Local and Global Feature Cross-attention Multimodal Place Recognition

Lu Xu[1,2], Shuaixin Li[1,2(✉)*], Xin Zhou[1,2,3], Xiaozhou Zhu[1,2], Wen Yao[1,2(✉)]

[1]Intelligent Game and Decision Laboratory, Beijing 100071, China
[2]Defense Innovation Institute, Chinese Academy of Military Science, Beijing 100071, China
[3]School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

yaowen@nudt.edu.cn, lsx_navigation@sina.com

Abstract

Recent advances in autonomous driving have greatly increased the focus on place recognition technology, a crucial aspect of environmental perception. However, current multimodal fusion techniques for combining camera and LiDAR data often neglect the distinct measurement attributes of these sensors, thereby restricting the efficiency of accurate place identification. In this work, we introduce LoG-PR, a multimodal place recognition approach that integrates 2D local and 3D global features and combines them through cross-attention to improve recognition accuracy. Our approach handles each modality separately, generating a comprehensive descriptor tensor through data fusion and key point enhancement, leveraging the unique strengths of both sensors. We simultaneously employ voxel-based and point-based methods for capturing global characteristics of the surveyed environmental point cloud, while utilizing residual neural networks to extract intricate local features from images. To effectively exploit the potential correlations between images and point clouds, we develop a cross-attention transformer that integrates complementary information from diverse modalities while preserving the original feature information. We test our method on the open-source benchmarks including Oxford RobotCar, NCLT, KITTI and MUN-FRL. The results demonstrate a significant performance improvement compared to existing SOTA methods, effectively enhancing the accuracy and robustness of place recognition. Our code is publicly accessible at: https://github.com/LuXu01/Log-PR.

***Keywords:** Place recognition, Multimodal, Local and global features, Deep learning*

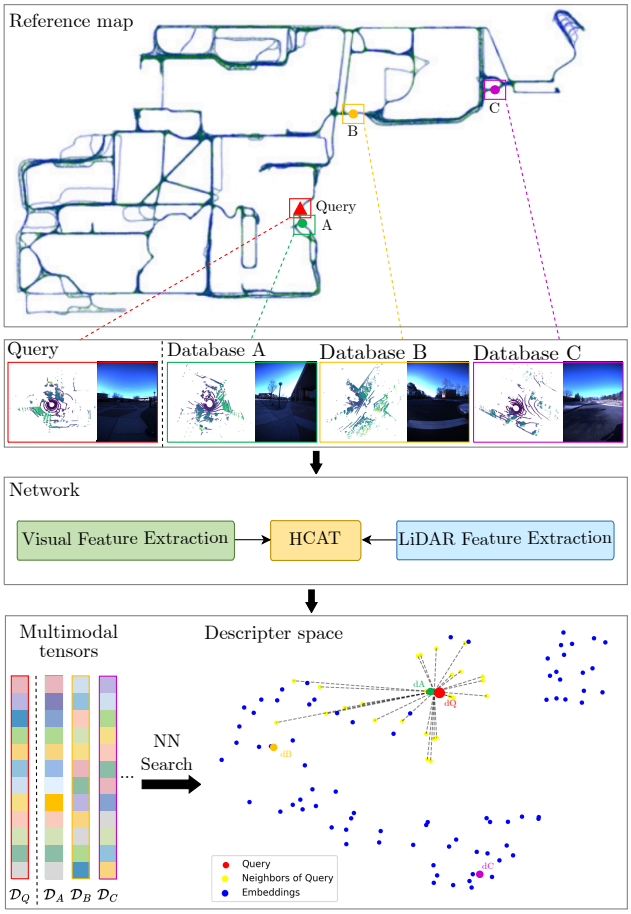

Figure 1. The pipeline of LoG-PR achieving place retrieval. An real-world data example is shown in the figure, where point A is the correct match to the query, while B and C are incorrect. Multimodal place perception data is fed into the LoG-PR network mapping the visual measurements as descriptors. Place retrieval is achieved by searching for the nearest neighbor tensor in the descriptor space. The bottom-right figure shows a visualization of the nearest neighbor search, with yellow dots marking the top 25 neighbors closest to the query.

*This work was supported by the National Natural Science Foundation of China (Grant No. 42201501).

1. Introduction

Place recognition (PR) stands as a prominent research topic within the realms of computer vision and artificial intelligence. It plays a vital role in simultaneous localization and mapping (SLAM) and global localization for robots. Through place recognition, vehicles are enabled to build the correct topological structure of the physical world, enhancing localization precision by matching the present view with that in the map database.

In autonomous vehicle systems, PR is typically classified into 2D visual perception and 3D stereo perception based on sensor usage. 2D visual perception primarily relies on onboard cameras to acquire environmental data, providing high-resolution color images that capture detailed information effectively at low cost. However, cameras are constrained to capturing objects within their field of view and are vulnerable to interference from adverse weather conditions such as rain, snow, or fog. Additionally, they are sensitive to variations in ambient light conditions. 3D stereo perception primarily relies on onboard LiDAR to acquire environmental data. LiDAR offers several advantages, such as the ability to detect objects at greater distances, accurately measure distances between objects and the sensor, generate precise point cloud data, and operate effectively regardless of lighting conditions. This makes it suitable for use in both low-light and high-light environments. However, LiDAR faces challenges in capturing fine details, and cannot distinguish object colors and textures. In summary, cameras and LiDAR each possess their own strengths and weaknesses. In many cases, 2D and 3D data are highly complementary, and their integration can compensate for the limitations of individual sensors, leading to more comprehensive environmental perception.

Multimodal data fusion is widely acknowledged in academia and industry as a crucial direction for advancing autonomous vehicle perception technology. While multisensor fusion can significantly enhance PR performance, existing multimodal methods lack a tightly integrated data fusion module, hindering better place retrieval. LiDAR point clouds provide a broad range of measurements, enabling a more comprehensive representation of the entire scanned scene through global 3D descriptors. Notably, in road scenarios, the global descriptor in point clouds exhibits approximate equivariance to laser scan rotations and translations, while considering the spatial object distribution [25]. On the other hand, cameras capture rich color and texture details of objects, providing a more detailed representation of specific objects in the scene. 2D Local descriptors focus on scenario details, aiding in the recognition of places through prominent landmarks. However, feature concatenation used in SOTA multimodal PR approaches, such as those in MinkLoc++ [17], CORAL [31] and the large scale PR method proposed by [39], may not be the most efficient or comprehensive way, as it extends the descriptor dimension through tensor stacking and ignores the implicit relationships between different modalities, making it challenging to emphasize key information efficiently and effectively.

In this work, we introduce LoG-PR, a multimodal fusion network for PR in road scenes (see Fig.1). Perceptual data from LiDAR and camera is fed into the feature extraction head to concurrently produce local 2D and global 3D features. Then, we design a cross-attention transformer module to achieve efficient and comprehensive data fusion by highlighting the most valuable key features. Through the combination of local and global features and the incorporation of the transformer, our approach exhibits a robust capability for place retrieval in scenarios involving viewpoint changes and scene transitions. In summary, the main contributions of our work are as follows:

- We present a novel multimodal place recognition method called LoG-PR that integrates local 2D and global 3D features, advancing the current SOTA in terms of the PR task in road scenes.
- We introduce a cross-attention transformer module accounting for various measurement properties of camera and LiDAR to achieve more efficient and comprehensive fusion of visual local features and point cloud global features.
- We conduct experiments on the Oxford RobotCar, NCLT, and KITTI datasets to evaluate the performance of our method across diverse environments. When our paper is accepted, a link to the code is made public.

2. Related Work

2.1. Image-based and Point Cloud-based Place Recognition

Traditionally, visual place recognition (VPR) depends on techniques of handcrafted features extraction from images, such as local features like SIFT [22] and SURF [3], or global features like HOG [8]. In recent years, deep learning technologies have been employed to address the challenges of VPR. Recent studies broadly categorized into global and local descriptors. Local descriptors are a collection of vectors that represent specific local regions in an image, containing detailed local information. Chen et al. [7] extract local features directly from the convolutional layers and utilized higher convolutional layers to pool these features, creating multiple local descriptors to represent each image. Global descriptors are single vectors representing the entire image, encapsulating the overall information of the image. Arandjelovic et al. [2] treating the output of convolutional

layers as $W \times H$ local descriptors of length K and aggregating them with a specially designed pooling layer to obtain a global descriptor.

Given the irreplaceable advantages of LiDAR data, such as extensive perception range and immunity to varying illumination conditions, PR using LiDAR has emerged as a highly focused research subject. LiDAR place recognition (LPR) can be broadly categorized into three types based on the input point cloud type, i.e. point-based, projection-based, and voxel-based. PointNetVLAD [35] is a point-based method. It integrates PointNet [32] and NetVLAD to tackle the PR problem. Ma et al. [29] introduce the OverlapTransformer, a projection-based method. It projects point clouds into range images and incorporates Transformer and NetVLAD to achieve yaw rotation invariance. MinkLoc3D [16] adopts a voxel-based approach. It uses sparse voxelized point cloud representation and sparse 3D convolutions to compute a discriminative 3D point cloud descriptor.

2.2. LiDAR and RGB Camera Fusion for Multimodal Place Recognition

Multimodal fusion for PR is a complex and challenging research area that involves integrating data from various sensors to overcome the limitations of individual modalities. PIC-Net [24] is a collaborative network of point clouds and images for large scale PR. It uses spatial attention VLAD to fuse discriminative points and pixels. MinkLoc++ [17] is a fusion method for LiDAR and monocular images for PR. This method concatenates the image descriptors and point cloud descriptors along the channel dimension, forming a multimodal descriptor. Many existing camera-LiDAR fusion methods simply combine the two sensors without considering their performance characteristics in different environments. AdaFusion [18] employs an attention branch network that adaptively assigns weights to the camera and LiDAR sensors based on current environmental conditions. LCPR [42] fuses LiDAR point clouds with multiview RGB images to generate discriminative and yaw rotation-invariant representations of the environment.

Previous research has achieved significant progress in PR, but certain challenges remain. Specific point feature extraction schemes have various limitations. For instance, methods such as PointNet, which are point-based, often neglect the local structural relationships among points. Conversely, voxel-based techniques employ sparse 3D convolution architectures to extract feature-rich information but may lose some points during quantization. Moreover, multimodal fusion approaches that rely on simple concatenation or summation of descriptors not only increase the dimension of descriptors but also disregard the inherent feature correlations among different modalities, thereby impacting the efficiency and accuracy of PR. To address these issues, we propose a local and global feature cross-attention multimodal place recognition approach. By leveraging cross-attention, we highlight the most valuable key points within fixed-dimensional descriptors and effectively utilize the complementary information provided by both modalities, capturing the intricate relationships between the 3D global features and the 2D local features.

3. Problem Statement

The reference database $\mathbf{M}_{ref} = \{\mathbf{m}_k, k = 1, \cdots, N\}$ is defined as a collection of pairs $\mathbf{m}_k = \{\mathcal{P}_k, \mathcal{I}_k\}$ of *single-shot* 3D LiDAR scans $\mathcal{P}_k$ and RGB images $\mathcal{I}_k$. Each pair of measurements is synchronized and marked with GPS coordinates at the sampling location. Let $\mathbf{q}$ represent a query pair comprising a single LiDAR scan $\mathcal{P}_q$ and a color image $\mathcal{I}_q$ captured at a particular timestamp. Then, the multimodal PR problem is defined as retrieving the best match $\mathbf{m}^* = \{\mathcal{P}^*, \mathcal{I}^*\}$ from $\mathbf{M}_{ref}$ which is the closest neighbor with $\mathbf{q}$ in the high-dimension descriptor space $\mathbb{R}^M, M = L_{\mathcal{P}} + L_{\mathcal{I}}$. We aim to design a network $\mathbf{\Phi}(\cdot)$ that encodes the multimodal measurement pair as an identified scene descriptor. This is allow us to retrieve the correct match $\mathbf{m}^* \in \mathbf{M}_{ref}$ through nearest-neighbor search in the reference database:

$$\mathbf{m}^* = \underset{\mathbf{m}_k \in \mathbf{M}_{ref}}{\operatorname{argmin}} \; d(\mathbf{\Phi}(\mathbf{q}), \mathbf{\Phi}(\mathbf{m}_k)), \tag{1}$$

where $d(\cdot)$ represents the Euclidean distance between descriptors of $\mathbf{q}$ and $\mathbf{m}_k$.

4. Proposed Method

4.1. Network Overview

The overview of the proposed network primarily comprises two modules, ie, the feature extraction module and the cross-attention fusion module, as depicted in Fig.2. The feature extraction module is subdivided into two parallel branches. The right branch focuses on extracting visual features using 2D convolution, while the left branch specializes in LiDAR feature extraction using 3D convolution. This dual-branch approach enables the network to capture rich representations from both modalities simultaneously. Following the feature extraction phase, the cross-attention fusion module comes into play. This module is designed to integrate the original visual and LiDAR descriptors obtained from the respective branches of the feature extraction module. By leveraging cross-attention transformer, it facilitates the highlighting the key points of these multimodal descriptors with fixed-length dimension and the more comprehensive data fusion, thereby enhancing the network's ability to process and interpret complex data from diverse sources effectively.

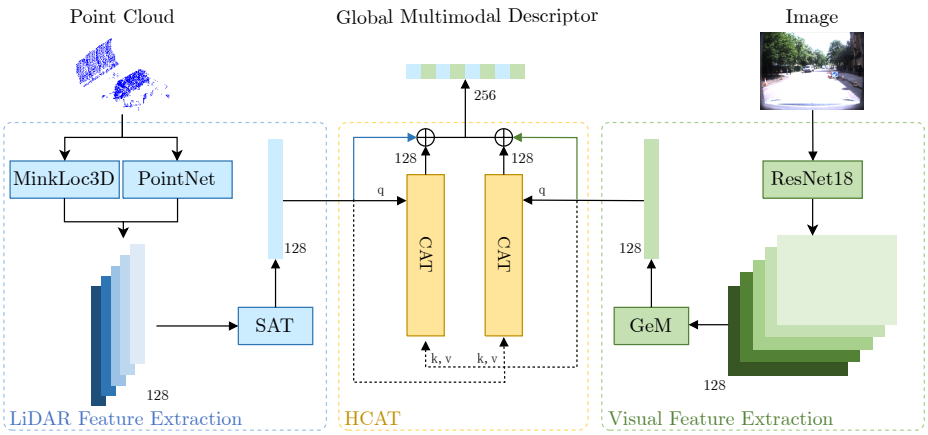

Figure 2. The LoG-PR network architecture comprises two primary components. The first component is responsible for extracting features from point clouds and images to generate $\mathcal{D}_{PC}$ and $\mathcal{D}_{RGB}$ respectively. The second component employs a hierarchical cross-modal attention transformer to fuse these descriptors. The numerical value (e.g., 128, 256) indicates the number of channels of the feature.

4.2. Feature Extraction

Considering that 2D images depict detailed target features of captured areas, while 3D sparse point clouds provide a comprehensive representation of scene contours and structural information, we design a feature descriptor that combines both aspects. This descriptor aims to simultaneously encompass target feature information and scene contour feature information. Based on the above ideas, as illustrated in Fig.2, the feature extraction module consists of left and right branches which take $\mathcal{P}_k$ and $\mathcal{I}_k$ as input and produce the point cloud and image descriptor tensor with dimensions $L_{\mathcal{P}}$ and $L_{\mathcal{I}}$, respectively.

In the image branch, we utilize ResNet18, chosen for its ability to capture image textures and semantic information effectively, addressing issues of network degradation during deep neural network training. Specifically, let $\mathbf{\Omega}(\cdot) : \mathbb{R}^{C\times H\times W} \mapsto \mathbb{R}^{L_{\mathcal{I}}\times H'\times W'}$ be the first 4 blocks of ResNet18, which outputs a $L_{\mathcal{I}} \times H' \times W'$ dimensional feature map $\mathbf{f}_k$ given the image input $\mathcal{I}_k \in \mathbb{R}^{C\times H\times W}$. Subsequently, as illustrated in Fig.2, we apply generalized mean pooling (GeM) on the feature map $\mathbf{f}_k$ to obtain the final RGB image descriptor $\mathcal{D}_{RGB}$:

$$\underset{L_{\mathcal{I}}\times 1}{\mathcal{D}_{RGB}} = \mathrm{GeM}(\mathbf{f}_k) = \left(\frac{1}{|H'\times W'|}\sum_{i=1}^{|H'\times W'|}(\mathbf{f}_k(i)^{\lambda})\right)^{\frac{1}{\lambda}}. \tag{2}$$

where $|H' \times W'|$ represents the number of spatial locations in $\mathbf{f}_k$, and λ is a parameter controlling the degree of pooling. In practice, we set $L_{\mathcal{I}} = 128$, $H' = 15$, $W' = 20$, and $\lambda = 3$ for the color image $\mathcal{I}_k \in \mathbb{R}^{3\times 320\times 240}$.

In the point cloud branch, we integrate PointNet and MinkLoc3D. PointNet excels in capturing fine-grained geometric features, while MinkLoc3D focuses on capturing coarse structural features. PointNet applies shared multilayer perceptrons (MLP) independently to each point $\mathbf{p}_i = [x_i, y_i, z_i] \in \mathcal{P}_k, i = 1, \cdots, S$ to extract features map with dimension $S \times L_{\mathcal{P}}$. The feature map is then aggregated using max pooling to generate a global feature vector $\mathbf{g}_k \in \mathbb{R}^{L_{\mathcal{P}}}$. MinkLoc3D, on the other hand, discretizes the input point cloud $\mathbf{p}_i = [x_i, y_i, z_i] \in \mathcal{P}_k$ into a sparse voxelized representation $\mathbf{v}_j = [\hat{x}_j, \hat{y}_j, \hat{z}_j] \in \mathcal{V}_k, j = 1, \cdots, S'$. A single discretized voxel $\mathbf{v}_j$ represents the average of all points that fall within the voxel. Then $\mathcal{V}_k$ is used by a sparse convolutional network to extract feature map with dimension $S' \times L_{\mathcal{P}}$. Notably, $S' < S$ since the sparse voxelization step will lose most points. However, PointNet can directly process $\mathcal{P}_k$, avoiding the loss of information that occurs during voxelization. Specifically, let $\mathbf{\Psi}(\cdot) : \mathbb{R}^{3\times S} \mapsto \mathbb{R}^{L_{\mathcal{P}}\times 1}$ and $\mathbf{\Upsilon}(\cdot) : \mathbb{R}^{3\times S'} \mapsto \mathbb{R}^{L_{\mathcal{P}}\times 1}$ represent the PointNet and MinkLoc3D respectively, which both output feature tensors $\mathbf{g}_k$ and $\mathbf{h}_k$ with the same feature dimension. The point-based and voxel-based feature tensors are simultaneously used to compensate quantization loss and maintain geometric details. We then directly add $\mathbf{g}_k$ and $\mathbf{h}_k$, denoting the result as $\mathbf{x}_k = \mathbf{g}_k + \mathbf{h}_k \in \mathbb{R}^{L_{\mathcal{P}}\times 1}$, and employ a self-attention transformer[41] on $\mathbf{x}_k$ to obtain the final point cloud descriptor $\mathcal{D}_{PC}$, fully capturing the complex relationship between the two feature map tensors. The SAT of the i-th element $x_i \in \mathbf{x}_k, i = 1, \cdots, L_{\mathcal{P}}$ is:

$$\mathrm{SAT}\left(x_i\right) = \sum_{j\in\mathbf{x}_k, j\neq i}\mathrm{Softmax}\left(\theta\left(x_i - x_j\right) + x_i \cdot x_j\right)\psi\left(x_j\right), \tag{3}$$

where θ represents the positional encoding function, and ψ is the value projection function. In practice, we set $L_{\mathcal{P}} = 128$.

4.3. Cross-attention Fusion

Traditional fusion methods often fail to adequately capture and leverage complex relationships between different

modalities, leading to suboptimal information interaction. In response, we present a novel approach: the hierarchical cross-attention transformer (HCAT) fusion module, designed to fuse point cloud and image features effectively. Fig.3 illustrates the detailed architecture of the HCAT module. This module aims to seamlessly integrate complementary information from diverse modalities while preserving their unique characteristics. Central to our approach is the utilization of a multihead attention transformer within a hierarchical structure, enabling the model to capture intricate relationships between features across different levels and perspectives. This comprehensive approach ensures robust performance across a wide range of complex environments and diverse scenes.

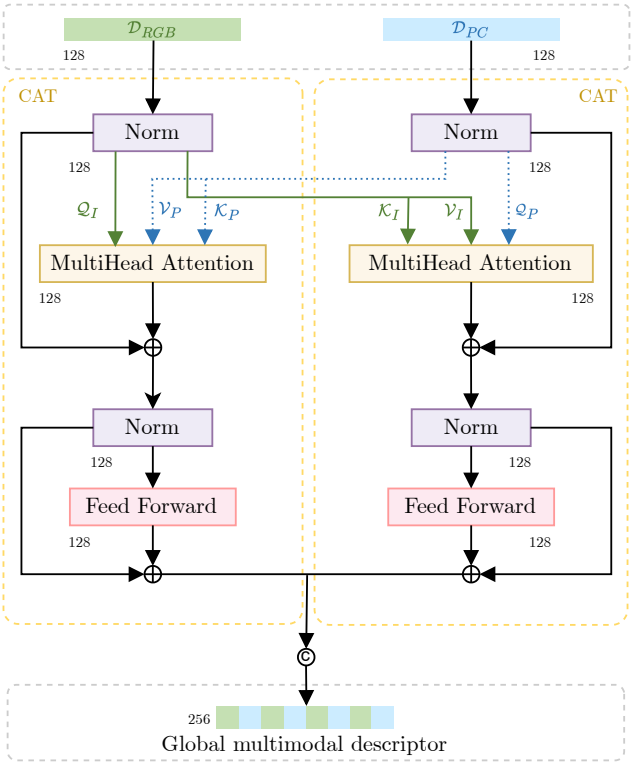

Figure 3. The detailed architecture of the hierarchical cross-attention transformer (HCAT) fusion module.

To enhance model stability and convergence speed, we introduce normalization layers to standardize point cloud and image features, ensuring consistent scaling across different dimensions. Subsequently, two cross-attention submodules are employed to compute cross-modal attention between normalized features. In the first sub-module, point cloud features and image features serve as input query and key-value sequences, respectively, into the cross-attention module. Similarly, the second sub-module performs symmetric operations, allowing each modality to focus on salient information areas in the other modality. The first cross-attention transformer can be represented as

$$Attention\left(\mathbf{Q}_{\mathcal{P}}, \mathbf{K}_{\mathcal{I}}, \mathbf{V}_{\mathcal{I}}\right) = \text{Softmax}\left(\frac{\mathbf{Q}_{\mathcal{P}}\mathbf{K}_{\mathcal{I}}^{T}}{\sqrt{d_k}}\right)\mathbf{V}_{\mathcal{I}} \tag{4}$$

where $\mathbf{Q}_{\mathcal{P}}$ represents the query sequence for point cloud features, while $\mathbf{K}_{\mathcal{I}}$ and $\mathbf{V}_{\mathcal{I}}$ denote the key and value sequences for image features, respectively. Similarly, the second cross-attention sub-module can be represented as

$$Attention\left(\mathbf{Q}_{\mathcal{I}}, \mathbf{K}_{\mathcal{P}}, \mathbf{V}_{\mathcal{P}}\right) = \text{Softmax}\left(\frac{\mathbf{Q}_{\mathcal{I}}\mathbf{K}_{\mathcal{P}}^{T}}{\sqrt{d_k}}\right)\mathbf{V}_{\mathcal{P}} \tag{5}$$

where $\mathbf{Q}_{\mathcal{I}}$ represents the query sequence for image features, and $\mathbf{K}_{\mathcal{P}}$ and $\mathbf{V}_{\mathcal{P}}$ denote the key and value sequences for point cloud features, respectively.

Following attention enhancement, the resulting feature tensors are added with the original feature tensor and passed through feedforward neural networks and additional normalization layers to maintain feature balance and promote information propagation. Finally, the enhanced features from both cross-attention submodules are concatenated along the feature dimension, culminating in the formation of the final global descriptor.

4.4. Loss Function

Similar to [17], we adopt a deep metric learning approach using a triplet loss framework composed of triplets. Each triplet comprises a mini-batch element containing a 3D point cloud paired with its corresponding RGB image. These mini-batches are structured into triplets consisting of an anchor, a positive example, and a negative example. The loss function is defined by Eq.6, where the first term corresponds to $\mathcal{D}_{PC}$, the second term represents $\mathcal{D}_{RGB}$, and the third term denotes $\mathcal{D}$.

$$\mathcal{L} = \alpha\mathcal{L}_{PC} + \beta\mathcal{L}_{RGB} + (1-\alpha-\beta)\,\mathcal{L}_{F} \tag{6}$$

The weights α and β are experimentally determined coefficients used to balance the contributions of various components ($\mathcal{L}_{PC}$, $\mathcal{L}_{RGB}$, $\mathcal{L}_{F}$) within the weighted loss function. These components reflect distinct aspects of the triplet margin-based loss function, defined as follows:

$$\mathcal{L}\left(a_i, p_i, n_i\right) = \max\left\{0, m - d\left(a_i, n_i\right) + d\left(a_i, p_i\right)\right\} \tag{7}$$

Here, a_i , p_i and n_i are descriptors for the anchor, positive example, and negative example, respectively, in the i-th triplet. m is a margin value used to define distance intervals in the triplet loss function. Within each batch, $\mathcal{L}_{PC}$ denotes the loss computed from triplets constructed using $\mathcal{D}_{PC}$; $\mathcal{L}_{RGB}$ signifies the loss computed from triplets constructed using $\mathcal{D}_{RGB}$; and $\mathcal{L}_{F}$ represents the loss computed from triplets constructed using $\mathcal{D}$.

4.5. Implementation Details

The preprocessing of LiDAR point cloud data involves the removal of the ground plane, which is considered irrelevant for our analysis. Subsequently, the point cloud undergoes downsampling to 4096 points via a voxel grid filter, followed by normalization to the range of $[-1, 1]$. The quantization step size for the 3D coordinates is set to 0.01. For RGB image data, we adopt data augmentation techniques such as color jitter and random erasing. $\mathcal{D}_{PC}$ and $\mathcal{D}_{RGB}$ are dimensionally set to $L = 128$. Consequently, $\mathcal{D}$ achieves the final dimensionality of $M = L_{\mathcal{P}} + L_{\mathcal{I}} = 256$.

Our network is implemented using PyTorch and trained on a single Nvidia RTX 3090 GPU and 64 GB RAM. Training spans 60 epochs, during which the learning rate undergoes reduction by a factor of 10 after 40 epochs. Evaluation of performance occurs every 10 epochs. The loss term coefficients in Eq.6 are $\alpha = 0.5$ and $\beta = 0.0$. The margin m in Eq.7 is set to $m = 0.2$.

5. Experiments

This section introduces the dataset and evaluation methods utilized, providing a comprehensive comparison of our method against SOTA PR techniques, including NetVLAD [2], PointNetVLAD [35], PIC-Net [24], CORAL [31], MinkLoc++ [17], among others. Additionally, we present findings from ablation experiments conducted to assess the individual contributions of various components within our proposed approach. Notably, all reported results for our method adhere to the architecture illustrated in Fig.2, ensuring consistency and transparency in our evaluation framework.

5.1. Datasets and Evaluation Methodology

Our proposed method is trained and evaluated on the Oxford RobotCar [30] and NCLT [5] datasets, and is tested directly on the KITTI [13] and MUN-FRL [34] datasets. These datasets are renowned benchmarks in the field of PR. These datasets offer diverse and challenging real-world environments, enabling comprehensive evaluation of our method's performance across varied scenarios.

Oxford RobotCar dataset. This dataset captures the central region of Oxford, UK. We utilize data processed by PointNetVLAD [35]. $\mathcal{P}_k$ is generated from continuous scans of the SICK LMS-151 2D LiDAR, covering a consecutive $20m$ range. Corresponding $\mathcal{I}_k$ is matched to each frame of $\mathcal{P}_k$ using timestamps. For each training instance, $\mathcal{I}_k$ is randomly sampled from the closest 15 images based on timestamps. During evaluation, only the closest $\mathcal{I}$ based on timestamps is utilized. Four randomly selected $150m \times 150m$ regions are designated for testing purposes, while the remainder serve as training areas.

NCLT dataset. This dataset encompasses routes within the North Campus of the University of Michigan. Notably, this dataset includes both indoor and outdoor scenes, presenting a more challenging and varied set of perspectives. $\mathcal{P}_k$ is generated from scans of the Velodyne HDL-32E 3D LiDAR. $\mathcal{I}_k$ is matched to $\mathcal{P}_k$ using timestamps, ensuring temporal alignment between modalities. Sequences 2012-01-08 and 2012-02-05 are used for training and evaluation. Four $100m \times 100m$ regions are randomly selected as testing areas, with the remaining areas designated for training.

KITTI dataset. This dataset is developed jointly by the Karlsruhe Institute of Technology and Technische Universität Darmstadt, includes a wide variety of real-world urban driving scenarios such as city streets, country roads, highways, and residential areas. The point cloud data is captured using a Velodyne HDL-64E 3D LiDAR sensor and synchronized with RGB images. To further enhance data diversity, the KITTI dataset features driving data collected at different times and under various weather conditions. For our experiments, we select sequence 00 as the test data, using the first 170 seconds as the database, with the remaining data serving as queries.

MUN-FRL dataset. This dataset is a distinctive multi-sensor aerial dataset specifically designed for studying navigation missions in GNSS-denied environments. It is collected in flight by the Memorial University of Newfoundland team using a DJI-M600 hexacopter UAV and the National Research Council of Canada's Bell 412 Advanced Systems Research Aircraft (ASRA), covering distances ranging from 300 meters to 5 kilometers. This diversity offers a rich testing environment for evaluating the generalization capability of the multimodal place recognition network developed in this paper. In our experiments, we select the Bell412-6 sequence to evaluate the place recognition task, using the first 1500 frames as the database and the remaining frames as query data.

Evaluation Metrics. We evaluate the methods using the *recall@N* metric, which quantifies the percentage of queries where at least one true positive (i.e., true match) appears among the top-*N* retrievals during K-nearest neighbors (KNN) search. Average recall at 1 (AR@1) evaluates retrieval performance by comparing the first returned result of $\mathbf{q}$ with the true match. In addition to AR@1, which assesses the accuracy of the top-ranked retrieval, we also introduce average recall at 1% (AR@1%). AR@1% extends this evaluation to consider the first 1% of the results returned for $\mathbf{q}$. This metric provides a more nuanced evaluation by considering retrieval performance within a larger subset of $\mathbf{M}_{ref}$.

5.2. Evaluation Results

The PR results on the Oxford RobotCar dataset are summarized in Tab.1. Our multimodal descriptors are compared not only with single-modal descriptors, but also with

Table 1. Evaluation results of PR on the Oxford RobotCar dataset.

Methods	Modality[1]	AR@1	AR@1%
NetVLAD [2]	V	51.49	65.21
Mixvpr [1]	V	92.70	-
EigenPlaces [4]	V	94.10	97.80
CricaVPR [23]	V	95.20	98.60
VXP [19]	V	-	98.79
PointNetVLAD [35]	L	63.10	80.70
PCAN [40]	L	70.72	86.40
DH3D-4096 [9]	L	73.28	84.26
LPD-Net [21]	L	86.28	94.92
HiTPR [15]	L	86.63	93.71
SOE-NET [38]	L	89.28	96.43
SVT-Net [11]	L	93.90	98.00
MinkLoc3D [16]	L	94.80	98.50
LoGG3D-Net [36]	L	94.90	97.90
LCDNet [6]	L	95.30	98.40
CASSPR [37]	L	95.60	98.50
VXP [19]	L	-	98.84
BEVPlace [26]	L	96.50	99.00
Lip-loc [33]	V+L	68.50	72.00
VXP [19]	V+L	-	84.39/76.93
CORAL [31]	V+L	88.93	93.13
PIC-Net [24]	V+L	-	97.70
MinkLoc++ [17]	V+L	96.55	99.07
LoG-PR(our)	V+L	**96.91**	**99.20**

[1] V:Visual, L:LiDAR, V+L:Visual+LiDAR.

multimodal descriptors based on 3D point clouds and RGB images. Our method achieves the best performance with AR@1 and AR@1% of 96.91% and 99.20%, respectively. Among unimodal methods, vision-based approaches, while effective in simpler environments, struggle with complex 3D structures and geometric variations. LIDAR methods, though superior in 3D spatial perception, fall short in capturing fine details and textures due to the absence of visual information. In contrast, multimodal methods excel at integrating visual and LiDAR features. The incorporation of a cross-attention module allows the multimodal approach to better leverage the complementary strengths of each modality, significantly enhancing model performance in complex scenes. Fig.4 shows an example of a successful retrieval as well as a mismatch case. The visualization of the descriptors shows that the query descriptor is very similar to the matched database descriptor, with only slight differences highlighted by the boxed areas. In contrast, the unmatched descriptors may have some similar regions but still exhibit noticeable differences.

Table 2. Evaluation results of PR on the NCLT dataset.

Methods	AR@1	AR@5	AR@20
PointNetVLAD [35]	74.6	82.3	87.5
MinkLoc3D [16]	80.2	86.4	92.6
OverlapTransformer [29]	86.1	89.9	93.0
CIMV [10]	87.1	92.5	95.7
SeqNet [12]	88.9	93.3	96.0
SeqLPD [20]	87.3	92.8	95.2
SeqOT [27]	91.7	94.7	96.8
CVTNet [28]	93.2	94.6	95.7
LoG-PR(our)	**94.2**	**95.6**	**97.3**

To further validate our network's performance, we conduct training and evaluation on the NCLT dataset. The PR results on the NCLT dataset are presented in Tab.2, where our method is compared with SOTA approaches including PointNetVLAD, MinkLoc3D, OverlapTransformer[29], CIMV[10], SeqNet[12], SeqLPD[20], SeqOT[27], and CVTNet[28]. Top 1 recall (AR@1), Top 5 recall (AR@5), and Top 20 recall (AR@20) are used as evaluation metrics. Notably, the NCLT dataset presents more challenging viewpoints, incorporating both indoor and outdoor scenes compared to the Oxford RobotCar dataset. Therefore, our method performs relatively worse on the NCLT dataset in terms of AR@1 compared to the Oxford RobotCar dataset. However, it still outperforms SOTA methods. Fig.5 shows an example of a successful retrieval case.

5.3. Generalization Evaluation

To assess the generalization ability of our method, we evaluate the model trained on the Oxford RobotCar dataset on the KITTI [14] dataset. For this purpose, we construct $\mathbf{M}_{ref}$ using the first 170 seconds of data from sequence 00, reserving the remaining portion for $\mathbf{q}$. Our comparative analysis includes benchmarks such as PointNetVLAD, LPD-Net [21], CORAL, and MinkLoc++, as presented in Tab.3. Notably, our method demonstrate superior performance compared to the other methods, showcasing its robustness and effectiveness in cross-dataset scenarios. It is essential to highlight the significant differences between the two datasets: while the point cloud data in the KITTI dataset is derived from a single 360° scan of a 3D LiDAR, the point cloud data in the Oxford RobotCar dataset is generated from continuous scans of a 2D LiDAR within a consecutive $20m$ range. These distinctions underscore the challenges posed by domain shift and highlight the importance of evaluating model generalization across diverse datasets. Fig.6 shows an example of a successful retrieval case.

In addition, we evaluate models trained on the OxfordRobotCar dataset using the MUN-FRL dataset. Specifi-

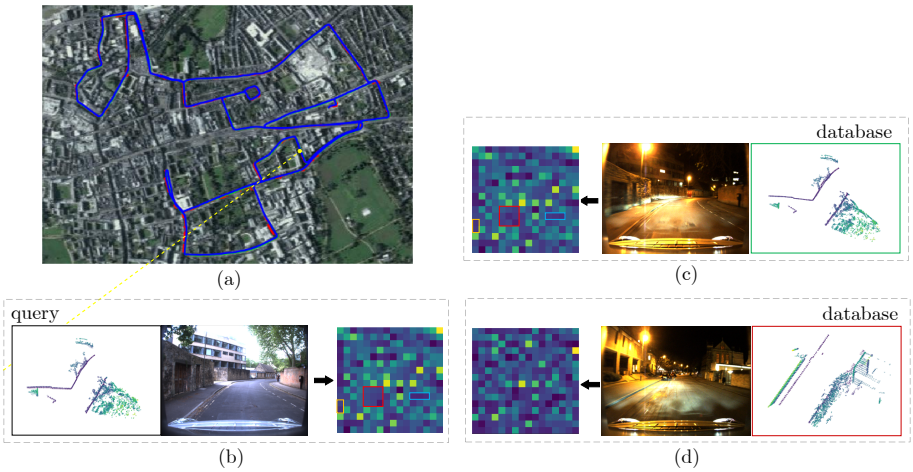

Figure 4. An example of a successful retrieval case at the yellow dot in the RobotCar dataset is depicted as follows: (a) shows the reference map, with the blue line indicating the trajectory; (b) displays $\mathbf{q}$ and its descriptor; (c) illustrates $\mathbf{m}^*$ and its descriptor; and (d) illustrates the mismatch and its descriptor.

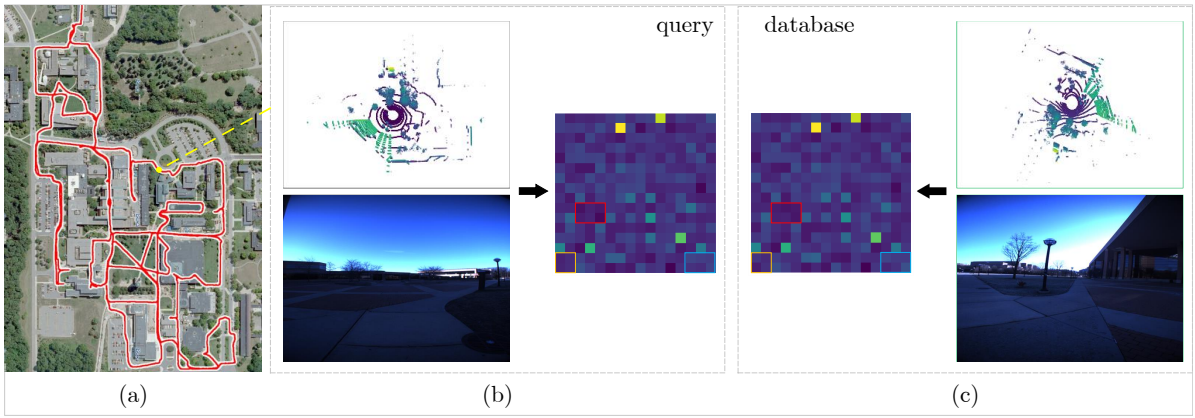

Figure 5. An example of a successful retrieval case at the yellow dot in the NCLT RobotCar dataset is depicted as follows: (a) shows the reference map, with the red line indicating the trajectory; (b) displays $\mathbf{q}$ and its descriptor; and (c) illustrates $\mathbf{m}^*$ and its descriptor.

Table 3. Evaluation results of PR on the KITTI dataset.

Methods	Modality	AR@1%
MinkLoc++ [17]	V	76.20
PointNetVLAD [35]	L	72.40
LPD-Net [21]	L	74.60
MinkLoc++ [17]	L	72.60
CORAL [31]	V+L	76.43
MinkLoc++ [17]	V+L	82.10
LoG-PR(our)	V+L	**82.49**

cally, we utilized the first 1500 frames from the Bell412-6 sequence to construct the database, with the remaining frames serving as query data. We compared our results with MinkLoc3D, CASSPR, and MinkLoc++, as shown in Tab.4. Our method consistently outperform these baselines. Notably, the MUN-FRL dataset is collected using the National Research Council (NRC) Bell412 Advanced Systems Research Aircraft (ASRA), while the OxfordRobotCar dataset is gathered from ground-based vehicles. This creates a substantial difference in perspective and scale between the two datasets, which inevitably impacts the test results. Nevertheless, our method remains superior in performance compared to the alternatives.

5.4. Ablation Study

Ablation studies are conducted to validate the effectiveness of incorporating the self-attention transformer (SAT) in the point cloud feature extraction stage (DPC) and the cross-attention transformer (CAT) in the fusion stage (DF). The tests are carried out using the Oxford RobotCar dataset.

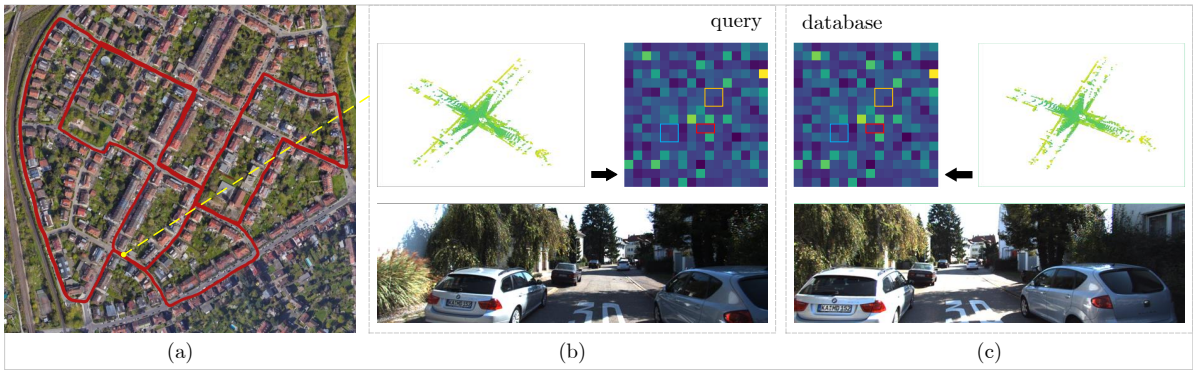

Figure 6. An example of a successful retrieval case at the yellow dot in the KITTI dataset is depicted as follows: (a) shows the reference map, with the red line indicating the trajectory; (b) displays $\mathbf{q}$ and its descriptor; and (c) illustrates $\mathbf{m}^*$ and its descriptor.

Table 4. Evaluation results of PR on the MUN-FRL dataset.

Network	AR@1%	AR@1	AR@5	AR@20
MinkLoc++	42.43	39.93	40.78	42.83
MinkLoc3D	62.67	61.83	62.23	62.96
CASSPR	65.49	62.01	63.81	66.37
LoG-PR(our)	**75.07**	**68.92**	**72.36**	**75.91**

In our analysis, DPC+DRGB+CAT indicates the absence of the self-attention transformer in the point cloud feature extraction stage, while DPC+DRGB+SAT indicates the absence of the cross-attention transformer, with features directly concatenated. The results, presented in Tab.5, highlight the impact of these modules on performance. With the addition of the self-attention transformer, our method demonstrates improvements of 0.29% on AR@1% and 1.15% on AR@1. Despite the performance of SOTA fusion methods nearing 100%, indicating limited room for improvement, our method still achieves increases of 0.02% on AR@1% and 0.16% on AR@1 with the addition of the cross-attention transformer.

We conduct a comprehensive evaluation to assess the effectiveness of our designed cross-attention transformer for feature fusion. This involves testing various methods, including utilizing different numbers of self-attention heads

Table 5. Ablation study on the self-attention transformer and the cross-attention transformer module.

Network	AR@1	AR@1%
DPC+DRGB+CAT	95.76	98.91
DPC+DRGB+SAT	96.75	99.18
DPC+DRGB+SAT+CAT	**96.91**	**99.20**

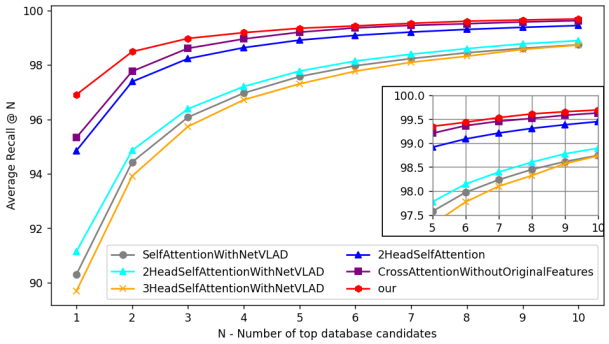

Figure 7. We evaluate the effectiveness of different feature fusion methods. The red hexagon represents the best result of our designed cross-attention transformer fusion module.

after concatenating point cloud features with image features, employing cross-attention without retaining the original features, and integrating NetVLAD feature aggregation. The results of these experiments are summarized in Fig.7. It is evident that retaining the original features proves effective for the cross-attention fusion. Among the methods employing different numbers of self-attention heads and NetVLAD feature aggregation, the approach without NetVLAD with 2 self-attention heads emerges as the top performer. This observation suggests that, while a lower number of self-attention heads may limit the model's ability to process and integrate input information, a higher number can increase model complexity, potentially leading to overfitting or higher computational costs. Optimal performance may be achieved when striking a balance between information richness and model complexity, as seen with the two attention heads configuration. Furthermore, the introduction of NetVLAD increases the model parameters, potentially leading to overfitting and consequently degrading model performance.

6. Conclusion

This paper introduces LoG-PR, a novel multimodal PR method that harnesses the complementary strengths of LiDAR point clouds and RGB images. We employ efficient 3D convolutional networks based on both point and sparse voxel to generate robust point cloud descriptors. Notably, the performance of our 3D modality is enhanced by integrating a self-attention transformer. Furthermore, we incorporate a cross-attention transformer to seamlessly fuse feature tensors extracted from image and point cloud, empowering us to effectively tackle challenging place retrieving tasks with efficiency. Our method is trained on the Oxford RobotCar and NCLT datasets and rigorously tested for generalization on the KITTI dataset. Experimental results underscore the superiority of our approach over existing techniques, reaffirming its efficacy and potential in advancing multimodal PR research. While our method demonstrates significant improvements in place recognition accuracy and robustness across various datasets, but the cross-attention transformer introduces additional computational. Future work could focus on optimizing the transformer architecture or exploring lightweight alternatives to further reduce inference time, making the method more suitable for resource-constrained platforms.

References

[1] A. Ali-Bey, B. Chaib-Draa, and P. Giguere. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2998–3007, 2023. 7

[2] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2, 6, 7

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008. 2

[4] G. Berton, G. Trivigno, B. Caputo, and C. Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11080–11090, 2023. 7

[5] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice. University of Michigan North Campus long-term vision and lidar dataset. *International Journal of Robotics Research*, 35(9):1023–1035, 2015. 6

[6] D. Cattaneo, M. Vaghi, and A. Valada. Lcdnet: Deep loop closure detection and point cloud registration for lidar slam. *IEEE Transactions on Robotics*, 38(4):2074–2093, 2022. 7

[7] Z. Chen, F. Maffra, I. Sa, and M. Chli. Only look once, mining distinctive landmarks from convnet for visual place recognition. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9–16. IEEE, 2017. 2

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 2

[9] J. Du, R. Wang, and D. Cremers. Dh3d: Deep hierarchical 3d descriptors for robust large-scale 6dof relocalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 744–762. Springer, 2020. 7

[10] J. M. Facil, D. Olid, L. Montesano, and J. Civera. Condition-invariant multi-view place recognition. *arXiv preprint arXiv:1902.09516*, 2019. 7

[11] Z. Fan, Z. Song, H. Liu, Z. Lu, J. He, and X. Du. Svt-net: Super light-weight sparse voxel transformer for large scale place recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 551–560, 2022. 7

[12] S. Garg and M. Milford. Seqnet: Learning descriptors for sequence-based hierarchical place recognition. *IEEE Robotics and Automation Letters*, 6(3):4305–4312, 2021. 7

[13] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 6

[14] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 7

[15] Z. Hou, Y. Yan, C. Xu, and H. Kong. Hitpr: Hierarchical transformer for place recognition in point cloud. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2612–2618. IEEE, 2022. 7

[16] J. Komorowski. Minkloc3d: Point cloud based large-scale place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1790–1799, 2021. 3, 7

[17] J. Komorowski, M. Wysoczańska, and T. Trzcinski. Minkloc++: lidar and monocular image fusion for place recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 2, 3, 5, 6, 7, 8

[18] H. Lai, P. Yin, and S. Scherer. Adafusion: Visual-lidar fusion with adaptive weights for place recognition. *IEEE Robotics and Automation Letters*, 7(4):12038–12045, 2022. 3

[19] Y.-J. Li, M. Gladkova, Y. Xia, R. Wang, and D. Cremers. Vxp: Voxel-cross-pixel large-scale image-lidar place recognition. *arXiv preprint arXiv:2403.14594*, 2024. 7

[20] Z. Liu, C. Suo, S. Zhou, F. Xu, H. Wei, W. Chen, H. Wang, X. Liang, and Y.-H. Liu. Seqlpd: Sequence matching enhanced loop-closure detection based on large-scale point cloud description for self-driving vehicles. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1218–1223. IEEE, 2019. 7

[21] Z. Liu, S. Zhou, C. Suo, P. Yin, W. Chen, H. Wang, H. Li, and Y.-H. Liu. Lpd-net: 3d point cloud learning for large-scale place recognition and environment analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2831–2840, 2019. 7, 8

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[23] F. Lu, X. Lan, L. Zhang, D. Jiang, Y. Wang, and C. Yuan. Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16772–16782, 2024. 7

[24] Y. Lu, F. Yang, F. Chen, and D. Xie. Pic-net: Point cloud and image collaboration network for large-scale place recognition. *arXiv preprint arXiv:2008.00658*, 2020. 3, 6, 7

[25] L. Lun, Z. Shuhang, L. Yixuan, F. Yongzhi, Y. Beinan, C. Siyuan, and S. Hui-Liang. BEVPlace: Learning LiDAR-based place recognition using bird's eye view images. *arXiv preprint arXiv:2302.14325*, 2023. 2

[26] L. Luo, S. Zheng, Y. Li, Y. Fan, B. Yu, S.-Y. Cao, J. Li, and H.-L. Shen. Bevplace: Learning lidar-based place recognition using bird's eye view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8700–8709, 2023. 7

[27] J. Ma, X. Chen, J. Xu, and G. Xiong. Seqot: A spatial–temporal transformer network for place recognition using sequential lidar data. *IEEE Transactions on Industrial Electronics*, 70(8):8225–8234, 2022. 7

[28] J. Ma, G. Xiong, J. Xu, and X. Chen. Cvtnet: A cross-view transformer network for lidar-based place recognition in autonomous driving environments. *IEEE Transactions on Industrial Informatics*, 2023. 7

[29] J. Ma, J. Zhang, J. Xu, R. Ai, W. Gu, and X. Chen. Overlaptransformer: An efficient and yaw-angle-invariant transformer network for lidar-based place recognition. *IEEE Robotics and Automation Letters*, 7(3):6958–6965, 2022. 3, 7

[30] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. 6

[31] Y. Pan, X. Xu, W. Li, Y. Cui, Y. Wang, and R. Xiong. Coral: Colored structural representation for bi-modal place recognition. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2084–2091. IEEE, 2021. 2, 6, 7, 8

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

[33] S. Shubodh, M. Omama, H. Zaidi, U. S. Parihar, and M. Krishna. Lip-loc: Lidar image pretraining for cross-modal localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 948–957, 2024. 7

[34] R. G. Thalagala, O. De Silva, A. Jayasiri, A. Gubbels, G. K. Mann, and R. G. Gosine. Mun-frl: A visual-inertial-lidar dataset for aerial autonomous navigation and mapping. *The International Journal of Robotics Research*, page 02783649241238358, 2024. 6

[35] M. A. Uy and G. H. Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4470–4479, 2018. 3, 6, 7, 8

[36] K. Vidanapathirana, M. Ramezani, P. Moghadam, S. Sridharan, and C. Fookes. Logg3d-net: Locally guided global descriptor learning for 3d place recognition. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2215–2221. IEEE, 2022. 7

[37] Y. Xia, M. Gladkova, R. Wang, Q. Li, U. Stilla, J. F. Henriques, and D. Cremers. Casspr: Cross attention single scan place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8461–8472, 2023. 7

[38] Y. Xia, Y. Xu, S. Li, R. Wang, J. Du, D. Cremers, and U. Stilla. Soe-net: A self-attention and orientation encoding network for point cloud based place recognition. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11348–11357, 2021. 7

[39] S. Xie, C. Pan, Y. Peng, K. Liu, and S. Ying. Large-scale place recognition based on camera-lidar fused descriptor. *Sensors*, 20(10):2870, 2020. 2

[40] W. Zhang and C. Xiao. 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*, pages 12436–12445, 2019. 7

[41] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 4

[42] Z. Zhou, J. Xu, G. Xiong, and J. Ma. Lcpr: A multi-scale attention-based lidar-camera fusion network for place recognition. *IEEE Robotics and Automation Letters*, 2023. 3