SCENEMOE-VPR: SCENE-ADAPTIVE MIXTURE-OF-EXPERTS FOR VISUAL PLACE RECOGNITION ACROSS HETEROGENEOUS SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual Place Recognition (VPR) aims to retrieve reference images depicting the same location as a query image. Most existing methods are trained for specific scenes and remain limited in their ability to generalize across heterogeneous scenes. Training VPR models on diverse datasets is a natural route toward robust recognition across heterogeneous environments. However, we observe that extending joint training with indoor data improves indoor recognition but degrades several outdoor benchmarks under fully shared adaptation, revealing a cross-environment trade-off. To address this issue, we propose **SceneMoE-VPR**, a heterogeneous Mixture-of-Experts framework for input-dependent adaptation. It keeps the DINOv2 backbone frozen and adapts its features through an always-active shared expert and sparsely routed heterogeneous experts designed to model viewpoint-induced geometric variation, local structure, directional layout, and multi-scale context. An image-level router dynamically selects experts for each input without scene labels, enabling input-dependent adaptation across heterogeneous environments. Under the same four-dataset training setting, SceneMoE-VPR consistently outperforms the fully shared baseline on all four benchmarks, improving average R@1 from 88.9% to 89.5%. It further achieves the highest average R@1 among the compared single-image methods with a compact 2,048-dimensional descriptor. These results demonstrate the effectiveness of conditional heterogeneous adaptation for multi-environment VPR across diverse environments. Code will be released publicly.

1 INTRODUCTION

Visual place recognition (VPR) aims to recognize the place depicted in a query image by matching it with reference images captured at previously visited locations (Lowry et al., 2015; Schubert et al., 2023), and serves as a fundamental component of visual localization systems for mobile robots (Xu et al., 2020) and autonomous vehicles (Bresson et al., 2017). Most VPR methods formulate this task as image retrieval (Cao et al., 2020), representing each image with a compact global descriptor and retrieving visually similar reference images through descriptor matching. Although recent VPR methods (Ali-Bey et al., 2024; Lu et al., 2025b;a; Zhang et al., 2026) have achieved strong performance on widely used benchmarks, particularly those collected in urban environments, their performance often varies substantially across different types of scenes. Achieving consistently strong performance across heterogeneous environments remains a significant challenge, yet is critical to the reliable long-term robotic deployment, whose localization systems must remain robust as they transition between these different operating environments. Developing a single VPR model that remains discriminative across diverse environments is therefore an important step toward universal and reliable visual place recognition.

Achieving both task-specific discriminability and cross-environment generalization remains challenging for existing VPR approaches. Training-free methods such as AnyLoc (Keetha et al., 2023) leverage pretrained DINOv2 features (Oquab et al., 2023) together with unsupervised feature aggregation, enabling broad transfer across diverse environments without VPR-specific training, but typically with lower recognition accuracy than task-specific supervised methods. In contrast, VPR-specialized methods learn discriminative place representations through supervised training (Arand-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

jelovic et al., 2016; Berton et al., 2022a; Ali-Bey et al., 2023), often on large-scale urban street-view datasets, and achieve strong performance on benchmarks similar to their training environments. To reduce the dependence on a single training distribution and improve generalization across environments, recent methods have increasingly explored joint training on multiple VPR datasets. QAA (Xiao et al., 2025) and SelaVPR++ (Lu et al., 2025a) exemplify this direction by incorporating multiple large-scale VPR datasets during training. While such multi-dataset training further improves performance across outdoor benchmarks, generalization to indoor environments remains considerably more challenging. A straightforward extension is therefore to incorporate indoor data into joint training. However, as shown in Table 2, adding indoor training data substantially improves indoor recognition, but the strong performance on several outdoor benchmarks cannot be fully preserved. This observation suggests that simply increasing training diversity does not necessarily lead to uniformly better performance across heterogeneous environments.

A natural way to alleviate this trade-off is to make the adaptation process input-dependent, allowing different samples to activate different parameter subsets while retaining a shared, frozen backbone. Mixture-of-Experts (MoE) architectures (Jacobs et al., 1991; Jordan & Jacobs, 1994) provide a suitable mechanism through dynamic routing and expert specialization. However, many conventional MoE architectures (Riquelme et al., 2021; Fedus et al., 2022) employ structurally identical experts that differ mainly in their parameters. While this supports input-dependent specialization, it may still be restrictive for multi-domain VPR, where heterogeneous inputs may benefit from different forms of spatial adaptation rather than different parameterizations of the same operator. This motivates us to construct an MoE with structurally diverse experts, allowing the routing mechanism to select different combinations of spatial transformations according to the input.

To this end, we propose SceneMoE-VPR, a scene-adaptive heterogeneous Mixture-of-Experts framework for learning a single VPR model across heterogeneous environments. Keeping the DINOv2 backbone frozen, SceneMoE-VPR introduces a parallel scene-adaptive network built upon heterogeneous MoE adapters that enable input-dependent feature adaptation while avoiding costly backbone fine-tuning. Specifically, at each layer, the adapter combines an always-active shared expert that captures common VPR knowledge with sparsely activated spatial experts that accommodate input-specific adaptation demands. Rather than employing structurally identical experts, we design four spatial experts with distinct architectures to model viewpoint-induced geometric variation, local structure, directional layout, and multi-scale context. For each image, an image-level router takes the dimension-reduced representation of the adapter input and selects the top-K heterogeneous experts. This image-level sparse routing enables input-dependent selection of expert combinations without requiring scene or dataset labels.

Our main contributions are summarized as follows:

- We empirically identify a limitation of fully shared adaptation for heterogeneous VPR: extending joint training with indoor data improves indoor recognition but can degrade outdoor performance, limiting the ability of a single model to generalize consistently across diverse environments. This observation suggests the need for input-dependent adaptation across diverse environments.
- We propose SceneMoE-VPR, a parameter-efficient heterogeneous Mixture-of-Experts framework that combines shared adaptation for common VPR knowledge with sparsely routed heterogeneous experts for input-specific adaptation. The experts capture complementary geometric, local, directional, and multi-scale spatial cues and are dynamically selected without scene labels.
- Under the same four-dataset joint-training setting, SceneMoE-VPR consistently outperforms the fully shared baseline across outdoor and indoor benchmarks. Ablations further validate the complementary roles of the heterogeneous experts and shared expert.

2 Related Work

2.1 Visual Place Recognition and Cross-Environment Generalization

Visual place recognition is commonly formulated as an image retrieval problem, where a query image is matched against a geo-tagged reference database using compact global descriptors. Early

deep VPR methods focused on learning task-specific representations and aggregation mechanisms, exemplified by NetVLAD (Arandjelovic et al., 2016). Subsequent methods improved either the training objective or feature aggregation to obtain more discriminative and efficient global descriptors. CosPlace (Berton et al., 2022a) reformulates large-scale VPR training as classification, EigenPlaces (Berton et al., 2023) improves robustness to viewpoint changes through viewpoint-aware training, while MixVPR (Ali-Bey et al., 2023), SALAD (Izquierdo & Civera, 2024), BoQ (Ali-Bey et al., 2024), and VLAD-BuFF (Khaliq et al., 2024) introduce increasingly expressive feature aggregation mechanisms.

These supervised methods achieve strong performance on established VPR benchmarks, but their generalization can remain sensitive to the distributions represented during training. To overcome such task- and environment-specificity, large-scale self-supervised vision models provide a source of general-purpose visual representations for VPR. AnyLoc (Keetha et al., 2023) demonstrates that such representations can support place recognition across a wide range of structured and unstructured environments without VPR-specific training or fine-tuning. More recently, several works have explored a complementary strategy by broadening the VPR training distribution through multi-dataset training. SelaVPR++ (Lu et al., 2025a) unifies the training protocols of several commonly used VPR datasets and jointly trains a parameter-efficient VPR model on their combination, while MegaLoc (Berton & Masone, 2025) combines multiple datasets and training strategies to learn a single retrieval model applicable across VPR, landmark retrieval, and visual localization. QAA (Xiao et al., 2025) more explicitly studies multi-dataset joint training toward universal VPR, attributing the difficulty of jointly learning from diverse datasets to the limited information capacity of feature aggregation. It therefore introduces a query-based reference codebook to improve aggregation capacity and achieve more balanced generalization across datasets. In contrast, our work focuses on how heterogeneous inputs should be adapted before aggregation, particularly when a shared pretrained representation is jointly optimized across multiple domains.

2.2 Mixture-of-Experts

Mixture-of-Experts (MoE) architectures (Jacobs et al., 1991) introduce multiple specialized expert networks together with a routing mechanism that dynamically assigns each input to a subset of experts, enabling conditional computation and efficient model scaling (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022; Du et al., 2022). This paradigm has been extended to vision, where sparse routing promotes expert specialization across visual inputs (Riquelme et al., 2021), with subsequent work further investigating routing strategies and expert assignment in vision MoEs (Puigcerver et al., 2024; Liu et al., 2024). Beyond model scaling, MoE has also been used to learn from heterogeneous visual data. Prior work explores dataset- or domain-specific experts for object detection and person re-identification (Jain et al., 2023; Dai et al., 2021), while recent methods show that expert specialization can emerge without dataset labels in multi-dataset 3D segmentation and visual adaptation (Chen et al., 2026; Mo et al., 2025). Our work differs from prior MoE approaches in both its adaptation setting and expert design. We study parameter-efficient multi-domain VPR with a frozen visual foundation model, using experts as lightweight adaptation branches rather than to increase backbone capacity. While conventional MoEs typically specialize experts through different parameters within a shared architecture, we instead construct heterogeneous experts from distinct spatial operators, introducing complementary inductive biases for adapting to diverse visual conditions. A sparse subset of these experts is then dynamically selected for each input without requiring domain labels.

3 Methods

3.1 Overview

As illustrated in Fig. 1, our framework adopts a ViT-based DINOv2 model as the visual backbone. Following the memory-efficient side-adaptation strategy of SelaVPR++ (Lu et al., 2025a), we freeze the pretrained backbone and refine its intermediate features through a parallel side adaptation network, thereby avoiding gradient backpropagation through the backbone. Unlike SelaVPR++, which employs MultiConv adapters in the side adaptation network, we introduce a mixture-of-experts (MoE) formulation that enables scene-aware feature adaptation through specialized experts. The

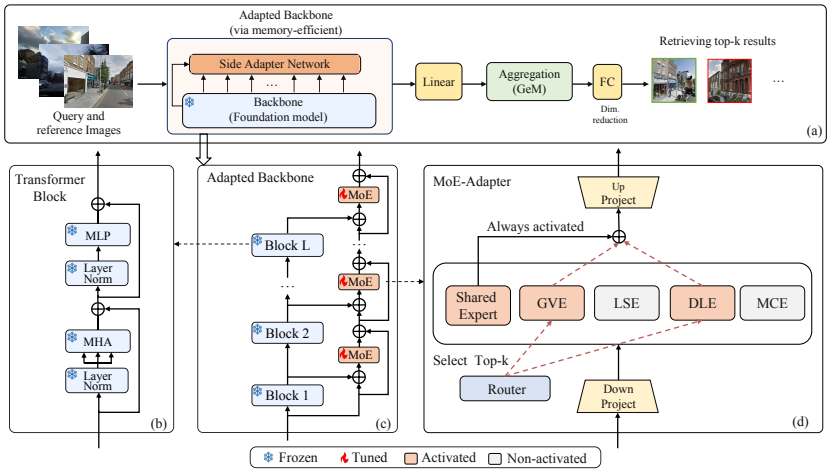


Figure 1: Overview of the proposed scene-adaptive VPR framework. We keep the pretrained DINOv2 backbone frozen and perform memory-efficient adaptation through a parallel side network with heterogeneous MoE adapters. Each adapter combines an always-active shared expert with four sparsely routed spatial experts, where an image-level router dynamically selects the top-k experts for each input. The resulting features are aggregated by GeM pooling to obtain compact global descriptors for place retrieval.

features produced by the model are subsequently aggregated using generalized mean (GeM) pooling (Radenović et al., 2018) and normalized to obtain a compact global descriptor for image retrieval.

To train a unified VPR model for both indoor and outdoor environments, we jointly train on three outdoor VPR datasets, GSV-Cities (Ali-Bey et al., 2022), SF-XL (Berton et al., 2022a), and MSLS (Warburg et al., 2020), together with the indoor ScanNet dataset (Dai et al., 2017). Each training iteration contains samples from all four datasets, with four images sampled for each place. Following SelaVPR++ (Lu et al., 2025a), we process SF-XL and MSLS into a place-based format compatible with GSV-Cities, enabling unified training across the three outdoor datasets. For ScanNet, we treat each valid scene as a place identity and construct training samples from its camera poses. We build a pose-compatibility graph in which two RGB frames are connected if their camera distance is within $[0.25, 10]$ m and their relative rotation is below $30°$. During training, we randomly sample a four-frame clique from each scene, ensuring that all frame pairs satisfy these geometric constraints.

3.2 Scene-Adaptive Mixture-of-Experts Adaptation

Visual place recognition involves substantially different spatial variations across images, ranging from viewpoint-induced geometric displacement to fine-grained local structures and large-scale scene layout. A single homogeneous adapter is therefore required to encode multiple, potentially competing, spatial priors within the same transformation. To increase the adaptation capacity without densely activating all parameters, we reformulate the side adaptation branch of SelaVPR++ (Lu et al., 2025a) as a heterogeneous Mixture-of-Experts (MoE) adapter. The adapter contains an always-active shared expert for common feature transformation and four sparsely activated spatial experts with complementary inductive biases.

3.2.1 Heterogeneous Expert Architecture

Let $\mathbf{S}_\ell \in \mathbb{R}^{B\times(N+1)\times D}$ denote the input to the ℓ-th side adapter, where $N = HW$ is the number of patch tokens and the first entry corresponds to the class token. We project it into a d-dimensional bottleneck space, $\mathbf{U}_\ell = \mathcal{P}_{\text{down}}(\mathbf{S}_\ell) \in \mathbb{R}^{B\times(N+1)\times d}$, and split the resulting representation as $\mathbf{U}_\ell = [\mathbf{u}_\ell^{\text{cls}}; \mathbf{U}_\ell^p]$. The patch tokens $\mathbf{U}_\ell^p$ are further reshaped into the spatial feature map $\mathbf{X}_\ell = \mathcal{R}(\mathbf{U}_\ell^p) \in \mathbb{R}^{B\times d\times H\times W}$. The shared expert operates on $\mathbf{U}_\ell$ to learn common feature transformations, while four heterogeneous experts process $\mathbf{X}_\ell$ to capture complementary spatial representations. Details of the shared and heterogeneous experts are given below.

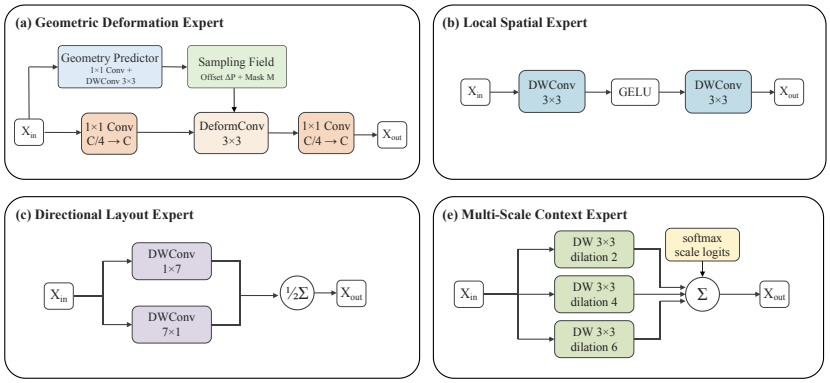


Figure 2: Architectures of the four heterogeneous spatial experts: (a) Geometric Variation Expert (GVE), (b) Local Spatial Expert (LSE), (c) Directional Layout Expert (DLE), and (d) Multi-Scale Context Expert (MCE).

Shared Expert. The shared expert is a lightweight two-layer MLP with bottleneck width $d/2$:

$$E_{\text{shared}}(\mathbf{U}_\ell) = \mathbf{W}_s^{(2)}\phi\Big(\mathbf{W}_s^{(1)}\mathbf{U}_\ell\Big), \tag{1}$$

where $\phi(\cdot)$ denotes GELU. Since this pathway is activated for every image, it captures transformations that are useful across different visual conditions and leaves the sparsely routed experts to specialize in spatially distinct variations.

Geometric Variation Expert. Changes in camera position and viewing direction can induce substantial spatial displacement of scene structures, which is difficult to capture with fixed-grid convolution. As shown in Fig. 2(a), we therefore introduce a Geometric Variation Expert (GVE) with adaptive spatial sampling. Given $\mathbf{X}$, a lightweight geometry predictor $\mathcal{G}_g$ estimates the sampling offsets $\Delta\mathbf{P}_g$ and modulation masks $\mathbf{M}_g$:

$$[\Delta\mathbf{P}_g,\, \mathbf{M}_g] = \mathcal{G}_g(\mathbf{X}_\ell). \tag{2}$$

The geometric variation Expert is then formulated as

$$E_{\text{GVE}}(\mathbf{X}_\ell) = \text{Conv}_{1\times1}^{\text{up}}\left(\text{DCNv2}_{3\times3}\left(\text{Conv}_{1\times1}^{\text{down}}(\mathbf{X}_\ell); \Delta\mathbf{P}_g, \sigma(\mathbf{M}_g)\right)\right), \tag{3}$$

where $\text{Conv}_{1\times1}^{\text{down}}$ projects the input into a compact feature space and $\text{Conv}_{1\times1}^{\text{up}}$ restores the original channel dimension. $\text{DCNv2}_{3\times3}$ denotes a 3×3 modulated deformable convolution (Zhu et al., 2019), whose sampling locations are controlled by $\Delta\mathbf{P}_g$ and whose modulation weights are given by $\sigma(\mathbf{M}_g)$, with $\sigma(\cdot)$ denoting the sigmoid function.

By predicting the sampling geometry from the original representation and performing deformable convolution in a compact feature space, the geometric variation Expert adapts to viewpoint-induced spatial variations with limited computational overhead.

Local Spatial Expert. While the geometric variation expert explicitly accounts for geometric displacement, robust place recognition also relies on local structural cues that remain discriminative across different observations. To capture such short-range spatial patterns, as shown in Fig. 2(b), we introduce a Local Spatial Expert (LSE) composed of two successive 3×3 depthwise convolutions with a GELU activation:

$$E_{\text{LSE}}(\mathbf{X}_\ell) = \text{DWConv}_{3\times3}^{(2)}\left(\phi\Big(\text{DWConv}_{3\times3}^{(1)}(\mathbf{X}_\ell)\Big)\right). \tag{4}$$

The stacked depthwise convolutions progressively aggregate neighboring information while preserving channel-wise responses, enabling the expert to emphasize fine-grained local structures with only a small parameter overhead. In this way, the local expert complements the geometry-aware geometric variation expert by providing stable local cues for place discrimination.

Directional Layout Expert. Local features alone may not sufficiently characterize the structural organization of a scene. In both indoor and outdoor environments, place appearance often contains

prominent horizontal and vertical structures. To explicitly encode such directional information, as shown in Fig. 2(c), we construct a Directional Layout Expert (DLE) using asymmetric depthwise convolutions:

$$E_{\mathrm{DLE}}(\mathbf{X}_\ell) = \frac{1}{2}\left[\mathrm{DWConv}_{1\times 7}(\mathbf{X}_\ell) + \mathrm{DWConv}_{7\times 1}(\mathbf{X}_\ell)\right]. \tag{5}$$

The elongated kernels extend the receptive field along individual spatial directions, enabling the expert to capture directional scene structures with a small parameter overhead.

Multi-Scale Context Expert. Place recognition also requires contextual information beyond individual local structures, particularly when similar local patterns appear at different locations. As shown in Fig. 2(d), we therefore introduce a Multi-scale Context Expert (MCE) using depthwise convolutions with different dilation rates. Its output is

$$E_{\mathrm{MCE}}(\mathbf{X}_\ell) = \sum_{\delta\in\{2,4,6\}} \pi_\delta \ \mathrm{DWConv}^{(\delta)}_{3\times 3}(\mathbf{X}_\ell), \qquad \pi = \mathrm{Softmax}(\alpha), \tag{6}$$

where $\alpha \in \mathbb{R}^3$ is a set of learnable scale logits and δ denotes the dilation rate. By combining multiple dilation rates, the resulting representation aggregates contextual information over different receptive-field scales while retaining the efficiency of depthwise convolution.

Together, these heterogeneous experts provide complementary spatial representations. Their contributions are dynamically determined by an image-level sparse router, as described next.

3.2.2 Image-level Sparse Routing

The router selects a sparse subset of heterogeneous experts for each image and determines their mixture weights. At each MoE adapter layer, the router predicts a single expert distribution for each image, which is shared by all patch tokens of that image. This image-level routing strategy maintains spatial consistency across patch features and is naturally aligned with the global descriptor used for place retrieval.

As described in the preceding subsection, the bottleneck representation $\mathbf{U}_\ell$ is split into the CLS token $\mathbf{u}_\ell^{\mathrm{cls}}$ and patch tokens $\mathbf{U}_\ell^p$. We construct an image-level routing representation by concatenating the class token with the mean-pooled patch tokens:

$$\mathbf{z}_\ell = \left[\mathbf{u}_\ell^{\mathrm{cls}}; \frac{1}{N}\sum_{i=1}^{N}\mathbf{U}^p_{\ell,i}\right] \in \mathbb{R}^{B\times 2d}, \tag{7}$$

where the class token provides global semantic information, while the mean-pooled patch representation summarizes spatially distributed visual evidence.

Each MoE adapter layer employs an independent linear router to compute the expert logits:

$$\mathbf{H}_\ell = \mathbf{z}_\ell\left(\mathbf{W}_r^{(\ell)}\right)^\top + \mathbf{b}_r^{(\ell)} \in \mathbb{R}^{B\times E}, \tag{8}$$

where $E = 4$ denotes the number of routed heterogeneous experts. The corresponding sparse routing weights are obtained as

$$\boldsymbol{\omega}_\ell = \mathrm{Softmax}\left(\mathrm{KeepTopK}\left(\mathbf{H}_\ell, k\right)\right), \tag{9}$$

where $\mathrm{KeepTopK}(\cdot, k)$ retains the k largest logits for each image and sets all remaining entries to $-\infty$. Consequently, $\boldsymbol{\omega}_\ell \in \mathbb{R}^{B\times E}$ contains only k nonzero routing weights for each image. The same routing weights are shared by all patch tokens of that image.

The routing operation is performed independently for each image and each MoE adapter layer. It does not require manually specified scene-category annotations and can dynamically select different combinations of heterogeneous experts at different representation depths.

3.3 Training Strategy

Following (Ali-Bey et al., 2022), we employ the Multi-Similarity (MS) loss (Wang et al., 2019) as the retrieval objective:

$$\mathcal{L}_{\mathrm{MS}} = \frac{1}{B}\sum_{q=1}^{B}\left\{\frac{1}{\alpha}\log\left[1+\sum_{p\in\mathcal{P}_q} e^{-\alpha(S_{qp}-\gamma)}\right] + \frac{1}{\beta}\log\left[1+\sum_{n\in\mathcal{N}_q} e^{\beta(S_{qn}-\gamma)}\right]\right\}, \tag{10}$$

where $\mathcal{P}_q$ and $\mathcal{N}_q$ denote the mined positive and negative sets for query q, and S_{qp} and S_{qn} are the corresponding cosine similarities.

To avoid unreliable cross-dataset pair mining, we compute the MS loss independently for each dataset and sum the resulting objectives:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{MS}}^{\mathrm{G}} + \mathcal{L}_{\mathrm{MS}}^{\mathrm{M}} + \mathcal{L}_{\mathrm{MS}}^{\mathrm{F}} + \mathcal{L}_{\mathrm{MS}}^{\mathrm{S}}, \tag{11}$$

where G, M, F, and S denote GSV-Cities, MSLS, SF-XL, and ScanNet, respectively.

4 Experiments

4.1 Datasets and Performance Evaluation

We evaluate on four VPR benchmarks covering heterogeneous environments, as summarized in Table 1. Pitts30k (Torii et al., 2013) and MSLS (Warburg et al., 2020) contain diverse urban scenes, Nordland (Sünderhauf et al., 2013) evaluates robustness to severe seasonal changes in natural environments, and Baidu Mall (Sun et al., 2017) contains challenging indoor scenes with viewpoint variations and perceptual aliasing. For Baidu Mall, we use the version processed by MegaLoc (Berton & Masone, 2025). Following standard VPR evaluation (Hausler et al., 2021; Ali-Bey et al., 2023), we report Recall@N (R@N). For Pitts30k and MSLS-val, a retrieved image within 25 m of the query is considered correct. For Nordland, we follow (Berton et al., 2022b) and use a tolerance of ± 10 frames. For Baidu Mall, we follow the evaluation protocol of the MegaLoc-processed benchmark (Berton & Masone, 2025).

Table 1: Summary of the evaluation datasets.

Dataset	Description	Database	Queries
Pitts30k	urban, viewpoint	10,000	6,816
MSLS-val	urban/suburban	18,871	740
Nordland	natural, seasonal	27,592	27,592
Baidu Mall	indoor	689	2,292

4.2 Implementation Details

We use DINOv2-Base as the frozen backbone and implement all experiments in PyTorch on an NVIDIA RTX A6000 GPU. Images are resized to 224×224 during training and 322×322 during inference, and the final descriptor dimension is 2,048. We train for 25 epochs using Adam with an initial learning rate of 4×10^{-4}, halved every five epochs. Each batch samples 64 places from each of SF-XL (Berton et al., 2022a), GSV-Cities (Ali-Bey et al., 2022), MSLS (Warburg et al., 2020), and ScanNet (Dai et al., 2017), with four images per place. Each MoE adapter contains four heterogeneous routed experts and selects the top-2 experts for each image.

4.3 Comparison with SelaVPR++

We adopt SelaVPR++ (Lu et al., 2025a) as the baseline and evaluate it under two training settings: training on the three outdoor datasets (SF-XL, MSLS, and GSV-Cities), and joint training with the additional indoor ScanNet dataset.

Table 2: Comparison with SelaVPR++ under different training distributions. Best and second-best results are shown in **bold** and underline, respectively. For SelaVPR++ trained with ScanNet, ↑ and ↓ indicate improvement and degradation relative to training on the three outdoor datasets only.

Method	Pitts30k			MSLS-val			Nordland			Baidu Mall			Avg.
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1
Training on SF-XL + MSLS + GSV-Cities													
SelaVPR++	92.9	96.8	97.8	93.8	97.3	**97.8**	**93.8**	**97.8**	**98.7**	74.3	90.7	95.1	88.7
Training on SF-XL + MSLS + GSV-Cities + ScanNet													
SelaVPR++	92.6$^{\downarrow}$	96.7$^{\downarrow}$	97.8	93.6$^{\downarrow}$	97.0$^{\downarrow}$	97.6$^{\downarrow}$	91.9$^{\downarrow}$	97.0$^{\downarrow}$	98.1$^{\downarrow}$	77.3$^{\uparrow}$	91.6$^{\uparrow}$	**96.1**$^{\uparrow}$	88.9
Ours	**93.4**	**96.9**	**98.1**	**94.2**	**97.4**	97.7	92.6	97.7	**98.7**	**77.6**	**91.8**	95.9	**89.5**

Table 3: Comparison with state-of-the-art VPR methods. Best and second-best results among single-image methods are shown in **bold** and underline, respectively. † denotes cross-image methods and is excluded from best/second-best ranking. ‡ denotes our reproduction using the same image resolutions as our method, i.e., 224×224 for training and 322×322 for inference. ⋆ denotes FoL-global results from its official GitHub repository using DINOv2-B at 322×322; we additionally evaluate FoL-global on Baidu Mall using the official implementation.

Method	Dim	Backbone	Pitts30k			MSLS-val			Nordland			Baidu Mall			Avg.		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD	32768	VGG16	81.9	91.2	93.7	53.1	66.5	71.1	6.4	10.1	12.5	66.1	86.9	92.7	51.9	63.7	67.5
SFRS	4096	VGG16	89.4	94.7	95.9	69.2	80.3	83.1	16.1	23.9	28.4	65.7	85.7	92.8	60.1	71.2	75.1
CosPlace	512	VGG16	88.4	94.5	95.7	82.8	89.7	92.0	58.5	73.7	79.4	53.3	73.3	82.0	70.8	82.8	87.3
MixVPR	4096	ResNet50	91.5	95.5	96.3	88.0	92.7	94.6	76.2	86.9	90.3	71.9	90.1	94.7	81.9	91.3	94.0
EigenPlaces	2048	ResNet50	92.5	96.8	97.6	89.1	93.8	95.0	71.2	83.8	88.1	69.5	87.3	92.1	80.6	90.4	93.2
CricaVPR†	4096	DINOv2-B	94.9	97.3	98.2	90.0	95.4	96.4	90.7	96.3	97.6	72.4	90.2	95.2	87.0	94.8	96.9
SuperVLAD†	3072	DINOv2-B	95.0	97.4	98.2	92.2	96.6	97.4	91.0	96.4	97.7	73.6	90.1	94.9	88.0	95.1	97.1
SALAD	8448	DINOv2-B	92.5	96.4	97.5	92.2	96.4	97.0	89.7	95.5	97.0	72.7	88.7	93.6	86.8	94.3	96.3
SelaVPR	/	DINOv2-L	92.8	96.8	97.7	90.8	96.4	97.2	87.3	93.8	95.6	75.8	90.1	94.9	86.7	94.3	96.4
BoQ‡	12288	DINOv2-B	92.6	96.6	97.3	92.8	96.0	96.6	89.7	95.8	97.5	64.5	84.1	91.2	84.9	93.1	95.7
FoL-global⋆	8448	DINOv2-B	92.1	96.4	97.6	91.1	95.7	96.4	72.7	85.5	89.6	74.3	90.3	95.3	82.6	92.0	94.7
Ours	2048	DINOv2-B	**93.4**	**96.9**	**98.1**	**94.2**	**97.4**	**97.7**	**92.6**	**97.7**	**98.7**	**77.6**	**91.8**	**95.9**	**89.5**	**96.0**	**97.6**

As shown in Table 2, incorporating ScanNet into SelaVPR++ substantially improves indoor recognition, increasing R@1 on Baidu Mall from 74.3% to 77.3%. However, this gain is accompanied by degraded outdoor performance, most notably on Nordland, where R@1 decreases from 93.8% to 91.9%. Pitts30k and MSLS-val also exhibit slight decreases. These results indicate that simply increasing training diversity does not necessarily yield uniform improvements across heterogeneous environments. By contrast, under the same four-dataset training setting, SceneMoE-VPR consistently outperforms SelaVPR++ on all four benchmarks, increasing the average R@1 from 88.9% to 89.5%. In particular, it partially recovers the degradation on Nordland while preserving the indoor improvement on Baidu Mall. These results suggest that scene-adaptive expert routing better alleviates the cross-environment trade-offs caused by fully shared adaptation.

4.4 Comparison with State-of-the-Art Methods

We compare SceneMoE-VPR with representative state-of-the-art VPR methods (Arandjelovic et al., 2016; Ge et al., 2020; Berton et al., 2022a; Ali-Bey et al., 2023; Berton et al., 2023; Izquierdo & Civera, 2024; Lu et al., 2024b; Ali-Bey et al., 2024; Wang et al., 2025), including the cross-image methods CricaVPR (Lu et al., 2024a) and SuperVLAD (Lu et al., 2024c). Additional comparisons on a broader set of evaluation benchmarks are provided in Appendix B.

As shown in Table 3, SceneMoE-VPR achieves the highest average R@1/R@5/R@10 of 89.5%/96.0%/97.6% among single-image methods. It also obtains the best R@1 on all four benchmarks, showing consistently strong recognition across both outdoor and indoor environments. Notably, these results are achieved with a compact 2,048-dimensional descriptor.

Figure 3 further compares recognition performance with descriptor dimensionality. SceneMoE-VPR achieves 93.4% average outdoor R@1 and 77.6% indoor R@1 with a 2,048-dimensional descriptor. In comparison, SALAD and FoL-global use 8,448-dimensional descriptors, while BoQ uses

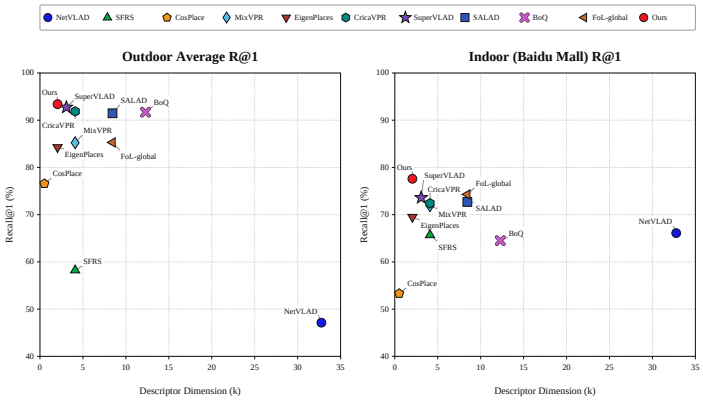


Figure 3: Recognition performance versus descriptor dimensionality. **Left:** average R@1 over Pitts30k, MSLS-val, and Nordland. **Right:** R@1 on Baidu Mall. SceneMoE-VPR achieves strong outdoor and indoor recognition with a compact 2,048-dimensional descriptor.

Table 4: Ablation of heterogeneous expert architectures. Avg. denotes the mean performance over the four benchmarks.

Variant	Pitts30k			MSLS-val			Nordland			Baidu Mall			Avg.		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
w/o GVE	92.8	97.1	98.0	93.9	97.2	97.7	91.2	96.8	98.0	76.7	91.3	95.0	88.7	95.6	97.2
w/o LSE	93.1	97.0	98.0	93.0	96.9	97.6	89.3	95.7	97.2	76.5	91.0	95.4	88.0	95.2	97.1
w/o DLE	92.9	96.9	97.8	94.1	97.6	97.7	88.8	95.7	97.4	75.6	90.2	94.8	87.9	95.1	96.9
w/o MCE	93.1	96.8	97.8	93.2	97.0	97.6	87.8	95.1	96.9	77.9	91.5	95.6	88.0	95.1	97.0
Full	93.4	96.9	98.1	94.2	97.4	97.7	92.6	97.7	98.7	77.6	91.8	95.9	**89.5**	**96.0**	**97.6**

a 12,288-dimensional descriptor. SceneMoE-VPR therefore provides a favorable balance between descriptor compactness and cross-environment recognition performance.

4.5 Ablation Study

Effect of Heterogeneous Experts. We evaluate each heterogeneous expert by removing it individually while keeping all other settings unchanged. As shown in Table 4, DLE has the largest impact on the overall R@1, with its removal reducing the average from 89.5% to 87.9%. The degradation is particularly evident on Nordland and Baidu Mall, consistent with the importance of directional scene layouts in natural and indoor environments. MCE is especially important for Nordland, where removing it causes the largest R@1 drop of 4.8 points, indicating the benefit of multi-scale context under severe appearance changes. LSE also contributes strongly to Nordland and MSLS-val, while GVE provides relatively consistent gains across benchmarks. Overall, the full model achieves the best average performance, demonstrating the complementary roles of the heterogeneous experts. More ablation studies and analyses are provided in the Appendix A (Tables 6, 5, and 7).

5 Conclusions

We study heterogeneous VPR training and observe that simply increasing training diversity can introduce cross-environment trade-offs when all inputs share the same adaptation function. To address this issue, we propose SceneMoE-VPR, which combines a frozen DINOv2 backbone with sparsely routed heterogeneous experts to enable input-dependent adaptation without scene labels. Under the same four-dataset joint-training setting, SceneMoE-VPR consistently improves over the fully shared baseline across all evaluated benchmarks, increasing the average R@1 from 88.9% to 89.5%, while expert ablations support the complementary roles of the proposed adaptation operators. These results suggest that conditional heterogeneous adaptation is a promising alternative to fully shared adaptation for multi-environment VPR across diverse environments.

AI USE STATEMENT

In this work, we used generative AI tools to improve the clarity and readability of the manuscript, including language editing, grammar correction, and wording refinement. Generative AI was not used to formulate research hypotheses, develop the proposed methodology, design experiments, generate or process datasets, formulate mathematical claims or proofs, interpret experimental results, or create scientific figures. All AI-assisted edits were carefully reviewed by the authors. The authors verified the technical accuracy of the manuscript, including its mathematical statements, experimental descriptions, results, and citations, and no AI-generated content was accepted without author review. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate the reproduction of our experimental results. The experimental setup, datasets, training objectives, hyperparameter settings, implementation details, and other relevant configurations are described in the main text and Appendix. All reported results are obtained following the experimental procedures described in the paper and Appendix. In addition, we provide detailed descriptions of the datasets and data processing procedures in the Appendix to enable consistent evaluation. The source code and configuration files will be made publicly available upon publication to further support reproducibility.

REFERENCES

Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022.

Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2998–3007, 2023.

Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Boq: A place is worth a bag of learnable queries. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17794–17803. IEEE, 2024.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5297–5307, 2016.

Gabriele Berton and Carlo Masone. Megaloc: One retrieval to place them all. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2852–2858. IEEE, 2025.

Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4878–4888, 2022a.

Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5386–5397. IEEE, 2022b.

Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11080–11090, 2023.

Gabriele Moreno Berton, Valerio Paolicelli, Carlo Masone, and Barbara Caputo. Adaptive-attentive geolocalization from few queries: A hybrid approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2918–2927, 2021.

Guillaume Bresson, Zayed Alsayed, Li Yu, and Sébastien Glaser. Simultaneous localization and mapping: A survey of current trends in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2(3):194–220, 2017.

Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *European conference on computer vision*, pp. 726–743. Springer, 2020.

Xuweiyi Chen, Wentao Zhou, Aruni RoyChowdhury, and Zezhou Cheng. Point-moe: Large-scale multi-dataset training with mixture-of-experts for 3d semantic segmentation. In *International Conference on Learning Representations*, volume 2026, pp. 28106–28122, 2026.

Mark Cummins and Paul Newman. Highly scalable appearance-only slam–fab-map 2.0. 2010.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16145–16154, 2021.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pp. 5547–5569. PMLR, 2022.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *European conference on computer vision*, pp. 369–386. Springer, 2020.

Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14136–14147. IEEE, 2021.

Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 17658–17668, 2024.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Yash Jain, Harkirat Behl, Zsolt Kira, and Vibhav Vineet. Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets. *Advances in Neural Information Processing Systems*, 36:69625–69637, 2023.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 9(2):1286–1293, 2023.

Ahmad Khaliq, Ming Xu, Stephen Hausler, Michael Milford, and Sourav Garg. Vlad-buff: Burst-aware fast feature aggregation for visual place recognition. In *European Conference on Computer Vision*, pp. 447–466. Springer, 2024.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Tianlin Liu, Mathieu Blondel, Carlos Riquelme, and Joan Puigcerver. Routers in vision mixture of experts: An empirical study. *arXiv preprint arXiv:2401.15969*, 2024.

Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *ieee transactions on robotics*, 32(1):1–19, 2015.

Feng Lu, Xiangyuan Lan, Lijun Zhang, Dongmei Jiang, Yaowei Wang, and Chun Yuan. Cricavpr: Cross-image correlation-aware representation learning for visual place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16772–16782, 2024a.

Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. In *International conference on learning representations*, volume 2024, pp. 14314–14335, 2024b.

Feng Lu, Xinyao Zhang, Canming Ye, Shuting Dong, Lijun Zhang, Xiangyuan Lan, and Chun Yuan. Supervlad: Compact and robust image descriptors for visual place recognition. *Advances in Neural Information Processing Systems*, 37:5789–5816, 2024c.

Feng Lu, Tong Jin, Xiangyuan Lan, Lijun Zhang, Yunpeng Liu, Yaowei Wang, and Chun Yuan. Selavpr++: Towards seamless adaptation of foundation models for efficient place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Feng Lu, Tong Jin, Canming Ye, Xiangyuan Lan, Yunpeng Liu, and Chun Yuan. Towards implicit aggregation: Robust image representation for place recognition in the transformer era. *Advances in Neural Information Processing Systems*, 38:85096–85124, 2025b.

Michael J Milford and Gordon F Wyeth. Mapping a suburb with a single camera using a biologically inspired slam system. *IEEE Transactions on Robotics*, 24(5):1038–1053, 2008.

Shentong Mo, Xufang Luo, and Dongsheng Li. pmoe: Prompting diverse experts together wins more in visual adaptation. In *International Conference on Learning Representations*, volume 2025, pp. 79496–79514, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Joan Puigcerver, Carlos Riquelme Ruiz, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, volume 2024, pp. 28435–28445, 2024.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Stefan Schubert, Peer Neubert, Sourav Garg, Michael Milford, and Tobias Fischer. Visual place recognition: A tutorial [tutorial]. *IEEE Robotics & Automation Magazine*, 31(3):139–153, 2023.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Xun Sun, Yuanfan Xie, Pei Luo, and Liang Wang. A dataset for benchmarking image-based localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7436–7444, 2017.

Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. In *Proc. of workshop on long-term autonomy, IEEE international conference on robotics and automation (ICRA)*, pp. 2013. Citeseer, 2013.

Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 883–890, 2013.

Changwei Wang, Shunpeng Chen, Yukun Song, Rongtao Xu, Zherui Zhang, Jiguang Zhang, Haoran Yang, Yu Zhang, Kexue Fu, Shide Du, et al. Focus on local: Finding reliable discriminative regions for visual place recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7536–7544, 2025.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5017–5025. IEEE, 2019.

Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2626–2635, 2020.

Jiuhong Xiao, Yang Zhou, and Giuseppe Loianno. Query-based adaptive aggregation for multi-dataset joint training toward universal visual place recognition. *arXiv preprint arXiv:2507.03831*, 2025.

Ming Xu, Niko Snderhauf, and Michael Milford. Probabilistic visual place recognition for hierarchical localization. *IEEE Robotics and Automation Letters*, 6(2):311–318, 2020.

Zheyuan Zhang, Jiwei Zhang, Boyu Zhou, Linzhimeng Duan, and Hong Chen. D^2-vpr: A parameter-efficient visual-foundation-model-based visual place recognition method via knowledge distillation and deformable aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13034–13042, 2026.

Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9300–9308. IEEE, 2019.

Table 5: Ablation study on the shared expert. Recall at K (R@K) is reported in percentage (%). Best and second-best results are shown in **bold** and underline, respectively.

Shared	Pitts30k-test			MSLS-val			Nordland			Baidu Mall			Avg.		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
No	92.8	**97.1**	97.9	93.2	97.2	97.6	90.4	96.3	97.6	**77.6**	**92.2**	95.7	88.5	95.7	97.2
Yes	**93.4**	96.9	**98.1**	**94.2**	**97.4**	**97.7**	**92.6**	**97.7**	**98.7**	**77.6**	91.8	**95.9**	**89.5**	**96.0**	**97.6**

Table 6: Ablation study on the weight λ_{LB} of the load-balancing loss. Recall at K (R@K) is reported in percentage (%). Best and second-best results are shown in **bold** and underline, respectively.

λ_{LB}	Pitts30k-test			MSLS-val			Nordland			Baidu Mall			Avg.		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
0	**93.4**	96.9	**98.1**	**94.2**	**97.4**	**97.7**	**92.6**	**97.7**	**98.7**	77.6	**91.8**	**95.9**	**89.5**	**96.0**	**97.6**
0.001	93.3	**97.2**	98.0	94.1	96.9	97.6	90.4	96.3	97.7	74.4	91.1	95.7	88.1	95.4	97.3
0.005	93.0	96.7	97.8	93.5	96.8	97.2	90.0	96.4	97.8	**79.0**	**91.8**	**95.9**	88.9	95.4	97.2

A MORE ABLATION STUDY

Effect of the Shared Expert. We further investigate the role of the always-active shared expert by removing it while keeping the routed heterogeneous experts unchanged. As shown in Table 5, introducing the shared expert improves the average R@1 from 88.5% to 89.5%. The gain is particularly pronounced on Nordland, where R@1 increases from 90.4% to 92.6%. Although removing the shared expert yields slightly higher R@5 on Pitts30k and Baidu Mall, the full model achieves higher average R@1/R@5/R@10 of 89.5%/96.0%/97.6%. These results suggest that the shared pathway complements the sparsely routed experts by capturing transformations useful across heterogeneous visual conditions.

Load Balancing. We investigate the effect of explicitly balancing expert utilization by varying the weight λ_{LB} of the auxiliary load-balancing loss. Following standard sparse MoE practice (Fedus et al., 2022), the loss is defined as

$$\mathcal{L}_{\mathrm{LB}} = N_E \sum_{i=1}^{N_E} f_i p_i, \tag{12}$$

where N_E is the number of routed experts, f_i denotes the fraction of inputs assigned to expert i, and p_i is the average routing probability of expert i over the batch.

As shown in Table 6, the model without explicit load balancing ($\lambda_{\mathrm{LB}} = 0$) achieves the best overall performance, with average R@1/R@5/R@10 of 89.5%/96.0%/97.6%. Introducing a small load-balancing weight ($\lambda_{\mathrm{LB}} = 0.001$) slightly improves Pitts30k R@5 from 96.9% to 97.2%, but degrades performance on the other benchmarks, reducing the average R@1 to 88.1%. Increasing the weight to 0.005 improves Baidu Mall R@1 from 77.6% to 79.0%, but still yields a lower average R@1 of 88.9% due to degraded outdoor performance. These results indicate that enforcing uniformly balanced expert utilization is not necessarily beneficial in our heterogeneous-scene setting. Different scenes may favor different expert combinations, making non-uniform routing better suited to input-dependent adaptation. We therefore set $\lambda_{\mathrm{LB}} = 0$ in the final model.

Router Input Feature. We investigate which intermediate representation is most suitable for expert routing while keeping the routing strategy and all other settings unchanged. Specifically, we compare three alternatives: the feature produced by the frozen DINOv2 backbone, the adapter input before down-projection, and the adapter input after down-projection. For all variants, the routing representation is constructed in the same manner by concatenating the [CLS] token with the mean-pooled patch tokens.

Table 7: Ablation study on the input feature used for expert routing. All variants construct the routing representation by concatenating the [CLS] token with the mean-pooled patch tokens. Recall at K (R@K) is reported in percentage (%). Best and second-best results are shown in **bold** and underline, respectively.

Router Input	Pitts30k-test			MSLS-val			Nordland			Baidu Mall			Avg.		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Frozen backbone feature	93.0	**97.0**	**98.2**	93.9	96.8	97.6	91.2	96.8	98.0	74.6	91.2	95.5	88.2	95.5	97.3
Adapter input (pre-down)	92.7	96.8	97.9	92.8	97.3	**98.1**	82.9	92.2	94.8	**77.7**	**91.8**	**96.1**	86.5	94.5	96.7
Adapter input (post-down)	**93.4**	96.9	98.1	**94.2**	**97.4**	97.7	**92.6**	**97.7**	**98.7**	77.6	**91.8**	95.9	**89.5**	**96.0**	**97.6**

As shown in Table 7, using the post-down adapter input achieves the best overall performance, with average R@1/R@5/R@10 of 89.5%/96.0%/97.6%. Routing directly from the frozen backbone features remains competitive on Pitts30k and MSLS-val, but its R@1 decreases to 91.2% on Nordland and 74.6% on Baidu Mall, resulting in a lower average R@1 of 88.2%. Using the pre-down adapter input yields the lowest average R@1 of 86.5%, mainly due to the substantial performance degradation on Nordland.

These results suggest that the post-down adapter representation provides a more suitable signal for expert selection. Since this representation has already been projected into the compact feature space processed by the heterogeneous experts, the routing decision is more closely aligned with the subsequent expert transformations. We therefore use the post-down adapter representation for routing in the final model.

B COMPARISONS ON MORE DATASETS

To provide a more comprehensive evaluation, we further compare our method with representative state-of-the-art VPR approaches on additional datasets, including Eynsham (Cummins & Newman, 2010), St Lucia (Milford & Wyeth, 2008), Pitts250k (Torii et al., 2013), and four SVOX (Berton et al., 2021) subsets under different environmental conditions. The complete quantitative results are reported in Table 8. Our method achieves consistently competitive performance across these benchmarks. In particular, it obtains the best results on Eynsham across all recall thresholds and reaches saturated performance on St Lucia. On the challenging SVOX subsets, our method also achieves competitive or best performance on several evaluation metrics, demonstrating strong robustness to diverse appearance and environmental variations.

C ADDITIONAL QUALITATIVE RESULTS

Expert Routing across Heterogeneous Environments. Figure 4 visualizes the Top-1 and Top-2 expert choices across the 12 adaptation layers on four test datasets. The routing patterns exhibit both shared and dataset-dependent behaviors. In several layers, different datasets select the same dominant expert, suggesting that some adaptation patterns are broadly useful across environments. In contrast, clear routing differences emerge at other layers, particularly in the secondary expert choices and several middle-to-deep layers. For example, the indoor Baidu Mall dataset and the outdoor datasets follow different expert combinations at some layers, while MSLS and Nordland also exhibit distinct routing preferences. These results indicate that the router does not assign each environment to a fixed expert; instead, it dynamically composes different expert combinations across representation depths. Such routing behavior is consistent with our motivation of retaining shared adaptation while allowing input-dependent specialization for heterogeneous VPR.

D DATASET DETAILS

We provide additional details of the datasets used for training and evaluation. As illustrated in Fig. 5, we jointly train on GSV-Cities, SF-XL, MSLS-train, and ScanNet, and evaluate on Pitts30k-test, MSLS-val, Nordland, and Baidu Mall. Together, these datasets cover heterogeneous outdoor

Table 8: Comparison with state-of-the-art VPR methods on additional datasets. Recall@K (%) is reported. The best result for each metric is highlighted in **bold**. "–" denotes unavailable or unreported results.

Method	Eynsham			St Lucia			Pitts250k		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD	77.7	87.8	90.5	64.6	80.3	85.8	85.9	93.1	95.0
SFRS	72.3	83.5	87.1	75.9	86.6	91.2	90.4	96.3	97.6
CosPlace	90.0	93.9	94.9	99.6	99.9	**100.0**	92.3	97.4	98.4
MixVPR	89.4	93.2	94.3	99.6	99.9	**100.0**	94.1	98.2	98.9
EigenPlaces	90.7	94.4	95.4	99.6	99.9	**100.0**	94.1	97.9	98.7
CricaVPR	91.6	95.0	95.8	99.9	99.9	99.9	–	–	–
SuperVLAD	92.1	95.6	96.4	–	–	–	**97.3**	**99.4**	**99.7**
SALAD	91.6	95.1	95.9	**100.0**	**100.0**	**100.0**	95.1	98.5	99.1
SelaVPR	90.6	95.3	96.2	99.8	**100.0**	**100.0**	95.7	98.8	99.2
BoQ	92.2	95.6	96.4	99.9	**100.0**	**100.0**	96.6	99.1	99.5
FoL-global	91.7	95.3	96.2	–	–	–	96.5	99.1	99.5
Ours	**92.5**	**95.8**	**96.6**	**100.0**	**100.0**	**100.0**	95.4	98.9	99.4

Method	SVOX Sun			SVOX Snow			SVOX			SVOX Overcast		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
NetVLAD	35.4	52.7	58.8	54.4	71.8	77.2	–	–	–	66.4	81.5	85.7
SFRS	54.8	68.3	74.1	76.0	86.1	89.4	–	–	–	81.1	88.4	91.2
CosPlace	78.5	89.7	93.1	92.0	98.4	98.9	–	–	–	92.2	97.7	97.9
MixVPR	84.8	93.2	94.7	96.8	98.4	98.9	–	–	–	96.2	98.3	98.9
EigenPlaces	86.4	95.0	96.4	93.1	97.6	98.2	–	–	–	93.1	97.8	98.3
CricaVPR	93.8	98.1	98.8	96.0	99.2	99.2	97.8	99.2	99.3	96.7	99.0	99.0
SuperVLAD	–	–	–	–	–	–	–	–	–	–	–	–
SALAD	97.2	99.4	**99.7**	98.9	**99.7**	99.8	98.2	99.3	99.4	98.3	**99.3**	99.3
SelaVPR	90.2	96.6	97.4	97.0	99.5	99.5	97.2	98.7	99.0	97.0	99.1	99.3
BoQ	97.5	99.4	99.5	**99.4**	**99.7**	**99.9**	**99.0**	**99.6**	**99.7**	**98.5**	**99.3**	99.4
FoL-global	**98.1**	99.4	99.5	99.1	**99.7**	99.8	98.4	99.4	99.6	97.9	99.2	99.3
Ours	97.5	**99.5**	99.5	98.9	**99.7**	99.8	98.6	99.3	99.5	98.3	**99.3**	**99.5**

and indoor environments with substantial variations in viewpoint, illumination, season, appearance, and scene structure. In addition to the main evaluation benchmarks, we further report comparisons on Pitts250k-test, Eynsham, St Lucia, and several SVOX evaluation settings, including the standard SVOX split and the Sun, Snow, and Overcast condition subsets.

GSV-Cities (Ali-Bey et al., 2022). GSV-Cities is a large-scale dataset designed for supervised visual place recognition, containing approximately 560K street-view images from 67K geographically distinct places across more than 40 cities worldwide. Its accurate geographic labels and diverse temporal and appearance conditions make it suitable for supervised VPR training. We use GSV-Cities as one of our outdoor training datasets.

SF-XL (Berton et al., 2022a). SF-XL is a large-scale visual geo-localization dataset covering the city of San Francisco, with more than 41 million images collected over a decade. It contains substantial viewpoint and long-term appearance variations. Following our unified multi-dataset training protocol, we process SF-XL into a place-based format and use it as an outdoor training dataset.

MSLS (Warburg et al., 2020). The Mapillary Street-Level Sequences (MSLS) dataset contains large-scale street-level imagery collected across multiple cities and includes urban, suburban, and natural environments under diverse weather, illumination, and viewpoint conditions. We use the

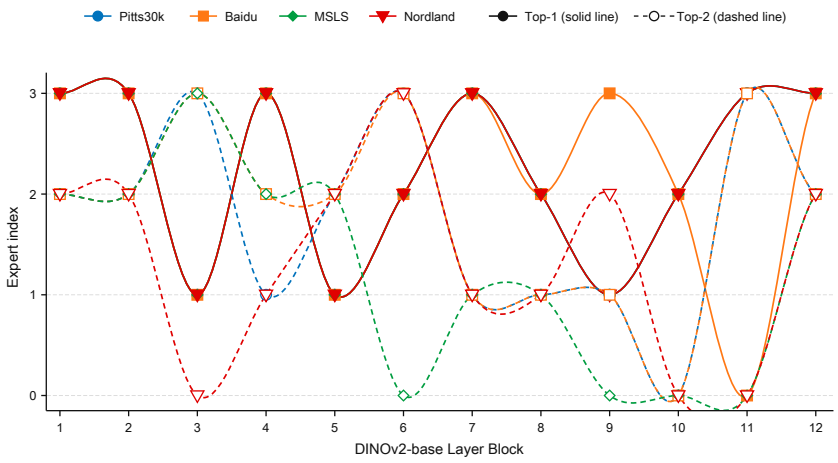


Figure 4: Expert routing patterns across heterogeneous test environments. Solid and dashed lines denote the Top-1 and Top-2 expert choices, respectively, across the 12 DINOv2 blocks. Routing is shared across datasets in some layers while diverging in others, illustrating layer-dependent and environment-dependent expert selection.

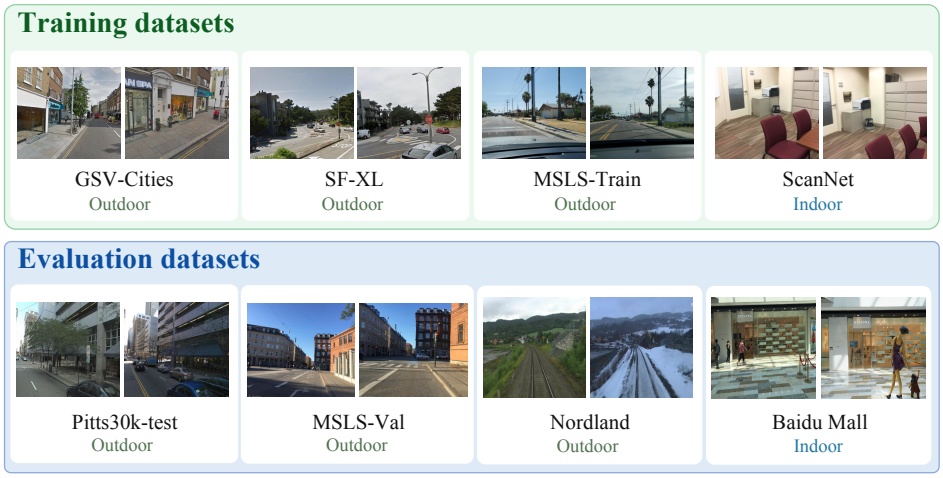


Figure 5: Overview of the training and evaluation datasets. We jointly train on three outdoor datasets (GSV-Cities, SF-XL, and MSLS-train) and the indoor ScanNet dataset, and evaluate on Pitts30k-test, MSLS-val, Nordland, and Baidu Mall.

MSLS training split for joint training and the standard validation split for evaluation. MSLS-val contains 18,871 database images and 740 query images.

ScanNet (Dai et al., 2017). ScanNet is a large-scale RGB-D dataset of real-world indoor environments, containing approximately 2.5 million views from 1,513 scans of 707 indoor spaces, together with camera poses and reconstructed 3D geometry. We use ScanNet as the indoor training dataset. In our training setup, each valid scene is treated as a place identity, and training samples are constructed from camera poses using the pose-compatibility constraints described in the main paper.

Pitts30k (Torii et al., 2013). Pitts30k is an urban VPR benchmark derived from Google Street View imagery in Pittsburgh and contains substantial viewpoint variations. We evaluate on the standard Pitts30k test split, consisting of 10,000 database images and 6,816 query images.

Nordland (Sünderhauf et al., 2013). Nordland consists of railway journeys recorded along the same route under different seasonal conditions. The largely aligned trajectories make seasonal appearance changes its primary challenge. Following the standard VPR protocol used in the main experiments, we evaluate with 27,592 database and 27,592 query images using a tolerance of ± 10 frames.

Baidu Mall (Sun et al., 2017). Baidu Mall is an indoor localization dataset collected in a large shopping mall with substantial viewpoint changes and perceptual aliasing. Following the MegaLoc-processed evaluation version used in our main experiments (Berton & Masone, 2025), we use 689 database images and 2,292 query images. The reference and query images are captured under different conditions and contain challenging repetitive structures and dynamic distractors such as people.

Pitts250k (Torii et al., 2013). Pitts250k is a large-scale urban visual place recognition benchmark constructed from geo-tagged street-view imagery in Pittsburgh. Compared with Pitts30k, it provides a substantially larger retrieval gallery and therefore poses a more challenging large-scale place retrieval problem. We evaluate on the standard Pitts250k test split, which contains 83,952 database images and 8,280 query images. The dataset exhibits substantial viewpoint changes as well as variations caused by the street-view image acquisition process.

Eynsham (Cummins & Newman, 2010). Eynsham is a place recognition benchmark collected along road trajectories around Eynsham, UK, and contains both urban and rural road scenes. The commonly used VPR benchmark version consists of 23,935 database images and 23,935 query images. The images are grayscale and contain considerable viewpoint variation, making the dataset particularly useful for evaluating robustness when appearance cues such as color are unavailable.

St Lucia (Milford & Wyeth, 2008). St Lucia consists of repeated vehicle-mounted camera traversals collected in the St Lucia suburb of Brisbane at different times. Following the standardized VPR evaluation protocol, the first and last traversals are used to form the database and query sets, respectively, with frames sampled along the driving trajectory. The resulting benchmark contains 1,549 database images and 1,464 query images. Repeated road structures, vegetation, shadows, traffic, and changes along the traversal make it a useful benchmark for evaluating place recognition under moderate appearance and viewpoint variations.

SVOX (Berton et al., 2021). Street View Oxford (SVOX) is an urban visual place recognition dataset designed to evaluate robustness under cross-domain and environmental changes. Its gallery is constructed from Google Street View imagery covering the city of Oxford, while additional queries are obtained under different weather and illumination conditions. In the standard test split, the shared gallery contains 17,166 images. We evaluate on the standard SVOX query split as well as three condition-specific subsets: SVOX Sun, SVOX Snow, and SVOX Overcast. These subsets share the same gallery but contain queries captured under substantially different illumination and weather conditions, providing a controlled evaluation of the model's robustness to environmental domain shifts.

E DETAILS OF THE COMPARED METHODS

We compare our approach with representative VPR methods covering CNN-based retrieval, classification-based metric learning, feature-mixing architectures, foundation-model-based representations, and parameter-efficient adaptation. Unless otherwise specified, we use publicly released implementations and pretrained checkpoints and follow the evaluation settings recommended by the corresponding authors.

NetVLAD. NetVLAD (Arandjelovic et al., 2016) introduces a differentiable version of the VLAD aggregation operator, enabling local convolutional activations to be aggregated into a compact image-level representation in an end-to-end trainable network. It is one of the most widely adopted learning-based VPR baselines and serves as a representative CNN-based global retrieval method. We use the PyTorch implementation[1] with the released VGG16 model trained on the Pitts30k dataset.

[1] https://github.com/Nanne/pytorch-NetVlad

SFRS. SFRS (Ge et al., 2020)[2] introduces a self-supervised strategy for learning fine-grained region similarities for large-scale image localization. It constructs image-to-region similarity supervision to exploit local correspondences between images, allowing the model to learn more discriminative representations without requiring additional manual annotations. The resulting representation is built upon a VGG16-based NetVLAD architecture and can be directly used for global image retrieval. We use the publicly released implementation and pretrained model provided in the OpenIBL framework.

CosPlace. CosPlace (Berton et al., 2022a)[3] formulates place recognition as a classification problem by dividing the geographic space into groups of non-overlapping classes. This formulation avoids explicit pair or triplet mining and enables efficient training on large-scale geo-localized imagery. We evaluate the official pretrained model using the configuration reported by the authors.

MixVPR. MixVPR (Ali-Bey et al., 2023)[4] performs global feature aggregation through a sequence of feature-mixing operations applied to backbone feature maps. Rather than relying on explicit local matching, it progressively models interactions across spatial feature locations and produces a single compact descriptor for image retrieval. We use the released model and its standard inference configuration in our comparison.

EigenPlaces. EigenPlaces (Berton et al., 2023)[5] improves the learning of viewpoint-robust place representations by constructing geographically defined classification tasks with an emphasis on viewpoint variation. The resulting model retains the efficiency of global descriptor retrieval while improving robustness to changes in camera orientation. We use the official pretrained weights and descriptor configuration provided by the authors.

CricaVPR. CricaVPR (Lu et al., 2024a)[6] introduces cross-image correlation-aware representation learning for visual place recognition. Instead of representing each image entirely independently, it uses an attention-based module to model correlations among multiple images, allowing cross-image variations such as viewpoint and appearance changes to contribute to representation learning. It further employs multi-scale convolution-enhanced adaptation to adapt a pretrained DINOv2 backbone to the VPR task. We use the official pretrained model and follow the evaluation configuration released by the authors. Since the method exploits cross-image information during representation construction, it is marked as a cross-image method in the main comparison table.

SuperVLAD. SuperVLAD (Lu et al., 2024c)[7] develops a compact VLAD-based representation for visual place recognition using a pretrained DINOv2 backbone. It combines VLAD-style feature aggregation with a cross-image encoder to enhance the robustness and discriminability of the resulting global descriptors while maintaining a relatively compact representation. In our comparison, we use the official pretrained model with the cross-image encoder enabled and follow the inference setting recommended by the authors. Accordingly, SuperVLAD is marked as a cross-image method in the main comparison table.

SALAD. SALAD (Izquierdo & Civera, 2024)[8] combines a DINOv2 backbone with an optimal-transport-based aggregation module for visual place recognition. Instead of directly pooling dense patch descriptors, SALAD learns assignments between local features and a set of representative clusters and jointly constructs a discriminative global representation. We use the publicly released checkpoint and the evaluation configuration provided by the authors.

SelaVPR. SelaVPR (Lu et al., 2024b)[9] adapts a pretrained DINOv2 model to VPR using lightweight global and local adaptation modules while keeping the backbone frozen. It produces

[2] https://github.com/yxgeee/OpenIBL
[3] https://github.com/gmberton/CosPlace
[4] https://github.com/amaralibey/MixVPR
[5] https://github.com/gmberton/EigenPlaces
[6] https://github.com/Lu-Feng/CricaVPR
[7] https://github.com/Lu-Feng/SuperVLAD
[8] https://github.com/serizba/salad
[9] https://github.com/Lu-Feng/SelaVPR

global descriptors for initial retrieval and dense local features for candidate re-ranking. In our comparison, we report the results of its official two-stage pipeline following the authors' evaluation protocol.

BoQ. BoQ (Ali-Bey et al., 2024)[10] introduces a query-based aggregation mechanism that learns a set of global queries to selectively aggregate local backbone features through cross-attention. Outputs from multiple BoQ blocks are combined to form a global descriptor for image retrieval. For our comparison, we reproduce the DINOv2-B variant using the official implementation with the same image resolutions as our method, i.e., 224×224 for training and 322×322 for inference.

FoL. FoL (Wang et al., 2025)[11] is a two-stage VPR framework that identifies reliable discriminative local regions to improve both global retrieval and local re-ranking. It introduces region-aware objectives to guide global representation learning and a weakly supervised local-feature strategy for re-ranking. For a consistent comparison with our single-stage global retrieval method, we report the FoL-global results using the official DINOv2-B model at 322×322. Results on Pitts30k, MSLS-val, and Nordland are taken from the official repository, while Baidu Mall is evaluated by us using the official implementation.

SelaVPR++. SelaVPR++ (Lu et al., 2025a)[12] extends the parameter-efficient adaptation paradigm of SelaVPR with a more memory-efficient framework for foundation-model-based VPR. It keeps the pretrained visual backbone frozen and introduces lightweight MultiConv adapters in a parallel side adaptation network to refine intermediate representations without propagating gradients through the backbone. SelaVPR++ further provides a unified training protocol for jointly learning from multiple VPR datasets. In our experiments, we use SelaVPR++ as the primary fully shared adaptation baseline and retrain it under the same multi-dataset setting as our method, including SF-XL, MSLS, GSV-Cities, and ScanNet, to enable a controlled comparison with SceneMoE-VPR.

[10] https://github.com/amaralibey/Bag-of-Queries
[11] https://github.com/chenshunpeng/FoL
[12] https://github.com/Lu-Feng/SelaVPRplusplus