

Figure 8: Augmented image with yaw variation, square mask, cylindrical mask, and line mask. The yellow line indicates a leftward sign shift, equivalent to horizontal pixel shifting from yaw variation. Yellow boxes highlight augmented regions, showing changes from the original image. In the line mask, the white box denotes noise from empty line pixels due to projection, while our yellow box illustrates augmentation similar to the original.

7 Implementaiton Details

ImLPR is trained on the last two transformer blocks, with MultiConv adapters inserted every three blocks. LiDAR scans are projected into images of size $H \times W = 126 \times 1022$. The image channels include reflectivity, range, and normal values. Reflectivity serves as a semantic feature to segment the scene. For LiDARs lacking reflectivity, the intensity channel is used during inference, fulfilling a similar semantic role. The singular value ratio is computed using $k = 8$ nearest neighbors for the HeLiPR dataset and $k = 25$ for the MulRan and NCLT datasets. After normalizing these channels to a $[0, 1]$ range, they are fed into the ImLPR network. Feature aggregation employs parameters $(m, l, e) = (128, 64, 256)$ to form the descriptor. For Patch-InfoNCE loss, 192 positive and 128 negative patch pairs are sampled per image with a temperature $\tau_l = 0.2$. Negative patches are mined using patch distance thresholds $v_{dist} = 3$ and $h_{dist} = 20$. Positive and negative patches are extracted from $1/8$ of the image batch for computational efficiency. To align two scans from range images, scans are first voxelized with a 0.4-meter resolution, followed by fine alignment. For TSAP loss, a temperature $\tau_g = 0.01$ is used, truncated to the top four ranked descriptors, with a batch size of 2048. The combined loss is balanced with $\lambda = 2.0$. Training runs for 100 epochs using the AdamW optimizer with a learning rate of 5×10^{-4} , a $1/10$ warmup period, and a cosine scheduler, executed on three Nvidia GeForce RTX 3090 GPUs, completing in 12 hours.

8 Data Augmentation in ImLPR

To improve ImLPR’s robustness for LiDAR place recognition, range image view (RIV) images with reflectivity, range, and normal channels undergo multiple data augmentations during training. Images are subjected to random yaw rotation, uniformly sampled from 0 to 2π radians, implemented by cyclically shifting columns horizontally. This enhances invariance to orientation changes common in LiDAR scans. To address cylindrical image continuity, the leftmost and rightmost 28 columns (2 patches) are concatenated to the right and left sides, respectively, preserving connectivity during training and inference. Reflectivity and normal channels are normalized to $[0, 1]$ by dividing by 255, while the range channel is normalized by dividing by 200, ensuring consistent input scales.

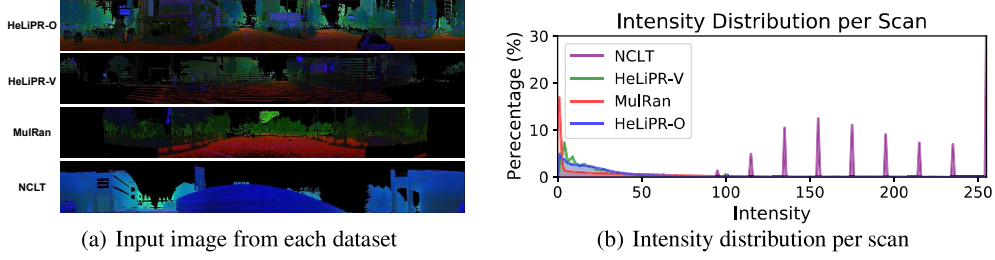


Figure 9: (a) Images from three datasets and two sensors. HeLiPR-O, used for training, exhibits high resolution with nearly no empty pixels. MulRan images show significant empty pixels in the leftmost and rightmost regions due to sensor occlusion. HeLiPR-V faces similar issues as MulRan, with sparse point clouds causing persistent empty pixels despite accumulation. Conversely, NCLT images have minimal empty pixels due to slow movement, but their intensity distribution is distinct, with discrete and reversed values, resulting in a bluer appearance. (b) Intensity distributions per scan for HeLiPR, MulRan, and NCLT, with NCLT displaying notable variations.

Three masking techniques are applied as shown in Fig. 8. Random patch masking uses square patches of random sizes, covering up to 40% of the image area based on a random mask ratio. Cylindrical masking applies a continuous mask, up to 30% of image width, with random start positions and cylindrical wrapping to handle boundary effects. These mitigate occlusion and sparse scene issues. Line-style masking introduces rectangular lines, placed randomly to simulate projection artifacts where multiple points map to a single pixel or horizontal lines appear empty due to RIV projection. These augmentations, paired with yaw-adjusted pose updates, enhance ImLPR’s resilience to orientation shifts, occlusions, and projection-related artifacts, contributing to its high performance in intra-session and inter-session place recognition.

9 Datasets

To validate ImLPR, we employ three datasets: HeLiPR, NCLT, and MulRan. Each dataset is represented using RIV images, as illustrated in Fig. 9(a).

9.1 HeLiPR Dataset

The HeLiPR dataset encompasses six places—Roundabout01-03, Town01-03, Bridge01-04, DCC04-06, KAIST04-06, and Riverside04-06—captured with four distinct LiDAR: Ouster OS2-128, Velodyne VLP-16C, Livox Avia, and Aeva Aeries II. Each sequence spans approximately 8.5 km, which is sufficiently long to identify multiple pairs for place recognition. For training, we use the Ouster OS2-128 sensor with the DCC04-06, KAIST04-06, and Riverside04-06 sequences, totaling 16,435 scans. The test set comprises the Roundabout01-03 and Town01-03 sequences from both the Ouster OS2-128 and Velodyne VLP-16C sensors, each containing 15,375 scans. Sequences are denoted as Sequence-Sensor (e.g., Roundabout-001 for Ouster, Roundabout-V01 for Velodyne) to differentiate sensor-specific data.

For the Roundabout-V sequence, we aggregate 5 seconds of data to create a submap, which is then projected into an RIV image. To prevent the continuous accumulation of identical scans, we excluded scans captured during stationary periods and accumulated only those from moving states. This approach mitigates the challenge of using lower-dimensional vertical images by employing a consistent dimensional vertical image, partially addressing the issue of sparse point clouds. As shown in Fig. 9(a), empty pixels still persist, posing challenges for LPR.

9.2 MulRan Dataset

The MulRan dataset encompasses four urban environments—DCC, KAIST, Riverside, and Sejong—each comprising three sequences (01-03) captured using an Ouster OS1-64 LiDAR. For this study, we utilize only the DCC01-03 sequences, which span an average distance of 4.9 km.

Table 6: Performance of High-Resolution LiDAR-Trained Models on Low-Resolution LiDAR

Method	NCLT (HDL-32E)								HeLiPR Roundabout-V (VLP-16C)								Total Avg	
	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	<u>0.817</u>	<u>0.920</u>	<u>0.830</u>	0.930	0.854	0.932	<u>0.834</u>	0.927	0.765	0.884	0.617	0.781	<u>0.753</u>	0.892	0.712	0.852	0.773	0.890

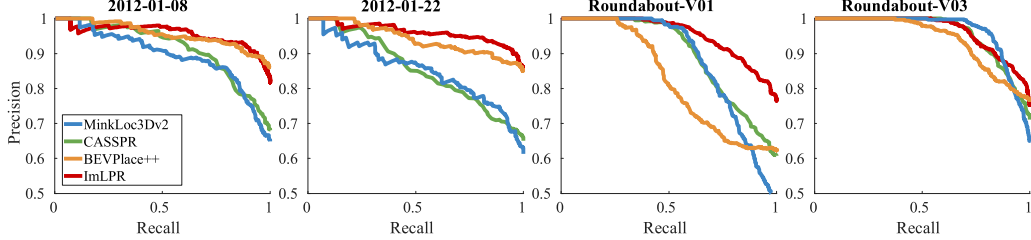


Figure 10: Precision-Recall curves for generalization evaluation on NCLT and HeLiPR Roundabout-V. ImLPR achieves consistent performance across both datasets. In contrast, BEVPlace++ exhibits strong performance on NCLT but lower performance on HeLiPR Roundabout-V. Additionally, the sparse 3D convolution-based methods, MinkLoc3Dv2 and CASSPR, show varying performance.

426 The dataset presents challenges due to low vertical density and occlusions caused by another sensor
 427 positioned behind the LiDAR, resulting in 2 – 30% of RIV image pixels being empty, as depicted in
 428 Fig. 9(a). Point clouds are projected into 1022×64 RIV images and resized to 1022×128 using
 429 linear interpolation. As the intensity values in the MulRan dataset exceed 255, we normalize them
 430 for use. A total of 4,328 scans are employed as test sets for generalization tasks and ablation studies.

431 9.3 NCLT Dataset

432 The NCLT dataset consists of 27 sequences captured in campus environments using a Velodyne
 433 HDL-32E LiDAR mounted on a Segway. Due to sparse point clouds, we aggregate scans following
 434 the approach used for the Roundabout-V sequence in HeLiPR. The Segway’s slower speed, compared
 435 to the vehicle-based system in HeLiPR, produces images with minimal empty pixels. Consequently,
 436 these images closely resemble those from the high-resolution training dataset, HeLiPR-O. However,
 437 the intensity distribution significantly deviates from other datasets, featuring discrete
 438 values and predominantly high-intensity points, unlike the typically low-intensity points in other
 439 datasets. This is evident in Fig. 9(b), where NCLT images appear blurrier than others. To address
 440 this, we invert and rescale the intensity distribution. Despite these adjustments, the inaccurate and
 441 discrete intensity channel continues to degrade the semantic quality of the RIV images, making place
 442 recognition more challenging. We employ the first three sequences, 2012-01-08, 2012-01-15, and
 443 2012-01-22, for generalization testing, utilizing a total of 6,368 scans as test sets.

444 10 Model Generalization Assessment - Sparse LiDAR

445 Following the intra-session place recognition evaluation conducted in the main paper, we further in-
 446 vestigate the generalization capability of the trained model using the HeLiPR-V and NCLT datasets.
 447 Consistent with the main paper, we evaluate performance through intra-session place recognition
 448 experiments, utilizing the HeLiPR-O trained model without additional training. We utilized the ac-
 449 cumulated scans as mentioned in §9. All methods are evaluated under the same configuration to
 450 ensure fairness.

451 10.1 Evaluation on NCLT Dataset

452 According to Table. 6, ImLPR and BEVPlace++ demonstrate comparable performance across the
 453 evaluated datasets. This similarity arises from inaccurate and discrete intensity values, which re-

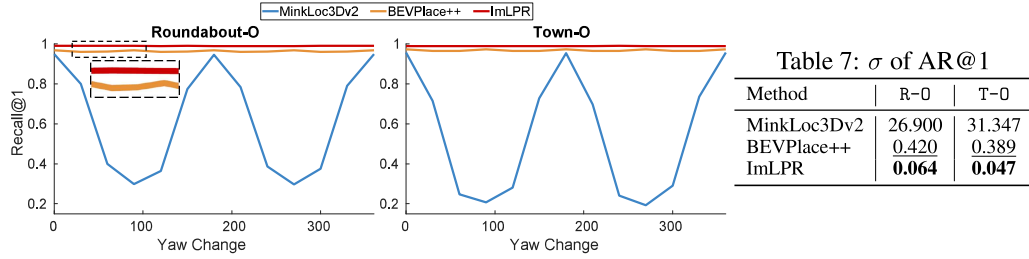


Figure 11: Average Recall@1 and its standard deviation (σ) for inter-session place recognition across yaw variations. ImLPR exhibits inherent yaw robustness, achieving the lowest σ .

duce pixel-level differences in RIV images and impede semantic reasoning. Nevertheless, ImLPR outperforms 3D sparse convolution-based methods, such as MinkLoc3Dv2, due to its effective use of geometric information from the range and normal channels. Despite ImLPR’s Recall@1 and F1 scores being slightly lower than those of BEVPlace++ for the 2012-01-08 sequence, the Precision-Recall curve shows that ImLPR’s performance is competitive with BEVPlace++ and achieves a higher Area Under the Curve (AUC) for the 2012-01-22 sequence, as depicted in Fig. 10. Moreover, despite the unreliable intensity channel, ImLPR achieves the highest average F1 score and the second-best average Recall@1, closely trailing BEVPlace++. These results underscore ImLPR’s robust generalization and its effectiveness in supporting place recognition when intensity data is unreliable.

10.2 Evaluation on HeLiPR-V Dataset

ImLPR outperforms other methods across most metrics, achieving the highest R@1 and F1 scores in Roundabout-V01 and Roundabout-V02, and performing competitively in Roundabout-V03, as shown in Table. 6. On average, ImLPR exhibits over 10% performance improvement over the second-best method, BEVPlace++, in both R@1 and F1 scores. Other methods display significant performance variations, largely due to sparse point clouds, which severely hinder their generalization. Notably, the second-best F1 scores across the three HeLiPR-V sequences are inconsistent, each achieved by a different method, underscoring the challenge of maintaining robust performance. The Precision-Recall curves in Fig. 10 reveal that for Roundabout-V03, ImLPR and MinkLoc3Dv2 achieve comparable AUC, followed by CASSPR and BEVPlace++. Conversely, for Roundabout-V01, ImLPR significantly surpasses other methods, with CASSPR, MinkLoc3Dv2, and BEVPlace++ following in that order. This demonstrates ImLPR’s consistent performance across the HeLiPR-V dataset.

Despite performance degradation across all datasets (MulRan, NCLT, and HeLiPR-V) due to domain shifts, sensor variations, inaccurate intensity values, and occlusions, ImLPR maintains robust performance under these challenging conditions. This resilience stems from its effective integration of geometric and visual cues, enabling reliable place recognition even with low-resolution or unstable inputs. These observations highlight the difficulty of achieving generalizability through conventional domain-specific training, while emphasizing ImLPR’s superior generalization capabilities. Furthermore, it suggests ImLPR’s potential as a versatile solution for generalized place recognition across diverse and challenging environments.

11 Robustness of Yaw Variation

11.1 Place Recognition Performance with Yaw Variation

Descriptors at the same location should remain consistent under transformations, particularly yaw angle changes, which are common when revisiting a scene. We evaluated this by applying yaw rotations to database scans from the Roundabout-0 and Town-0 datasets and comparing them to original queries. The results, averaged from inter-session place recognition, are shown in Fig. 11.

As shown in Fig. 11, MinkLoc3Dv2 shows performance degradation with varying yaw angles, indicating reduced robustness to such changes. BEVPlace++ uses a rotation-equivariant module to address yaw invariance. This module rotates the original image at fixed angular intervals x° , processes each rotated image through a CNN, and applies max pooling across the resulting features. This method improves robustness, though performance varies slightly when yaw variations do not exactly align with the discrete angles used for max pooling. However, ImLPR produces yaw variation robust descriptors without additional computation for it, due to its model’s architecture.

In RIV images, yaw variations manifest as horizontal shifts. DINOv2, a vision transformer, processes patch features through attention mechanisms, modeling global relationships between patches. Since a horizontal shift reorders patches without altering their content, the attention mechanism produces nearly translation-equivariant feature representations, as the relational structure remains largely consistent. As a result, vision transformers generate patch features that are shifted from the original patch features, along with consistent global tokens for both the original and shifted images. Similarly, the MultiConv adapter employs 2D convolutions, which are translation-equivariant, meaning a horizontal shift in the input image results in a corresponding shift in patch features, preserving their values. These refined patch features are aggregated using SALAD’s optimal transport method, which is invariant to horizontal shifts, as proven in §11.2. Data augmentation strategies, such as masking and random transformations, further enhance this stability. However, positional encoding introduces a minor limitation. As yaw rotation shifts the image, identical patches receive different positional encodings, causing slight differences in vision transformer before and after rotation. Despite this, ImLPR exhibits less performance variation than BEVPlace++, which incorporates explicit yaw invariance, as shown in Table. 7. This indicates that ImLPR, leveraging the model’s properties and RIV representation, provides robust yaw variation robustness.

11.2 Horizontal Shift Invariance in SALAD’s Optimal Transport Aggregation

In this section, we show that SALAD’s optimal-transport (OT) aggregation yields an identical global descriptor even when the input feature map is horizontally (column-wise) shifted. Let $\mathbf{F}' \in \mathbb{R}^{C \times H' \times W'}$ be the un-flattened feature map derived from adapter-refined patch features, with $n = H'W'$ patches. A convolutional module maps \mathbf{F}' to a score map $\mathbf{S}' \in \mathbb{R}^{m \times H' \times W'}$, which is flattened to $\mathbf{S} \in \mathbb{R}^{n \times m}$, representing assignment probabilities of n patches to m learnable cluster centers. The convolutional module is translation-equivariant, meaning a column shift in \mathbf{F}' (e.g., $\mathbf{F}'_{c,p,q} \rightarrow \mathbf{F}'_{c,p,(q-s)/W'}$) results in an identical column shift in \mathbf{S}' : $\mathbf{S}'_{m,p,q} \rightarrow \mathbf{S}'_{m,p,(q-s)/W'}$. After flattening, this translates to a row shift in \mathbf{S} : $\mathbf{S}_{i,:} \rightarrow \mathbf{S}_{(i-s')/n,:}$, where $s' = s \cdot H'$. This ensures that the score values for each patch are preserved, only their row indices are shifted, contributing to the shift invariance of the subsequent OT aggregation.

To compute OT in SALAD’s aggregation process, the Sinkhorn algorithm is employed. It is an efficient method for solving entropically regularized OT problems. The score matrix $\mathbf{S} \in \mathbb{R}^{n \times m}$, which represents assignment probabilities of n patches to m cluster centers, is first augmented with a dustbin column to form $\mathbf{S}_{\text{aug}} \in \mathbb{R}^{n \times (m+1)}$. This augmentation accounts for patches that may not be assigned to any cluster. The cost matrix is then defined as $\mathbf{M} = \mathbf{S}_{\text{aug}}/\lambda$, where λ is the regularization parameter controlling the entropy of the transport plan. A column shift in the input score map \mathbf{S}' propagates to a row shift in \mathbf{M} , and the Sinkhorn algorithm iteratively updates dual variables $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{v} \in \mathbb{R}^{m+1}$ to satisfy marginal constraints, using uniform log-weights $\log \mathbf{a}$ and $\log \mathbf{b}$. Since $\log \mathbf{a}$ is uniform across patches, the variable \mathbf{u} shifts identically under a column shift. Specifically, for patch i :

$$u_i = \log a_i - \log \sum_{k=1}^{m+1} \exp(M_{i,k} + v_k). \quad (5)$$

Shifting rows (e.g., $\mathbf{M}_{i,k} \rightarrow \mathbf{M}_{(i-s')/n,k}$) results in $u_i \rightarrow u_{(i-s')/n}$. The OT matrix is computed as:

$$\log R_{i,k} = M_{i,k} + u_i + v_k. \quad (6)$$

Table 8: Ablation study for three channel in RIV image

	Image Channel			Roundabout-0		Town-0		DCC		Average	
	CH1	CH2	CH3	AR@1	AF1	AR@1	AF1	AR@1	AF1	AR@1	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

Thus, the rows of $\log \mathbf{R}$ (and $\mathbf{R} = \exp(\log \mathbf{R})$) shift identically to \mathbf{S} , preserving the assignment weights. After removing the dustbin column, $\mathbf{R} \in \mathbb{R}^{n \times m}$. The aggregated cluster features are computed as a weighted sum of intermediate feature embeddings $\bar{\mathbf{F}} \in \mathbb{R}^{n \times l}$, obtained by applying a convolutional layer to \mathbf{F}' , flattened to n patches with l feature dimensions:

$$V_{j,k} = \sum_{i=1}^n R_{i,j} \bar{F}_{i,k}, \quad j = 1, \dots, m, \quad k = 1, \dots, l, \quad (7)$$

producing $\mathbf{V} \in \mathbb{R}^{m \times l}$. The feature matrix $\bar{\mathbf{F}}$ is also derived from \mathbf{F}' via a convolutional module, inheriting the same column shift: $\bar{\mathbf{F}}'_{l,p,q} \rightarrow \bar{\mathbf{F}}'_{l,p,(q-s)/W'}$, or $\bar{\mathbf{F}}_{i,:} \rightarrow \bar{\mathbf{F}}_{(i-s')/n,:}$ after flattening. Since both $\bar{\mathbf{F}}_{i,k}$ and $R_{i,j}$ shift identically, and summation is commutative, the aggregated features are invariant to column shifts:

$$\sum_{i=1}^n R_{(i-s')/n,j} \bar{F}_{(i-s')/n,k} = \sum_{i=1}^n R_{i,j} \bar{F}_{i,k}. \quad (8)$$

The global descriptor concatenates the normalized global token, which is processed independently via an MLP and is unaffected by shifts, with the flattened cluster features, ensuring horizontal shift invariance.

12 Ablation Studies

To evaluate the contribution of each component in ImLPR for place recognition, we conduct a series of ablation studies. Consistent with other evaluations, we report performance using Average Recall@1 and F1-score for inter-session place recognition.

12.1 Image Channels

ImLPR processes input images with three channels: reflectivity (intensity), range, and normal. To assess the impact of each channel on performance, we incrementally include channels in our experiments, starting with reflectivity alone (ExpA-1), followed by the addition of range (ExpA-2), and finally incorporating normal (ExpA-3). Note that the Patch-InfoNCE loss requires the range channel, so it is not evaluated independently in these experiments. The results, presented in Table 8, demonstrate that adding each channel increases the information available, leading to improved performance across all metrics on all datasets. Using reflectivity alone yields sufficiently accurate results; however, incorporating the range channel significantly enhances performance by providing geometric features. The normal channel, which is inferable from geometric features, improves performance, but less than the range channel improvement. This improvement stems from the normal channel's ability to capture relationships with neighboring points in 3D space, beyond mere pixel-wise interactions, enabling better distinction of boundaries and regions within the image. Thus, starting with reflectivity, which shares semantic characteristics with vision images, each added channel effectively contributes to the overall performance of ImLPR in place recognition.

12.2 MultiConv Adapter and the Number of Trained Block in DINOv2

Foundation models are leveraged to capitalize on their robust performance derived from training on large-scale datasets. To preserve their learned representations and avoid catastrophic forgetting,

Table 9: Ablation study for adapter and trained block

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

Table 10: Ablation study for yaw change and mask types

	Yaw Change	Line & Square Mask	Cylindrical Mask	Roundabout-0		Town-0		DCC		Roundabout-V		Average	
				AR@1	AF1	AR@1	AF1	AR@1	AF1	AR@1	AF1	AR@1	AF1
ExpC-1				0.937	0.972	0.943	0.973	0.905	0.966	0.781	0.893	0.892	0.951
ExpC-2	✓			0.980	0.990	0.958	0.980	0.896	0.983	0.821	0.927	0.914	0.970
ExpC-3	✓	✓		<u>0.984</u>	<u>0.992</u>	<u>0.975</u>	<u>0.989</u>	0.894	<u>0.985</u>	<u>0.837</u>	<u>0.928</u>	<u>0.923</u>	<u>0.974</u>
ExpC-4	✓	✓	✓	0.990	0.996	0.989	0.995	0.942	0.973	0.888	0.948	0.952	0.978

it is critical to maintain the pre-trained model’s integrity during adaptation. Fine-tuning multiple transformer blocks can adapt the model to a specific training domain but risks overwriting the general knowledge acquired from large datasets. In this section, we evaluate the impact of the number of trained transformer blocks and the role of the MultiConv adapter in this context.

As shown in Table. 9, the absence of the MultiConv adapter leads to a significant decline in place recognition performance. Due to the domain gap between natural vision images and RIV images, the model without both training and the adapter (ExpB-1) fails to perform effective place recognition. Similarly, fine-tuning without the adapter (ExpB-2) yields lower performance compared to models equipped with the adapter. This performance degradation stems from the model’s limited parameter capacity, which hinders its ability to extract robust features for descriptors from RIV images. In contrast, incorporating the MultiConv adapter achieves high performance while keeping the remaining transformer blocks frozen. This demonstrates that the MultiConv adapter efficiently mitigates domain shifts, leveraging the pre-trained representations of the foundation model to enhance place recognition.

Furthermore, when increasing the number of trained transformer blocks with the adapter in place, datasets such as Roundabout-0 and Town-0 exhibit consistent performance across all configurations. This suggests that the MultiConv adapter, with fine-tuning of a few transformer blocks, achieves effective domain adaptation to the LiDAR domain. However, for datasets like DCC and Roundabout-V, performance decreases as more transformer blocks are trained. This decline is attributed to the increased number of trainable parameters, which leads to overfitting to the training dataset, thereby reducing the foundation model’s ability to produce effective feature representations. Therefore, the MultiConv adapter is essential for balancing domain adaptation and the retention of pre-trained knowledge, enabling effective performance with minimal fine-tuning of transformer blocks.

12.3 Effect of Data Augmentation

The impact of data augmentation on enhancing ImLPR’s place recognition performance is assessed in Table. 10, which evaluates various augmentation configurations for inter-session place recognition. Without augmentations (ExpC-1), the model delivers baseline performance, exhibiting moderate capability but limited resilience to orientation shifts and projection artifacts inherent in RIV images.

Introducing random yaw rotation (ExpC-2) markedly enhances performance, indicating that column shifts effectively support robustness against orientation variability in LiDAR scans. The addition of combined line and square masking (ExpC-3) further improves resilience. Line and square masking are applied concurrently as a unified augmentation strategy, leveraging their similar pixel-level

Table 11: Ablation study for feature dimension and computation time

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

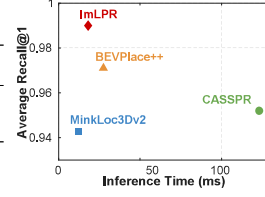


Figure 12: (Left) Table. 11 displays feature dimensions and inference times for ViT models, showing comparable performance despite larger dimensions. (Right) The plot illustrates inference time versus average Recall@1 on Roundabout-0 and Town-0, with ImLPR achieving the highest performance and an inference time comparable to MinkLoc3Dv2 for descriptors.

effects to introduce empty pixels in RIV images and simulate sparse or occluded regions. This approach mirrors the regularization effect of dropout by training the model to perform place recognition with only a subset of pixels, thereby enhancing its robustness to incomplete or sparse image data.

Optimal performance is achieved when all augmentations—random yaw rotation, line, square, and cylindrical masking—are combined (ExpC-4). The cylindrical mask significantly strengthens the model’s capacity to manage occlusions and sparse scenes. Similar as other masks, cylindrical mask enhances robustness by training the model to perform place recognition using only partial RIV images. This configuration yields superior results across all datasets, with particularly notable improvements on DCC and Roundabout-V, Which feature occlusions in their scans. These findings highlight the critical role of data augmentation in training, with the synergistic combination of yaw rotation and masking techniques enhancing ImLPR’s ability to address orientation variability, projection challenges, and scene sparsity in place recognition.

12.4 Dimension of DINOv2

The impact of varying the feature dimension of the DINOv2 backbone on ImLPR’s inter-session place recognition performance and computational efficiency is analyzed in Table. 11. This ablation study evaluates three configurations of DINOv2—ViT-S/14, ViT-B/14, and ViT-L/14—corresponding to feature dimensions of 384, 768, and 1024, respectively. These configurations are compared to assess their effectiveness in place recognition and their computational cost, with the last two transformer blocks trained for all models to ensure consistency in fine-tuning.

12.4.1 Performance Analysis of Place Recognition

The performance of ImLPR across different feature dimensions is presented in Table. 1, which compares place recognition results on multiple inter-session datasets. All three configurations—ViT-S/14 (ExpD-1), ViT-B/14 (ExpD-2), and ViT-L/14 (ExpD-3)—demonstrate comparable performance, achieving robust place recognition. This consistency suggests that increasing the feature dimension does not necessarily lead to proportional gains in performance. Notably, the smallest model, ViT-S/14, exhibits strong robustness, particularly on Roundabout-0, Town-0 and Roundabout-V, where it delivers highly competitive results. These findings suggest that larger models do not always translate to superior performance, as larger models can be more challenging to train effectively for the complex and variable characteristics of RIV images, potentially leading to overfitting on the training data.

12.4.2 Computational Cost

The computational efficiency of DINOv2 configurations is crucial for practical deployment in place recognition. As shown in Table. 11, descriptor extraction runtime increases with feature dimension: ViT-S/14 (ExpD-1) is the fastest, followed by ViT-B/14 (ExpD-2), while ViT-L/14 (ExpD-3) has the highest computational cost. As shown in Fig. 12, we further compare ImLPR with other baselines using average Recall@1 on Roundabout-0 and Town-0. BEVPlace++ requires 27.5ms

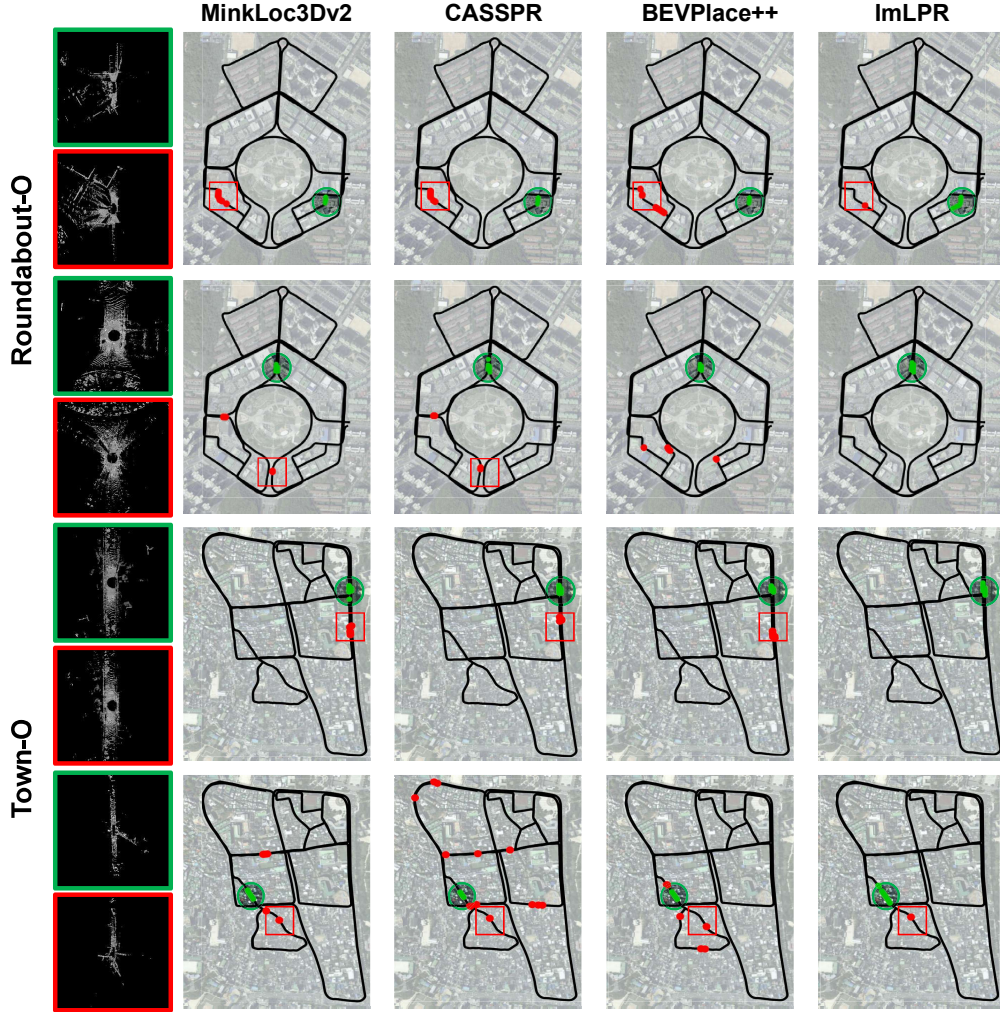


Figure 13: We retrieved 20 locations from Roundabout-001 and Town-001 using queries from Roundabout-002 and Town-002. Each image, overlaid with trajectory and satellite data, marks the query with a green circle, true positives with green points, and false positives with red points. Additionally, BEV images depict the sensing environments, highlighting the query location in green and false retrieved locations with a red square.

641 for descriptor extraction despite a smaller feature dimension of 128, significantly exceeding ViT-
642 S/14 and ViT-B/14, due to its use of multiple ResNet instances for yaw invariance. MinkLoc3Dv2,
643 relying on 3D sparse convolution, achieves the fastest extraction but with the lowest Recall@1 per-
644 formance. Conversely, CASSPR’s attention networks and per-point feature extraction result in the
645 longest computation time. In contrast, ImLPR’s Vision Transformer-based backbones, which also
646 ensure yaw invariance, deliver lower latency at 18.1ms with the highest Recall@1, as shown in
647 Fig. 12, highlighting the superiority of ImLPR’s real-time place recognition applications where low
648 latency is essential.

649 The trade-off between performance and computational cost is a key consideration in selecting the
650 optimal backbone. The results indicate that excessively large models, such as ViT-L/14, may in-
651 troduce unnecessary complexity without commensurate performance gains, potentially due to over-
652 fitting or increased training difficulty. By contrast, ViT-S/14 strikes an ideal equilibrium between
653 model size, computational speed, and place recognition accuracy, making it well-suited for practical
654 deployment.

655 13 Qualitative Results

656 In this section, we examine the retrieval distribution for inter-session place recognition, where a pos-
657 itive match is defined as within 50 meters, using Roundabout-001 and Town-001 as the database
658 and queries from Roundabout-002 and Town-002. As shown in Fig. 13, ImLPR retrieves almost
659 candidates nearer to the query than other methods, demonstrating that its descriptors maintain small
660 feature distances for both the top-1 match and nearby locations. This shows RIV images effectively
661 map geometric space into the feature embedding space. Conversely, other methods produce nu-
662 merous false positives at specific locations due to high scene appearance similarity with the query,
663 resulting in similar descriptors and failure to differentiate the places. For instance, in Town-001’s
664 upper case, BEV images reveal nearly identical scene contours, illustrating the difficulty of distin-
665 guishing locations using only geometric structure. This underscores the importance of multi-channel
666 approaches like ImLPR, which improve LPR by integrating diverse inputs beyond geometry.