DensePR: Dense 3D Geometry-aware Visual Place Recognition

***Abstract*—Place recognition, the task of localizing a vehicle by matching a current query against a global database, has recently focused on integrating geometric information with visual cues. One prevalent approach directly fuses visual features with encoded LiDAR point clouds, though these sensors inherently yield sparse geometric representations. Meanwhile, other methods aim to build structured scene representations to explicitly capture spatial geometries. Since these methods often suffer from reliance on LiDAR priors and limited scalability in novel environments, recent advances in monocular geometry estimation offer a promising alternative. In such approaches, both visual and geometric representations originate from the shared image source, ensuring inherent spatial alignment. The point clouds generated through this approach provide rich and dense geometric details, yet are often accompanied by inherent noise. Consequently, directly applying existing 3D encoders leads to degraded spatial representations and substantial computational costs due to the density, demanding a tailored encoding architecture. In this paper, we propose DensePR, a novel dense 3D geometry-based place recognition framework. To tackle the aforementioned limitations, we introduce the offset-shared voxel deformable attention module, which effectively and efficiently encodes noisy and dense point clouds by sharing spatial offsets across adjacent query points. Moreover, our multi-scope feature fusion module comprehensively captures the relationships across panoramic views to perform robust cross-modal fusion between 2D visual and 3D geometric cues. Experimental results on public outdoor and indoor benchmarks demonstrate that DensePR significantly outperforms existing fusion-based methods.**

***Index Terms*—Place Recognition, Monocular Geometry.**

I. Introduction

PLACE recognition (PR) is a crucial component of autonomous navigation, enabling vehicles to identify visited locations through a retrieval process between the current query observation and a global database. While cameras are widely used for visual place recognition (VPR) leveraging their rich semantic information [1], [2], visual representations remain highly sensitive to appearance variations caused by illumination and weather changes.

To ensure reliable recognition in complex environments, several recent studies have focused on integrating complementary camera and LiDAR features. While early methods like MinkLoc++ [3] and AdaFusion [4] focus on simple descriptor-level fusion, recent works explore fine-grained integration; LCPR [5] performs multi-scale feature fusion, and PRFusion [6] introduces local and global fusion modules. However, relying on an additional LiDAR sensor limits its applicability to camera-only platforms. Furthermore, they largely assume ideal cross-modal alignment, neglecting inherent calibration and projection errors, and fail to explicitly account for the varying spatial importance across the 360-degree views.

Meanwhile, some studies focus on building structured scene representations to directly capture spatial geometries for place recognition. For instance, BEV^{2}PR [7] extracts structural information by projecting image features into a bird's-eye-view space. GSPR [8] utilizes 3D Gaussian Splatting [9] to construct a unified scene representation from sequential multi-view images. While their advancements demonstrate that structurally-aware representations provide a robust performance, they still face practical limitations. BEV^{2}PR inevitably discards height information essential for distinguishing complex 3D environments. Moreover, GSPR relies on LiDAR priors and requires scene-specific Gaussian reconstruction, limiting scalability in large or evolving environments.

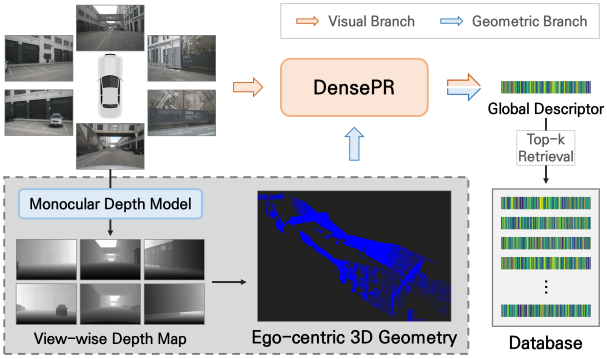

Fig. 1: **Overview of the proposed DensePR.** Unlike previous fusion-based methods relying on sparse and expensive LiDAR sensors, we adopt a pre-trained monocular depth model as an effective alternative to construct ego-centric 3D geometry from surround-view images. Subsequently, DensePR fuses features from both visual and geometric branches to extract a robust global descriptor for dense geometry-aware place recognition.

In this context, recent advances in monocular geometry estimation [10], [11] offer a promising alternative to construct a dense and fully structured 3D scene without relying on LiDAR. These methods enable the direct reconstruction of dense 3D point clouds from camera images, thereby ensuring inherent alignment with visual representations. However, most existing 3D encoders [12]–[14] are designed for sparse LiDAR data, which typically exhibit precise geometric measurements and relatively low point density. Directly applying such encoders reveals fundamental limitations. First, inherent depth noise leads to erroneous spatial correlations, misguiding attention between unrelated points. Second, the high density of the reconstructed points results in high computational cost.

To address these fundamental challenges, we propose a dense 3D geometry-aware visual place recognition method, namely DensePR. Figure 1 presents an overview of this framework leveraging 3D geometry from a monocular depth estimation model. DensePR is built upon two core components. First, we design the offset-shared voxel deformable attention (OVDA) module. The previous deformable attention method for sparse point clouds [13], which predicts independent offsets for each point, tends to be highly susceptible to local noise when applied to dense data. To address this issue, OVDA is specifically engineered to share deformed sampling offsets among voxels within the same 3D window. By effectively

forcing local regions to capture consistent structural context, this offset-sharing mechanism not only enforces structural stability against geometric noise but also reduces computational complexity. Second, to robustly integrate 2D visual and 3D geometric features, we introduce a multi-scope feature fusion module comprising two complementary streams. Specifically, local projection fusion (LPF) robustly associates 3D structures with 2D textures by dynamically aggregating local structural context, compensating for fine-grained misalignments caused by voxel quantization and calibration errors. Concurrently, direction-aware global fusion (DGF) captures long-range spatial correlations across the continuous panoramic view, dynamically assigning adaptive weights to each viewing direction based on its relative importance during global integration.

Our main contributions are summarized as follows:

- We propose DensePR, a place recognition method leveraging the structural context of dense point clouds from multi-view images via monocular geometry estimation.
- We design the OVDA module for encoding dense point clouds. It ensures structural stability and computational efficiency by sharing offsets among adjacent voxels.
- We introduce a multi-scope feature fusion strategy: LPF handles local integration by accounting for cross-modal spatial discrepancies, while DGF performs global fusion by reflecting the varying importance across views.
- We validate DensePR on diverse benchmarks, achieving state-of-the-art performance. Extensive ablation studies demonstrate the effectiveness of each proposed module.

II. Related Work

A. Fusion-based Place Recognition

Early fusion-based place recognition methods focus on descriptor-level integration. For instance, MinkLoc++ [3] concatenates modality-wise descriptors, while AdaFusion [4] dynamically weights them. Recently, LCPR [5] employs multi-scale fusion for cross-modal correlations, and PRFusion [6] combines metric attention with strict pixel-point projections. Despite their progress, these LiDAR-based methods assume perfectly aligned modalities, neglecting inevitable calibration and projection misalignments. Furthermore, they uniformly treat surrounding views, overlooking the varying spatial contributions across 360-degree directions. To address this, we propose a multi-scope fusion strategy comprising misalignment-aware local projection fusion (LPF) and view-adaptive direction-aware global fusion (DGF).

B. Geometry-aware Place Recognition

Early geometry-aware methods [15] apply direct sparse odometry [16] to extract semi-dense structures, but require continuous sequences and struggle in textureless regions. Alternatively, projecting 3D structures into 2D planes, such as LiDAR elevation images in CORAL [17] or bird's-eye-view features in BEV^{2}PR [7], inevitably discards crucial height information. Recently, GSPR [8] utilizes 3D Gaussian Splatting [9], but its scalability is constrained by reliance on LiDAR priors and scene-specific reconstruction. In contrast, DensePR leverages monocular geometry estimation [10], [11] to construct fully dense 3D geometries from multi-view images. Unlike prior works, these dense representations provide comprehensive spatial contexts that preserve both fine-grained structural details and height information, leading to highly robust place recognition.

C. Deformable Attention

Widely studied across various vision tasks, deformable attention enables flexible spatial aggregation by computing attention between queries and adaptively deformed keys. Following its success in 2D vision [18], [19] and 2D-to-3D lifting [20], PointSDA [13] recently applied it to sparse LiDAR point clouds. However, PointSDA predicts independent offsets for each individual query. When processing fully dense point clouds, this query-specific formulation incurs heavy computational overhead and renders the model vulnerable to local geometric noise. To resolve these issues, we propose the offset-shared voxel deformable attention (OVDA). By enforcing voxels within the same local 3D window to share predicted offsets, OVDA achieves strong robustness against geometric noise while ensuring high computational efficiency.

III. Method

A. Overall Architecture

As depicted in Figure 2, we propose DensePR, a framework that integrates visual and geometric features from multi-view images into a global descriptor for place recognition. As input, we take $N_{cam} = 6$ images $\mathcal{I} = \{I_1, I_2, \ldots, I_{N_{cam}}\}$ covering a 360-degree field of view. Our architecture employs two branches: a visual encoder [21] for 2D visual features, and a geometric encoder designed to process 3D structures. Specifically, we first construct a unified ego-centric dense point cloud from multi-view images via monocular depth estimation and extrinsic transformation (Section III-B). Because this 3D geometry is directly derived from the corresponding 2D images, it intrinsically ensures spatial alignment within each respective view without requiring explicit cross-modal alignment. This point cloud is then fed into the geometric branch, where it is encoded by our offset-shared voxel deformable attention (OVDA) module (Section III-C) to ensure robustness against depth noise. Finally, the multi-scope feature fusion module (Section III-D) integrates these multimodal features, which are then aggregated via a NetVLAD [1] layer to produce the global descriptor (Section III-E).

B. Multi-view Geometry Generation

To construct a consistent geometric representation of the 360-degree surroundings, we transform individual 2D images into a unified ego-centric 3D space utilizing multi-camera images. First, for each image I_i in the surround-view images, we estimate a pixel-wise depth map $D_i(u, v)$ using a pre-trained monocular depth estimation model [10]. Then, we unproject each pixel coordinate $\mathbf{u} = [u, v]^\top$ into a 3D point $\mathbf{p}_{cam,i}$ in its respective camera coordinate system using the camera's intrinsic matrix $\mathbf{K}_i$:

$$\mathbf{p}_{cam,i}(u, v) = D_i(u, v)\mathbf{K}_i^{-1}[u, v, 1]^\top. \tag{1}$$

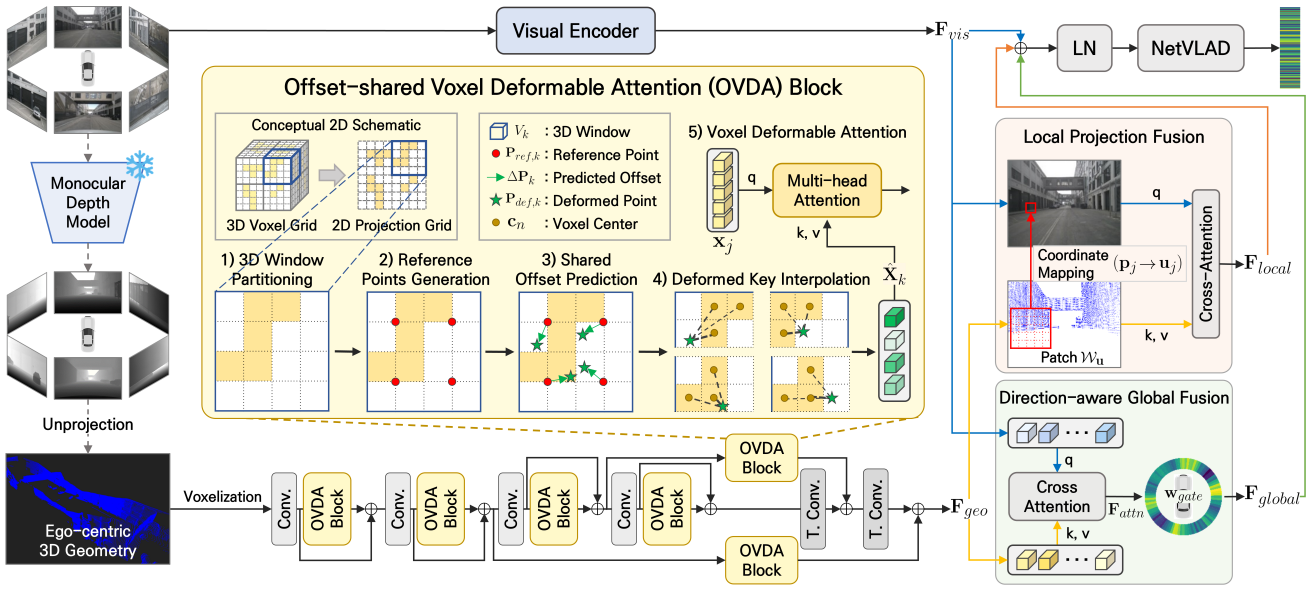

Fig. 2: The overall architecture of DensePR. Multi-view images are processed to extract visual features ($\mathbf{F}_{vis}$) and construct ego-centric 3D geometry via a monocular depth estimation model. The voxelized 3D geometry is fed into a backbone incorporating proposed OVDA blocks to extract geometric features ($\mathbf{F}_{geo}$). For clear presentation, the 3D operations within the OVDA block are conceptually illustrated on a 2D projection grid. Finally, $\mathbf{F}_{vis}$ and $\mathbf{F}_{geo}$ are adaptively combined via local projection fusion (LPF) and direction-aware global fusion (DGF), followed by subsequent aggregation via NetVLAD for place recognition.

To form a unified ego-centric 3D dense geometry, we project the points from all camera views into an ego-coordinate system. By applying the camera-to-ego extrinsic matrix $\mathbf{T}_{cam\rightarrow ego,i} = [\mathbf{R}_i \mid \mathbf{t}_i]$, we aggregate them into a global point cloud $\mathbf{P}_{raw} = \bigcup_{i=1}^{N_{cam}} \{\mathbf{R}_i \cdot \mathbf{p}_{cam,i} + \mathbf{t}_i\}$. To enhance computational efficiency, RANSAC-based plane fitting [22] is employed to remove the floor plane. The resulting refined dense 3D point cloud $\mathbf{P}_{final}$ is provided as the input for the geometric encoder.

C. Offset-shared Voxel Deformable Attention (OVDA)

To effectively encode 3D geometric information, we employ a U-Net style architecture with sparse convolutions [23], where the input point cloud $\mathbf{P}_{final}$ is voxelized in the Cartesian coordinate system. However, attention mechanisms in previous 3D point cloud networks [13], [14] tailored for sparse LiDAR inputs are not well-suited for processing the noisy and high-density point clouds derived from monocular geometry estimation. Directly applying such methods leads to erroneous spatial correlation and misguided attention between unrelated points due to inherent depth noise, while the high point density results in substantial computational cost.

To address these challenges, we propose offset-shared voxel deformable attention (OVDA), a specialized module tailored for dense point clouds and designed to achieve structural stability and computational efficiency. By integrating OVDA blocks at each stage where skip connections exist within our encoder, we process the input through the following steps:

1) 3D Window Partitioning: To capture local structural context, the OVDA block first organizes the sparse voxel space into localized 3D windows. For a given encoder stage, let $\mathbf{C} \in \mathbb{R}^{N\times 3}$ and $\mathbf{X} \in \mathbb{R}^{N\times C}$ be the input coordinates and geometric features for N sparse voxels. Although N and C vary across different encoder stages, we describe the operation for a single block for simplicity. Specifically, we partition the voxel grid into a set of non-overlapping 3D windows $\{V_k\}_{k=1}^{M}$, where M denotes the total number of partitioned 3D windows. Each 3D window V_k has a spatial resolution of $S_x \times S_y \times S_z$.

2) Reference Points Generation: Within each k-th 3D window V_k, we initialize a set of uniformly spaced 3D reference points $\mathbf{P}_{ref,k} \in \mathbb{R}^{N_{ref}\times 3}$ to serve as spatial anchors. To ensure these anchors represent the center of each partitioned spatial region, we define $N_{ref} = g^3$, where g is the number of points along each axis. For indices $a, b, c \in \{0, \ldots, g-1\}$, the local coordinates of a reference point $\mathbf{p}_{a,b,c}$ within the 3D window is defined as:

$$\mathbf{p}_{a,b,c} = \left(\frac{S_x}{g}(a+0.5), \frac{S_y}{g}(b+0.5), \frac{S_z}{g}(c+0.5)\right). \quad (2)$$

For notation simplicity, we flatten the triplet indices (a, b, c) into a single linear index $m \in \{1, \ldots, N_{ref}\}$, representing the m-th reference point in the k-th 3D window as $\mathbf{p}_{ref,k,m}$. These points are then shifted to the global coordinate. Serving as a structured spatial prior, these anchors provide base positions for the predicted deformable offsets.

3) Shared Offset Prediction: To robustly capture geometric structures from noisy point clouds, we introduce a shared offset prediction mechanism. In a previous deformable attention method for point clouds [13], independent offsets are predicted for every individual query. However, offsets predicted from individual noisy points are often unreliable, and sampling features at such numerous, independent locations incurs substantial computational cost.

To overcome this, our approach shifts to a 3D window-wise perspective. Instead of processing each voxel individually, our block aggregates local contexts to predict only a single, shared

set of offsets $\Delta\mathbf{P}_k \in \mathbb{R}^{N_{ref}\times 3}$ for each local 3D window V_k. Specifically, for each voxel $j \in V_k$ with its feature $\mathbf{x}_j$, lightweight linear layers generate a candidate offset proposal $\Delta\mathbf{p}_j \in \mathbb{R}^{N_{ref}\times 3}$. Concurrently, a scoring network maps $\mathbf{x}_j$ to a scalar confidence score s_j to evaluate the geometric reliability of each proposal. To dynamically emphasize reliable geometric anchors while reducing the influence of noisy points, the final shared offsets $\Delta\mathbf{P}_k$ for the 3D window V_k are computed directly as a weighted sum of all voxel-wise proposals using a 3D window-level softmax function:

$$\Delta\mathbf{P}_k = \sum_{j\in V_k} \alpha_j \Delta\mathbf{p}_j, \quad \text{where} \quad \alpha_j = \frac{\exp(s_j)}{\sum_{l\in V_k}\exp(s_l)}. \quad (3)$$

By applying these shared offsets across all voxels and attention heads within the local 3D window, we inherently mitigate the impact of individual point noise and reduce the computational cost of the feature sampling process. Finally, the deformed sampling points are obtained by shifting the initial reference points, computed as $\mathbf{P}_{def,k} = \mathbf{P}_{ref,k} + \Delta\mathbf{P}_k$.

4) Deformed Key Interpolation: Once the deformed sampling points $\mathbf{P}_{def,k} = \{\mathbf{p}_{def,k,m}\}_{m=1}^{N_{ref}}$ are determined, we retrieve their geometric features via inverse distance weighted (IDW) interpolation. Since these points lie in continuous 3D coordinates, we use a distance-weighted sum of the neighboring voxels to compute stable features. For each deformed point $\mathbf{p}_{def,m} \in \mathbf{P}_{def,k}$ (dropping index k for brevity), we identify its 3 nearest neighbors $\mathcal{N}_m$ and compute the interpolated feature $\hat{\mathbf{x}}_m$ using their coordinates $\mathbf{c}_n$ and features $\mathbf{x}_n$:

$$\hat{\mathbf{x}}_m = \sum_{n\in\mathcal{N}_m} w_{m,n}\mathbf{x}_n, \quad w_{m,n} = \frac{(d_{m,n}+\epsilon)^{-1}}{\sum_{l\in\mathcal{N}_m}(d_{m,l}+\epsilon)^{-1}}, \quad (4)$$

where $d_{m,n} = \|\mathbf{p}_{def,m} - \mathbf{c}_n\|_2$ and ϵ is a small constant. By interpolating only at the N_{ref} shared locations per each 3D window, we drastically reduce the complexity of K-NN searches and memory access compared to per-voxel sampling. The resulting features $\hat{\mathbf{X}}_k = \{\hat{\mathbf{x}}_m\}_{m=1}^{N_{ref}}$ serve as the deformed reference features, which are used to generate keys and values for the subsequent deformable attention stage.

5) Voxel Deformable Attention: In the final stage, we perform multi-head attention between the original sparse features as queries and the deformed geometric keys. For each voxel $j \in V_k$, a query $\mathbf{q}_j$ is projected from its feature $\mathbf{x}_j$, while the shared keys $\mathbf{k}_m$ and values $\mathbf{v}_m$ are projected from the interpolated 3D window features $\hat{\mathbf{X}}_k = \{\hat{\mathbf{x}}_m\}_{m=1}^{N_{ref}}$. Crucially, because the deformed sampling positions $\mathbf{P}_{def,k}$ are shared, all voxels within the 3D window V_k attend to the exact same set of N_{ref} key-value pairs using standard multi-head attention. By sharing the memory-intensive feature sampling and key-value projection steps across the entire 3D window, our module significantly reduces computational complexity. Finally, a linear projection aggregates the attention outputs to update the voxel feature $\mathbf{x}_j$, which are then aggregated across all 3D windows to form the final output of the OVDA block.

D. Multi-scope Feature Fusion

Let $\mathbf{F}_{vis} \in \mathbb{R}^{C_{vis}\times H\times W_{total}}$ and $\mathbf{F}_{geo} \in \mathbb{R}^{N_{geo}\times C_{geo}}$ denote the feature representations extracted from the visual and geometric encoders, respectively. Naively concatenating these heterogeneous features is suboptimal due to cross-modal spatial misalignments caused by minor extrinsic errors and varying importance across views. To effectively integrate them, we propose a multi-scope feature fusion module consisting of two components: local projection fusion (LPF) and direction-aware global fusion (DGF).

1) Local Projection Fusion (LPF): To associate 3D geometric structures with 2D visual features, we project the 3D coordinate $\mathbf{p}_j$ of each geometric feature onto the 2D panoramic image plane. For each geometric feature $\mathbf{f}_j^{geo}$ at 3D position $\mathbf{p}_j$, the projected 2D coordinate $\mathbf{u}_j$ on the i-th camera plane is computed as $\mathbf{u}_j = \Pi(\mathbf{K}_i\mathbf{R}_i^\top(\mathbf{p}_j - \mathbf{t}_i))$, where Π denotes the perspective projection function. This formulation maps ego-centric points back to the i-th camera frame using the inverse rotation $\mathbf{R}_i^\top$ and translation $\mathbf{t}_i$. While this projection establishes an explicit spatial mapping, voxel quantization and minor extrinsic errors can lead to slight spatial misalignments. To address this, we employ a patch-based cross-attention mechanism. For each visual pixel at coordinate $\mathbf{u}$, we define a local spatial patch $\mathcal{W}_\mathbf{u}$ centered at $\mathbf{u}$. Specifically, we perform cross-attention by setting the visual pixel feature $\mathbf{f}_\mathbf{u}^{vis}$ as the query. The keys and values are derived from the geometric features whose projected coordinates fall within this patch, defined as the set $\mathcal{S}_\mathbf{u} = \{\mathbf{f}_j^{geo} \mid \mathbf{u}_j \in \mathcal{W}_\mathbf{u}\}$. This patch-based attention allows the network to dynamically aggregate the most relevant structural context from its local neighborhood, generating the locally fused feature $\mathbf{F}_{local}$.

2) Direction-aware Global Fusion (DGF): DGF aims to globally integrate multimodal context along the continuous panoramic views. Specifically, we flatten $\mathbf{F}_{vis}$ into $H\times W_{total}$ spatial tokens to serve as the query, while the N_{geo} geometric tokens from $\mathbf{F}_{geo}$ act as the keys and values for a multi-head cross-attention module. This produces a global attention feature map $\mathbf{F}_{attn} \in \mathbb{R}^{C_{vis}\times H\times W_{total}}$, inherently capturing long-range spatial correlations across the panoramic width W_{total}. To dynamically assign higher weights to structurally important viewing directions within this global context, we introduce a circular direction-wise gating module. We perform vertical average pooling on $\mathbf{F}_{attn}$ and apply 1D circular convolutions to derive the direction-wise gating weights $\mathbf{w}_{gate} \in \mathbb{R}^{1\times W_{total}}$, capturing panoramic context. These weights are then element-wise multiplied with the attention output to yield the globally fused feature $\mathbf{F}_{global}$. This mechanism allows DensePR to focus on viewing directions that are more informative, while reducing the influence of ambiguous regions.

E. Global Descriptor Generation

The final stage of DensePR aggregates the multi-scope representations into a global descriptor. We integrate the visual feature $\mathbf{F}_{vis}$, the local geometric-aware feature $\mathbf{F}_{local}$, and the globally weighted feature $\mathbf{F}_{global}$ via a residual connection and layer normalization, formulated as $\mathbf{F}_{fused} = \text{LayerNorm}(\mathbf{F}_{vis} + \mathbf{F}_{local} + \mathbf{F}_{global})$. Subsequently, $\mathbf{F}_{fused}$ is aggregated through a NetVLAD [1] pooling layer.

F. Network Training

We adopt the training strategy from LCPR [5] and optimize our network in an end-to-end manner using the triplet margin loss. For each query q, we construct a mini-batch containing one positive sample p^q and multiple negative samples $\{n_i^q\}$. To mitigate overfitting and enhance training efficiency, we select the positive sample that has the smallest feature distance to the query. Furthermore, we retain informative negative samples that satisfy the filtering condition $\|\mathbf{f}_q - \mathbf{f}_{n_i^q}\|_2 < \tau$. Here, τ denotes a distance threshold used for hard negative mining, which is determined based on the positive sample distances and a predefined margin m. The network is then trained to enforce this margin between the positive and the $N_{neg} = 5$ retained negative pairs.

IV. Experiments

A. Experimental Setup

1) Datasets: For our primary evaluation, we utilize the nuScenes dataset [24]. Among the 1,000 driving sequences, we specifically leverage three subsets: Boston Seaport (BS), SG-OneNorth (SON), and SG-Queenstown (SQ). Based on the data preparation protocols from LCPR [5], the BS split is partitioned into a database along with training, validation, and test query sets. In contrast, the SON and SQ splits are divided solely into a database and a test query set. To evaluate performance under more challenging conditions, we follow [8], [25] by subsampling the database set and query set at 3m and 9m intervals, respectively.

Furthermore, to evaluate the model's capability in complex indoor spaces, we construct a new benchmark by subsampling the NAVER LABS indoor localization dataset [26]. This dataset contains environmental challenges, including illumination changes, occlusions from moving pedestrians, textureless surfaces, and repetitive architectural textures. Therefore, this dataset provides a suitable environment to evaluate the robustness of our dense geometry-based approach under visually challenging conditions. We leverage the database sequences captured by its six cameras. To minimize spatial redundancy within these sequences, we subsampled the data using a 2m translation and 5° rotation threshold. This process results in 2,455 training samples from the 1F session of a department store (2019-04-16_14-35-00) under standard lighting, and 1,608 evaluation samples from the B1 session of the department store (2019-04-16_15-35-46) under low-light conditions. Unlike the nuScenes dataset, we omit the RANSAC-based floor removal, as the floor plane is occasionally absent.

2) Evaluation Protocols: Following LCPR [5], we employ Recall@N (R@N) as the evaluation metric. For the nuScenes dataset, we conduct inter-session evaluation by matching queries against a database from a separate traversal of the same route. The ground-truth distance threshold for a successful match is set to 9m.

For the NAVER LABS indoor dataset, we conduct intra-session evaluation to assess loop closure detection capabilities. In this setup, each sequence serves as both query and database. Samples within 30 seconds of the query are excluded to prevent trivial matches. For this dataset, we utilize a distance threshold of 5m to account for the dense indoor environment.

3) Implementation Details: DensePR is implemented using PyTorch and MinkowskiEngine [23] by applying voxelization with a voxel size of [0.1m, 0.1m, 0.1m]. The visual branch employs an ImageNet-pretrained ResNet18 [21] backbone. The input consists of $N_{cam} = 6$ surround-view images, each resized to 384×512 pixels. For the multi-view geometry generation described in Section III-B, we utilize MoGe-2 [10] with a DINOv2-ViT-L backbone. For global aggregation, a NetVLAD [1] layer with 64 clusters generates a 256-dimensional descriptor. In the OVDA module, we set the localized 3D window resolution to $[32, 32, 32]$ and employ $N_{ref} = 27$ reference points per 3D window. We optimize the network using Adam with a learning rate of 1×10^{-5} on a single NVIDIA RTX A6000 GPU. The triplet loss configurations are tailored to each dataset. For the nuScenes dataset, we use a triplet margin $m = 0.5$ with positive and negative thresholds of 9m and 18m, respectively. In the case of the NAVER LABS indoor dataset, we apply $m = 0.1$ with thresholds of 5m and 10m.

B. Quantitative Results

To ensure a fair comparison, all methods extract a 256-dimensional descriptor. Additionally, the visual branches of NetVLAD [1], MixVPR [27], MinkLoc++ [3], LCPR [5], and our DensePR are implemented with a ResNet18 backbone.

1) Evaluation on Large-scale Outdoor Environments: In Table I, we present a comprehensive performance comparison on the nuScenes dataset. DensePR consistently outperforms state-of-the-art methods across all evaluated splits. Specifically, on the BS split, DensePR achieves a significant improvement in R@1, reaching 66.48% and outperforms the previous best fusion method AdaFusion [4] by 4.49%. Furthermore, on the SON split, DensePR yields the highest performance, significantly surpassing both single-modality methods [1], [28], [29] and multimodal fusion approaches [3]–[5] with an R@1 of 90.66%. These results indicate that our strategy of fusing visual and dense geometric features effectively captures more robust and distinct representations than conventional unimodal and LiDAR-based fusion approaches.

2) Evaluation on Challenging Indoor Environments: As shown in Table II, DensePR exhibits robust recognition capabilities on the NAVER LABS indoor dataset. It substantially outperforms AdaFusion [4] by an 11.3% margin in R@1 (87.0% vs 75.7%). Notably, existing LiDAR-dependent approaches [3], [4], [28], [29] exhibit limited performance, some of which underperform the vision-only method, NetVLAD [1]. This degradation is primarily attributed to the sparsity of the 16-channel LiDAR sensor, which fails to provide sufficient structural details. In contrast, DensePR effectively bypasses this hardware limitation by directly extracting dense geometric information from images, thereby ensuring highly robust recognition without relying on LiDAR.

C. Ablation Studies

We conduct comprehensive ablation studies to analyze the contributions of our proposed modules. All results are reported

TABLE I: Performance comparison for inter-session place recognition on the nuScenes dataset.

Method	Modality[1]	BS split			SON split			SQ split		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD [1]	V	58.63	76.02	81.63	84.85	94.19	95.71	80.49	89.33	90.55
MixVPR [27]	V	59.40	77.13	82.27	85.92	94.88	96.08	79.92	88.11	89.95
MinkLoc3D [28]	L	41.94	60.03	66.34	75.25	88.64	90.40	63.11	75.91	81.10
MinkLoc3Dv2 [29]	L	42.64	62.55	70.55	76.77	88.89	91.67	64.33	76.83	81.71
MinkLoc++ [3]	V+L	48.39	70.83	78.82	76.52	88.38	92.17	68.90	81.71	84.76
AdaFusion [4]	V+L	61.99	76.16	80.36	81.31	92.17	92.93	84.76	90.24	91.77
LCPR [5]	V+L	53.16	69.57	76.02	79.04	92.42	95.71	73.48	88.41	**92.07**
DensePR (Ours)	V+G	**66.48**	**84.57**	**90.18**	**90.66**	**96.97**	**97.73**	**85.37**	**90.55**	91.16

[1] V: Visual, L: LiDAR, G: points generated from monocular geometry estimation.

TABLE II: Performance comparison for intra-session place recognition on the NAVER LABS indoor dataset.

Method	Modality[1]	R@1	R@5	R@10
NetVLAD [1]	V	74.4	86.9	89.0
MinkLoc3D [28]	L	67.8	82.5	85.1
MinkLoc3Dv2 [29]	L	69.4	83.3	86.4
MinkLoc++ [3]	V+L	71.3	84.5	87.7
AdaFusion [4]	V+L	75.7	88.4	90.2
DensePR (Ours)	V+G	**87.0**	**95.2**	**96.6**

[1] V: Visual, L: LiDAR, G: points generated from monocular geometry estimation.

TABLE III: Ablation on the main components in DensePR.

OVDA	LPF	DGF	R@1	R@5	R@10
-	-	-	76.56	87.91	91.11
✓	-	-	78.00	88.95	91.76
✓	✓	-	78.43	88.06	91.47
✓	-	✓	79.69	89.94	92.61
-	✓	✓	77.60	88.37	92.10
✓	✓	✓	**80.83**	**90.70**	**93.02**

as the mean score across the three splits of the nuScenes dataset (i.e., BS, SON, and SQ).

1) Effect of Main Components: In Table III, we demonstrate the contributions of the proposed components in DensePR. The baseline (row 1) indicates the integration of both visual and geometric branches without our proposed modules. Incorporating OVDA improves the baseline by 1.44% in R@1 by extracting stable features from dense 3D geometry. To effectively integrate this geometric representation with 2D visual features, our multi-scope feature fusion module utilizes LPF for local spatial association and DGF for global panoramic context, each contributing distinctively to the overall performance. Notably, integrating OVDA into the configuration employing both fusion modules leads to a significant performance gain from 77.60% to 80.83%.

2) Analysis of Attention Mechanisms: We compare our OVDA module with other 3D point cloud attention mechanisms, as illustrated in Table IV. To ensure a fair comparison, we use the same place recognition network while varying only the attention mechanism. Specifically, we evaluate the overall performance and the computational efficiency within the attention blocks. Swin3D [14] performs self-attention among all points within a cubic 3D window, leading to a substantial computational burden for dense point clouds. Compared to Swin3D, our OVDA significantly reduces the FLOPs by 39.4%

TABLE IV: Ablation on different attention mechanisms.

Attention Mechanism	R@1	R@5	R@10	FLOPs[1] (G)	#Param[2] (M)
Swin3D [14]	77.60	88.37	92.10	4.879	**2.23**
PointSDA [13]	77.25	87.20	90.30	3.258	2.34
OVDA (Ours)	**80.83**	**90.70**	**93.02**	**2.955**	2.75

[1] FLOPs exclude projection layers to highlight computations unique to each strategy.
[2] #Param reflects the total parameter count within the attention blocks.

TABLE V: Robustness to local geometric noise, simulated by applying Gaussian noise (σ) to a random 10% of points. Values in $(\cdot)$ indicate R@1 drops relative to the baseline ($\sigma = 0.0$).

Attention Mechanism	Local Noise Scale (σ)			
	0.0	0.1	0.2	0.3
PointSDA [13]	77.25	76.32 (↓ 0.93)	75.54 (↓ 1.71)	74.80 (↓ 2.45)
OVDA (Ours)	**80.83**	**80.37 (↓ 0.46)**	**80.09 (↓ 0.74)**	**79.59 (↓ 1.24)**

(4.879G → 2.955G). Furthermore, PointSDA [13] calculates individual offsets for each query point, making it vulnerable to noisy points and resulting in a low performance of 77.25% in R@1. Despite a marginal increase in parameter count within the attention blocks (2.75M), OVDA robustly handles dense point clouds while maintaining low computational cost, achieving the highest R@1 score of 80.83%.

3) Robustness to Local Geometric Noise: In Table V, the proposed method exhibits robustness against local geometric noise. When substantial noise ($\sigma = 0.3$) is injected into the 3D coordinates of 10% of the points, PointSDA [13] suffers an R@1 degradation of 2.45%, whereas our OVDA suppresses this drop to merely 1.24%. This gap stems from their offset prediction mechanisms. PointSDA predicts independent offsets for each query, making it highly vulnerable to corrupted coordinates. Conversely, our OVDA computes shared offsets for each 3D window via a confidence-weighted sum of voxel proposals. By dynamically emphasizing informative geometric structures, this mechanism minimizes the influence of individual noisy points and preserves robust spatial aggregation.

4) LPF Patch Size Analysis: In Table VI, we evaluate the performance across various patch sizes ($\mathcal{W}_\mathbf{u}$) under random rotational calibration noise. Interestingly, applying a localized 1×1 patch, as in PRFusion++ [6], underperforms the baseline that excludes the LPF. This degradation is primarily attributed to spatial misalignments from voxelization and extrinsic calibration errors among the multi-view cameras. This vulnerability is critically exposed by the injected rotational noise, where

TABLE VI: Robustness to rotational calibration noise across different LPF patch sizes. R@1 is reported and values in red denote performance drops compared to the 0° reference.

LPF	Patch Size ($\mathcal{W}_\mathbf{u}$)	Rotational Noise		
		0°	1°	3°
-	-	79.69	77.54 (↓ 2.15)	74.80 (↓ 4.89)
✓	1×1	79.52	77.70 (↓ 1.82)	74.37 (↓ 5.15)
	3×3	80.63	79.40 (↓ 1.23)	76.32 (↓ 4.31)
	5×5	**80.83**	**80.01 (↓ 0.82)**	**77.38** (↓ 3.45)
	7×7	80.18	79.30 (↓ 0.88)	76.90 (↓ **3.28**)

TABLE VII: Ablation on Direction-wise Weighting in DGF.

Direction-wise Weighting	R@1	R@5	R@10
-	79.28	89.51	92.40
✓	**80.83**	**90.70**	**93.02**

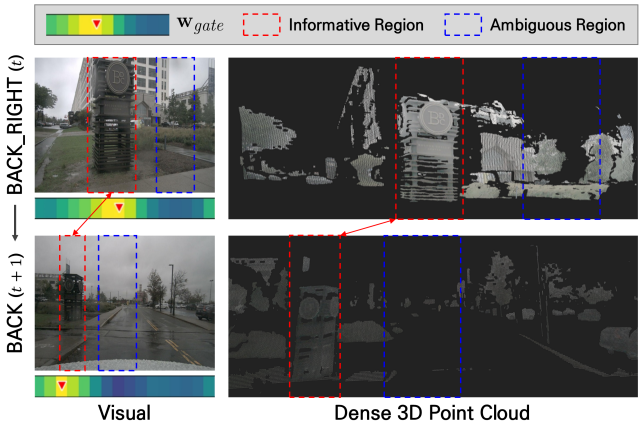

Fig. 3: Visualization of the Direction-wise gating weights ($\mathbf{w}_{gate}$) generated by the DGF module. The heatmaps illustrate that DGF dynamically assigns higher weights to highly discriminative viewing directions (red boxes), effectively reducing the influence of ambiguous areas (blue boxes) within the global context.

employing a 1×1 patch results in the highest performance drop of 5.15% compared to the baseline under 3° rotational noise. By expanding the patch size, the network effectively mitigates these misalignments by aggregating extensive local spatial context. Consequently, we adopt the 5×5 patch for our default configuration as it yields the best performance of 80.83% while providing robustness against calibration noise.

5) Analysis of Direction-wise Weighting in DGF: We investigate the contribution of the direction-wise gating within DGF. As shown in Table VII, applying the gating weights ($\mathbf{w}_{gate}$) yields consistent performance improvements across all evaluation metrics. This result indicates that viewing directions should be weighted differently according to their discriminative importance. To qualitatively demonstrate this effect, we visualize the predicted $\mathbf{w}_{gate}$ for consecutive frames ($t \rightarrow t+1$) in Figure 3. The heatmaps illustrate that DGF actively emphasizes highly discriminative viewing directions, successfully capturing informative regions such as distinct landmarks (red boxes). Conversely, it effectively mitigates the impact of ambiguous or featureless areas, such as plain roads (blue boxes). As a result, this adaptive weighting mechanism yields a more discriminative global representation.

TABLE VIII: Quantitative analysis of metric depth rescaling for place recognition.

Depth Model	Rescaling	Depth Quality		PR Score		
		A.Rel↓	δ_1↑	R@1	R@5	R@10
MoGe-2 [10]	-	0.232	0.621	80.83	90.70	93.02
	Random[1]	0.401	0.354	79.36	90.08	93.04
	w/ LiDAR[2]	**0.156**	**0.824**	**81.36**	**91.85**	**93.61**

[1] Multiplied by a random scale factor within $[0.5, 2.0]$ per timestamp.
[2] Rescaled using LiDAR following the method of Marsal *et al.* [30].

TABLE IX: Performance comparison of a lightweight depth model for place recognition.

Exp. #	Depth Model	Supervision	Depth Quality		PR Score	
			A.Rel↓	RMSE↓	R@1	R@5
1	BinsFormer [31] (Res50)	LiDAR	0.255	8.482	78.39	89.37
2		MoGe-2 [10]	0.253	7.843	78.64	89.27
3		MoGe-2* [10]	**0.230**	7.334	79.26	**90.74**
4	MoGe-2 [10]	*Zero-shot*	0.232	**7.060**	**80.83**	90.70

MoGe-2*: Rescaling method of Marsal *et al.* [30] applied to raw MoGe-2 outputs.

D. Impact of Depth Estimation Quality on Place Recognition

We analyze the impact of monocular depth estimation quality on the overall place recognition (PR) performance. Depth estimation quality is evaluated using absolute mean relative error (A.Rel), root mean square error (RMSE), and the ratio of inlier pixels (δ_i) with threshold 1.25^i. In Table IX, BinsFormer [31] is trained using non-overlapping samples from the BS split that are excluded from our PR experiments. Consequently, all depth quality evaluations in this section are reported as the average performance of the SON and SQ splits.

1) Correlation between Metric Depth Rescaling and PR: To investigate the impact of the absolute metric scale, we evaluate DensePR using both random scale perturbations ($[0.5, 2.0]$) and LiDAR-based rescaling. As shown in Table VIII, although random scaling severely degrades the depth inlier ratio (δ_1) to 0.354, the performance decrease is marginal, with R@5 dropping by only 0.62%. Conversely, LiDAR-based rescaling only slightly improves R@1 to 81.36% despite significantly boosting δ_1 to 0.824. These results demonstrate that DensePR prioritizes relative structural geometry over absolute metric distances, validating its robust effectiveness without reliance on precise scale.

2) Lightweight Depth Model for Geometry Generation: Considering practical applicability, we evaluate PR performance using multi-view geometry generated by a lightweight depth model, specifically BinsFormer [31], utilizing a ResNet50 backbone. As shown in Table IX, BinsFormer supervised by the rescaled MoGe-2 outputs (Exp. 3) during the training phase significantly outperforms the versions trained with sparse LiDAR (Exp. 1) or raw MoGe-2 outputs (Exp. 2). Notably, BinsFormer in Exp. 3 reduces A.Rel to 0.230 and achieves an R@5 of 90.74%, slightly surpassing the zero-shot performance of the MoGe-2 (90.70%, Exp. 4). These results demonstrate that refined training supervision enables

lightweight depth estimation models to achieve competitive place recognition performance comparable to large-scale models without any LiDAR dependency during inference.

V. Conclusion

In this paper, we propose DensePR, a framework for place recognition that leverages the fusion of visual features and dense ego-centric 3D geometry generated via monocular depth estimation. By introducing the OVDA block, DensePR effectively processes these dense and noisy 3D structures. Furthermore, DensePR integrates visual and geometric features through LPF and DGF to build robust global descriptors. Extensive experiments on public datasets demonstrate that our approach, utilizing dense 3D geometry, outperforms existing fusion-based methods.

However, the explicit extraction of 3D geometry via an independent depth model inevitably introduces computational overhead compared to conventional baselines. Additionally, the generated depth maps are inherently susceptible to motion blur and occlusion, which can degrade the overall recognition performance. Nevertheless, we have demonstrated that dense 3D representations offer a significant advantage over inherently sparse geometries for place recognition, highlighting promising directions for future research to address these remaining computational and environmental constraints.

References

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[2] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," *IEEE Robot. Automat. Lett.*, vol. 9, no. 2, pp. 1286–1293, 2023.

[3] J. Komorowski, M. Wysoczańska, and T. Trzcinski, "Minkloc++: lidar and monocular image fusion for place recognition," in *Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.

[4] H. Lai, P. Yin, and S. Scherer, "Adafusion: Visual-lidar fusion with adaptive weights for place recognition," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 12 038–12 045, 2022.

[5] Z. Zhou, J. Xu, G. Xiong, and J. Ma, "Lcpr: A multi-scale attention-based lidar-camera fusion network for place recognition," *IEEE Robot. Automat. Lett.*, vol. 9, no. 2, pp. 1342–1349, 2024.

[6] S. Wang, Q. Kang, R. She, K. Zhao, Y. Song, and W. P. Tay, "Prfusion: Toward effective and robust multi-modal place recognition with image and point cloud fusion," *IEEE Trans. Intell. Transport. Syst.*, vol. 25, no. 12, pp. 20 523–20 534, 2024.

[7] F. Ge, Y. Zhang, S. Shen, W. Hu, Y. Wang, and J. Gao, "Bev^{2}pr: Bev-enhanced visual place recognition with structural cues," in *IEEE Int. Conf. Intell. Robots Syst.*, 2024, pp. 13 274–13 281.

[8] Z. Qi, J. Ma, J. Xu, Z. Zhou, L. Cheng, and G. Xiong, "Gspr: Multimodal place recognition using 3d gaussian splatting for autonomous driving," in *IEEE Int. Conf. Intell. Robots Syst.*, 2025, pp. 8864–8871.

[9] B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis, *et al.*, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–14, 2023.

[10] R. Wang, S. Xu, Y. Dong, Y. Deng, J. Xiang, Z. Lv, G. Sun, X. Tong, and J. Yang, "Moge-2: Accurate monocular geometry with metric scale and sharp details," *arXiv preprint arXiv:2507.02546*, 2025.

[11] H. Lin, S. Chen, J. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang, "Depth anything 3: Recovering the visual space from any views," *arXiv preprint arXiv:2511.10647*, 2025.

[12] X. Lai, Y. Chen, F. Lu, J. Liu, and J. Jia, "Spherical transformer for lidar-based 3d recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17 545–17 555.

[13] X. Sheng, Z. Shen, and G. Xiao, "Pointsda: Spatio-temporal deformable attention network for point cloud video modeling," *IEEE Robot. Automat. Lett.*, vol. 9, no. 12, pp. 10 946–10 953, 2024.

[14] Y.-Q. Yang, Y.-X. Guo, J.-Y. Xiong, Y. Liu, H. Pan, P.-S. Wang, X. Tong, and B. Guo, "Swin3d: A pretrained transformer backbone for 3d indoor scene understanding," *Comput. Vis. Media*, vol. 11, no. 1, pp. 83–101, 2025.

[15] A. Oertel, T. Cieslewski, and D. Scaramuzza, "Augmenting visual place recognition with structural cues," *IEEE Robot. Automat. Lett.*, vol. 5, no. 4, pp. 5534–5541, 2020.

[16] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 40, no. 3, pp. 611–625, 2017.

[17] Y. Pan, X. Xu, W. Li, Y. Cui, Y. Wang, and R. Xiong, "Coral: Colored structural representation for bi-modal place recognition," in *IEEE Int. Conf. Intell. Robots Syst.*, 2021, pp. 2084–2091.

[18] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," in *Int. Conf. Learn. Represent.*, 2021, pp. 1–11.

[19] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4794–4803.

[20] H. Li, H. Zhang, Z. Zeng, S. Liu, F. Li, T. Ren, and L. Zhang, "Dfa3d: 3d deformable attention for 2d-to-3d feature lifting," in *IEEE Int. Conf. Comput. Vis.*, 2023, pp. 6684–6693.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[22] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[23] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3075–3084.

[24] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11 621–11 631.

[25] X. Xu, S. Lu, J. Wu, H. Lu, Q. Zhu, Y. Liao, R. Xiong, and Y. Wang, "Ring++: Roto-translation invariant gram for global localization on a sparse scan map," *IEEE Trans. Robot.*, vol. 39, no. 6, pp. 4616–4635, 2023.

[26] D. Lee, S. Ryu, S. Yeon, Y. Lee, D. Kim, C. Han, Y. Cabon, P. Weinzaepfel, N. Guérin, G. Csurka, *et al.*, "Large-scale localization datasets in crowded indoor spaces," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3227–3236.

[27] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "Mixvpr: Feature mixing for visual place recognition," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 2998–3007.

[28] J. Komorowski, "Minkloc3d: Point cloud based large-scale place recognition," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1790–1799.

[29] J. Komorowski, "Improving point cloud based place recognition with ranking-based loss and large batch training," in *Int. Conf. Pattern Recognit.*, 2022, pp. 3699–3705.

[30] R. Marsal, A. Chapoutot, P. Xu, and D. Filliat, "A simple yet effective test-time adaptation for zero-shot monocular metric depth estimation," in *IEEE Int. Conf. Intell. Robots Syst.*, 2025, pp. 7858–7865.

[31] Z. Li, X. Wang, X. Liu, and J. Jiang, "Binsformer: Revisiting adaptive bins for monocular depth estimation," *IEEE Trans. Image Processing*, vol. 33, pp. 3964–3976, 2024.