ImLPR: Image-based LiDAR Place Recognition using Vision Foundation Models

Minwoo Jung
Department of Mechanical Engineering
Seoul National University, South Korea
moonshot@snu.ac.kr

Lanke Frank Tarimo Fu
Department of Engineering Science
University of Oxford, United Kingdom
fu@robots.ox.ac.uk

Maurice Fallon
Department of Engineering Science
University of Oxford, United Kingdom
mfallon@robots.ox.ac.uk

Ayoung Kim*
Department of Mechanical Engineering
Seoul National University, South Korea
ayoungk@snu.ac.kr

Abstract: LiDAR Place Recognition (LPR) is a key component in robotic localization, enabling robots to align current scans with prior maps of their environment. While Visual Place Recognition (VPR) has embraced Vision Foundation Models (VFMs) to enhance descriptor robustness, LPR has relied on task-specific models with limited use of pre-trained foundation-level knowledge. This is due to the lack of 3D foundation models and the challenges of using VFM with LiDAR point clouds. To tackle this, we introduce ImLPR, a novel pipeline that employs a pre-trained DINOv2 VFM to generate rich descriptors for LPR. To the best of our knowledge, ImLPR is the first method to utilize a VFM for LPR while retaining the majority of pre-trained knowledge. ImLPR converts raw point clouds into novel three-channel Range Image Views (RIV) to leverage VFM in the LiDAR domain. It employs MultiConv adapters and Patch-InfoNCE loss for effective feature learning. We validate ImLPR on public datasets and outperform state-of-the-art (SOTA) methods across multiple evaluation metrics in both intra- and inter-session LPR. Comprehensive ablations on key design choices such as channel composition, RIV, adapters, and the patch-level loss quantify each component's impact. We release ImLPR as open source for the robotics community.

Keywords: LiDAR Place Recognition, Deep Learning, Vision Foundation Model

1 Introduction

Place recognition, encompassing both VPR and LPR, provides an initial localization estimate when seeking to determine if a location has been previously visited, using a prior map of the location. Early learning-based methods relied on specialized architectures trained on domain-specific datasets [1–6]. While VPR has advanced by using VFMs [7–10], pre-trained on large-scale datasets [11–13], LPR has been limited by the lack of suitable 3D foundation models and the modality gap when adapting VFMs to LiDAR data.

Efforts applying foundation models on LiDAR data can be characterized in two ways: developing full 3D foundation models or converting LiDAR data into a format that can be used with VFM. Existing 3D foundation models [14, 15], designed for object detection or indoor scene analysis, are unsuitable for LPR due to their focus on specific tasks and the heavy computation required when working with 3D data with large sets of parameters. On the other hand, converting 3D point clouds into 2D images is a more straightforward approach for leveraging VFM for LPR. Two types of image projection are typically used to create a 2D representation from a 3D point cloud, Bird's-Eye-View (BEV) and RIV, as shown in Fig. 1. However, a substantial domain gap exists between

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

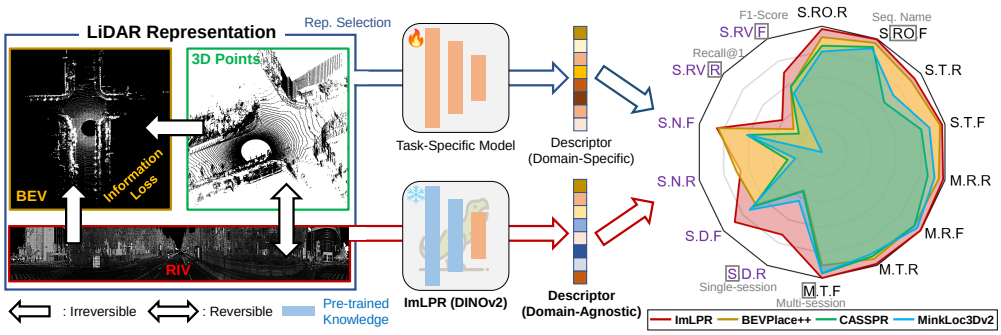


Figure 1: Without using a foundation model, traditional LPR relies on domain-specific training with 3D point clouds, BEV, or RIV. To leverage a VFM, ImLPR uses RIV images to capture geometric information while avoiding the information loss that the BEV projection induces. The radar chart (right) highlights ImLPR's superior performance in trained (black) and unseen (purple) domains.

natural images used to pre-trained VFMs and projected LiDAR images, particularly when using single-channel [6] or overly large multi-channel inputs [16]. Previous approaches that rely on such representations struggle to effectively leverage pre-trained three-channel vision models.

To tackle these issues, we present ImLPR, a novel pipeline for LPR that adapts DINOv2 [12], a SOTA VFM with robust and transferable feature extraction capabilities. We have designed our system to around a RIV which encodes point clouds using reflectivity, range, and normal ratio. This allows RGB-pretrained VFM to extract discriminative LiDAR features Our empirical results show that each proposed channel has a significant impact on LPR performance, highlighting the importance of channel design. Notably, the three-channel configuration proves more effective for RIV than for BEV. Furthermore, by inserting and training lightweight adapters, we preserve most of the pre-trained DINOv2 weights while adapting it effectively for LPR using RIV input images. Additionally, we propose the Patch-InfoNCE loss, a patch-level contrastive loss, enhancing discriminability and robustness by ensuring consistent feature representations across corresponding RIV patches. Combining LiDAR's geometric precision and semantic information with DINOv2's representational power, ImLPR surpasses SOTA LPR methods. Our contributions include:

- ImLPR is the first LPR pipeline using a VFM while retaining the majority of pre-trained knowledge. Our key innovation lies in a tailored three-channel RIV representation and lightweight convolutional adapters, which seamlessly bridge the 3D LiDAR and 2D vision domain gap. Freezing most DINOv2 layers preserves pre-trained knowledge during training, ensuring strong generalization and outperforming task-specific LPR networks.
- We introduce the Patch-InfoNCE loss, a patch-level contrastive loss, to enhance the local discriminability and robustness of learned LiDAR features. We demonstrate that our patch-level contrastive learning strategy achieves a performance boost in LPR.
- ImLPR demonstrates versatility on multiple public datasets, outperforming SOTA methods. Furthermore, we also validate the importance of each component of the ImLPR pipeline.

2 Related Work

2.1 Traditional LiDAR Place Recognition

Numerous works have studied LPR. Early methods like PointNetVLAD [17] used PointNet for feature extraction, while others [18, 19] refined descriptors with MLPs. Efficiency-driven approaches, such as OverlapTransformer [20] and CVTNet [21], projected point clouds into range images for 2D convolutions. Sparse convolution methods [3, 4] optimized 3D processing, while BEV projections [22] have been used to capture the global contour of a scene. Recent efforts, like SE(3)-equivariant networks [23], leverage rotation-invariant features, and SeqMatchNet [24] uses contrastive learning with sequence matching for robustness. However, these methods, trained on specific datasets, struggle to adapt to scene variations and exhibit reduced performance when tested in

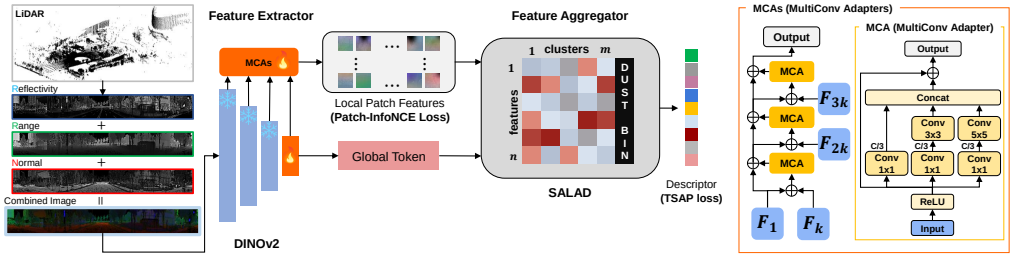


Figure 2: The point cloud is projected into a RIV image containing reflectivity, range, and normal channels. A pre-trained DINOv2 model, adapted via MultiConv adapters (MCAs), extracts rich patch-level features. Patch-InfoNCE loss enhances local feature discriminability, while SALAD aggregates features into a global descriptor. This effectively leverages the VFM for use with LiDAR.

different domains. Their domain-specific descriptors struggle with diverse outdoor conditions, hindering broad applicability. Inspired by the progress of VPR [7, 8], we address these limitations by introducing VFMs to the task of LPR, leveraging the rich, robust descriptors of foundation models to overcome the traditional domain-specific constraints described above.

2.2 Vision and 3D Foundation Models for 3D Data

VFMs [11–13] excel in VPR due to their robust feature representations [7, 9]. In particular, DINOv2 was pre-trained on massive datasets containing 140 million images. In contrast, 3D foundation models like PTv3 [14] and Sonata [15] are trained on smaller datasets with 23-139 thousand point clouds, and focus on object-level tasks, limiting their generality for outdoor LPR. Recently, researchers have attempted to apply VFM to point clouds [25, 26]. However, they focus on object detection or registration, and still rely on visual images, diverging from LPR. Instead, to adapt LiDAR data for use with image-based networks, traditional methods project point clouds into 2D formats like BEV [22] or RIV [16], and then use these images with ImageNet-pretrained CNNs such as ResNet [27]. However, compared to VFM, these methods lack the expressiveness needed for diverse outdoor scenes and are limited by their small training datasets and simple architectures. Additionally, domain differences between LiDAR images and visual images hinder the direct application of VFMs. Fine-tuning the entire model such as LIP-Loc [28], can lead to catastrophic forgetting, reducing generalization by forgetting pre-trained knowledge. To address these challenges, we utilize DINOv2 with a MultiConv adapter to bridge domain differences while preserving almost the entire pre-trained knowledge within the VFM network. We adopt the RIV projection, which captures denser geometric detail and is more similar to normal visual images compared to the sparser BEV images [6, 22].

3 Methodology

3.1 Overview of ImLPR

ImLPR first renders the scan as an RGB-style image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and then computes a global representation $\mathbf{g} \in \mathbb{R}^t$ through the mapping function $\Omega = h \circ f$. The VFM-based encoder $f : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H'W'\times C}$ ($H' = \lfloor H/14 \rfloor$, $W' = \lfloor W/14 \rfloor$) extracts patch features, which are then pooled into the descriptor $\mathbf{g}$ by the aggregator $h : \mathbb{R}^{H'W'\times C} \to \mathbb{R}^t$. Metric learning optimizes Ω such that, if the spatial distance $\mathcal{D}(q,i) < \mathcal{D}(q,j)$, then the descriptor distance $d_g(q,i) < d_g(q,j)$, enforcing this relationship at both the feature and global descriptor levels.

3.2 Input Data Processing

To convert raw LiDAR point clouds into structured images, we map each point $p_i = (x_i, y_i, z_i) \in \mathbf{P}$ to RIV pixel coordinates (u_i, v_i) via:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} 0.5\,[\,1 - \arctan(y_i, x_i)/\pi\,] \cdot W \\ [\,1 - (\arcsin(z_i/r_i) + f_{\text{up}})/f\,] \cdot H \end{pmatrix}, \tag{1}$$

where W and H denote the image width and height, r_i is the range from the origin to p_i, f_{up} is the upper bound of the vertical field of view (FOV), and f denotes the total vertical FOV. To further enrich the image with geometric information, we include the three LiDAR channels—reflectivity, range, and normal ratio—in each pixel. Reflectivity offers semantic distinction between objects with different surface properties, while range encodes geometric features of the scene, such as object distance and depth variation. The normal ratio is derived from the singular value ratio of the covariance matrix for the k-nearest neighboring points of p_i. Singular Value Decomposition (SVD) is applied to this covariance matrix, and the logarithmic ratio of the largest to smallest singular values encodes the normal information. This scalar effectively summarizes local geometric variations—such as surface planarity and edge-like structures—into a single channel, offering a computationally efficient alternative to multi-channel normal vectors [16]. This multi-channel image is used as input for ImLPR, as depicted in Fig. 2.

3.3 Feature Extraction with Adapted DINO ViT

We adopt a DINO ViT-S/14 model pre-trained on RGB images and tailor it to LiDAR RIV images by fine-tuning only the final two transformer blocks of the model. Similar to SelaVPR++ [29], we bridge the gap between the DINO pre-training RGB dataset and new LiDAR data by postprocessing the per-patch DINO features using lightweight MultiConv adapter layers. Concretely, we insert these adapters at regular intervals among the frozen layers, allowing them to refine intermediate patch representations while keeping most of the network frozen. This design preserves the rich pre-trained capabilities of the model while facilitating LiDAR-specific adaptation with lower computational overhead compared to full end-to-end fine-tuning. Since MultiConv adapters are placed at k regular intervals, this further optimizes memory and computation. Formally, let x_l represent the intermediate DINO feature of the l-th transformer block, comprising both patch features x_l^{patch} and token features x_l^{token}. We define the adapter-refined patch features y_i as:

$$y_i = \begin{cases} \text{Adapter}\big(x_i^{\text{patch}} + x_{ik}^{\text{patch}}\big) + x_i^{\text{patch}}, & i = 1 \\ \text{Adapter}\big(y_{i-1}^{\text{patch}} + x_{ik}^{\text{patch}}\big) + y_{i-1}^{\text{patch}}, & 2 \le i \le \lfloor L/k \rfloor \end{cases} \tag{2}$$

where L is the total number of transformer blocks. The token features x_l^{token} remain untouched. By focusing on the refined patch features, our method efficiently leverages strong generalizable visual representations from the pre-trained model while accommodating LiDAR-specific structures.

3.4 Feature Aggregation with Optimal Transport

We aggregate the refined patch features into a global representation using an optimal transport approach similar to SALAD [8]. Let $\mathbf{F} \in \mathbb{R}^{H'W' \times C}$ be the local features derived from the adapter-refined patch features. A convolutional layer first maps $\mathbf{F}$ to a score matrix $\mathbf{S} \in \mathbb{R}^{n \times m}$, representing the assignment probabilities of local features to m learnable cluster centers. We then apply optimal transport, specifically the Sinkhorn algorithm [30], to obtain an optimized assignment matrix $\mathbf{R} \in \mathbb{R}^{n \times m}$ by iteratively normalizing the rows and columns of the exponentiated score matrix. The aggregated feature $\mathbf{V} \in \mathbb{R}^{m \times l}$ is computed as $V_{j,k} = \sum_{i=1}^{n} R_{i,j}\, \bar{F}_{i,k}$, where $\bar{\mathbf{F}}$ denotes the intermediate feature embeddings obtained by applying a convolution layer to the original feature $\mathbf{F}$.

To capture global structural context, we generate a compact global embedding $\mathbf{G} \in \mathbb{R}^e$ via linear layers. The global token passes through two linear layers to compress and capture global context. The final descriptor combines flattened local features and the global embedding as $g = [\mathbf{V}.flatten(), \mathbf{G}] \in \mathbb{R}^{m \times l + e}$. This fusion of clustering-based local features and global embedding captures both detailed local patterns and robust global structure.

3.5 Patch-level Contrastive Learning

To encourage learning both locally (patch feature-level) and globally (descriptor-level), we introduce a patch-level contrastive loss in addition to the global objective. This promotes effective gradient

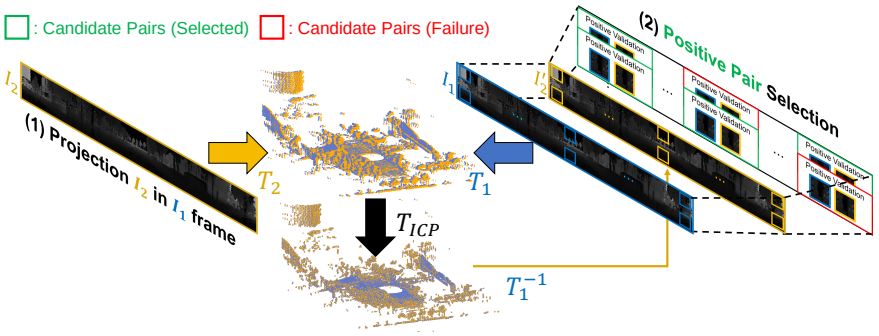


Figure 3: Patch correspondence pipeline aligns two RIV images by transforming their point clouds to a shared coordinate system using known poses and Iterative Closest Point (ICP). Positive patch pairs are validated based on spatial overlap and range similarity, while negatives ensure large spatial separation. This process supplies pairs for Patch-InfoNCE loss to train discriminative local features.

flow throughout the network. For patch-level supervision, we mine positive and negative patch pairs from two RIV range images based on geometric consistency.

Positive Patch Mining: We transform both range images into 3D point clouds using their RIV geometry and align them in a common world frame using ground truth poses refined by ICP. Points from the second scan are reprojected onto the first RIV to identify candidate patch correspondences, as illustrated in Fig. 3. Positive pairs are selected based on pixel overlap and range similarity. A patch pair (p_1, p_2) is considered positive if the ratio of valid overlapping pixels $\mathcal{P}_{\text{ov}}$, containing range values in both patches, exceeds $\rho_{\text{valid}} = 0.5$. For all pixels in $\mathcal{P}_{\text{ov}}$, we compute the per-pixel range differences $d_i = r_{1,i} - r_{2,i}$. An adaptive threshold τ_r is computed using the median absolute deviation (MAD): $\tau_r = \text{median}(d_i) + k \cdot \text{MAD}$, ensuring robustness to measurement noise. A pair is confirmed positive if the mean range difference $\Delta_r = \frac{1}{|\mathcal{P}_{\text{ov}}|} \sum_{i \in \mathcal{P}_{\text{ov}}} |r_{1,i} - r_{2,i}|$ is less than τ_r. This ensures that spatially consistent, range-coherent positive patch pairs are selected.

Negative Patch Mining: To select negative patch pairs, we enforce large vertical and horizontal distances (v_{dist} and h_{dist}) from the positive patches to avoid spatial overlap. The cylindrical geometry is considered, ensuring that patches near the 0° and 360° boundaries are treated as spatially distant.

To supervise patch-level contrastive learning, we propose Patch-InfoNCE, an extension of the InfoNCE loss [13] for local features. Unlike global descriptor losses, Patch-InfoNCE operates on local embeddings from features $\mathbf{F_1}, \mathbf{F_2} \in \mathbb{R}^{H'W' \times C}$ of two adjacent RIV images. It supports multiple positive patch pairs (p_1, p_2) per image and computes their cosine similarity. For each positive pair $F_1^{p_1}$ and $F_2^{p_2}$, we select hard negatives: n_1 from $\mathbf{F_1}$ relative to p_2, and n_2 from $\mathbf{F_2}$ relative to p_1. The Patch-InfoNCE loss is defined as:

$$\mathcal{L}_P = -\frac{1}{|\mathcal{P}|} \sum_{(p_1,p_2)\in\mathcal{P}} \log\left(\frac{\exp(s_{\text{p}}/\tau_l)}{\exp(s_{\text{p}}/\tau_l) + \sum_j \exp(s_{\text{n},j}/\tau_l)}\right), \tag{3}$$

where $\mathcal{P}$ is the set of positive pairs, τ_l is the temperature, and s_{p}, $s_{\text{n},j}$ denote the cosine similarities of positive and negative pairs, respectively. This loss encourages consistent local features at corresponding spatial locations across RIV images, enabling discriminative patch-level embeddings for LPR. The final objective combines both local and global terms as $\mathcal{L}_{\text{final}} = \mathcal{L}_P + \lambda \mathcal{L}_{\text{TSAP}}$, where λ balances their ratio. Details of $\mathcal{L}_{\text{TSAP}}$ [4] are provided in the appendix. This dual-loss formulation captures both fine-grained context and global semantics, improving overall LPR performance.

4 Experiment

4.1 Experimental Setup

Implementation Details: ImLPR was implemented in PyTorch using a DINO ViT-S/14, training only the last two transformer blocks and inserting MultiConv adapters every three blocks. The size of the RIV image used is $H \times W = 126 \times 1022$. Additional details are available in the appendices.

Table 1: Performance Comparison in Intra-Session Place Recognition

Method	Round01-O		Round02-O		Round03-O		Town01-O		Town02-O		Town03-O		Average	
	R@1	F1	R@1	F1	R@1	F1	R@1	F1	R@1	F1	R@1	F1	R@1	F1
LoGG3D-Net	0.425	0.740	0.257	0.475	0.542	0.806	0.101	0.274	0.135	0.256	0.151	0.299	0.269	0.475
MinkLoc3d v2	0.933	0.968	0.856	0.937	0.908	0.959	0.847	0.925	0.834	0.922	0.907	0.957	0.881	0.945
CASSPR	0.940	0.969	0.883	0.943	0.942	0.972	0.803	0.892	0.789	0.882	0.862	0.935	0.870	0.932
BEVPlace++	0.975	0.988	0.926	0.962	0.962	0.983	0.946	0.975	0.942	0.973	0.961	0.980	0.952	0.977
ImLPR	**0.992**	**0.996**	**0.974**	**0.988**	**0.992**	**0.997**	**0.981**	**0.990**	**0.950**	**0.977**	**0.966**	**0.984**	**0.976**	**0.989**

(Round: Roundabout, **Bold**: Best and Underline: Second-Best

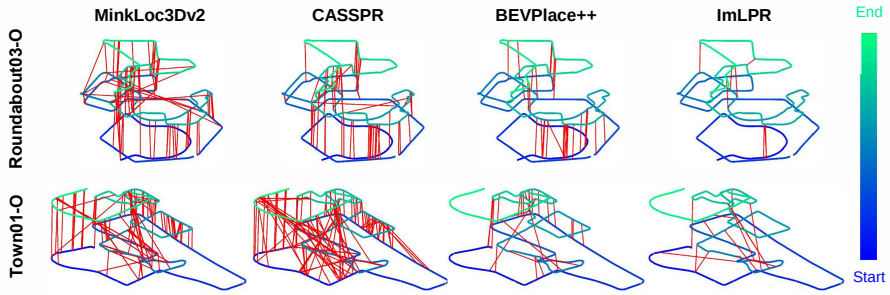


Figure 4: Trajectory from two sequences, with red lines marking false positives (fewer red lines indicate better performance). It highlights ImLPR having fewer errors compared to SOTA methods.

Datasets & Evaluation Metrics: ImLPR is evaluated on HeLiPR (O: OS2-128 and V: VLP-16C) [31], MulRan (OS1-64) [32], and NCLT (HDL-32E) [33] datasets, with scans sampled at 3m intervals. From HeLiPR, we use sequences Roundabout01-03 and Town01-03, where the suffix O or V indicates the LiDAR used. From MulRan, we use the DCC01-03 sequences for evaluation. We use nine HeLiPR sequences with the Ouster for training: DCC04-06, KAIST04-06, and Riverside04-06. To handle varying point cloud density, 3D sparse convolution methods downsample to 8192 points, while other parameters retain default settings. ImLPR is compared against LoGG3D-Net [3], MinkLoc3Dv2 [4], CASSPR [5], and BEVPlace++ [6], using Recall@1 (R@1) and maximum F1 score. All methods are trained under the same conditions. Positive pairs are defined as matches within 10 m, assuming a four-lane highway. During training, negative pairs are sampled beyond 30 m, while during evaluation, pairs outside the 10 m range are treated as negatives.

4.2 Intra-session Place Recognition

Scans within 60 seconds of the query are excluded to prevent intra-session matching (i.e., matching within the same trajectory), and processing begins only after 90 seconds to ensure a sufficient database. As shown in Table. 1 and Fig. 4, ImLPR surpasses all baselines on Roundabout-O and Town-O. LoGG3D-Net exhibits lower recall due to the domain shift between training and testing. BEVPlace++ ranks second, capturing scan contours but underperforms ImLPR in narrow alleys and small-scale environments, as ImLPR leverages RIV representations for fine-grained feature extraction. 3D sparse convolution methods like MinkLoc3Dv2 and CASSPR rank third, limited by the absence of large-scale pre-trained models. ImLPR achieves the highest R@1 and F1 score, demonstrating robust and discriminative intra-session LPR performance.

4.3 Inter-session Place Recognition

Table. 2 and Table. 3 show inter-session results (i.e., between scans collected at different times) for Roundabout-O and Town-O. Each result is shown in the format Database-Query. ImLPR outperforms other methods across both environments, adapting to session-induced variations. BEVPlace++ achieves the second-highest Recall@1, but its lower F1 score indicates reduced precision. We believe this is due to the simple network consisting of 2D convolution without additional mechanisms such as attention. Furthermore, its BEV representation, projecting all data into a single plane, limits its ability to distinguish subtle session-specific variations. As shown in Fig. 5, BEVPlace++ has a lower F1-Recall curve than MinkLoc3Dv2 and CASSPR, despite higher recall. Since the F1-Recall curve captures the trade-off between precision and recall across thresholds, a lower curve suggests a higher false positive rate. MinkLoc3Dv2 and CASSPR follow ImLPR in F1 score with

Table 2: Performance Comparison in Inter-Session Place Recognition (HeLiPR Roundabout-0)

Method	Round01-02 R@1	Round01-02 F1	Round01-03 R@1	Round01-03 F1	Round02-01 R@1	Round02-01 F1	Round02-03 R@1	Round02-03 F1	Round03-01 R@1	Round03-01 F1	Round03-02 R@1	Round03-02 F1	Average R@1	Average F1
LoGG3D-Net	0.610	0.847	0.610	0.847	0.577	0.801	0.637	0.844	0.617	0.855	0.681	0.859	0.622	0.842
MinkLoc3Dv2	0.930	0.970	0.955	0.981	0.931	0.968	0.954	0.979	0.965	0.985	0.952	0.981	0.948	0.977
CASSPR	0.939	0.970	0.961	0.981	0.945	0.972	0.968	0.985	0.962	0.982	0.964	0.986	0.957	0.979
BEVPlace++	0.963	0.918	0.970	0.947	0.956	0.948	0.973	0.963	0.979	0.964	0.970	0.949	0.969	0.948
ImLPR	**0.984**	**0.992**	**0.991**	**0.996**	**0.989**	**0.995**	**0.991**	**0.996**	**0.994**	**0.997**	**0.992**	**0.997**	**0.990**	**0.996**

Table 3: Performance Comparison in Inter-Session Place Recognition (HeLiPR Town-0)

Method	Town01-02 R@1	Town01-02 F1	Town01-03 R@1	Town01-03 F1	Town02-01 R@1	Town02-01 F1	Town02-03 R@1	Town02-03 F1	Town03-01 R@1	Town03-01 F1	Town03-02 R@1	Town03-02 F1	Average R@1	Average F1
LoGG3D-Net	0.282	0.546	0.281	0.478	0.280	0.538	0.332	0.568	0.272	0.482	0.330	0.587	0.296	0.533
MinkLoc3Dv2	0.948	0.975	0.962	0.981	0.960	0.980	0.958	0.980	0.950	0.976	0.960	0.980	0.956	0.979
CASSPR	0.916	0.958	0.912	0.957	0.939	0.970	0.947	0.975	0.924	0.963	0.934	0.967	0.929	0.965
BEVPlace++	0.960	0.954	0.966	0.960	0.974	0.953	0.983	0.972	0.971	0.972	0.982	0.989	0.973	0.967
ImLPR	**0.989**	**0.995**	**0.991**	**0.996**	**0.988**	**0.996**	**0.988**	**0.994**	**0.985**	**0.993**	**0.990**	**0.995**	**0.989**	**0.995**

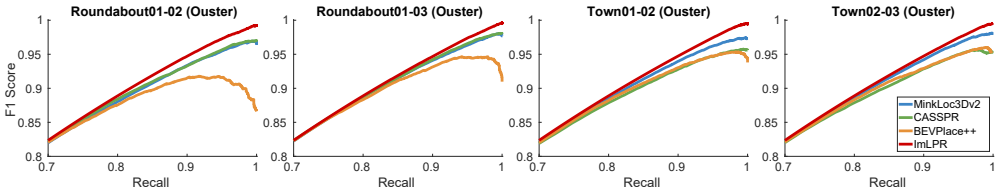


Figure 5: F1-Recall curve for inter-session place recognition. ImLPR consistently outperforms all baselines across all sequence pairs. BEVPlace++ exhibits a low F1 hindered by the projection-induced information loss. This highlights ImLPR's superior balance of precision and recall.

inconsistent results across sequences. This variability highlights the benefit of using large-scale pretrained models for consistent results. LoGG3D-Net ranks lowest due to domain mismatches. ImLPR's RIV representations precisely capture fine-grained details as shown in Fig. 6(b), and its network design ensures top inter-session LPR performance.

Table 4: Generalization Assessment

Method	DCC AR@1	DCC AF1	NCLT AR@1	NCLT AF1	Roundabout-V AR@1	Roundabout-V AF1	Average AR@1	Average AF1
LoGG3D-Net	0.094	0.214	0.149	0.434	0.131	0.394	0.125	0.347
MinkLoc3Dv2	0.722	0.870	0.622	0.808	0.513	0.764	0.619	0.814
CASSPR	0.683	0.851	0.652	0.810	0.630	0.793	0.655	0.818
BEVPlace++	0.678	0.839	**0.850**	0.922	0.655	0.790	0.728	0.850
ImLPR	**0.865**	**0.943**	0.834	**0.927**	**0.712**	**0.852**	**0.804**	**0.907**

AR@1: Average Recall@1, AF1: Average F1 score

Table 5: Performance Impact of RIV Image Channels

Method	CH1	CH2	CH3	Roundabout-0 AR@1	Roundabout-0 AF1	Town-0 AR@1	Town-0 AF1	DCC AR@1	DCC AF1	Average AR@1	Average AF1
ExpA-1	✔			0.946	0.973	0.956	0.979	0.773	0.884	0.892	0.945
ExpA-2	✔	✔		0.975	0.989	0.980	0.990	0.901	0.951	0.952	0.977
ExpA-3	✔	✔	✔	**0.979**	**0.990**	**0.985**	**0.993**	**0.945**	**0.970**	**0.970**	**0.984**

CH1: Reflectivity, CH2: Range, CH3: Normal Ratio

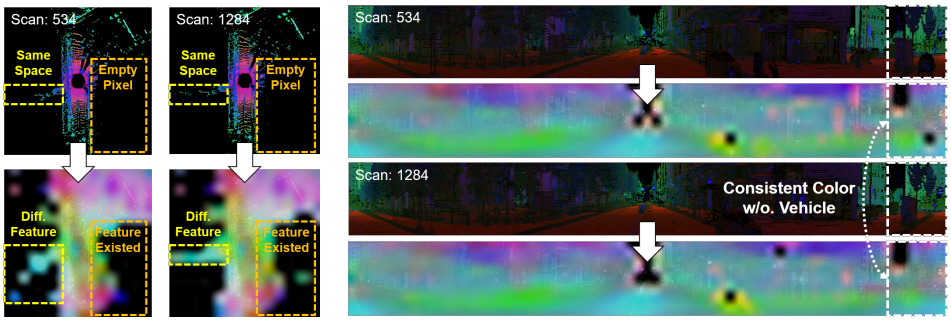
4.4 Ablation Studies

Model Generalization Assessment: To assess the generality of ImLPR, we conducted intra-session LPR on MulRan DCC, NCLT (2012-01-08, 01-15, 01-22), and HeLiPR Roundabout-V without additional training. As these datasets lack a reflectivity channel, normalized intensity was used. For NCLT and Roundabout-V, all methods were validated using accumulated scans as both query and database. The average performance across three sequences is reported in Table. 4, with dataset-specific details provided in the appendix. In DCC, while all methods experienced performance declines, ImLPR demonstrated the most robust performance, excelling in handling sensor differences and temporal shifts. In NCLT and Roundabout-V, it effectively generalized to unseen sensors and environments. Although discrete and noisy intensity from NCLT degraded the semantic quality of RIV images (Fig. 8(b)), ImLPR still attained the highest F1 score by leveraging geometric cues. Unlike competing methods that showed inconsistent rankings across datasets—reflecting sensitivity to specific domains—ImLPR consistently maintained best performance as shown in Fig. 1 and Table. 4. This stability highlights its strong generalization, enabled by RIV and VFM with adapter.

Image Channels: ImLPR uses three image channels—reflectivity, range, and normal ratio—to represent LiDAR point clouds. To assess their individual contributions, we evaluate three configurations: reflectivity only (ExpA-1), reflectivity and range (ExpA-2), and all three channels (ExpA-3), excluding Patch-InfoNCE loss since it relies on the range channel. We average inter-session place recognition results across all sequences, as shown in Table. 5. Each added channel improves AR@1 and AF1. Reflectivity provides a strong semantic baseline, while range adds geometric depth and significantly boosts performance. The normal ratio further improves accuracy by capturing local 3D structure, though its effect is smaller due to the similarity with range, as both channels encode

Table 6: Ablation Study on Input Representation, Loss Function, and Adapter for Inter-Session LPR

Method	Input	Adapter	Loss	Roundabout-O		Town-O		DCC		Roundabout-V		Average	
				AR@1	AF1	AR@1	AF1	AR@1	AF1	AR@1	AF1	AR@1	AF1
ExpB-1	BEV	✓	$\mathcal{L}_{\text{TSAP}}$	0.945	0.977	0.923	0.968	0.901	0.966	0.869	0.942	0.911	0.965
ExpB-2	RIV	✓	$\mathcal{L}_{\text{TSAP}}$	0.979	0.990	0.985	0.993	**0.945**	0.970	0.844	0.920	0.938	0.968
ExpB-3	RIV	✗	$\mathcal{L}_P, \mathcal{L}_{\text{TSAP}}$	0.850	0.925	0.833	0.913	0.850	0.943	0.645	0.809	0.795	0.898
ExpB-4	RIV	✓	$\mathcal{L}_P, \mathcal{L}_{\text{TSAP}}$	**0.990**	**0.996**	**0.989**	**0.995**	0.942	**0.973**	**0.888**	**0.948**	**0.952**	**0.978**



(a) Feature Visualization with BEV (b) Feature Visualization with RIV

Figure 6: Feature visualizations comparing BEV and RIV from Roundabout01-O positives. (a) BEV distorts feature shapes (yellow) and misidentifies empty pixels as features (orange), causing inconsistency. (b) RIV shows consistent features between scans, even with a missing car (white).

geometric information. These results demonstrate that each channel contributes distinct, complementary features that collectively strengthen the performance of ImLPR.

BEV and RIV Representations with Patch-InfoNCE Loss: We evaluate the effectiveness of BEV and RIV representations using three model variants: BEV without Patch-InfoNCE loss (ExpB-1), RIV without the loss (ExpB-2), and RIV with the loss (ExpB-4), since Patch-InfoNCE is not compatible with BEV. As shown in Table. 6, RIV consistently outperforms BEV, even without Patch-InfoNCE, and further improves with it by enhancing patch-level learning. Visualizations in Fig. 6 illustrate these differences. BEV struggles with feature representation: the yellow box shows distorted DINO features—diagonal in scan 534 and linear in scan 1284. The orange box highlights empty pixels misidentified as foreground, revealing DINOv2's limitations. In contrast, RIV produces consistent features across both scans. In Fig. 6(b), white boxes remain stable even as a car disappears in scan 1284, with the region correctly excluded. These results indicate that VFM performs more reliably with RIV, which better retains spatial and semantic patterns.

MultiConv Adapter: Table. 6 shows the benefit of incorporating a MultiConv adapter (ExpB-3) versus not using one (ExpB-4). Even when the same number of last transformer blocks are fine-tuned, the presence of the adapter leads to significant performance improvements, notably increasing both average AR@1 and AF1. This confirms the adapter's effectiveness in addressing sensor domain shifts while retaining the generalizable representations learned during pre-training. Additional and detailed ablation studies are provided in the appendix.

5 Conclusion

In this paper, we introduced ImLPR, the first novel pipeline that adapts the DINOv2 VFM to bridge the LiDAR and vision domains. The proposed approach utilized a RIV representation with three channels, integrated MultiConv adapters, and employed Patch-InfoNCE loss for patch-level contrastive learning. Evaluations on the public datasets demonstrate that ImLPR outperforms SOTA methods in both intra-session and inter-session LPR, achieving superior performance. Our approach redefines LPR by moving beyond traditional 3D sparse convolution and BEV-based descriptors. Instead, we leverage VFM with RIV images, rather than BEV, for enhanced feature representation. Future work could explore integrating ImLPR with other foundation models, such as segmentation models, or developing hierarchical pipelines. These approaches could combine ImLPR's patch-based retrieval with a segmentation-driven re-ranking process, where local patch correspondences are refined to improve matching accuracy, enhancing overall LPR performance.

6 Limitation

While ImLPR achieves SOTA results on a variety of datasets, its application remains limited to homogeneous LPR, where the same type of LiDAR sensor is used for both query and database. It is not yet suited for the emerging task of heterogeneous LPR [34], which requires matching across different LiDAR sensor types with varying specifications and fields of view. Accumulating a batch of scans, as adopted in our experiments, partially mitigates this limitation by improving robustness across sessions, but significant discrepancies in sensor geometry prevent it from serving as a complete solution. Furthermore, although we fine-tuned the DINOv2 vision foundation model with MultiConv adapters to better align with the LiDAR domain, effectively addressing heterogeneous LPR or broader generalization still demands training on larger and more diverse datasets. Despite these constraints, ImLPR's VFM-based approach extends beyond traditional place recognition methods by adapting pre-trained DINOv2 features to LiDAR RIV inputs. This VFM-based approach enables future development of generalizable place recognition systems through integration with other foundation models, such as those for segmentation [11] and multi-modal processing [13].

ACKNOWLEDGMENT

This project has been partly funded by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00461409), the UKRI project Mobile Robotic Inspector (EP/Z531212/1), and a Royal Society Univ. Research Fellowship (Fallon).

References

[1] 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] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14141–14152, 2021.

[3] 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.

[4] J. Komorowski. Improving point cloud based place recognition with ranking-based loss and large batch training. In *2022 26th international conference on pattern recognition (ICPR)*, pages 3699–3705. IEEE, 2022.

[5] 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.

[6] L. Luo, S.-Y. Cao, X. Li, J. Xu, R. Ai, Z. Yu, and X. Chen. Bevplace++: Fast, robust, and lightweight lidar global localization for unmanned ground vehicles. *arXiv preprint arXiv:2408.01841*, 2024.

[7] K. Garg, S. S. Puligilla, S. Kolathaya, M. Krishna, and S. Garg. Revisit anything: Visual place recognition via image segment retrieval. In *European Conference on Computer Vision*, pages 326–343. Springer, 2024.

[8] S. Izquierdo and J. Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 17658–17668, 2024.

[9] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 9 (2):1286–1293, 2023.

[10] J. Nie, D. Xue, F. Pan, S. Cheng, W. Liu, J. Hu, and Z. Ning. Mixvpr++: Enhanced visual place recognition with hierarchical-region feature-mixer and adaptive gabor texture fuser. *IEEE Robotics and Automation Letters*, 2024.

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[12] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[14] X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, and H. Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4840–4851, 2024.

[15] X. Wu, D. DeTone, D. Frost, T. Shen, C. Xie, N. Yang, J. Engel, R. Newcombe, H. Zhao, and J. Straub. Sonata: Self-supervised learning of reliable point representations. *arXiv preprint arXiv:2503.16429*, 2025.

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

[17] M. A. Uy and G. H. Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. pages 4470–4479, 2018.

[18] 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. pages 2831–2840, 2019.

[19] J. Guo, P. V. Borges, C. Park, and A. Gawel. Local descriptor for robust place recognition using lidar intensity. 4(2):1470–1477, 2019.

[20] 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.

[21] 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*, 20(3):4039–4048, 2023.

[22] 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.

[23] C. E. Lin, J. Song, R. Zhang, M. Zhu, and M. Ghaffari. Se (3)-equivariant point cloud-based place recognition. In *Conference on Robot Learning*, pages 1520–1530. PMLR, 2023.

[24] S. Garg, M. Vankadari, and M. Milford. Seqmatchnet: Contrastive learning with sequence matching for place recognition & relocalization. In *Conference on Robot Learning*, pages 429–443. PMLR, 2022.

[25] G. Puy, S. Gidaris, A. Boulch, O. Siméoni, C. Sautier, P. Pérez, A. Bursuc, and R. Marlet. Three pillars improving vision foundation model distillation for lidar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21519–21529, 2024.

[26] N. Vödisch, G. Cioffi, M. Cannici, W. Burgard, and D. Scaramuzza. Lidar registration with visual foundation models. *arXiv preprint arXiv:2502.19374*, 2025.

[27] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[28] 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 (WACV) Workshops*, pages 948–957, January 2024.

[29] F. Lu, T. Jin, X. Lan, L. Zhang, Y. Liu, Y. Wang, and C. Yuan. Selavpr++: Towards seamless adaptation of foundation models for efficient place recognition. *arXiv preprint arXiv:2502.16601*, 2025.

[30] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. volume 26. Curran Associates, Inc., 2013.

[31] M. Jung, W. Yang, D. Lee, H. Gil, G. Kim, and A. Kim. Helipr: Heterogeneous lidar dataset for inter-lidar place recognition under spatiotemporal variations. *The International Journal of Robotics Research*, 43(12):1867–1883, 2024.

[32] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim. Mulran: Multimodal range dataset for urban place recognition. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 6246–6253. IEEE, 2020.

[33] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice. University of michigan north campus long-term vision and lidar dataset. *The International Journal of Robotics Research*, 35(9): 1023–1035, 2016.

[34] M. Jung, S. Jung, H. Gil, and A. Kim. Helios: Heterogeneous lidar place recognition via overlap-based learning and local spherical transformer. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Atlanta, May. 2025.

[35] J. Revaud, J. Almazán, R. S. Rezende, and C. R. d. Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5107–5116, 2019.

[36] A. Ali-Bey, B. Chaib-draa, and P. Giguere. Boq: A place is worth a bag of learnable queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17794–17803, 2024.

[37] F. Lu, L. Zhang, X. Lan, S. Dong, Y. Wang, and C. Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. *arXiv preprint arXiv:2402.14505*, 2024.

Appendices

A Experimental Setup

Implementation Details: ImLPR is trained by fine-tuning the final two transformer blocks, with MultiConv adapters integrated every three blocks. LiDAR point clouds are converted into images with dimensions $H \times W = 126 \times 1022$. The image channels comprise reflectivity, range, and normal ratio. Reflectivity acts as a semantic cue for scene segmentation. For LiDAR sensors without reflectivity data, the intensity channel is utilized during inference, serving a comparable semantic purpose. The singular value ratio is calculated using $k = 8$ nearest neighbors for the HeLiPR-O dataset and $k = 25$ for the MulRan, NCLT and HeLiPR-V datasets. To ensure continuity in cylindrical images, the leftmost and rightmost 28 columns (equivalent to 2 patches) are appended to the opposite sides, maintaining seamless connectivity during both training and inference. Feature aggregation uses parameters $(m, l, e) = (128, 64, 256)$ to construct the descriptor. For the Patch-InfoNCE loss, 192 positive and 128 negative patch pairs are sampled per image, with a temperature $\tau_l = 0.2$. Negative patches are selected based on patch distance thresholds $v_{\text{dist}} = 3$ and $h_{\text{dist}} = 20$. The Patch-InfoNCE loss is computed using only 1/8 of the positive image pairs within the batch to optimize computational efficiency. To establish correspondence between two LiDAR scans represented as RIV images, the RIV images are first converted into 3D point clouds and voxelized at a 0.4-meter resolution, followed by precise alignment using the ICP alignment. For the TSAP loss, a temperature $\tau_g = 0.01$ is applied, truncated to the top four ranked descriptors, with a batch size of 2048. The combined loss is weighted with $\lambda = 2.0$. Training proceeds for 100 epochs using the AdamW optimizer, with a learning rate of 5×10^{-4}, a 1/10 warmup phase, and a cosine scheduler. When performed on three NVIDIA GeForce RTX 3090 GPUs, this is completed in 12 hours.

Evaluation Metrics: We evaluate the performance using Recall@1 (R@1) and the maximum F1 score (F1). Recall@1 measures the percentage of queries where the top-1 retrieved scan is a correct match (within 10 meters of the ground truth), defined as:

$$\text{Recall@1} = \frac{\text{Number of queries with correct top-1 match}}{\text{Total number of query candidates}} \tag{4}$$

The F1 score represents the optimal harmonic mean of precision and recall across all thresholds, calculated as:

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

where TP, FP, and FN denote the number of True Positives, False Positives, and False Negatives for each threshold. The maximum F1 score is the highest F1 value obtained by varying the threshold for positive matches. Additionally, we report average results, including Average Recall@1 (AR@1) and Average maximum F1 score (AF1), computed as the mean values across multiple sequences to provide a comprehensive assessment of ImLPR's performance.

B Truncated SmoothAP Loss with Global Image Descriptor

To train ImLPR with the global descriptor, we adopt the Truncated SmoothAP (TSAP) loss [4]. Unlike the triplet and contrastive loss, TSAP optimizes Average Precision (AP) through a continuous approximation, focusing ranking evaluation on top candidates to reduce computation. For a query descriptor g_q with positives in $\mathcal{P}^+$ and all batch descriptors in $\mathcal{D}$, TSAP is formulated as:

$$\mathcal{L}_{\text{TSAP}} = \frac{1}{m}\sum_{q=1}^{m}(1 - \text{AP}_q), \quad \text{AP}_q = \frac{1}{|\mathcal{P}^+|}\sum_{i\in\mathcal{P}^+}\frac{1 + \sum_{j\in\mathcal{P}^+, j\neq i} H\big(d_g(q,i) - d_g(q,j); \tau_g\big)}{1 + \sum_{j\in\mathcal{D}, j\neq i} H\big(d_g(q,i) - d_g(q,j); \tau_g\big)}, \tag{6}$$

where $H\big(x; \tau_g\big) = \big(1 + e^{x/\tau_g}\big)^{-1}$ approximates a step function, with τ_g as a temperature parameter. Summation over $\mathcal{D}$ uses a top-ranked subset to reduce computation. In large-batch training, TSAP

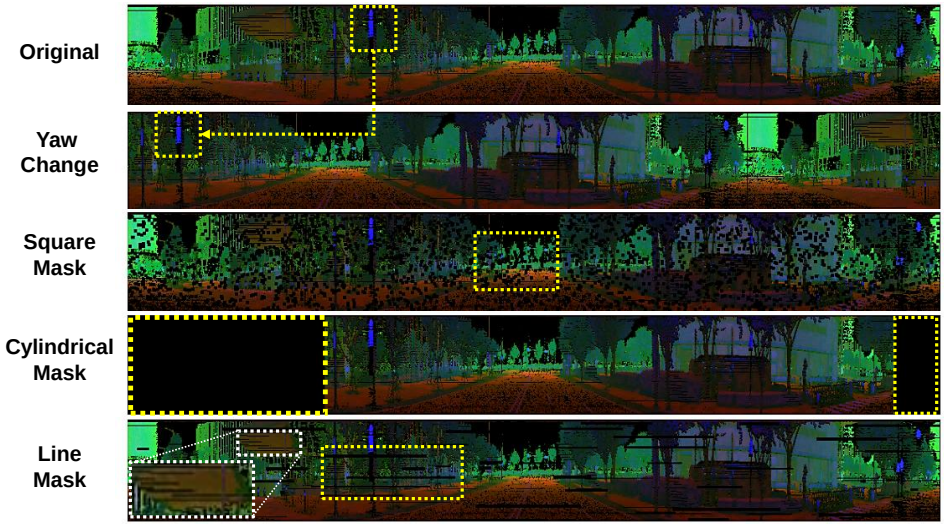


Figure 7: Augmented image featuring yaw variation, square mask, cylindrical mask, and line mask. The yellow line represents a leftward sign displacement, corresponding to horizontal pixel shifts induced by yaw variation. Yellow boxes emphasize augmented areas, highlighting deviations from the original image. In the line mask, the white box indicates noise from vacant line pixels caused by point cloud projection, whereas our yellow box showcases augmentation similar to the original.

matches retrieval metrics such as Average Recall, boosting global descriptor performance. We also apply multi-staged backpropagation [35] for efficient large-batch optimization.

C Data Augmentation in ImLPR

To enhance the robustness of ImLPR, RIV images are subjected to various data augmentations during training. Images are randomly rotated through yaw, sampled uniformly from 0 to 2π radians, using cyclic horizontal column shifts. This promotes robustness to orientation variations prevalent in LiDAR scans. The reflectivity and normal ratio channels are scaled to the [0, 1] range by dividing by 255, while the range channel is normalized by dividing by the LiDAR's maximum range, ensuring uniform input scales.

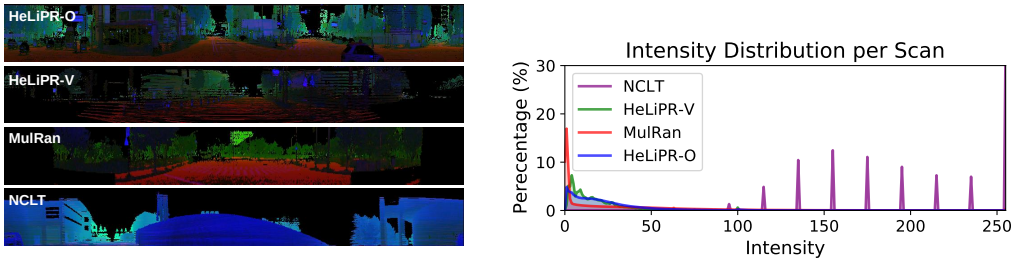
Three masking strategies are employed, as illustrated in Fig. 7. Random patch masking applies square patches of varying sizes, occluding up to 40% of the image area based on a randomly determined mask ratio. Cylindrical masking introduces a contiguous mask, spanning up to 30% of the image width, with random starting positions and cylindrical wrapping to account for boundary effects. These techniques address challenges such as occlusions and sparse scene representations. Line-style masking incorporates randomly placed rectangular lines to emulate projection artifacts, where multiple points collapse into a single pixel or horizontal lines appear vacant due to RIV projection. These augmentations bolster ImLPR's robustness, significantly contributing to its superior performance in both intra-session and inter-session place recognition.

D Datasets

To evaluate ImLPR, we utilize three public datasets: HeLiPR, NCLT, and MulRan. An example scan from each dataset, represented as a RIV, is depicted in Fig. 8(a).

D.1 HeLiPR Dataset

The HeLiPR dataset comprises six distinct environments—`Roundabout01-03`, `Town01-03`, `Bridge01-04`, `DCC04-06`, `KAIST04-06`, and `Riverside04-06`—captured using four LiDAR sensors: Ouster OS2-128, Velodyne VLP-16C, Livox Avia, and Aeva Aeries II. Each sequence covers approximately 8.5 km, providing sufficient coverage for identifying multiple place recognition



(a) Input image from each dataset (b) Intensity distribution per scan

Figure 8: (a) Visualization of RIV images from four datasets across various LiDARs. HeLiPR with Ouster (HeLiPR-O) delivers high-resolution point clouds with minimal empty pixels. MulRan exhibits pronounced empty regions at the leftmost and rightmost edges, attributed to sensor occlusion. Similarly, HeLiPR with Velodyne (HeLiPR-V) suffers from sparse point clouds, resulting in persistent empty pixels despite scan accumulation. In contrast, NCLT scans, benefiting from slower platform motion, display few empty pixels; however, their intensity distribution is markedly distinct, characterized by discrete and inverted values, yielding a predominantly blue appearance. (b) Intensity distributions for each scans from HeLiPR, MulRan, and NCLT, with NCLT demonstrating a significant distributional difference.

pairs. For training, we employ the Ouster OS2-128 sensor across the `DCC04-06`, `KAIST04-06`, and `Riverside04-06` sequences, yielding a total of 16,435 scans. The test set includes the `Roundabout01-03` sequences from both the Ouster OS2-128 and Velodyne VLP-16C sensors, and the `Town01-03` sequences from only the Ouster OS2-128 sensor, with a total of 15,375 scans for the Ouster OS2-128 and 7,728 scans for the Velodyne VLP-16C. Sequences are labeled as `Sequence-Sensor` (e.g., `Roundabout01-O` for Ouster, `Roundabout01-V` for Velodyne) to distinguish sensor-specific data.

For the `Roundabout-V` sequence, we aggregate 5 seconds of scan data to form a submap, which is subsequently projected into a RIV image. To avoid redundant accumulation of static scans, we exclude scans acquired during stationary periods, retaining only those captured during motion. This strategy ensures consistent vertical image dimensions, partially mitigating the challenges posed by sparse point clouds in lower-dimensional representations. However, as shown in Fig. 8(a), the inherent sparsity of the LiDAR data results in persistent empty pixels, making this dataset a challenging testbed for evaluating the robustness of LPR methods to incomplete representations.

D.2 MulRan Dataset

The MulRan dataset covers four urban environments—`DCC`, `KAIST`, `Riverside`, and `Sejong`—each including three sequences (`01-03`) acquired with an Ouster OS1-64 LiDAR. For this dataset, we focus exclusively on the `DCC01-03` sequences, which span an average distance of 4.9km. The dataset poses challenges due to the lower vertical resolution of this LiDAR sensor and occlusions induced by a secondary sensor positioned behind the LiDAR, leading to 20% of RIV image pixels remaining empty, as shown in Fig. 8(a). Point clouds are projected into RIV images of dimension 1022×64 and subsequently resized to 1022×128 via linear interpolation. Given that MulRan intensity values exceed 255, we apply normalization to ensure compatibility. A total of 4,328 scans are utilized as test sets for generalization experiments and ablation studies.

D.3 NCLT Dataset

The NCLT dataset comprises 27 sequences recorded in a college campus using a Velodyne HDL-32E LiDAR mounted on a Segway platform. Owing to the sparse nature of the point clouds, we aggregate scans using the methodology applied to HeLiPR-V dataset. The Segway's reduced velocity, relative to the vehicle used for HeLiPR, yields RIV images with fewer empty pixels, closely resembling the high-resolution images of the HeLiPR-O dataset. Nevertheless, the intensity distribution markedly differs from other datasets, exhibiting discrete values and a prevalence of high-intensity points, in contrast to the predominantly low-intensity points typical of other datasets. This difference results in

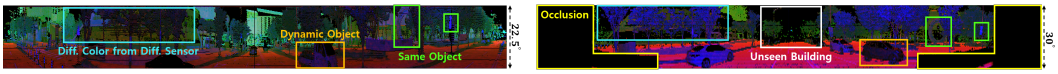


Figure 9: RIV images from HeLiPR (Left) and MulRan (Right) DCC sequence, taken 0.3m apart, vary in appearance: occluded area (sensor location), hidden buildings (range), warped proportions (FoV and sensor location), and shifted colors (intensity distributions).

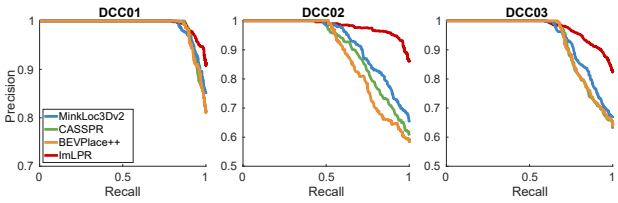


Table 7: Results in MulRan dataset

Method	MulRan DCC (OS1-64) 01 R@1	01 F1	02 R@1	02 F1	03 R@1	03 F1	Average AR@1	Average AF1
LoGG3D-Net (D)	0.117	0.310	0.075	0.144	0.090	0.187	0.094	0.214
CASSPR (D)	0.813	0.943	0.607	0.793	0.629	0.818	0.683	0.851
MinkLoc3dv2 (D)	0.849	0.939	0.651	0.829	0.666	0.841	0.722	0.870
MinkLoc3Dv2 (R)	0.853	0.932	0.632	0.814	0.688	0.832	0.724	0.859
BEVPlace++ (D)	0.811	0.937	0.585	0.766	0.639	0.813	0.678	0.839
BEVPlace++ (R)	0.822	0.922	0.588	0.773	0.655	0.811	0.688	0.835
ImLPR (D)	0.909	0.965	**0.861**	**0.945**	**0.825**	0.918	**0.865**	0.943
ImLPR (R)	**0.918**	**0.968**	0.844	0.940	**0.825**	**0.925**	0.862	**0.944**

Figure 10: Performance of High-Resolution LiDAR-Trained Models on MulRan dataset (OS1-64)

blurrier NCLT RIV images as depicted in Fig. 8. To mitigate this, we invert and rescale the intensity distribution. Despite these adjustments, the discrete and inaccurate intensity channel continues to impair the semantic quality of RIV images, posing challenges for LPR. For generalization testing, we utilize the first three sequences—`2012-01-08`, `2012-01-15`, and `2012-01-22`—encompassing a total of 6,368 scans as test sets.

E Evaluation on Generalization Capability

Following the summarized generalization experiments in the main paper, here we further assess the performance of the trained model in detail using three datasets. Consistent with the primary evaluation, we conduct intra-session place recognition using the HeLiPR-O trained model without additional fine-tuning. Scan accumulation follows the methodology outlined in §D. To ensure a fair comparison, all methods are assessed using the same accumulated scans.

E.1 Evaluation on MulRan Dataset

We evaluate the generalization ability of ImLPR on the MulRan `DCC01-03` sequences. Although the training and test sets have spatial overlap, they differ substantially due to a four-year temporal gap, differences in LiDAR hardware (OS1-64), and environmental changes such as occlusions. These differences are illustrated in Fig. 9. For evaluation, LiDAR scans are projected into 1022×64 RIV images and resized to 1022×126 to normalize vertical ray differences.

To ensure the robustness of our evaluation protocol and rule out the possibility of performance inflation due to overlap, we additionally replace the original training sequences (`DCC04-06`) with a geographically seperate sequence, `Roundabout04-06`, from the HeLiPR with Ouster. We retrain three representative methods, MinkLoc3Dv2 (point-based), BEVPlace++ (BEV-based), and ImLPR (RIV-based), on both training sets: the original (denoted as D) and the revised one (denoted as R).

As shown in Table. 7 and Fig. 10, all methods exhibit only minor differences in performance between the two training sets, validating that the evaluation results are not overly dependent on spatial overlap. Moreover, all models experience performance degradation on the MulRan dataset compared to high-resolution training data, highlighting the impact of domain shift. Among them, ImLPR consistently achieves the highest performance, regardless of training configuration. These results demonstrate that ImLPR generalizes robustly across datasets with different LiDAR sensors and environmental conditions. MinkLoc3Dv2 benefits from sparse convolution that avoids feature generation in occluded regions, ranking second. BEVPlace++ ranks lower, as BEV images tend to retain features even in empty or occluded areas. ImLPR, by contrast, effectively handles low resolution and occlusions through its robust RIV representation and powerful feature extractor. These results demonstrate that ImLPR exhibits strong generalization on the MulRan dataset, even under substantial domain shifts.

Table 8: Performance of High-Resolution LiDAR-Trained Models on Low-Resolution LiDAR

Method	NCLT (HDL-32E)								HeLiPR Roundabout-V (VLP-16C)								Average	
	2012-01-08		2012-01-15		2012-01-22		Average		01		02		03		Average			
	R@1	F1	R@1	F1	R@1	F1	R@1	F1	R@1	F1	R@1	F1	R@1	F1	R@1	F1	R@1	F1
LoGG3D-Net	0.173	0.568	0.141	0.416	0.132	0.318	0.149	0.434	0.038	0.358	0.052	0.133	0.302	0.690	0.131	0.394	0.140	0.414
MinkLoc3dv2	0.649	0.836	0.603	0.773	0.613	0.815	0.622	0.808	0.469	0.771	0.421	0.634	0.648	0.887	0.513	0.764	0.567	0.786
CASSPR	0.681	0.839	0.623	0.789	0.651	0.801	0.652	0.810	0.607	0.784	0.571	0.728	0.711	0.867	0.630	0.793	0.641	0.801
BEVPlace++	**0.861**	**0.928**	**0.838**	0.915	0.850	0.923	**0.850**	0.922	0.624	0.769	0.575	0.731	**0.767**	0.870	0.655	0.790	0.753	0.856
ImLPR	0.817	0.920	0.830	**0.930**	**0.854**	**0.932**	0.834	**0.927**	**0.765**	**0.884**	**0.617**	**0.781**	0.753	**0.892**	**0.712**	**0.852**	**0.773**	**0.890**

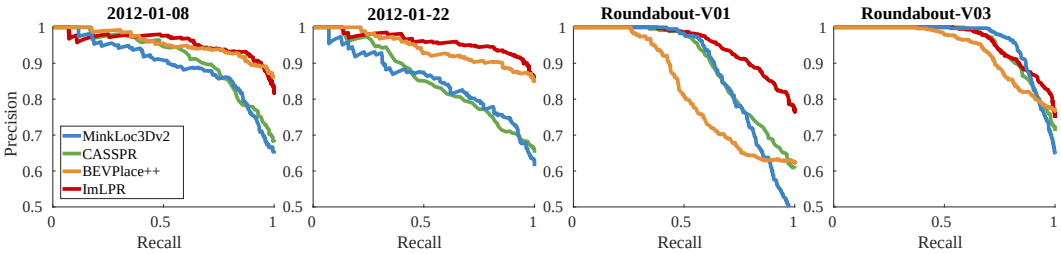


Figure 11: Precision-Recall curves for generalization assessment on NCLT and HeLiPR-V datasets. ImLPR demonstrates robust and consistent performance across both datasets. Conversely, BEVPlace++ yields strong results on NCLT but diminished performance on HeLiPR `Roundabout-V`. Sparse 3D convolution methods, MinkLoc3Dv2 and CASSPR, exhibit variable performance across the evaluated datasets.

E.2 Evaluation on NCLT Dataset

As reported in Table. 8, ImLPR and BEVPlace++ exhibit comparable performance across the evaluated datasets. This similarity stems from the discrete and inaccurate intensity values, which diminish pixel-level distinctions in RIV images and hinder semantic interpretation. Nevertheless, ImLPR surpasses 3D sparse convolution methods, such as MinkLoc3Dv2, by leveraging geometric information from the range and normal ratio channels effectively. Although ImLPR's Recall@1 and F1 score for the `2012-01-08` sequence are marginally lower than those of BEVPlace++, the Precision-Recall curve, shown in Fig. 11, reveals that ImLPR remains competitive with BEVPlace++ and achieves a superior Area Under the Curve (AUC) for the `2012-01-22` sequence. Furthermore, despite the unreliable intensity channel, ImLPR attains the highest average F1 score and the second-highest Average Recall@1, closely following BEVPlace++. These findings highlight ImLPR's robust generalization capabilities and its efficacy in facilitating LPR under challenging intensity conditions.

E.3 Evaluation on HeLiPR-V Dataset

ImLPR surpasses competing methods across nearly all of our evaluation metrics, securing the highest Recall@1 and F1 score for the `Roundabout01-V` and `Roundabout02-V` sequences, and delivering competitive performance in `Roundabout03-V`, as detailed in Table. 8. On average, ImLPR achieves a performance improvement more than 10% better than the second-ranked method, BEVPlace++, in both Recall@1 and F1 score. In contrast, other methods show significant performance variability. This is primarily due to the different point cloud distributions which were unobserved during training, which greatly hinder their generalization. Notably, the second-highest F1 scores across the three HeLiPR-V sequences are inconsistent, each attained by a different method, highlighting the difficulty that existing methods have in achieving reliable performance. The Precision-Recall curves presented in Fig. 11 indicate that for `Roundabout03-V`, ImLPR and MinkLoc3Dv2 yield comparable AUC, followed by CASSPR and BEVPlace++. For `Roundabout01-V`, however, ImLPR markedly outperforms other methods, with CASSPR, MinkLoc3Dv2, and BEVPlace++ trailing in that order. These results underscore ImLPR's consistently high performance across the HeLiPR-V dataset.

Despite performance degradation across the MulRan, NCLT, and HeLiPR-V datasets due to domain shifts, sensor disparities, inaccurate intensity values, and occlusions, ImLPR demonstrates robust performance. This resilience stems from two key factors. First, ImLPR's adept fusion of geometric and semantic features included in RIV channels enables reliable LPR even with low-resolution or

Table 9: Intra-session PR with different threshold in `Roundabout01-O`

Method	Threshold 1m		Threshold 3m		Threshold 5m		Threshold 7m		Threshold 10m	
	R@1	F1	R@1	F1	R@1	F1	R@1	F1	R@1	F1
MinkLoc3Dv2	0.732	0.963	0.933	0.968	0.798	0.888	0.839	0.915	0.933	0.968
BEVPlace++	0.913	0.959	0.904	0.951	0.909	0.953	0.921	0.959	0.975	0.988
ImLPR	**0.965**	**0.987**	**0.933**	**0.970**	**0.975**	**0.988**	**0.960**	**0.980**	**0.992**	**0.996**

unstable inputs. Second, its exceptional generalization performance, enabled by VFM, overcomes the limitations of traditional domain-specific training, which struggles to achieve generalizability. These observations underscore ImLPR's potential as a robust framework for generalized LPR across diverse and challenging environments.

F Performance Variations According to Different Thresholds

To rigorously evaluate spatial precision in place recognition, we analyze the impact of varying the distance threshold used to define positive correspondences. While a default threshold of 10 m, which roughly corresponds to the width of a four-lane road, is a reasonable setting, this choice often results in uniformly high scores across methods. limiting discriminative insight. We therefore systematically tighten the thresholds, and evaluate intra-session performance on `Roundabout01-Ouster` using three representative methods: MinkLoc3Dv2 (3D point cloud-based), BEVPlace++ (BEV image-based), and ImLPR (RIV image-based).

As presented in Table. 9, ImLPR exhibits consistently superior performance across all threshold levels. Notably, under the most stringent 1 m setting, it achieves an R@1 of 0.965 and F1 score of 0.987, substantially outperforming both BEVPlace++ and MinkLoc3Dv2. This indicates that ImLPR is capable of not only recognizing the correct place but also localizing it with high spatial accuracy. BEVPlace++ maintains relatively stable performance across thresholds but shows limited gains under stricter criteria, highlighting its reduced sensitivity to fine-grained spatial alignment. In contrast, MinkLoc3Dv2 exhibits fluctuations due to the sparsity of true positives at lower thresholds, which can degrade metric stability in absence of sufficient positive samples. These results reinforce the robustness of ImLPR under varying spatial tolerances, confirming its capacity for precise place recognition.

G Robustness to Yaw Variation

G.1 Place Recognition Performance under Yaw Variation

Descriptors for identical locations should maintain consistency despite sensor rotations, particularly yaw angle variations, which frequently occur during scene revisit. To assess this, we applied yaw rotations to database scans from the `Roundabout-O` and `Town-O` sequences and compared them against unrotated query scans. The Average Recall@1, obtained from inter-session place recognition experiments across all evaluated sequences, is depicted in Fig. 12.

As illustrated in Fig. 12, MinkLoc3Dv2 exhibits performance degradation with varying yaw angles, reflecting limited robustness to such transformations. BEVPlace++ incorporates a rotation-equivariant module to mitigate yaw variance. This module rotates the input image at a fixed angular interval of $\theta = 45^\circ$, processes each rotated image through a convolutional neural network, and applies max pooling across the resultant features. While this approach enhances the yaw variation robustness, performance fluctuates slightly when the variations deviate from the predefined discrete angles used in max pooling. In contrast, ImLPR generates yaw-robust descriptors without additional computational overhead, owing to its architectural design.

For a RIV image, yaw variation is equivalent to a horizontal shift. DINOv2, the vision transformer we use, processes patch features via attention mechanisms that capture global inter-patch relation-

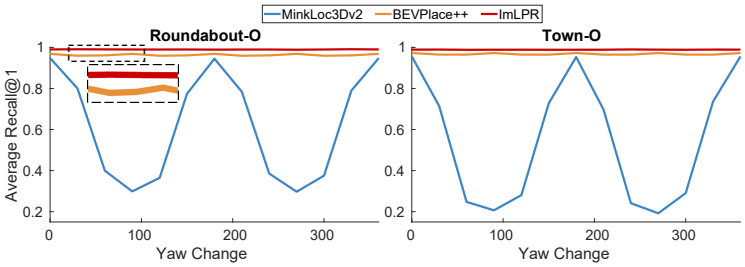


Table 10: σ of AR@1

Method	Round-O	Town-O
MinkLoc3Dv2	26.900	31.347
BEVPlace++	0.420	0.389
ImLPR	**0.064**	**0.047**

Figure 12: Average Recall@1 and its standard deviation (σ) for inter-session place recognition across yaw variations. ImLPR exhibits inherent yaw robustness, achieving the lowest σ.

ships. Since horizontal shifts reorder patches without modifying their content, the attention mechanism yields nearly translation-equivariant feature representations, preserving the relational structure. Consequently, vision transformers produce similar patch features that are shifted from the original patch features, alongside consistent global tokens for both original and shifted images. Similarly, the MultiConv adapter employs 2D convolutions, which are inherently translation-equivariant, ensuring that a horizontal shift in the input image results in a corresponding shift in patch features while maintaining their values. These refined patch features are aggregated using SALAD's optimal transport approach, which is invariant to horizontal shifts, as demonstrated in §G.2. Data augmentation strategies, including masking and random transformations, further bolster this stability. However, positional encoding introduces a minor constraint: as yaw rotation shifts the image, identical patches receive different positional encodings, leading to subtle differences in the vision transformer's output before and after rotation. Despite this, ImLPR achieves more consistent performance than BEVPlace++ under yaw variation, as shown in Table. 10. This highlights the robustness of ImLPR's model architecture and RIV representation in handling yaw variations effectively.

G.2 Horizontal Shift Invariance in SALAD's Optimal Transport Aggregation

In this section, we prove that SALAD's optimal transport (OT) aggregation generates an identical global descriptor despite horizontal (column-wise) shifts in the input feature map. Let $\mathbf{F} \in \mathbb{R}^{H'W' \times C}$ be the flattened local features derived from adapter-refined patch features, originating from an un-flattened feature map $\mathbf{F}' \in \mathbb{R}^{H' \times W' \times C}$ with $n = H'W'$ patches. A translation-equivariant convolutional layer maps $\mathbf{F}'$ to a score map $\mathbf{S}' \in \mathbb{R}^{H' \times W' \times m}$, which is flattened to $\mathbf{S} \in \mathbb{R}^{n \times m}$, representing assignment probabilities of n patches to m learnable cluster centers. A column shift in $\mathbf{F}'$ (e.g., $\mathbf{F}'_{p,q,c} \rightarrow \mathbf{F}'_{p,\mathrm{mod}((q-s),W'),c}$) induces an identical column shift in $\mathbf{S}'$: $\mathbf{S}'_{p,q,m} \rightarrow \mathbf{S}'_{p,\mathrm{mod}((q-s),W'),m}$. Here, $p \in [1, H']$ and $q \in [1, W']$ denote the row and column indices, and the function $\mathrm{mod}(a, b) = ((a-1) \mod b) + 1$ ensures the result lies in $[1, b]$. After flattening, this translates to a row shift in $\mathbf{S}$: $\mathbf{S}_{i,:} \rightarrow \mathbf{S}_{\mathrm{mod}((i-s'),n),:}$, where $s' = s \times H'$. This preserves the score values for each patch, only reordering their row indices, ensuring shift invariance in the subsequent OT aggregation.

For OT aggregation, the Sinkhorn algorithm [30] solves an entropically regularized OT problem. The score matrix $\mathbf{S} \in \mathbb{R}^{n \times m}$ represents assignment probabilities to m cluster centers. The cost matrix is defined as $\mathbf{M} = -\mathbf{S}/\lambda$, where λ is the regularization parameter controlling the entropy of the transport plan. A column shift in $\mathbf{S}$ results in a row shift in $\mathbf{M}$. The Sinkhorn algorithm iteratively updates dual variables $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^m$ to satisfy marginal constraints, using uniform log-weights $\log \mathbf{a}$ and $\log \mathbf{b}$. Since $\log \mathbf{a}$ is uniform across patches, $\mathbf{u}$ shifts consistently under a row shift. Specifically, for patch i:

$$u_i = \log a_i - \log \sum_{k=1}^{m} \exp(M_{i,k} + v_k). \tag{7}$$

A row shift (e.g., $\mathbf{M}_{i,k} \rightarrow \mathbf{M}_{\mathrm{mod}((i-s'),n),k}$) results in $u_i \rightarrow u_{\mathrm{mod}((i-s'),n)}$. The OT matrix is computed as:

$$\log R_{i,k} = M_{i,k} + u_i + v_k. \tag{8}$$

Table 11: Ablation study for adapter and trained block

Method	Adapter	Trained Block	Roundabout-O AR@1	Roundabout-O AF1	Town-O AR@1	Town-O AF1	DCC AR@1	DCC AF1	Roundabout-V AR@1	Roundabout-V AF1	Average AR@1	Average AF1
ExpC-1		0	0.321	0.530	0.379	0.553	0.376	0.555	0.249	0.464	0.331	0.526
ExpC-2		2	0.850	0.925	0.833	0.913	0.850	0.943	0.645	0.809	0.795	0.898
ExpC-3	✔	0	0.972	0.987	0.965	0.983	0.928	0.965	0.799	0.906	0.916	0.960
ExpC-4	✔	2	**0.990**	**0.996**	0.989	0.995	**0.942**	**0.973**	**0.888**	**0.948**	**0.952**	**0.978**
ExpC-5	✔	6	**0.990**	0.995	**0.990**	**0.996**	0.941	**0.973**	0.868	0.940	0.947	0.976
ExpC-6	✔	10	0.986	0.994	0.988	0.995	0.935	0.972	0.870	0.936	0.945	0.974

Consequently, the rows of $\log \mathbf{R}$ shift identically to $\mathbf{S}$, preserving assignment weights, yielding $\mathbf{R} \in \mathbb{R}^{n \times m}$. The aggregated cluster features are computed as a weighted sum of intermediate feature embeddings $\bar{\mathbf{F}} \in \mathbb{R}^{n \times l}$, derived by applying a convolutional layer to $\mathbf{F}'$ and flattened from $\bar{\mathbf{F}}' \in \mathbb{R}^{H' \times W' \times l}$:

$$V_{j,k} = \sum_{i=1}^{n} R_{i,j} \bar{F}_{i,k}, \quad j = 1, \ldots, m, \quad k = 1, \ldots, l, \tag{9}$$

producing $\mathbf{V} \in \mathbb{R}^{m \times l}$. The feature matrix $\bar{\mathbf{F}}$ inherits the same column shift: $\bar{\mathbf{F}}'_{p,q,l} \to \bar{\mathbf{F}}'_{p,\mathrm{mod}((q-s),W'),l}$, or $\bar{\mathbf{F}}_{i,:} \to \bar{\mathbf{F}}_{\mathrm{mod}((i-s'),n),:}$ after flattening. Since $\bar{\mathbf{F}}_{i,k}$ and $R_{i,j}$ shift identically, and summation is commutative, the aggregated features are invariant to column shifts:

$$\sum_{i=1}^{n} R_{\mathrm{mod}((i-s'),n),j} \bar{F}_{\mathrm{mod}((i-s'),n),k} = \sum_{i=1}^{n} R_{i,j} \bar{F}_{i,k}. \tag{10}$$

The global descriptor concatenates the flattened cluster features $\mathbf{V} \in \mathbb{R}^{m \times l}$ with the global embedding $\mathbf{G} \in \mathbb{R}^{e}$, processed independently via linear layers and unaffected by shifts, forming $g = [\mathbf{V}.\mathrm{flatten}(), \mathbf{G}] \in \mathbb{R}^{m \times l + e}$, ensuring horizontal shift invariance.

H Ablation Studies

To assess the impact of individual components in ImLPR for place recognition, we perform a series of ablation studies. We evaluate performance using AR@1 and AF1 for inter-session place recognition. These studies systematically analyze the contributions of key elements to ImLPR's overall effectiveness. The results provide insights into the significance of each component in achieving robust and efficient LPR performance.

H.1 MultiConv Adapter and the Number of Trained Block in DINOv2

VFMs, such as DINOv2, leverage extensive pre-training on large-scale datasets to deliver robust feature representations. To preserve these learned representations during adaptation for LPR, it is essential to avoid catastrophic forgetting while effectively bridging the domain gap between natural vision images and RIV images. Fine-tuning multiple transformer blocks risks overwriting the model's general knowledge, whereas minimal fine-tuning with strategic adaptations can maintain performance. Following the main evaluation, this section examines how the inclusion of the MultiConv adapter and varying the number of trained transformer blocks influence the performance of ImLPR.

The results, presented in Table. 11, highlight the critical role of the MultiConv adapter. Without the adapter and without fine-tuning (ExpC-1), the model fails to perform effective place recognition due to the significant domain gap between natural images and RIV images. Fine-tuning without the adapter (ExpC-2) also results in reduced performance, as the limited number of trainable parameters hinders the model's ability to extract robust features from RIV images. In contrast, incorporating the MultiConv adapter while keeping most transformer blocks frozen (ExpC-3 and ExpC-4) significantly improves performance. This demonstrates the adapter's ability to efficiently address domain shifts, leveraging DINOv2's pre-trained knowledge for enhanced place recognition.

Additionally, we analyze the effect of varying the number of trained transformer blocks alongside the MultiConv adapter. For Roundabout-O and Town-O, performance remains stable across config-

Table 12: Ablation study for yaw change and mask types

Method	Yaw Change	Line & Square Mask	Cylindrical Mask	Roundabout-O AR@1	Roundabout-O AF1	Town-O AR@1	Town-O AF1	DCC AR@1	DCC AF1	Roundabout-V AR@1	Roundabout-V AF1	Average AR@1	Average AF1
ExpD-1				0.937	0.972	0.943	0.973	0.905	0.966	0.781	0.893	0.892	0.951
ExpD-2	✔			0.980	0.990	0.958	0.980	0.896	0.983	0.821	0.927	0.914	0.970
ExpD-3	✔	✔		0.984	0.992	0.975	0.989	0.894	0.985	0.837	0.928	0.923	0.974
ExpD-4	✔	✔	✔	**0.990**	**0.996**	**0.989**	**0.995**	**0.942**	**0.973**	**0.888**	**0.948**	**0.952**	**0.978**

urations, indicating that effective domain adaptation can be achieved with the adapter and minimal fine-tuning of just a few transformer blocks. However, for DCC and Roundabout-V, performance declines as more blocks are trained. This degradation likely results from overfitting to the training dataset due to an increased number of trainable parameters, which compromises the model's ability to generalize. These findings highlight the critical role of preserving pre-trained knowledge and the necessity of the MultiConv adapter in effectively balancing domain adaptation, ensuring robust LPR performance with minimal fine-tuning.

H.2 Effect of Data Augmentation

We evaluate the impact of data augmentation on ImLPR's inter-session place recognition performance, with results presented in Table. 12. Without any augmentations (ExpD-1), the model achieves baseline performance but struggles with orientation variability and projection artifacts in RIV images, limiting its robustness. Applying random yaw rotation (ExpD-2) significantly improves performance by introducing column shifts that enhance the model's resilience to orientation changes in LiDAR scans. Further improvement is observed with the addition of line and square masking (ExpD-3). Both masks simulate sparse regions by introducing empty pixels in RIV images, mimicking the effect of dropout and training the model to perform place recognition with partially empty patch, thus boosting robustness to incomplete inputs.

The best performance is achieved when combining all augmentations—random yaw rotation, line, square, and cylindrical masking (ExpD-4). The cylindrical mask also enhances the model's ability to handle the extreme occlusions by training it to rely on partial RIV images. This configuration delivers superior results across all datasets, with notable gains on DCC and Roundabout-V, which include occlusions in their scans. These results underscore the vital role of data augmentation in enhancing ImLPR's robustness, with the combined yaw rotation and masking strategies effectively addressing orientation variability, projection artifacts, and scene sparsity in LPR.

H.3 Dimension of DINOv2

The impact of varying the DINOv2 backbone's feature dimension on ImLPR's inter-session place recognition performance and computational efficiency is assessed in Table. 13. All experiments here use a single NVIDIA GeForce RTX 3090 GPU to measure computation time. This ablation study evaluates three DINOv2 configurations—ViT-S/14, ViT-B/14, and ViT-L/14—with feature dimensions of 384, 768, and 1024, respectively, analyzing their place recognition efficacy and computational cost. The final two transformer blocks are fine-tuned for all models to ensure consistent adaptation, and computational efficiency is compared against other SOTA methods.

H.3.1 Performance Analysis

The place recognition performance of ImLPR across different feature dimensions is detailed in Table. 13, which presents results for multiple datasets. All three configurations—ViT-S/14 (ExpE-1), ViT-B/14 (ExpE-2), and ViT-L/14 (ExpE-3)—exhibit similar performance, achieving robust place recognition outcomes. This similarity indicates that higher feature dimensions do not always yield proportional improvements in accuracy. Notably, the smallest model, ViT-S/14, demonstrates exceptional robustness, particularly on Roundabout-O, Town-O, and Roundabout-V, where it delivers highly competitive results. These findings indicate that larger models may not always outperform smaller ones, as their increased complexity can hinder effective training on diverse RIV images and lead to overfitting on the training data.

Table 13: Ablation study for feature dimension and computation time

Method	Feature Dim.	Runtime (ms)	Roundabout-0 AR@1	Roundabout-0 AF1	Town-0 AR@1	Town-0 AF1	DCC AR@1	DCC AF1	Roundabout-V AR@1	Roundabout-V AF1	Average AR@1	Average AF1
ExpE-1	384	18.1	**0.990**	**0.996**	**0.989**	**0.995**	0.942	0.973	**0.888**	**0.948**	**0.952**	**0.978**
ExpE-2	768	23.3	0.989	0.995	0.988	0.994	0.943	0.973	0.861	0.942	0.945	0.976
ExpE-3	1024	58.0	0.989	0.995	0.986	0.994	**0.957**	**0.979**	0.873	0.944	0.951	**0.978**

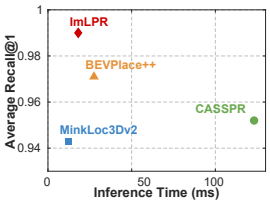


Figure 13: (Left) Table. 13 displays feature dimensions and inference times for ViT models, showing comparable performance despite larger dimensions. (Right) The plot illustrates inference time versus Average Recall@1 on Roundabout-0 and Town-0, with ImLPR achieving the highest performance and an inference time comparable to MinkLoc3Dv2 for descriptors.

Table 14: Average Results of Intra-session PR with Visual Place Recognition models

Method	DINOv2 Backbone	Runtime (ms)	Roundabout-0 AR@1	Roundabout-0 AF1	Town-0 AR@1	Town-0 AF1	Average AR@1	Average AF1
SALAD	ViT-B/14	21.3	0.841	0.919	0.771	0.874	0.806	0.896
BoQ	ViT-B/14	22.7	0.934	0.967	0.896	0.947	0.915	0.957
SelaVPR (global)	ViT-L/14	79.5	0.955	0.978	0.941	0.970	0.948	0.974
ImLPR	ViT-S/14	**18.1**	**0.986**	**0.994**	**0.966**	**0.984**	**0.976**	**0.989**

H.3.2 Computational Cost

The computational efficiency is critical for real-world place recognition applications. As reported in Table. 13, the time required for descriptor extraction scales with feature dimension: ViT-S/14 (ExpE-1) is the most efficient, followed by ViT-B/14 (ExpE-2), with ViT-L/14 (ExpE-3) incurring the highest computational cost. All configurations achieve extraction times below 100ms, demonstrating their suitability for real-time performance in LPR. As depicted in Fig. 13, we further compare ImLPR against baseline methods using Average Recall@1 on Roundabout-0 and Town-0. BEVPlace++ requires 27.5ms for descriptor extraction, despite a smaller feature dimension of 128, surpassing the runtime of ViT-S/14 and ViT-B/14, due to its reliance on multiple ResNet instances for yaw invariance. MinkLoc3Dv2, utilizing 3D sparse convolution, achieves the fastest extraction time but yields the lowest Average Recall@1. In contrast, CASSPR's attention-based networks and per-point feature extraction lead to the longest computation time. ImLPR's Vision Transformer-based backbone inherently supports robustness to yaw variation and achieve a low latency of 18.1ms. Furthermore, as shown in Fig. 13, ImLPR achieves the highest Average Recall@1, demonstrating its suitability for real-time place recognition applications requiring minimal latency.

The balance between performance and computational cost is a critical factor in selecting an optimal backbone. The results suggest that overly large models, such as ViT-L/14, may introduce unnecessary complexity without corresponding performance benefits, potentially due to overfitting or challenges in training. Conversely, ViT-S/14 offers an optimal trade-off between model size, computational efficiency, and the performance of place recognition, making it highly suitable for practical deployment.

I Comparison with Visual Place Recognition Models

ImLPR employs a VFM and adopts image-based aggregation strategies, SALAD, similar to those used in VPR. To assess whether RIV images—composed of reflectivity, range, and normal ratio channels—can be effectively processed by conventional VPR models, we compare ImLPR with recent VPR methods, including SALAD [8], BoQ [36], and SelaVPR [37]. To ensure fair comparison, all models were trained using the same RIV inputs and identical data augmentation strategies. We also used the same loss function (TSAP loss) across all methods, and all feature encoders and training hyperparameters were kept at their default settings.

Table. 14 reports average intra-session place recognition performance on Roundabout-0 and Town-0. While existing VPR methods achieve high accuracy in traditional visual tasks, they show

Table 15: Performance Comparison of 2D and 3D Foundation Models

Method	Params (M)	Runtime (ms)	Intra-session LPR				Inter-session LPR			
			Roundabout-O		Town-O		Roundabout-O		Town-O	
			AR@1	AF1	AR@1	AF1	AR@1	AF1	AR@1	AF1
PTv3+SALAD	46.3	55.0	0.927	0.964	0.938	0.969	0.969	0.985	0.970	0.985
ImLPR	**25.7**	**12.9**	**0.986**	**0.994**	**0.966**	**0.984**	**0.990**	**0.996**	**0.989**	**0.995**

limited performance when directly applied to RIV images. SALAD and BoQ, which use ViT-B/14 backbones trained with the TSAP loss, underperform compared to ImLPR across all metrics, indicating that naive application of VPR pipelines to LiDAR-based images does not adequately handle the sensor domain gap. SelaVPR, which integrates adapters and uses a larger ViT-L/14 backbone, performs better than SALAD and BoQ but still falls short of ImLPR.

Despite using a smaller backbone (ViT-S/14), ImLPR outperforms all baselines by a notable margin and operates with the lowest runtime (18.1 ms). This highlights both the computational efficiency and accuracy of the proposed method. The performance gap emphasizes the importance of explicitly addressing the domain shift between camera and LiDAR data. In particular, the use of adapters allows for better alignment with LiDAR-specific structures, as evidenced in both SelaVPR and ImLPR, while our Patch-InfoNCE loss further enhances RIV feature learning by leveraging range-aware supervision, which is unavailable in image-based models. These results demonstrate that high performance on LiDAR images cannot be achieved by merely applying VPR models composed of DINOv2 and an aggregation network to RIV inputs. Instead, effective place recognition requires proper adaptation beyond simple fine-tuning, as well as LiDAR-aware feature learning.

J Comparison with 3D Foundation Model

In this section, we evaluate ImLPR against a 3D foundation model, specifically PointTransformerv3 (PTv3). Other 3D foundation models, such as Sonata, could potentially serve as feature extractors; however, Sonata lacks a pre-trained model trained on outdoor datasets and requires color information, which is incompatible with raw LiDAR scans. To align PTv3 with outdoor LPR, we re-train its pre-trained model, originally trained on the nuScenes dataset[1], with the HeLiPR-O dataset. This retraining uses x, y, z coordinates and reflectivity from the point cloud. For consistency with ImLPR, we aggregate PTv3's output features using the SALAD. To ensure computational efficiency, we turn on the flash attention for PTv3 and both methods are tested on a single NVIDIA GeForce RTX 4090 GPU. Additionally, we apply the same augmentation strategy used during the original training of PTv3 on the nuScenes dataset.

As shown in Table. 15, ImLPR is more lightweight, with approximately half the parameters of PTv3 and a runtime four times faster. Although PTv3 surpasses baselines like CASSPR and MinkLoc3Dv2 in Table. 1, Table. 2, and Table. 3 thanks to its ability of 3D foundation model, ImLPR consistently outperforms PTv3 in both intra-session and inter-session LPR while maintaining superior computational efficiency. These results demonstrate that ImLPR's adaptation of LiDAR data to the vision domain delivers a more effective and efficient solution for LPR. By leveraging DINOv2's extensive pre-trained knowledge from millions of images, ImLPR outperforms PTv3, which is trained on thousands of point clouds. This underscores the advantage of a VFM with a well-crafted approach, enabling it to surpass 3D foundation models even within the LiDAR domain for LPR.

K Additional Qualitative Results

In this section, we analyze the retrieval distribution for inter-session place recognition, defining a positive match as within 50 meters, using `Roundabout01-O` and `Town01-O` as the database and queries from `Roundabout02-O` and `Town02-O`. As illustrated in Fig. 14, ImLPR retrieves nearly all candidates closer to the query compared to other methods, indicating that its descriptors preserve

[1]https://www.nuscenes.org/

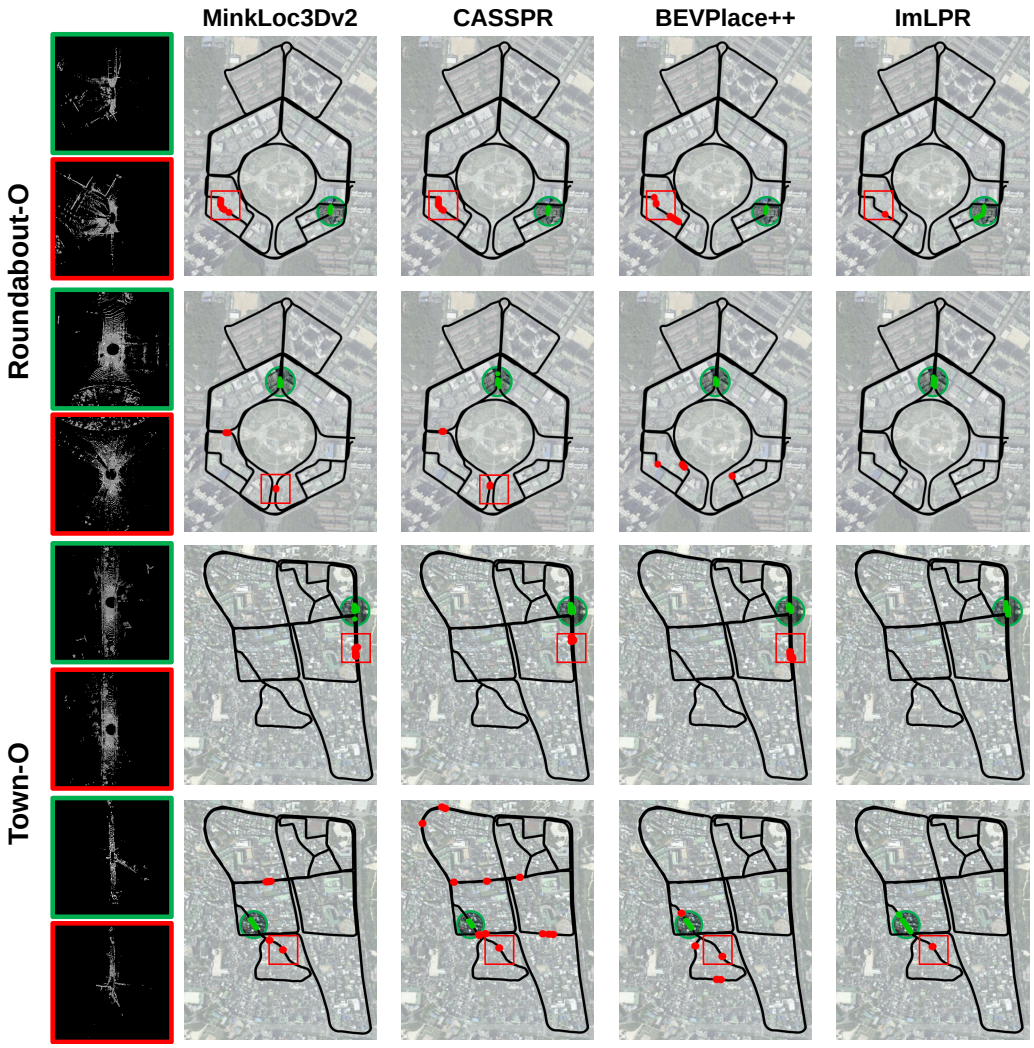


Figure 14: We retrieved 20 locations from `Roundabout01-O` and `Town01-O` using queries from `Roundabout02-O` and `Town02-O`. Each image, overlaid with trajectory and satellite imagery, indicates the query with a green circle, true positives with green points, and false positives with red points. Additionally, BEV images illustrate the sensing environments, marking the query location in green and incorrect retrieved locations with a red square.

small feature distances for both the top-1 match and proximate locations. This demonstrates that RIV images effectively project geometric space into the descriptor embedding space. In contrast, other approaches generate multiple false positives at certain locations due to strong scene appearance similarity with the query, leading to closely related descriptors and an inability to distinguish places. For example, in the upper case of `Town01-O`, BEV images exhibit highly similar scene contours, highlighting the challenge of differentiating locations based solely on geometric structure. This emphasizes the value of multi-channel strategies like ImLPR, which enhance LPR by incorporating diverse inputs beyond mere geometry.