimensional feature embeddings rich in semantic information, preserving more detailed information without spatial or channel compression, unlike methods [3,10,12–14] that rely on low-level or label-level semantic information or their post-processed data.

Table 1. F_1 max scores on raw KITTI dataset.

#	Methods	00	02	05	06	07	08	Mean
1	M2DP [24]	0.708	0.717	0.602	0.787	0.560	0.073	0.575
	SC [22]	0.750	0.782	0.895	0.968	0.662	0.607	0.777
	LI [23]	0.668	0.762	0.768	0.913	0.629	0.478	0.703
	PNV [7]	0.779	0.727	0.541	0.852	0.631	0.037	0.595
	OT [39]	0.952	0.853	0.909	0.987	0.330	0.256	0.715
	DiSCO [40]	0.964	0.892	0.964	0.990	0.897	0.903	0.935
	LoGG3D-Net [8]	0.953	0.888	**0.976**	0.977	**1.000**	0.843	0.939
	BEVPlace [6]	0.979	0.900	0.974	0.991	0.906	0.894	0.941
2	SSC [3]	0.951	0.891	0.951	0.985	0.875	**0.940**	0.932
	SGPR [10]	0.820	0.751	0.751	0.655	0.868	0.750	0.766
	Locus [14]	0.957	0.745	0.968	0.948	0.921	0.900	0.907
	RINet [13]	0.978	**0.947**	0.917	0.978	0.967	0.869	0.943
	SC_LPR [12]	0.900	0.870	0.920	0.910	0.870	0.650	0.850
3	**SG-LPR(Ours)**	**0.980**	0.918	**0.976**	**1.000**	**1.000**	0.898	**0.962**

The best scores are marked in bold, and the second-best scores are underlined. #1, #2, and #3 denote experiments on non-semantic methods, semantic-based methods, and our SG-LPR, respectively.

5.2.2. Qualitative Results

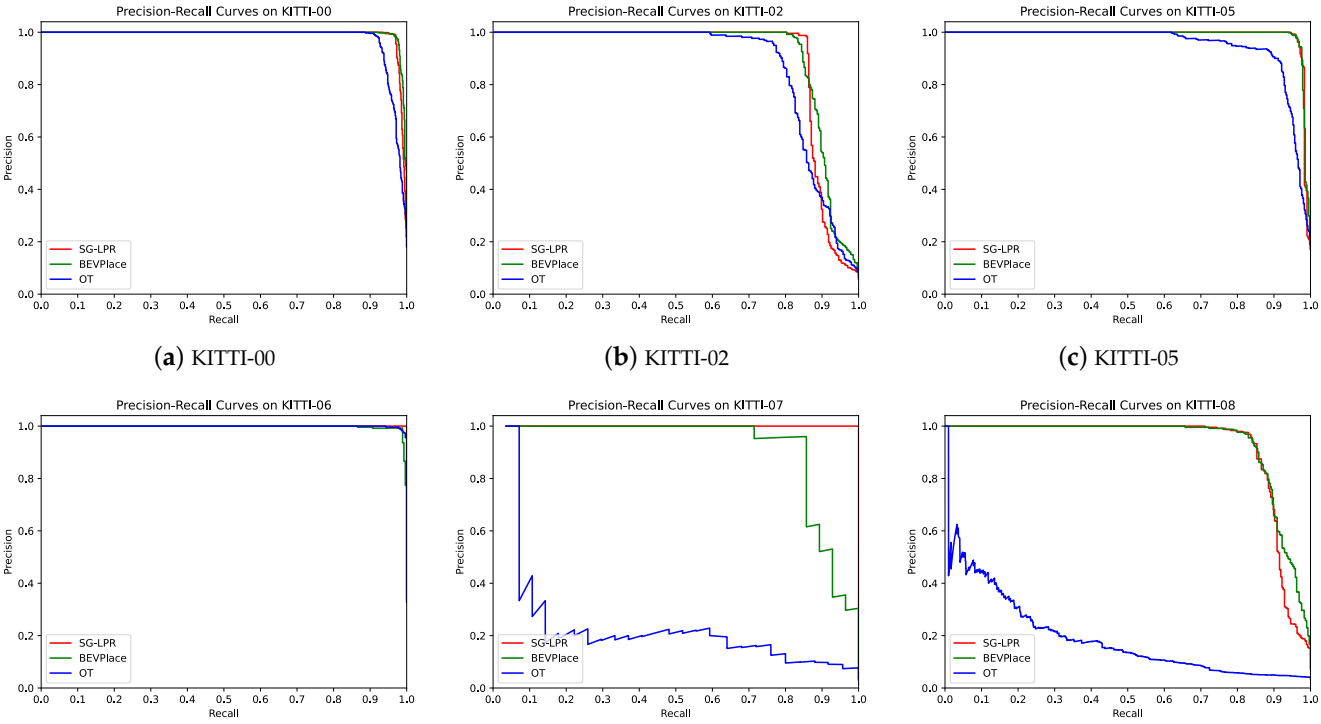
To intuitively demonstrate the performance of the proposed method, we compared its PR curves with those of OT [39] and BEVPlace [6], as shown in Figure 5a–f. The results indicate that our method exhibits more stable performance across six sequences in the KITTI dataset. Furthermore, Figure 6 illustrates the qualitative performance of our model's top-1 retrieval results along the trajectory in various KITTI sequences. In this visualization, red, black, and blue points represent true positives, false negatives, and true negatives, respectively. The visualization results clearly indicate that our model can achieve accurate detection across different types of sequences. Notably, we observed that most failure cases occur at intersections or in repetitive scenes, as shown in Figure 6b,f. Such scenes often exhibit significant geometric similarity, therefore imposing higher demands on model performance. Additionally, Figure 7 shows the input BEV images of our model, the ground-truth semantic maps, and the predicted semantic maps from the semantic segmentation auxiliary task branch. As evident from Figure 7, the proposed model not only demonstrates excellent performance in the primary LPR task but also achieves remarkable results in the semantic segmentation auxiliary task.

5.3. Robustness Test

We follow the experimental setups of methods such as RINet [13] and SGPR [10] to evaluate the robustness of the proposed model on multiple sequences of the KITTI dataset. This experiment primarily focuses on investigating the impact of viewpoint variations on the model's robustness. In real-world scenarios, mobile robots may observe the same place from different perspectives, leading to variations in the observed data. Since the model's training data cannot encompass all possible viewpoints, these variations may cause the model to misidentify the same place as different places, a challenge known as the viewpoint variation problem, which significantly impacts the model's robustness.

To simulate this situation, we randomly rotate the point clouds from the KITTI dataset around the z-axis and project it into 2D BEV images. The experimental results are presented in Table 2. The results for BEVPlace [6] are reproduced by us, and the numerical metrics for the other comparison methods are taken directly from the original paper of RINet [13]. The results show that, while our method outperforms others on sequences 05 and 07, it is slightly outperformed by RINet on other sequences and on average, particularly on sequences 02 and 08, where the maximum F_1 scores are significantly lower than those of RINet. This discrepancy is due to the fact that our model does not incorporate a specific design for rotation invariance. Instead, we apply random rotation augmentation to the

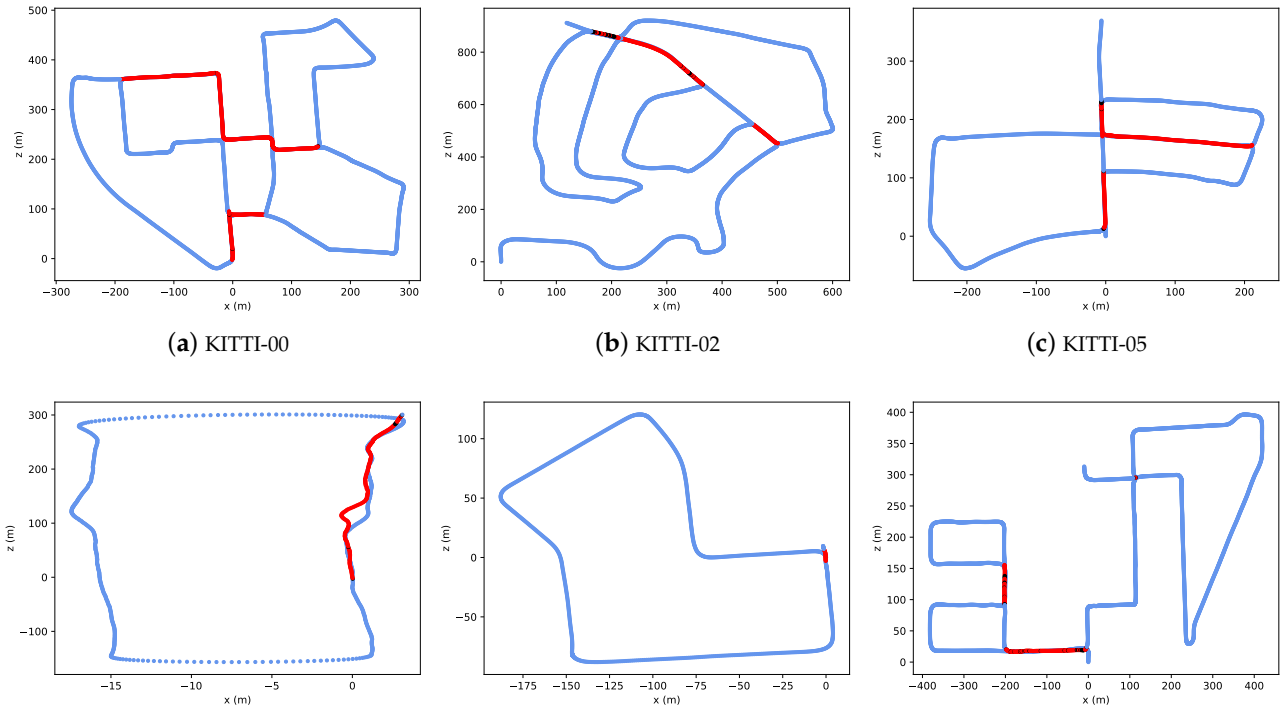
input data during training. In contrast, RINet is a rotation-invariant network, and its input consists of handcrafted features that are inherently rotation-invariant.



(**a**) KITTI-00 (**b**) KITTI-02 (**c**) KITTI-05

(**d**) KITTI-06 (**e**) KITTI-07 (**f**) KITTI-08

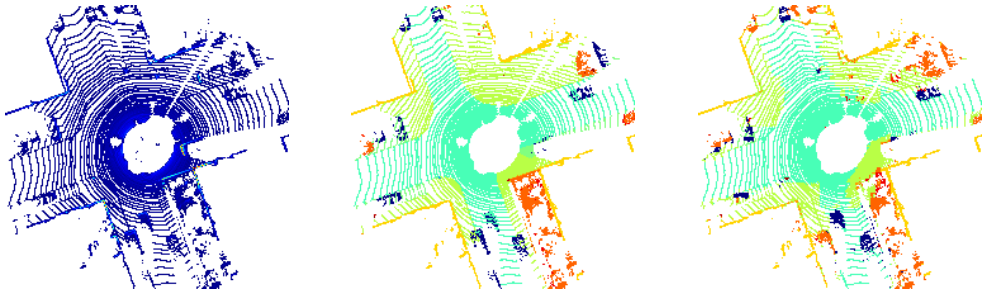
Figure 5. The Precision–Recall curves on multiple sequences of KITTI dataset.



(**a**) KITTI-00 (**b**) KITTI-02 (**c**) KITTI-05

(**d**) KITTI-06 (**e**) KITTI-07 (**f**) KITTI-08

Figure 6. Qualitative performance at top-1 retrieval of SG-LPR on multiple KITTI sequences along the trajectory. Red: true positives, black: false negatives, blue: true negatives.



(**a**) (**b**) (**c**)

Figure 7. Qualitative performance for the auxiliary semantic segmentation task in SG-LPR. (**a**) shows the original BEV image generated from 3D LiDAR point cloud, (**b**) displays the ground-truth semantic map constructed from Semantic-KITTI [46], and (**c**) illustrates the predicted semantic map produced by SG-LPR, guided by auxiliary semantic segmentation task during training.

Table 2. F_1 max scores on random rotated KITTI dataset around z-axis.

#	Methods	00	02	05	06	07	08	Mean	Cmp *
1	M2DP [24]	0.276	0.282	0.341	0.316	0.204	0.201	0.270	−0.305
	SC [22]	0.719	0.734	0.844	0.898	0.606	0.546	0.725	−0.052
	LI [23]	0.667	0.764	0.772	0.912	0.633	0.470	0.703	0.000
	PNV [7]	0.083	0.090	0.490	0.094	0.064	0.086	0.151	−0.444
	DiSCO [40]	0.960	0.891	0.952	0.985	0.894	0.892	0.929	−0.006
	BEVPlace [6]	0.979	0.900	0.974	0.991	0.906	0.894	0.941	0.000
2	SSC [3]	0.955	0.889	0.952	0.986	0.876	0.943	0.934	+0.002
	SGPR [10]	0.772	0.716	0.723	0.640	0.748	0.678	0.713	−0.053
	Locus [14]	0.944	0.726	0.960	0.927	0.911	0.877	0.891	−0.016
	RINet [13]	**0.992**	**0.942**	0.954	**1.000**	0.990	**0.962**	**0.973**	+0.030
	SC_LPR [12]	0.900	0.870	0.920	0.910	0.870	0.650	0.850	0.000
3	**SG-LPR(Ours)**	0.969	0.913	**0.976**	0.993	**1.000**	0.880	0.955	−0.007

* Cmp is the comparison with the standard results shown in Table 1.

In conclusion, rotation invariance design is crucial for enhancing the model's robustness to viewpoint variations. We plan to address this limitation in future versions of our model.

5.4. Generalization Ability

Due to the point clouds in the NCLT [50] and KITTI datasets being captured by different LiDAR sensors, as well as the significant differences in the scenes, the generalization performance test on the NCLT dataset primarily evaluates the model's ability to adapt to varied scenes and sensor types. We follow the experimental setup of BEVPlace++ [41] and use Recall@1 as the performance evaluation metric, with the sequence "2012-01-15" used to construct the database and the remaining sequences as the query set. It is important to note that the model was trained solely on the KITTI dataset. The experimental results, shown in Table 3, report numerical metrics for the comparison methods (M2DP [24], BoW3D [51], CVTNet [44], LoGG3D-Net [8], BEVPlace [6], and BEVPlace++) as presented in the original BEVPlace++ paper. The results demonstrate that our model outperforms other methods across multiple sequences of the NCLT dataset, showcasing its strong generalization capability. This is primarily due to the model's ability to effectively capture implicit semantic features in the scenes. As illustrated in Figure 8, our model successfully focuses on regions rich in semantic information in both the KITTI and NCLT datasets, further enhancing feature-encoding precision.

Table 3. Generalization performance on NCLT dataset using recall at top-1 metric.

Methods	2012-02-04	2012-03-17	2012-06-15	2012-09-28	2012-11-16	2013-02-23	Mean
M2DP [24]	0.632	0.580	0.424	0.406	0.493	0.279	0.469
BoW3D [51]	0.149	0.107	0.065	0.050	0.052	0.075	0.083
CVTNet [44]	0.892	0.880	0.812	0.749	0.771	0.803	0.818
LoGG3D-Net [8]	0.699	0.196	0.110	0.087	0.109	0.256	0.243
LCDNet [36]	0.605	0.542	0.442	0.349	0.317	0.109	0.394
BEVPlace [6]	0.935	0.927	0.874	0.878	0.889	0.862	0.894
BEVPlace++ [41]	**0.953**	**0.942**	0.902	0.889	0.913	0.878	0.913
SG-LPR (Ours)	0.947	0.936	**0.931**	**0.916**	**0.914**	**0.913**	**0.926**

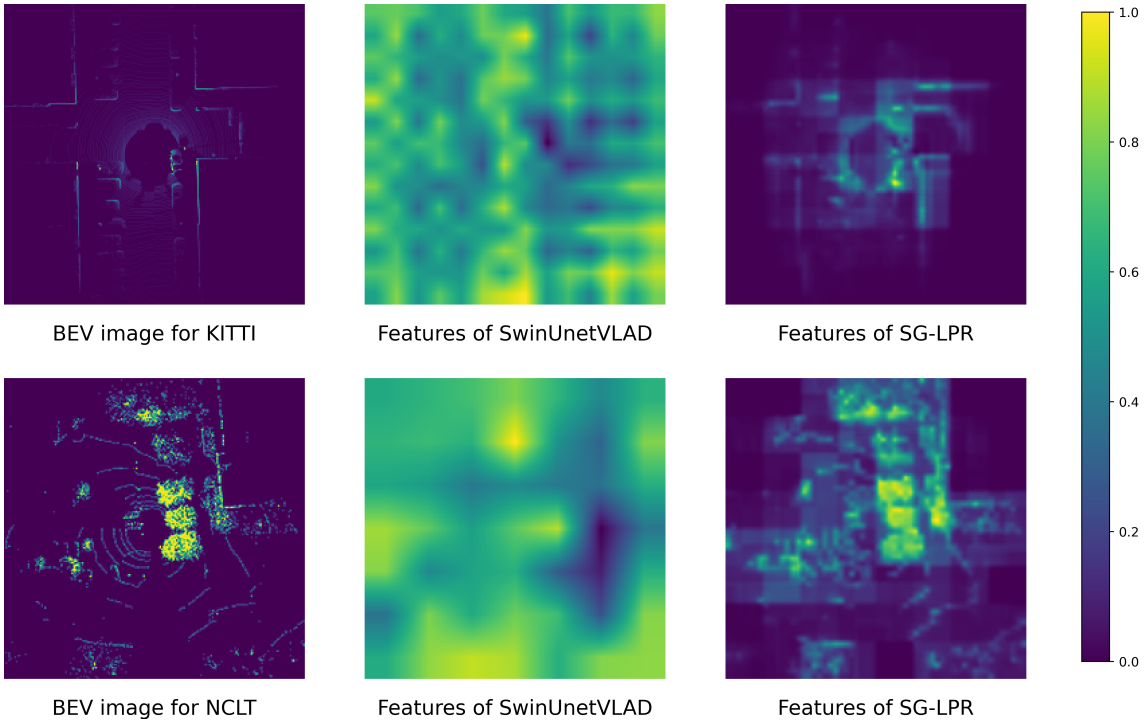


Figure 8. Feature heatmaps comparison of SG-LPR outputs with and without the semantic segmentation auxiliary task. SwinUnetVLAD is the SG-LPR variant without the semantic segmentation auxiliary task branch. The heatmaps illustrate the differences in feature activation patterns, highlighting the influence of the auxiliary task on the model's ability to capture regions with high-level semantic features.

5.5. Ablation Study

5.5.1. Ablation of Key Components in LPR Task Module

We conduct three sets of experiments on multiple sequences from the KITTI dataset to assess the impact of the spatial-channel attention module (essentially CBAM [48]) and the NetVLAD [30] module on the LPR main task branch. The experimental results are documented in Table 4: #1 indicates the removal of the CBAM and NetVLAD modules from the LPR main task branch of the SG-LPR model, #2 retains only the NetVLAD module, and #3 represents the complete SG-LPR model. The ablation results demonstrate that the NetVLAD layer significantly contributes to the enhancement of the model's performance, while the inclusion of CBAM further improves this performance. These results validate that our structural design for the LPR main task branch is both reasonable and effective.

Table 4. Ablation study on the Effectiveness of CBAM and NetVLAD modules using F_1 max scores metric.

#	Convs *	CBAM	NetVLAD	00	02	05	06	07	08	Mean
1	✓			0.926	0.836	0.878	0.991	0.224	0.716	0.762
2	✓		✓	0.961	0.906	0.936	0.983	0.885	0.819	0.915
3	✓	✓	✓	**0.980**	**0.918**	**0.976**	**1.000**	**1.000**	**0.898**	**0.962**

* Convs represents multi-layer convolution.

5.5.2. Ablation of Semantic Segmentation Auxiliary Task Branch

We conduct two sets of experiments on multiple sequences from the KITTI dataset to validate the effectiveness of the semantic segmentation auxiliary task module in our model. The results are detailed in Table 5. Experiment #1 indicates the removal of the semantic segmentation task branch from SG-LPR, termed SwinUnetVLAD, while experiment #2 retains this branch. The findings reveal that retaining the semantic segmentation task branch results in a 12.2% improvement in the average maximum F_1 score compared to its removal. This enhancement can be attributed to the feature extraction module's difficulty in effectively mining semantic information from the original BEV images without the semantic segmentation task branch. Adhering to the "segmentation-while-describing" design principle, integrating the semantic segmentation auxiliary task branch during training enables the model to focus on regions rich in semantic information that are relevant to the main LPR task, as demonstrated by the feature heatmaps from both the KITTI and NCLT datasets in Figure 8. This enhances the model's capacity to represent place. Ultimately, this ablation study underscores the importance of the semantic segmentation auxiliary task branch in our model.

Table 5. Ablation study on the presence or absence of the semantic segmentation auxiliary task branch using F_1 max scores metric.

#	seg_branch	00	02	05	06	07	08	Mean
1		0.932	0.842	0.903	0.975	0.604	0.785	0.840
2	✓	**0.980**	**0.918**	**0.976**	**1.000**	**1.000**	**0.898**	**0.962**

5.5.3. Ablation of Different Types of Input for the LPR Task Branch

According to the model architecture illustrated in Figure 2, the LPR task branch receives feature embeddings that are rich in high-level semantic information, output by the feature extraction module. In contrast, methods such as SGPR [10] and RINet [13] input low-level semantic information at the label or handcrafted feature level. This distinction contributes to our model's superior performance compared to other state-of-the-art methods. To validate this assertion, we conducted an ablation study assessing different input types for the LPR task branch, with results detailed in Table 6. Specifically, experiment #1 involves inputting only raw BEV images (RB), experiment #2 involves inputting ground-truth semantic maps at the label level (RS), experiment #3 simultaneously inputs both raw BEV images and ground-truth semantic maps (RB+RS), and experiment #4 inputs high-level semantic features (HS). The experimental results indicate that providing only low-level raw data or initial semantic maps at the label level leads to suboptimal model performance, whereas significant improvements are observed with the inclusion of high-level semantic information. This finding reinforces our argument. It is also important to note that the subpar performance when inputting low-level raw data may arise from the relatively simple structural design of the LPR branch in our model, which limits its ability to effectively extract meaningful features from raw data for place representation.

Table 6. Ablation study on different types of input for the LPR task branch using F_1 max scores metric.

#	RB	RS	HS	00	02	05	06	07	08	Mean
1	✓			0.934	0.845	0.872	0.942	0.465	0.738	0.799
2		✓		0.955	0.859	0.926	0.975	0.936	0.721	0.895
3	✓	✓		0.972	0.851	0.936	0.981	0.955	0.762	0.910
4			✓	**0.980**	**0.918**	**0.976**	**1.000**	**1.000**	**0.898**	**0.962**

RB represents the raw BEV image, where pixel values correspond to point density. RS denotes the raw semantic map, with pixel values indicating semantic category labels, reflecting low-level semantic information. HS signifies high-level semantic features, which are derived from the output of the feature extraction module of SG-LPR.

5.5.4. Ablation of Loss Function Terms

To investigate the impact of various components in the joint loss function on model performance, we conduct four additional experiments, with detailed results presented in Table 7. Experiment #1 employs only the triplet loss component λ_{lt}, which effectively corresponds to the removal of the semantic segmentation auxiliary task branch. Experiments #2 to #4 examine the influence of the presence or absence of the cross-entropy loss function $\mathcal{L}_{ce}$ and the Dice loss function $\mathcal{L}_{dice}$ while retaining the semantic segmentation task branch. The experimental results indicate that the model achieves optimal performance when all components of the joint loss function are included. This improvement is attributed to the cross-entropy loss facilitating rapid convergence and enhancing the overall category prediction capability of the semantic segmentation auxiliary task branch, while the Dice loss sharpens the model's focus on critical regions, such as small objects or difficult-to-segment areas. The combination of both loss functions ensures that the model addresses overall classification accuracy during training while also prioritizing the segmentation quality of each category, therefore implicitly enhancing the feature representation capability of the LPR task module.

Table 7. Ablation study on the impact of the presence or absence of various loss components in the joint loss function on model performance on raw KITTI using F_1 max scores metric.

#	λ_{lt}	λ_{ce}	λ_{dice}	00	02	05	06	07	08	Mean
1	✓			0.932	0.842	0.903	0.975	0.604	0.785	0.840
2	✓	✓		0.957	0.906	0.950	0.973	1.000	0.852	0.940
3	✓		✓	0.965	0.907	0.950	0.985	1.000	0.809	0.936
4	✓	✓	✓	**0.980**	**0.918**	**0.976**	**1.000**	**1.000**	**0.898**	**0.962**

5.5.5. Ablation of the Number of Semantic Categories

During the experiments, to generate the ground-truth semantic maps, we adopted the parameter settings of RINet [13], merging the 34 original semantic labels from the Semantic-KITTI [46] point cloud data into 20 categories. The merging principle involves grouping dynamic objects and their corresponding static counterparts into the same category; for example, "bus" and "moving bus" were combined into one category. To investigate the impact of different numbers of semantic categories on model performance, we conduct six experiments, setting the category numbers to 0, 6, 15, 20, 25, and 34, respectively. Here, 34 represents all categories in Semantic-KITTI; 25 groups all dynamic objects into a single category; and 15 assigns dynamic objects as unannotated, effectively excluding them from the scene. The setting of 6 merges categories with similar attributes, such as combining "person", "moving-person", "bicyclist", and "moving-bicyclist" into a single category. A setting of 0 indicates the absence of semantic labels, effectively removing the semantic segmentation auxiliary task branch.

The experimental results on different sequences of the KITTI dataset are shown in Figure 9. The results demonstrate that when the number of categories is 0, i.e., without the semantic segmentation task branch, the model's performance is the lowest. However, with

the semantic segmentation task branch included, variations in the number of categories have only a minor impact on overall model performance. This finding is consistent with the results in Figure 7, where the semantic segmentation outcomes do not perfectly match the ground-truth semantic maps, yet they do not significantly degrade the LPR performance. Additionally, we observe that increasing the number of categories leads to higher computational costs during training. Balancing efficiency and performance, we ultimately adopted the category configuration used in [13].

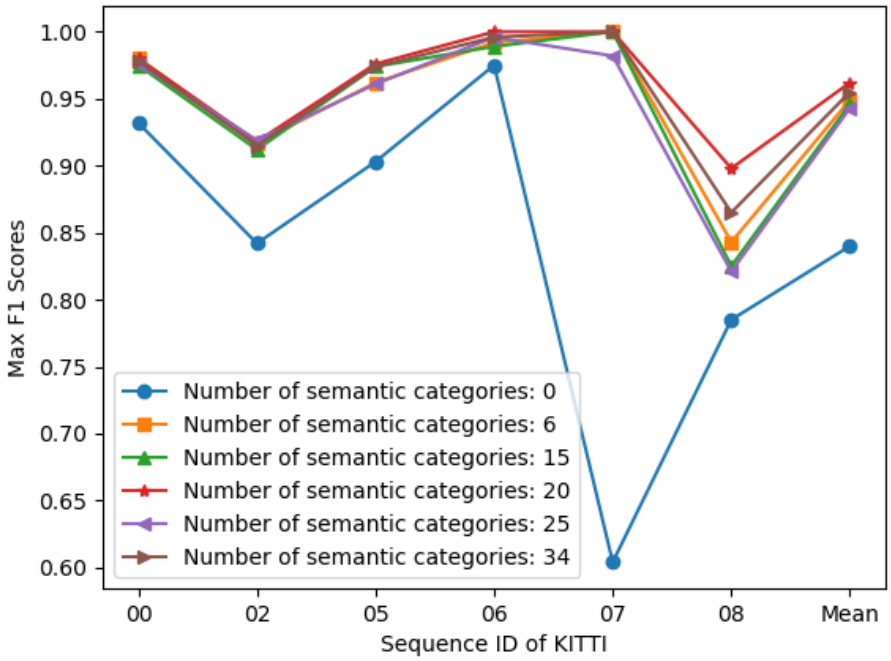


Figure 9. Ablation study on the number of semantic categories used for the training of our SG-LPR model.

6. Conclusions

In this work, we propose a semantic-guided LPR model that introduces a simple semantic segmentation auxiliary task to implement an end-to-end "segmentation-while-describing" process, spanning from raw input data to place feature descriptors. The LPR main task and the semantic segmentation auxiliary task are jointly trained, with the latter only contributing during training to guide the model in learning how to extract high-level semantic features from the scene. During the testing phase, the semantic segmentation branch is frozen, preventing any additional time overhead. Notably, our model does not explicitly segment or process semantic information, therefore avoiding extra intermediate data handling or storage and simplifying the workflow for semantic-based LPR methods. Additionally, by leveraging the strengths of the Swin Transformer and U-Net, we enhance the model's ability to capture both global contextual information and fine-grained features. A series of experiments conducted on the KITTI and NCLT datasets validate the effectiveness, robustness, and generalization capacity of the proposed method. Compared to state-of-the-art techniques, our approach demonstrates significant performance improvements.

However, there remain several areas for future enhancement: (1) Rotation invariance design: Due to the absence of a dedicated rotation invariance mechanism, the model's robustness to viewpoint variations has yet to reach its full potential; (2) Utilizing semantic information in unannotated scenes: During training, we rely on ground-truth semantic labels for joint training with the auxiliary segmentation task. However, many publicly available datasets and real-world applications lack accurate semantic annotations. Therefore, a key direction for future work is exploring how to effectively utilize semantic information

in scenes without labeled data; (3) Further enhancing generalization ability: Although our model achieves the best average Recall@1 score on the NCLT dataset, there is still substantial room for improvement, especially compared to the KITTI dataset. Enhancing the model's adaptability to different types of scenes and sensor configurations is another important avenue for future research.

Author Contributions: Conceptualization, W.J. and H.X.; methodology, W.J., H.X. and S.S.; software, W.J.; validation, W.J.; formal analysis, W.J., H.X., S.S. and C.M.; investigation, W.J. and L.X.; resources, W.J., H.X. and L.X.; data curation, W.J.; writing—original draft preparation, W.J.; writing—review and editing, W.J., H.X., S.S., C.M., L.X., Y.N. and B.D.; visualization, W.J. and H.X.; supervision, S.S., C.M., L.X., Y.N. and B.D.; project administration, L.X., Y.N. and B.D.; funding acquisition, L.X., Y.N. and B.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LiDAR	Light Detection and Ranging
LPR	LiDAR-based Place Recognition
BEV	Bird's Eye View
SG-LPR	Semantic-guided LiDAR-based Place Recognition

References

1. Shi, P.; Zhang, Y.; Li, J. LiDAR-based place recognition for autonomous driving: A survey. *arXiv* **2023**, arXiv:2306.10561.
2. Yin, P.; Zhao, S.; Cisneros, I.; Abuduweili, A.; Huang, G.; Milford, M.; Liu, C.; Choset, H.; Scherer, S. General place recognition survey: Towards the real-world autonomy age. *arXiv* **2022**, arXiv:2209.04497.
3. Li, L.; Kong, X.; Zhao, X.; Huang, T.; Li, W.; Wen, F.; Zhang, H.; Liu, Y. SSC: Semantic scan context for large-scale place recognition. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 2092–2099.
4. Du, J.; Wang, R.; Cremers, D. Dh3d: Deep hierarchical 3d descriptors for robust large-scale 6dof relocalization. In Proceedings of the European Conference on Computer Vision (ECCV), Virtual Venue, 23–28 August 2020; pp. 744–762.
5. Komorowski, J. Minkloc3d: Point cloud based large-scale place recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Virtual Venue, 3–8 January 2021; pp. 1790–1799.
6. Luo, L.; Zheng, S.; Li, Y.; Fan, Y.; Yu, B.; Cao, S.Y.; Li, J.; Shen, H.L. BEVPlace: Learning LiDAR-based place recognition using bird's eye view images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 4–6 October 2023; pp. 8700–8709.
7. Uy, M.A.; Lee, G.H. PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, 18–22 June 2018; pp. 4470–4479.
8. Vidanapathirana, K.; Ramezani, M.; Moghadam, P.; Sridharan, S.; Fookes, C. LoGG3D-Net: Locally guided global descriptor learning for 3D place recognition. In Proceedings of the 2022 International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2022; pp. 2215–2221.
9. Arce, J.; Vödisch, N.; Cattaneo, D.; Burgard, W.; Valada, A. PADLoC: LiDAR-based deep loop closure detection and registration using panoptic attention. *IEEE Robot. Autom. Lett.* **2023**, *8*, 1319–1326. [CrossRef]
10. Kong, X.; Yang, X.; Zhai, G.; Zhao, X.; Zeng, X.; Wang, M.; Liu, Y.; Li, W.; Wen, F. Semantic graph based place recognition for 3d point clouds. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 25–29 October 2020; pp. 8216–8223.
11. Yin, P.; Xu, L.; Feng, Z.; Egorov, A.; Li, B. PSE-Match: A viewpoint-free place recognition method with parallel semantic embedding. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 11249–11260. [CrossRef]
12. Kong, D.; Li, X.; Xu, Q.; Hu, Y.; Ni, P. SC_LPR: Semantically consistent LiDAR place recognition based on chained cascade network in long-term dynamic environments. *IEEE Trans. Image Process.* **2024**, *33*, 2145–2157. [CrossRef] [PubMed]
13. Li, L.; Kong, X.; Zhao, X.; Huang, T.; Li, W.; Wen, F.; Zhang, H.; Liu, Y. RINet: Efficient 3D LiDAR-based place recognition using rotation invariant neural network. *IEEE Robot. Autom. Lett.* **2022**, *7*, 4321–4328. [CrossRef]

14. Vidanapathirana, K.; Moghadam, P.; Harwood, B.; Zhao, M.; Sridharan, S.; Fookes, C. Locus: LiDAR-based place recognition using spatiotemporal higher-order pooling. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 5075–5081.
15. Dai, D.; Wang, J.; Chen, Z.; Bao, P. SC-LPR: Spatiotemporal context based LiDAR place recognition. *Pattern Recognit. Lett.* **2022**, *156*, 160–166. [CrossRef]
16. Chen, X.; Läbe, T.; Milioto, A.; Röhling, T.; Vysotska, O.; Haag, A.; Behley, J.; Stachniss, C. OverlapNet: Loop closing for LiDAR-based SLAM. *arXiv* **2021**, arXiv:2105.11344.
17. Wu, H.; Zhang, Z.; Lin, S.; Mu, X.; Zhao, Q.; Yang, M.; Qin, T. MapLocNet: Coarse-to-Fine Feature Registration for Visual Re-Localization in Navigation Maps. *arXiv* **2024**, arXiv:2407.08561.
18. Ming, Y.; Yang, X.; Zhang, G.; Calway, A. Cgis-net: Aggregating colour, geometry and implicit semantic features for indoor place recognition. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; pp. 6991–6997.
19. Ming, Y.; Ma, J.; Yang, X.; Dai, W.; Peng, Y.; Kong, W. AEGIS-Net: Attention-Guided Multi-Level Feature Aggregation for Indoor Place Recognition. In Proceedings of the ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, 14–19 April 2024; pp. 4030–4034.
20. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual Venue, 11–17 October 2021; pp. 10012–10022.
21. Yin, H.; Xu, X.; Lu, S.; Chen, X.; Xiong, R.; Shen, S.; Stachniss, C.; Wang, Y. A survey on global lidar localization: Challenges, advances and open problems. *Int. J. Comput. Vis.* **2024**, *132*, 3139–3171. [CrossRef]
22. 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), Madrid, Spain, 1–5 October 2018; pp. 4802–4809.
23. Wang, Y.; Sun, Z.; Xu, C.Z.; 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), Las Vegas, NV, USA, 25–29 October 2020; pp. 5769–5775.
24. 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), Daejeon, South Korea, 9–14 October 2016; pp. 231–237.
25. Magnusson, M.; Andreasson, H.; Nuchter, A.; Lilienthal, A.J. Appearance-based loop detection from 3D laser data using the normal distributions transform. In Proceedings of the 2009 IEEE International Conference on Robotics and Automation (ICRA), Kobe, Japan, 12–17 May 2009; pp. 23–28.
26. Bosse, M.; Zlot, R. Place recognition using keypoint voting in large 3D lidar datasets. In Proceedings of the 2013 IEEE International Conference on Robotics and Automation (ICRA), Karlsruhe, Germany, 6–10 May 2013; pp. 2677–2684.
27. 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), Marina Bay Sands, Singapore, 29 May–3 June 2017; pp. 5266–5272.
28. Zou, X.; Li, J.; Wang, Y.; Liang, F.; Wu, W.; Wang, H.; Yang, B.; Dong, Z. PatchAugNet: Patch feature augmentation-based heterogeneous point cloud place recognition in large-scale street scenes. *ISPRS J. Photogramm. Remote Sens.* **2023**, *206*, 273–292. [CrossRef]
29. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
30. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5297–5307.
31. Zhang, W.; Xiao, C. PCAN: 3D attention map learning using contextual information for point cloud based retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12436–12445.
32. Xia, Y.; 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), virtual venue, 19–25 June 2021; pp. 11348–11357.
33. Sun, Q.; Liu, H.; He, J.; Fan, Z.; Du, X. Dagc: Employing dual attention and graph convolution for point cloud based place recognition. In Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR), New York, NY, USA, 8–11 June 2020; pp. 224–232.
34. Hui, L.; Yang, H.; Cheng, M.; Xie, J.; Yang, J. Pyramid point cloud transformer for large-scale place recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual Venue, 11–17 October 2021; pp. 6098–6107.
35. Fan, Z.; Song, Z.; Liu, H.; Lu, Z.; He, J.; Du, X. SVT-Net: Super light-weight sparse voxel transformer for large scale place recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Virtual Venue, 22 February–1 March 2022; Volume 36, pp. 551–560.

36. Cattaneo, D.; Vaghi, M.; Valada, A. Lcdnet: Deep loop closure detection and point cloud registration for lidar slam. *IEEE Trans. Robot.* **2022**, *38*, 2074–2093. [CrossRef]
37. Xia, Y.; Gladkova, M.; Wang, R.; Li, Q.; Stilla, U.; Henriques, J.F.; Cremers, D. Casspr: Cross attention single scan place recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 4–6 October 2023; pp. 8461–8472.
38. Wu, T.; Fu, H.; Liu, B.; Xue, H.; Ren, R.; Tu, Z. Detailed analysis on generating the range image for lidar point cloud processing. *Electronics* **2021**, *10*, 1224. [CrossRef]
39. Ma, J.; Zhang, J.; Xu, J.; Ai, R.; Gu, W.; Chen, X. OverlapTransformer: An efficient and yaw-angle-invariant transformer network for LiDAR-based place recognition. *IEEE Robot. Autom. Lett.* **2022**, *7*, 6958–6965. [CrossRef]
40. Xu, X.; Yin, H.; Chen, Z.; Li, Y.; Wang, Y.; Xiong, R. Disco: Differentiable scan context with orientation. *IEEE Robot. Autom. Lett.* **2021**, *6*, 2791–2798. [CrossRef]
41. Luo, L.; Cao, S.; Li, X.; Xu, J.; Ai, R.; Yu, Z.; Chen, X. BEVPlace++: Fast, Robust, and Lightweight LiDAR Global Localization for Unmanned Ground Vehicles. *arXiv* **2024**, arXiv:2408.01841.
42. Cao, F.; Yan, F.; Wang, S.; Zhuang, Y.; Wang, W. Season-invariant and viewpoint-tolerant LiDAR place recognition in GPS-denied environments. *IEEE Trans. Ind. Electron.* **2020**, *68*, 563–574. [CrossRef]
43. Lu, S.; Xu, X.; Tang, L.; Xiong, R.; Wang, Y. DeepRING: Learning roto-translation invariant representation for LiDAR based place recognition. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 1904–1911.
44. Ma, J.; Xiong, G.; Xu, J.; Chen, X. CVTNet: A cross-view transformer network for LiDAR-based place recognition in autonomous driving environments. *IEEE Trans. Ind. Inform.* **2023**, *20*, 4039–4048. [CrossRef]
45. Zhang, J.; Zhang, Y.; Rong, L.; Tian, R.; Wang, S. MVSE-Net: A Multi-View Deep Network With Semantic Embedding for LiDAR Place Recognition. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 17174–17186. [CrossRef]
46. Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; Gall, J. SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9297–9307.
47. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like pure transformer for medical image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 205–218.
48. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
49. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
50. Carlevaris-Bianco, N.; Ushani, A.K.; Eustice, R.M. University of Michigan North Campus long-term vision and lidar dataset. *Int. J. Robot. Res.* **2016**, *35*, 1023–1035. [CrossRef]
51. Cui, Y.; Chen, X.; Zhang, Y.; Dong, J.; Wu, Q.; Zhu, F. Bow3d: Bag of words for real-time loop closing in 3d lidar slam. *IEEE Robot. Autom. Lett.* **2022**, *8*, 2828–2835. [CrossRef]