D²-VPR: A Parameter-efficient Visual-foundation-model-based Visual Place Recognition Method via Knowledge Distillation and Deformable Aggregation

Zheyuan Zhang[1, 2], Jiwei Zhang[1, 2], Boyu Zhou[1, 2], Linzhimeng Duan[1, 2], Hong Chen[1, 2*]

[1]School of Computer Science, Beijing University of Posts and Telecommunications
[2]Key Laboratory of Interactive Technology and Experience System, Ministry of Culture and Tourism,
Beijing University of Posts and Telecommunications
{zzy1998, chenhong76}@bupt.edu.cn

Abstract

Visual Place Recognition (VPR) aims to determine the geographic location of a query image by retrieving its most visually similar counterpart from a geo-tagged reference database. Recently, the emergence of the powerful visual foundation model, DINOv2, trained in a self-supervised manner on massive datasets, has significantly improved VPR performance. This improvement stems from DINOv2's exceptional feature generalization capabilities but is often accompanied by increased model complexity and computational overhead that impede deployment on resource-constrained devices. To address this challenge, we propose D^2-VPR, a Distillation- and Deformable-based framework that retains the strong feature extraction capabilities of visual foundation models while significantly reducing model parameters and achieving a more favorable performance-efficiency trade-off. Specifically, first, we employ a two-stage training strategy that integrates knowledge distillation and fine-tuning. Additionally, we introduce a Distillation Recovery Module (DRM) to better align the feature spaces between the teacher and student models, thereby minimizing knowledge transfer losses to the greatest extent possible. Second, we design a Top-Down-attention-based Deformable Aggregator (TDDA) that leverages global semantic features to dynamically and adaptively adjust the Regions of Interest (ROI) used for aggregation, thereby improving adaptability to irregular structures. Extensive experiments demonstrate that our method achieves competitive performance compared to state-of-the-art approaches. Meanwhile, it reduces the parameter count by approximately 64.2% (compared to CricaVPR).

Code — https://github.com/tony19980810/D2VPR
Extended version — https://arxiv.org/abs/2511.12528

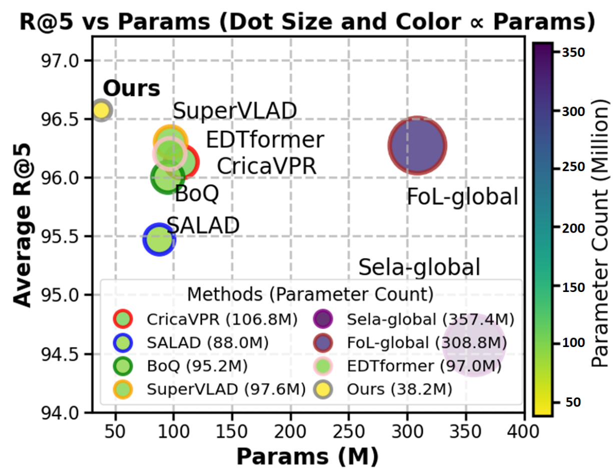

Figure 1: The comparison of average R@5 against parameter count on Pitts30k, MSLS-val, and SPED (with image size of 224×224) demonstrates that our model achieves competitive performance despite significantly reduced parameter count, striking an effective trade-off.

Introduction

Visual Place Recognition (VPR) is to identify the location where a query image is captured by finding the most visually similar image in a geo-tagged reference database (Chen et al. 2017; Arandjelovic et al. 2016). It serves as a cornerstone for numerous real-world applications, including autonomous robot navigation (Lowry et al. 2015), augmented reality (Ventura et al. 2014), and location-based services (Sarlin et al. 2019). VPR offers advantages over other sensing modalities (e.g., LiDAR or RADAR) such as lower cost, easier deployment and passive data acquisition, making it an attractive option for many applications (Miao et al. 2024). However, VPR still faces major challenges from long-term appearance variations (e.g., due to seasonal, weather, or illumination changes), perceptual aliasing (where distinct places appear deceptively similar) and viewpoint shifts (Torii et al. 2015).

Recently, with further breakthroughs in foundational vision models, VPR approaches (Lu et al. 2024c; Ali-bey, Chaib-draa, and Giguère 2024; Izquierdo and Civera 2024; Lu et al. 2024a,b; Jin et al. 2025; Wang et al. 2025) using DINOv2 (Oquab et al. 2023) as the backbone, to some extent, have mitigated the aforementioned challenges. DINOv2, a self-supervised vision transformer pretrained on 142 million diverse web images (spanning varied lighting, seasons, and viewpoints) and filtered via embedding clustering for diversity and quality, leverages large-scale exposure for feature generalization. Combined with a self-supervised teacher-student distillation strategy, it produces patch-level features

[*]Hong Chen is the corresponding author.

robust to long-term appearance changes, perceptual aliasing, and viewpoint shifts. Yet, while yielding more generalizable features, the massive parameter counts of visual foundation models restrict the deployment and application of methods built on them to resource-constrained devices. For instance, LPS-VPR (Nie et al. 2024)—a method using a convolutional neural network (CNN) architecture—has an overall parameter count of 32M. In contrast, existing DINOv2-based VPR methods mostly rely on DINOv2-base (with 86M parameters for the backbone alone) or DINOv2-large (with 300M parameters for the backbone alone).

To preserve the strong feature representation of vision foundation models while substantially reducing parameter count, a natural and direct strategy is to adopt the smaller visual foundation model, e.g., DINOv2-small (with 22M parameters for the backbone alone). However, when replicating the training pipeline of CricaVPR (Lu et al. 2024a) using DINOv2-small as the backbone, our experiments (see Ablation Study) show a pronounced drop in performance compared to using DINOv2-base. This explains why current methods have not selected the smaller variant as the backbone. To overcome this limitation, we still retain the more lightweight DINOv2-small as our backbone. However, instead of directly conducting training related to the VPR task, we adopt a two-stage training strategy—knowledge distillation followed by fine-tuning—which dramatically reduces the parameter count while preserving the rich representational power of the visual foundation model. To minimize knowledge loss during transfer, we introduce a distillation recovery module that aligns teacher and student features through the fusion of shallow and deep representations. Furthermore, to enhance the capacity of spatial-pooling-based aggregators (e.g., CricaVPR's) to represent irregular geometric structures, we design a flexible deformable aggregator that dynamically adapts pooling regions to better capture complex spatial relationships. Our deformable aggregator, inspired by neural top-down attention (Lou and Yu 2025), combines semantic and global features to dynamically deform pooling regions so they precisely fit and emphasize irregular, key local areas. Afterwards, these focused local representations are fed back to reinforce the global features, creating a bidirectional interaction that enables the network to both guide its attention based on semantics and refine its understanding of overall context. Incorporating the improvements mentioned above, our method demonstrates strong competitiveness in both parameter efficiency and performance. While significantly reducing the number of parameters, it retains the robust feature representation capabilities of visual foundation models, showing competitive performance compared to existing state-of-the-art (SOTA) models on popular benchmarks, as illustrated in Figure 1.

To summarize, our work makes the following contributions: **1)** We have designed a two-stage training strategy that combines knowledge distillation and fine-tuning to train a parameter-efficient VPR model based on visual foundation models. Additionally, we propose the Distillation Recovery Module (DRM) to minimize knowledge loss during the knowledge distillation process to the greatest extent possible. **2)** We design a Top-Down-attention-based Deformable Aggregator (TDDA), which controls the deformation of local ROIs through global semantic information and demonstrates better adaptability to irregular structures and regions compared with existing aggregators centered on spatial pooling. **3)** Extensive experiments show that D^2-VPR can deliver competitive performance compared with SOTA methods on popular benchmarks while achieving significant reductions in parameter count.

Related Works

VPR methods fall into two main categories. One-stage VPR generates a single global descriptor per image by aggregating local features—early methods used handcrafted features such as SURF (Bay et al. 2008), while more recent approaches employ deep learning architectures like NetVLAD (Arandjelovic et al. 2016), MixVPR (Ali-bey, Chaib-draa, and Giguère 2023), and AnyLoc (Keetha et al. 2023). These methods facilitate efficient, end-to-end retrieval. Two-stage VPR first retrieves candidates using global descriptors and then refines the top-K results through local feature matching. Representative examples include Patch-NetVLAD (Hausler et al. 2021), which improves precision at the cost of additional computation. Since our goal is to achieve a better trade-off between parameter-efficiency and performance, our method focuses on one-stage VPR.

VPR Based on Visual Foundation Models. With the emergence of visual foundation models trained on massive amounts of data via unsupervised learning, a new wave of VPR methods has adopted DINOv2 (Oquab et al. 2023) as the backbone to build more robust descriptors. For instance, DINO-Mix (Huang et al. 2024) combines DINOv2 with a multi-layer-perceptron-based aggregation, significantly improving performance under challenging illumination and seasonal variations. SALAD (Izquierdo and Civera 2024) fine-tunes DINOv2 and leverages optimal-transport-based aggregation to set new SOTA results for one-stage VPR. EffoVPR (Tzachor et al. 2024) extracts internal attention features from frozen DINOv2 layers to create compact yet discriminative descriptors. EDTformer (Jin et al. 2025) uses a lightweight decoder transformer with low-rank adaptation to refine DINOv2 features efficiently, while SciceVPR (Wan et al. 2025) enhances cross-image correlation and multilayer fusion to boost retrieval robustness. Although these VPR techniques, backed by powerful foundation models, mitigate challenges such as appearance changes, viewpoint variations, and perceptual aliasing to a certain extent, their large parameter size and heavy computational demand greatly limit deployment on edge or resource-constrained devices.

Parameter-efficient VPR aims to balance model compactness and retrieval performance. LPS-VPR (Nie et al. 2024) uses a pooling-centric saliency encoder to fuse multi-scale CNN features, while EPSA-VPR (Nie et al. 2025) introduces patch saliency–weighted aggregation on ResNet-50 for compact, robust descriptors. Though lightweight, these methods lag in performance as their backbone features lack the strong generalization of visual foundation models. TeTRA-VPR (Grainge et al. 2025), with similar goals to ours, also uses knowledge distillation for efficient foundation-model-based VPR. However, the key difference

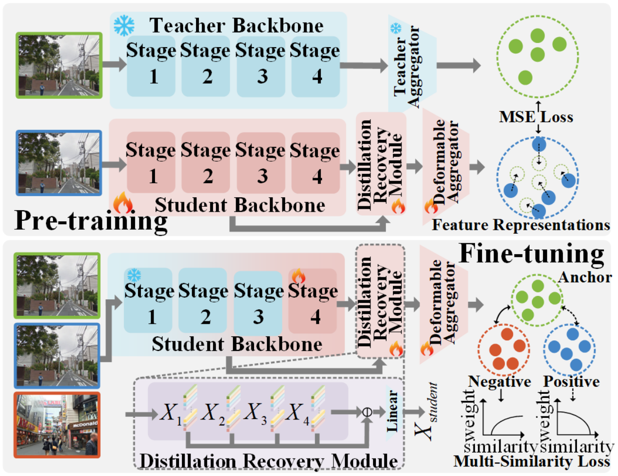

Figure 2: Two training stages of our VPR model.

lies in that their core approach is to binarize the model parameters, which still impairs the performance of visual foundation models, leading to a slight performance drop between the binarized model and the original one.

Our approach aims to bridge the gap between existing efficient VPR models and those based on large-scale pre-trained visual foundation models by integrating their respective strengths. It maintains a lightweight architecture while effectively harnessing the powerful representations learned by foundation models, thereby achieving a more favorable balance between parameter-count and retrieval performance.

Methodology

Two-stage Training Strategy

As shown in Figure 2, the training process of our model is divided into two stages: a pre-training stage centered on knowledge distillation (Mean Squared Error (MSE) loss is used here), and a fine-tuning stage centered on the multi-similarity loss (Wang et al. 2019).

Knowledge Distillation Based Pre-training aims to compress and transfer the knowledge of a teacher model based on a visual foundation model to a student model, thereby achieving effective parameter initialization, particularly for the backbone. This process enables the lightweight student model to approximate the teacher model's semantic understanding and representation capacity while maintaining a lower computational burden. Specifically, we employ CricaVPR (Lu et al. 2024a) with a DINOv2-base backbone as the teacher model, and select DINOv2-small as the backbone of our D^2-VPR (student model). To minimize knowledge loss during knowledge transfer, a distillation recovery module is introduced to align the feature dimensions of both models. At this stage, we optimize the entire parameter set of the student model by minimizing the MSE loss between the output features of the teacher and student models, ensuring the effective transfer of representational knowledge.

Fine-tuning stage is to further enhance the representational capacity of the student model by adapting it to the VPR task. This is achieved by updating a limited number of parameters, primarily from the deeper layers closer to the output, while keeping most of the parameters of the backbone fixed. Specifically, we adopt the multi-similarity loss (Wang et al. 2019) as the optimization loss as shown in Equation 1, which is widely used in metric learning for VPR tasks. During this stage, the parameters of the first three stages of the backbone are frozen to prevent the catastrophic forgetting of the knowledge transferred during pre-training, and only the last stage of the backbone and the following layers are updated.

$$\mathcal{L}_{\text{MS}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{\alpha}\log\left(1+\sum_{j\in P_i}\exp\left(-\alpha(s_{ij}-\lambda)\right)\right)\right. \\ \left.+\frac{1}{\beta}\log\left(1+\sum_{k\in N_i}\exp\left(\beta(s_{ik}-\lambda)\right)\right)\right], \tag{1}$$

where N is the number of training samples, $s_{ij} = \langle F_i, F_j \rangle$ denotes the cosine similarity between features, P_i and N_i represent the mined positive and negative sets for anchor i, α and β are weighting hyperparameters, and λ is a margin.

Model Architecture

As shown in Figure 2, the overall model architecture is divided into three parts. 1) The backbone, a small visual foundation model, is responsible for extracting features from images. 2) The distillation recovery module fuses the features of the backbone from shallow to deep layers and aligns them with the feature dimensions of the teacher model, so as to avoid losses in knowledge transfer to the greatest extent. 3) The top-down-attention-based deformable aggregator, on the other hand, further compresses the extracted features to form a compact feature representation.

Backbone. We adopt a vision transformer backbone (DINOv2-small) to extract features from the input image. The output $X_4 \in \mathbb{R}^{B\times(1+P)\times D/2}$ from the final transformer layer (stage 4) consists of a global class token and patch-level tokens:

$$X_4 = \{f_{\text{cls}}, f_1, f_2, \ldots, f_P\}, \tag{2}$$

where $f_{\text{cls}} \in \mathbb{R}^{D/2}$ is the class token, $f_i \in \mathbb{R}^{D/2}$ are the spatial patch tokens, and $D/2$ is the feature dimension (D for tokens' dimension of teacher backbone).

Distillation Recovery Module. To align the feature output dimensions of the teacher model and the student model, thus reducing knowledge loss during pre-training, we introduce a distillation recovery module at the end of the student backbone as shown in Figure 2. This module recovers feature dimensions consistent with those of the teacher model by fusing features from shallow to deep layers. Specifically, since the teacher model uses DINOv2-base with an output feature dimension (D) that is twice that of our student model ($D/2$), we bridge this gap by concatenating hierarchical features from all four stages of the student backbone in a shallow-to-deep order and projecting them into the teacher's feature space via a linear transformation:

$$X_{\text{student}} = \text{Linear}\left(\text{Concat}(X_1, X_2, X_3, X_4)\right), \tag{3}$$

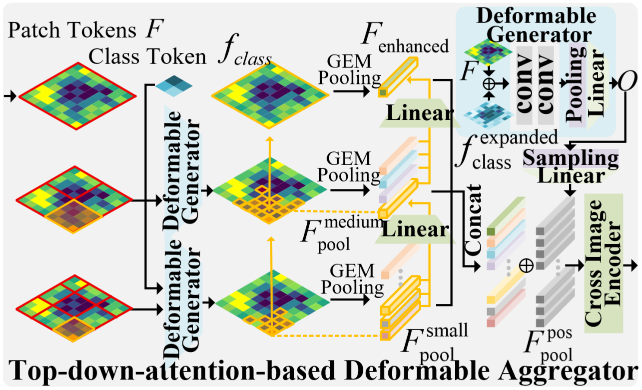

Figure 3: Top-down-attention-based deformable aggregator.

where X_i denotes the output token sequence from the i-th stage of the student backbone. This alignment mechanism constructs a shared knowledge distillation space and enables the student model to mimic the teacher's semantic representations while maintaining low computational complexity. Finally, $X_{\text{student}} \in \mathbb{R}^{B\times(1+P)\times D}$ is rearranged into a spatial feature map $F \in \mathbb{R}^{B\times D\times H\times W}$ and class token as $f_{\text{class}} \in \mathbb{R}^{B\times D}$ for subsequent aggregation.

Top-down-attention-based Deformable Aggregator. Top-down neural attention (Gilbert and Sigman 2007) plays a crucial role in human visual perception. It suggests that the brain first constructs a rapid and abstract interpretation of a scene, which is then employed to guide and refine the processing of incoming sensory signals. This ultimately results in more accurate recognition of the positions, shapes, and categories of objects. As shown in Figure 3, the design of our TDDA follows this concept by utilizing global and semantic information to drive corresponding deformations of ROIs, enabling the model to better focus on irregular geometric regions, and is divided into the following components: multi-scale pyramid ROI, top-down deformable region pooling, hierarchical down-top fusion, and deformation-aware position embedding.

Multi-Scale Pyramid ROI. To align with homogeneous knowledge distillation (the teacher and student models have the same architecture) and thereby improve distillation efficiency (Gou et al. 2021), we adopt the same multi-scale pooling strategy as employed in the teacher's aggregator (CricaVPR). Specifically, we divide the input image into multiple spatial regions at different granularities, including one global region, four medium-scale regions, and nine fine-grained small regions as indicated by the red line division in Figure 3. Each region is defined by an ROI (x_1, y_1, x_2, y_2) over the backbone feature map and is processed through the subsequent deformable region pooling mechanism.

Top-down Deformable Region Pooling. To improve the model's perception flexibility toward irregular and non-rigid spatial regions, we propose a top-down-attention-based deformable region pooling method. This mechanism allows each region to adaptively adjust its receptive field based on the fusion of local content (ROI patch token) and global semantic context (class token), improving robustness to viewpoint, scale, and structural variations. Specifically, for an ROI, we construct a sampling grid $\mathcal{G}_{\text{base}} \in \mathbb{R}^{H'\times W'\times 2}$ over this region, where H' and W' represent the resolution of the sampling grid. To incorporate global context, the class token f_{class} is spatially broadcast and concatenated with the full feature map along the channel dimension:

$$F_{\text{fused}} = \text{Concat}(F, f_{\text{class}}^{\text{expanded}}), \tag{4}$$

where $F_{\text{fused}} \in \mathbb{R}^{B\times 2D\times H\times W}$. Then, a deformable generator (contains two CNN layers as shown in the top-right corner of Figure 3) is applied to predict four transformation parameters for each spatial location:

$$O = \text{Deformable-Generator}(F_{\text{fused}}), \tag{5}$$

where $O \in \mathbb{R}^{B\times 4\times H\times W}$ contains horizontal and vertical offsets $(\Delta x, \Delta y)$, and raw scaling factors (s_w, s_h). To obtain localized transformation parameters for each region, we sample the offset and scaling fields at the base grid positions using bilinear interpolation:

$$\tilde{O}_{\text{ROI}} = \text{GridSample}(O, \mathcal{G}_{\text{base}}), \tag{6}$$

where $\tilde{O}_{\text{ROI}} \in \mathbb{R}^{B\times H'\times W'\times 4}$ denotes the transformation parameters ($\tilde{\Delta x}, \tilde{\Delta y}, \tilde{s_w}, \tilde{s_h}$) at each sampled point of the ROI region. The sampled parameters are subsequently used to deform the base grid coordinates as follows:

$$\begin{aligned} x_{\text{deform}} &= x_{\text{center}} + (x_{\text{rel}} \cdot \tilde{s_w} + \tilde{\Delta x}) \cdot \frac{w}{2}, \\ y_{\text{deform}} &= y_{\text{center}} + (y_{\text{rel}} \cdot \tilde{s_h} + \tilde{\Delta y}) \cdot \frac{h}{2}, \end{aligned} \tag{7}$$

where $(x_{\text{rel}}, y_{\text{rel}}) \in [-1, 1]$ are the normalized coordinates from the base grid, and $(x_{\text{center}}, y_{\text{center}}, w, h)$ denote the center and size of the ROI. x_{deform} and y_{deform} constitute the deformable sampling grid $\mathcal{G}_{\text{deform}}$, which is then applied to the original feature map using bilinear interpolation:

$$F_{\text{region}} = \text{GridSample}(F, \mathcal{G}_{\text{deform}}). \tag{8}$$

Each sampled region feature map $F_{\text{region}} \in \mathbb{R}^{B\times D\times H'\times W'}$ (as shown by the orange irregular region in Figure 3) is then aggregated into a compact vector representation using Generalized Mean Pooling (GeM) (Radenović, Tolias, and Chum 2018), which introduces a learnable parameter p to balance between average and max pooling:

$$F_{\text{pool}} = \left(\frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} F_{\text{region}}^{(p)}[:, :, i, j] \right)^{1/p}. \tag{9}$$

Hierarchical Down-top Fusion. After obtaining the deformed region pooled features driven by the semantic class tokens, we design a down-top fusion strategy that progressively enhances the global feature representation by incorporating fine-grained local deformable information as shown by the orange arrow in Figure 3, thereby forming a bidirectional interaction mechanism. Specifically, each deformable region—whether at the medium or global level—is refined

Dataset	Description	Number	
		Database	Queries
Pitts30k	urban, panorama	10,000	6,816
Pitts250k	urban, panorama	83952	8280
MSLS-val	urban, suburban	18,871	740
Nordland	natural, seasonal	27,592	27,592
AmsterTime	very long-term	1,231	1,231
SPED	various scenes	607	607

Table 1: Summary of the test datasets in experiments.

by aggregating its spatially neighboring lower-level regions based on deformed region centers, followed by a linear projection and residual addition. This unified fusion process can be formulated as:

$$F_{\text{enhanced}} = F_{\text{pool}}^{\text{top}} + \text{Linear}\left(\frac{1}{N}\sum_{i=1}^{N} F_{\text{pool}}^{i}\right), \tag{10}$$

where $F_{\text{pool}}^{\text{top}}$ is the top target region feature to be enhanced (e.g., medium or global), $\{F_{\text{pool}}^{i}\}_{i=1}^{N}$ are its associated down (e.g., medium or small) neighboring region pooled features.

Deformation-aware Position Embedding. Deformable region pooling disrupts the spatial regularity of fixed-grid regions, making it difficult for the model to retain explicit awareness of the original location and size of each region (before deformation, the ROI positions and sizes are fixed, and the ordered feature arrangement enables the model to easily perceive this information). To address this, we embed deformation-specific geometric information for each region using its post-deformation center and size. Specifically, for a region, $\tilde{O}_{\text{ROI}}$ is then projected through a linear layer and added to the region feature to inject geometric context:

$$F_{\text{pool}}^{\text{pos}} = F_{\text{pool}} + \text{Linear}(\tilde{O}_{\text{ROI}}). \tag{11}$$

After injecting deformation-aware positional information into each region feature, all region descriptors from multiple scales are concatenated to form a unified feature set. Following (Lu et al. 2024a), we feed this sequence into a cross-image transformer encoder. The output sequence is then flattened and subjected to L2 normalization to obtain the final descriptor, yielding compact representations.

Experiments

Benchmarks

We conduct experiments on multiple VPR benchmark datasets that exhibit viewpoint variations, environmental condition changes, and perceptual aliasing challenges. Table 1 summarizes these datasets: Pitts30k and Pitts250k (Torii et al. 2013) primarily feature significant viewpoint changes; MSLS (Warburg et al. 2020) spans urban, suburban and natural scenes captured several years with diverse visual variations; SPED (Zaffar et al. 2021) comprises surveillance camera imagery. Additionally, we include challenging datasets: Nordland (seasonal variations) (Sünderhauf, Neubert, and Protzel 2013) and AmsterTime (long-term changes) (Yildiz et al. 2022).

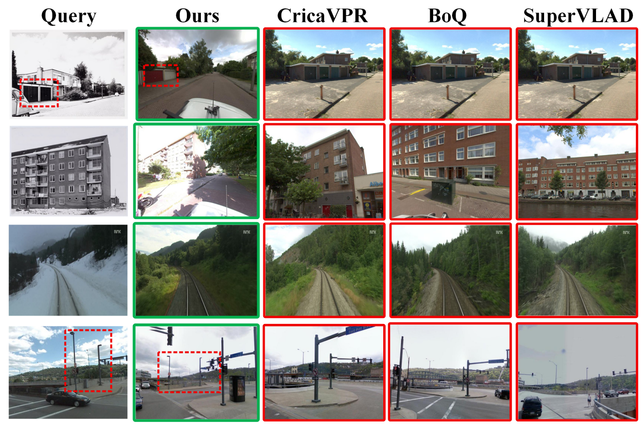

Figure 4: Qualitative VPR comparison results. Our method demonstrates competitive performance compared to these DINOv2-based SOTA models under these challenging cases: long-term appearance changes (first row), drastic lighting variations (second row), perceptual aliasing (third row), and viewpoint changes (fourth row). Green indicates the right match while red is for the wrong one. Key matching regions are highlighted with red dashed boxes.

We employ Recall@N (R@N) as the evaluation metric, measuring the percentage of queries where at least one top-N retrieved database image is within a ground truth threshold. Following standard protocols (Warburg et al. 2020; Torii et al. 2013), thresholds are: 25m + 40° for MSLS; 25m for Pitts30k, Pitts250k and SPED; ± 10 frames for Nordland; and unique counterpart matching for AmsterTime.

Implementation Details

Our training follows a two-stage strategy. In the knowledge distillation stage, we use CricaVPR with DINOv2-base as the teacher model, and our D^2-VPR (student model) uses DINOv2-small (initialized with pretrained weights) as the backbone. During this stage, all parameters of the student model are trained. We use the ADAM (Kingma and Ba 2014) optimizer with a batch size of 8 and a learning rate of 2.5e-5. In the fine-tuning stage, we freeze the first 3/4 layers of the backbone and train the remaining parameters. This stage uses the ADAMW (Loshchilov, Hutter et al. 2017) optimizer with a batch size of 128 and a learning rate of 2e-4. Both stages are trained on the GSV-Cities dataset (Ali-bey, Chaib-draa, and Giguere 2022), a large-scale urban location dataset collected via Google Street View. Each batch consists of 4 images, and the input image size for both training and evaluation is set to 224×224. All experiments are conducted on an RTX 3090 GPU, with PyTorch 2.3.0 and Python 3.10. Our model outputs 10752-dimensional global features, and following (Lu et al. 2024a), we apply principal component analysis (PCA) for dimensionality reduction to 4096 dimensions.

Comparisons with State-of-the-art Methods

Comparison with Similar-scale Baselines. We conduct evaluations against several similar-scale approaches, includ-

Method	Pitts30k			MSLS-val			SPED			Pitts250k			AmsterTime			Nordland		
	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	R@1	R@5	R@10
Cosplace	90.9	95.7	96.7	87.4	94.1	94.9	-	-	-	92.3	97.4	98.4	47.7	69.8	75.8	71.9	83.8	88.1
MixVPR	91.5	95.5	96.3	87.2	93.1	94.3	-	-	-	94.1	98.2	**98.9**	40.2	59.1	64.6	76.2	86.9	90.3
EigenPlaces	**92.5**	**96.8**	**97.6**	89.1	93.8	95.0	-	-	-	94.1	97.9	98.7	48.9	69.5	76.0	71.2	83.8	88.1
LPS-VPR	91.8	96.0	96.8	89.9	94.2	95.0	84.8	**93.9**	95.7	94.1	98.0	**98.9**	-	-	-	-	-	-
BoQ†	92.0	95.6	96.6	86.6	92.3	93.5	82.7	91.3	94.2	**94.4**	97.9	98.7	42.0	60.5	66.5	**78.9**	88.5	91.5
EPSA-VPR	-	-	-	89.3	93.8	95.4	84.5	93.7	**96.0**	93.9	97.9	98.7	-	-	-	-	-	-
Clus-VPR	90.8	95.2	96.6	82.7	88.5	92.4	-	-	-	92.4	96.9	97.6	-	-	-	-	-	-
D^2-VPR$_{\text{no encoder}}$	91.7	95.8	96.8	**90.7**	**95.4**	**96.4**	**86.0**	92.9	94.1	**94.4**	**98.3**	**98.9**	**49.1**	**70.7**	**76.1**	77.1	**88.6**	**91.9**

Table 2: Comparison with similar-scale SOTA methods on popular benchmarks. The best results are highlighted in **bold** and the second best are underlined. BoQ† is re-evaluated using the officially provided weights. The input image size is set to 224×224, and the feature dimension is reduced from 16384 to 4096 using PCA for a fairer comparison. - for not reported. Results of Cosplace and MixVPR are reported from EigenPlaces. Here, D^2-VPR does not use cross-image encoder.

Method	Dim.	Backbone	Image Size	Param. (M)
Cosplace CVPR' 2022	2048	ResNet50	No Resize	27.70
MixVPR WACV' 2023	4096	ResNet50	No Resize	**10.88**
EigenPlaces ICCV' 2023	2048	ResNet50	No Resize	27.70
LPS-VPR RAL' 2024	2048	ResNet50	640×480	29.71
BoQ CVPR' 2024	4096	ResNet50	224×224	23.84
EPSA-VPR JVCI' 2025	1024	ResNet50	-	27.71
Clus-VPR TAI' 2025	4096	CWTNet	640×480	53.12
D^2-VPR$_{\text{no encoder}}$	4096	Dinov2	224×224	27.21

Table 3: Detailed information of the similar-scale SOTA methods in the comparison. Clus-VPR and EPSA-VPR report only the parameters of their aggregators. Therefore, we use their full ResNet50 version and report 23.51M parameters for the backbone. Note that MixVPR and BoQ utilize a cropped ResNet-50 as their backbone, with the parameter count of the backbone network being less than 23.51M.

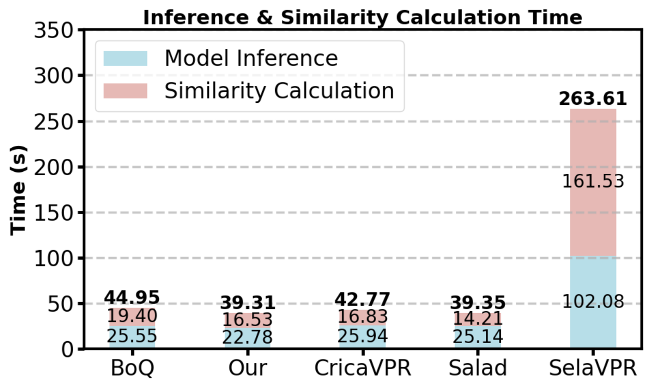

Figure 5: Method comparison of inference and computational speed on AmsterTime. The inference time includes the inference of both the database and the query set. The batch size is 32. SelaVPR performs calculations in a two-stage manner. PCA is not used here.

ing LPS-VPR (Nie et al. 2024), Clus-VPR (Xu et al. 2024), EPSA-VPR (Nie et al. 2025), BoQ (Ali-bey, Chaib-draa, and Giguère 2024), Cosplace (Berton, Masone, and Caputo 2022), MixVPR (Ali-bey, Chaib-draa, and Giguère 2023) and EigenPlaces (Berton et al. 2023). These methods adopt relatively lightweight CNN architectures. As shown in Tables 2 and 3, our proposed method achieves the best performance on most datasets, particularly on MSLS-val (exceeding the second-best by 1.2 in R@5), AmsterTime, Nordland (exceeding the second-best by 1.2 in R@5), and Pitts250k, while maintaining a competitive parameter size and using smaller input image resolution of 224×224.

Comparison with Larger-scale Baselines. We also comprehensively compare our proposed method with SOTA one-stage VPR methods in Tables 4 and 5, including CricaVPR (Lu et al. 2024a), SALAD (Izquierdo and Civera 2024), BoQ (Ali-bey, Chaib-draa, and Giguère 2024), SuperVLAD (Lu et al. 2024c), SelaVPR (Lu et al. 2024b), FoL (Wang et al. 2025), and EDTformer (Jin et al. 2025). Note that we have unified the evaluation image resolution to 224×224 rather than a higher resolution, which aligns with our goal of deploying the model on resource-constrained devices. When using the cross-image encoder, the results demonstrate competitive performance, particularly on Pitts30k, MSLS-val, Pitts250k, and AmsterTime. Figure 4 provides visualizations comparing the retrieval performance of our method with other baselines. However, due to the inherent limitations of the cross-image encoder, we also report comparisons without it. In this setting, our method achieves performance comparable to SALAD on Pitts30k, MSLS-val, and Pitts250k, and performs better than FoL-global on Nordland, indicating that our approach still maintains a reasonable level of competitiveness, despite requiring substantially fewer parameters than the competing methods.

Comparison of Inference and Computational Speed. As shown in Figure 5, our method achieves competitive inference speed and overall processing time. In terms of similarity calculation time, our approach ranks as the second fastest (here our method does not use PCA and retains the original feature dimension of 10752)—behind SALAD (Izquierdo and Civera 2024), whose descriptor di-

Method	Pitts30k			MSLS-val			SPED			Pitts250k			AmsterTime			Nordland		
	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	R@1	R@5	R@10
CricaVPR*	**94.9**	**97.3**	**98.2**	90.0	95.4	96.4	**91.9**	95.7	96.7	97.5	**99.4**	**99.7**	**64.7**	**82.8**	**87.5**	**90.7**	**96.3**	**97.6**
SALAD	91.6	95.7	97.1	90.5	95.5	96.2	90.8	95.2	96.9	94.6	98.1	98.9	53.4	75.0	79.9	81.2	91.1	94.0
BoQ	93.1	96.5	97.5	**91.6**	95.9	96.8	91.8	95.6	96.5	95.9	98.8	99.4	57.6	76.9	81.9	85.8	93.6	95.9
SuperVLAD*	94.1	**97.3**	98.0	90.7	96.0	96.8	90.9	95.6	96.5	96.1	99.0	99.5	60.0	80.3	84.4	88.6	94.7	96.5
Sela-global	90.2	96.1	97.1	87.7	95.8	96.6	84.5	91.8	93.9	92.8	98.0	98.9	41.5	62.1	69.3	72.3	89.4	94.4
FoL-global	92.6	96.9	97.7	90.4	95.7	96.9	90.6	**96.2**	96.9	95.3	98.8	99.4	54.1	76.0	81.0	74.7	86.9	91.0
EDTformer	92.9	96.8	97.8	91.5	**96.4**	96.6	90.9	95.4	96.7	95.5	98.7	99.3	58.2	80.8	84.8	81.0	91.2	94.1
D^2-VPR$_{\text{no encoder}}$	91.7	95.8	96.8	90.7	95.4	96.4	86.0	92.9	94.1	94.4	98.3	98.9	49.1	70.7	76.1	77.1	88.6	91.9
D^2-VPR$^*_{\text{encoder}}$	**94.9**	**97.3**	98.0	**91.6**	96.1	**97.2**	90.9	96.0	**97.4**	**97.8**	**99.4**	**99.7**	62.9	80.8	85.3	86.6	94.1	96.0

Table 4: We compare our method with larger-scale SOTA methods on popular benchmarks. To ensure a fair comparison, the evaluation resolution is set to 224×224. Except for CricaVPR, all other methods are originally evaluated at higher resolutions, so we re-evaluate them using their source code and provided weights. Therefore, the results may differ from those reported in the original papers. Methods with the suffix '-global' correspond to the first stage of the two-stage approach.* for using cross-image encoder.

Method	Dim.	Dinov2	Image Size	Param. (M)
CricaVPR CVPR' 2024	4096	Base	224×224	106.76
SALAD CVPR' 2024	8448	Base	224×224	87.99
BoQ CVPR' 2024	12288	Base	224×224	95.21
SuperVLAD NIPS' 2024	3072	Base	224×224	97.61
Sela-global ICLR' 2024	1024	Large	224×224	357.43
FoL-global AAAI' 2025	8448	Large	224×224	308.83
EDTformer TCSVT' 2025	4096	Base	224×224	96.96
D^2-VPR no encoder	4096	Small	224×224	**27.21**
D^2-VPR encoder	4096	Small	224×224	38.24

Table 5: Detailed information of the larger-scale SOTA methods in the comparison.

mensionality is smaller (8,848)—and remains comparable to CricaVPR (Lu et al. 2024a).

We observe that the improvement in inference time is not as substantial as expected. This is primarily because the deformable aggregator introduces multiple sampling operations that involve grid generation and bilinear interpolation. These operations cannot be effectively parallelized or fused like standard convolution, thereby constraining the inference speed. We plan to explore more efficient and hardware-friendly designs for this component in future work to improve the overall acceleration.

Ablation Study

We conduct ablation studies on both training strategies and module designs, as shown in Table 6.

Training Strategy. Compared to the baseline (direct fine-tuning without distillation), the introduction of knowledge distillation leads to a significant performance gain on MSLS-val (+3.3%, +1.5%, and +1.1% on R@1/5/10), demonstrating effective knowledge transfer from the teacher model. Applying fine-tuning afterward yields further improvements (+3.2%, +1.1%, and +0.8%), confirming its crucial role in adapting the model to the final retrieval task.

Configurations	MSLS-val		
	R@1	R@5	R@10
Baseline	85.1	93.5	95.3
+distillation	88.4	95.0	96.4
+finetune	91.6	96.1	97.2
Baseline	89.9	95.0	95.7
+DRM	90.5	95.5	96.2
+TDDA	91.6	96.1	97.2

Table 6: Ablation study on the MSLS-val benchmark.

Module Design. Adding the distillation recovery module leads to improvement (+0.6%, +0.5%, +0.5%), showing that it effectively fuses shallow and deep backbone features and reduces knowledge-transfer loss. Introducing the deformable aggregator yields the best performance (+1.1%, +0.6%, +1.0%), demonstrating its ability to flexibly handle irregular ROIs and enhance overall retrieval accuracy.

Conclusion

In this work, we present D^2-VPR, a lightweight visual-foundation-model-based framework that combines knowledge distillation and deformable aggregation to retain the strong representation capabilities of visual foundation models while significantly reducing computational cost. Through a two-stage training strategy and the proposed distillation recovery module, our method effectively bridges the feature gap between teacher and student models. The top-down-attention-based deformable aggregator further enhances adaptability by dynamically adjusting aggregation regions based on global semantics. Extensive experiments demonstrate that D^2-VPR achieves a favorable balance between performance and efficiency.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2024YFF0907404.

References

Ali-bey, A.; Chaib-draa, B.; and Giguere, P. 2022. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513: 194–203.

Ali-bey, A.; Chaib-draa, B.; and Giguère, P. 2023. MixVPR: Feature Mixing for Visual Place Recognition. In *WACV*, 2998–3007.

Ali-bey, A.; Chaib-draa, B.; and Giguère, P. 2023. MixVPR: Feature Mixing for Visual Place Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2998–3007.

Ali-bey, A.; Chaib-draa, B.; and Giguère, P. a. 2024. BoQ: A Place is Worth a Bag of Learnable Queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17794–17803.

Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 5297–5307.

Bay, H.; Ess, A.; Tuytelaars, T.; and Gool, L. V. 2008. Speeded-up robust features (SURF). *Computer vision and image understanding*, 110(3): 346–359.

Berton, G.; Masone, C.; and Caputo, B. 2022. Rethinking Visual Geo-Localization for Large-Scale Applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4878–4888.

Berton, G.; Trivigno, G.; Caputo, B.; and Masone, C. 2023. EigenPlaces: Training Viewpoint Robust Models for Visual Place Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 11080–11090.

Chen, Z.; Jacobson, A.; Sünderhauf, N.; Upcroft, B.; Liu, L.; Shen, C.; Reid, I.; and Milford, M. 2017. Deep learning features at scale for visual place recognition. In *2017 IEEE international conference on robotics and automation (ICRA)*, 3223–3230. IEEE.

Gilbert, C. D.; and Sigman, M. 2007. Brain States: Top-Down Influences in Sensory Processing. *Neuron*, 54(5): 677–696.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6): 1789–1819.

Grainge, O.; Milford, M.; Bodala, I.; Ramchurn, S. D.; and Ehsan, S. 2025. TeTRA-VPR: A Ternary Transformer Approach for Compact Visual Place Recognition. arXiv:2503.02511.

Hausler, S.; Garg, S.; Xu, M.; Milford, M.; and Fischer, T. 2021. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *CVPR*, 14141–14152.

Huang, G.; Zhou, Y.; Hu, X.; Zhang, C.; Zhao, L.; and Gan, W. 2024. DINO-Mix enhancing visual place recognition with foundational vision model and feature mixing. *Scientific Reports*, 14(1): 22100.

Izquierdo, S.; and Civera, J. 2024. Optimal Transport Aggregation for Visual Place Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jin, T.; Lu, F.; Hu, S.; Yuan, C.; and Liu, Y. 2025. EDTformer: An Efficient Decoder Transformer for Visual Place Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

Keetha, N.; Mishra, A.; Karhade, J.; Jatavallabhula, K. M.; Scherer, S.; Krishna, M.; and Garg, S. 2023. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Loshchilov, I.; Hutter, F.; et al. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5(5): 5.

Lou, M.; and Yu, Y. 2025. OverLoCK: An Overview-first-Look-Closely-next ConvNet with Context-Mixing Dynamic Kernels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 128–138.

Lowry, S.; Sünderhauf, N.; Newman, P.; Leonard, J. J.; Cox, D.; Corke, P.; and Milford, M. J. 2015. Visual place recognition: A survey. *ieee transactions on robotics*, 32(1): 1–19.

Lu, F.; Lan, X.; Zhang, L.; Jiang, D.; Wang, Y.; and Yuan, C. 2024a. CricaVPR: Cross-image Correlation-aware Representation Learning for Visual Place Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lu, F.; Zhang, L.; Lan, X.; Dong, S.; Wang, Y.; and Yuan, C. 2024b. Towards Seamless Adaptation of Pre-trained Models for Visual Place Recognition. In *ICLR*.

Lu, F.; Zhang, X.; Ye, C.; Dong, S.; Zhang, L.; Lan, X.; and Yuan, C. 2024c. SuperVLAD: Compact and Robust Image Descriptors for Visual Place Recognition. In *Advances in Neural Information Processing Systems*, volume 37, 5789–5816.

Miao, J.; Jiang, K.; Wen, T.; Wang, Y.; Jia, P.; Wijaya, B.; Zhao, X.; Cheng, Q.; Xiao, Z.; Huang, J.; Zhong, Z.; and Yang, D. 2024. A Survey on Monocular Re-Localization: From the Perspective of Scene Map Representation. *IEEE Transactions on Intelligent Vehicles*, 1–33.

Nie, J.; Xue, D.; Pan, F.; Ning, Z.; Liu, W.; Hu, J.; and Cheng, S. 2024. Efficient saliency encoding for visual place recognition: Introducing the lightweight pooling-centric saliency-aware VPR method. *IEEE Robotics and Automation Letters*, 9(7): 6035–6042.

Nie, J.; Zhào, Q.; Xue, D.; Pan, F.; and Liu, W. 2025. EPSA-VPR: A lightweight visual place recognition method with an Efficient Patch Saliency-weighted Aggregator. *Journal of Visual Communication and Image Representation*, 104440.

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

Radenović, F.; Tolias, G.; and Chum, O. 2018. Fine-tuning CNN image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7): 1655–1668.

Sarlin, P.-E.; Cadena, C.; Siegwart, R.; and Dymczyk, M. 2019. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *CVPR*.

Sünderhauf, N.; Neubert, P.; and Protzel, P. 2013. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. 2013.

Torii, A.; Arandjelovic, R.; Sivic, J.; Okutomi, M.; and Pajdla, T. 2015. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1808–1817.

Torii, A.; Sivic, J.; Pajdla, T.; and Okutomi, M. 2013. Visual place recognition with repetitive structures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 883–890.

Tzachor, I.; Lerner, B.; Levy, M.; Green, M.; Shalev, T. B.; Habib, G.; Samuel, D.; Zailer, N. K.; Shimshi, O.; Darshan, N.; et al. 2024. Effovpr: Effective foundation model utilization for visual place recognition. *arXiv preprint arXiv:2405.18065*.

Ventura, J.; Arth, C.; Reitmayr, G.; and Schmalstieg, D. 2014. Global localization from monocular slam on a mobile phone. *IEEE transactions on visualization and computer graphics*, 20(4): 531–539.

Wan, S.; Wei, Y.; Kang, L.; Shen, T.; Wang, H.; and Yang, Y.-H. 2025. SciceVPR: Stable Cross-Image Correlation Enhanced Model for Visual Place Recognition. *arXiv preprint arXiv:2502.20676*.

Wang, C.; Chen, S.; Song, Y.; Xu, R.; Zhang, Z.; Zhang, J.; Yang, H.; Zhang, Y.; Fu, K.; Du, S.; et al. 2025. Focus on Local: Finding Reliable Discriminative Regions for Visual Place Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7536–7544.

Wang, X.; Han, X.; Huang, W.; Dong, D.; and Scott, M. R. 2019. Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5022–5030.

Warburg, F.; Hauberg, S.; Lopez-Antequera, M.; Gargallo, P.; Kuang, Y.; and Civera, J. 2020. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2626–2635.

Xu, Y.; Shamsolmoali, P.; Zareapoor, M.; and Yang, J. 2024. ClusVPR: Efficient Visual Place Recognition with Clustering-based Weighted Transformer. *IEEE Transactions on Artificial Intelligence*.

Yildiz, B.; Khademi, S.; Siebes, R. M.; and Van Gemert, J. 2022. Amstertime: A visual place recognition benchmark dataset for severe domain shift. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 2749–2755. IEEE.

Zaffar, M.; Garg, S.; Milford, M.; Kooij, J.; Flynn, D.; McDonald-Maier, K.; and Ehsan, S. 2021. Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *International Journal of Computer Vision*, 1–39.