Focus on Local: Finding Reliable Discriminative Regions for Visual Place Recognition

Changwei Wang[1,3], Shunpeng Chen[2], Yukun Song[2], Rongtao Xu[4], Zherui Zhang[2], Jiguang Zhang[4], Haoran Yang[5], Yu Zhang[5], Kexue Fu[1,3], Shide Du[6], Zhiwei Xu[7], Longxiang Gao[1,3,*], Li Guo[2], Shibiao Xu[2]

[1]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences)
[2]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[3]Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science
[4]MAIS, Institute of Automation, Chinese Academy of Sciences
[5]Tongji University
[6] College of Computer and Data Science, Fuzhou University
[7]Shandong University
gaolx@sdas.org

Abstract

Visual Place Recognition (VPR) is aimed at predicting the location of a query image by referencing a database of geo-tagged images. For VPR task, often fewer discriminative local regions in an image produce important effects while mundane background regions do not contribute or even cause perceptual aliasing because of easy overlap. However, existing methods lack precisely modeling and full exploitation of these discriminative regions. In this paper, we propose the Focus on Local (FoL) approach to stimulate the performance of image retrieval and re-ranking in VPR simultaneously by mining and exploiting reliable discriminative local regions in images and introducing pseudo-correlation supervision. First, we design two losses, Extraction-Aggregation Spatial Alignment Loss (SAL) and Foreground-Background Contrast Enhancement Loss (CEL), to explicitly model reliable discriminative local regions and use them to guide the generation of global representations and efficient re-ranking. Second, we introduce a weakly-supervised local feature training strategy based on pseudo-correspondences obtained from aggregating global features to alleviate the lack of local correspondences ground truth for the VPR task. Third, we suggest an efficient re-ranking pipeline that is efficiently and precisely based on discriminative region guidance. Finally, experimental results show that our FoL achieves the state-of-the-art on multiple VPR benchmarks in both image retrieval and re-ranking stages and also significantly outperforms existing two-stage VPR methods in terms of computational efficiency.

Code — https://github.com/chenshunpeng/FoL

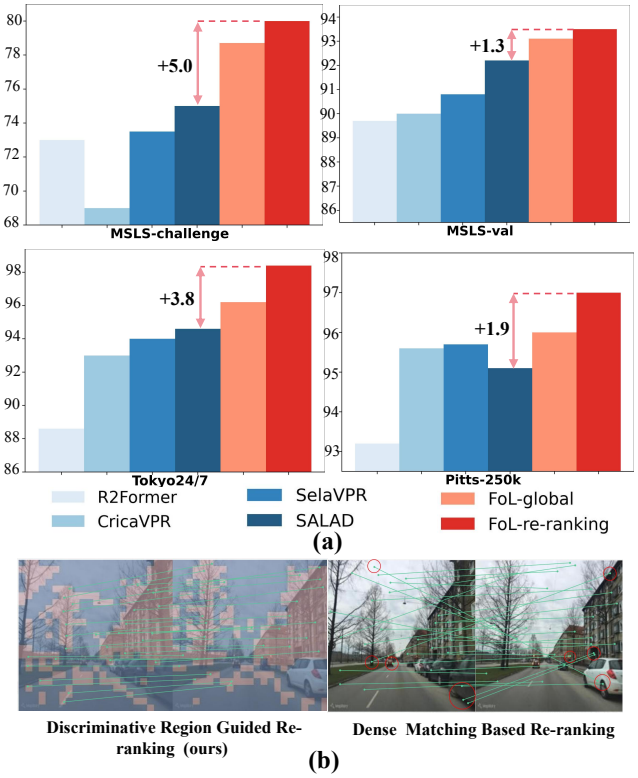


Figure 1: (a) The results show that our FoL achieves state-of-the-art performance in the image retrieval phase alone and significantly outperforms recent methods after re-ranking. (b) Our FoL proposes to do local matching only in the discriminative region, which not only improves the accuracy of re-ranking but also greatly improves the efficiency.

Introduction

Visual Place Recognition (VPR), also known as visual geolocalization (Berton, Masone, and Caputo 2022a), is used to retrieve images of the most similar locations from geo-tagged databases to estimate the rough geographic loca-

*Longxiang Gao is the corresponding author.

tion of a query image and is an essential part of many robotics (Chen et al. 2017) and computer vision (Middelberg et al. 2014; Wang et al. 2019b; Guo et al. 2020; Yuan et al. 2024; Li et al. 2024) tasks. Early approaches viewed VPR as a single-stage image retrieval problem (Datta et al. 2008), using global features (*a.k.a.* global descriptors) to represent the location image, similarity search in that feature space, and determining the localized location by nearest neighbor matching. However, these global features ignore spatial information, making single-stage VPR methods based on these features prone to perceptual aliasing (Izquierdo and Civera 2024), *i.e.*, it is difficult to distinguish highly similar images taken from different locations.

To utilize the spatial information, some recent VPR methods have proposed two-stage promising solutions (Hausler et al. 2021; Zhu et al. 2023; Lu et al. 2024c), in which they retrieve the top k candidates in the database using global features, and then rerank these candidates by matching local features. Although the two-stage methods have yielded encouraging results in terms of accuracy through the introduction of spatial information checking, there are still three problems with these methods that have not been well addressed: ***i)*** They lacked modeling and utilization of discriminatory regions, which are critical for VPR task; ***ii)*** The lack of local correspondence labels in the existing VPR dataset leads to ineffective local matching supervision of the re-ranking stage; ***iii)*** Matching on dense local features in the re-ranking stage leads to huge runtime overhead and memory usage.

To alleviate the above problems, this work advocates "focus on local" and proposes FoL, an efficient two-stage VPR method that makes better use of spatial local information. First, we design two losses to explicitly model regions of the VPR task that are discriminative and reliable. Specifically, we propose Extraction-Aggregation Spatial Alignment Loss (SAL) to achieve mutually reinforcing self-supervised learning by aligning the regions of interest of the feature extractor (*i.e.*, the self-attention map in the transformer) with the regions that are not discarded by the feature aggregator. Besides, we introduce the Foreground-Background Contrast Enhancement Loss (CEL) to encourage the network to generate regions that are both reliable, meaning they yield consistent attention across images, and discriminative, ensuring they are well differentiated from the background. As shown in Figure 1 (a), under the influence of discriminative region modeling, our FoL generates better global features, thus obtaining higher retrieval accuracy. Second, we introduce a weakly supervised local feature training strategy to improve the performance of local matching (Xu et al. 2024; Wu et al. 2023) in the re-ranking stage. This strategy is based on pseudo-correspondences derived from aggregated global features, which helps to solve the problem of scarcity of ground truth for local correspondences (Wang et al. 2023, 2022a). Third, we suggest an efficient re-ranking pipeline that is efficiently and precisely based on discriminative region guidance, as shown in Figure 1 (b).

To summarize, our work brings the following contributions: **1)** We design two well-designed losses for explicitly modeling reliable discriminative regions and enabling both image retrieval and local matching stages to benefit from them. **2)** We introduce a weakly supervised local feature training strategy based on pseudo-correspondence to improve re-ranking performance. **3)** We suggest using the discriminative region for guiding the local features matching in the re-ranking stage, thereby furthering the accuracy and efficiency. **4)** Extensive experiments showing our FoL has a state-of-the-art performance on a wide range of VPR benchmarks with a competitive inference speed.

Related Works

One-Stage VPR: The one-stage VPR approaches produce global representations for image matching by aggregating early handcrafted local features (Bay, Tuytelaars, and Van Gool 2006) or deep learning-based features from convolutional neural networks (Arandjelovic et al. 2016), MLPs (Ali-bey, Chaib-draa, and Giguère 2023), and transformers (Keetha et al. 2023). With the recent rise of visual foundation models (VFMs), several approaches have adopted the foundation model DINOv2 (Oquab et al. 2023) as the backbone of the VPR task and have shown promising performance. One representative work is CricaVPR (Lu et al. 2024b), which integrates adapters into DINOv2, employing a self-attention mechanism to associate multiple images within a batch, leveraging variations across images as cues to guide representation learning. Meanwhile, SALAD (Izquierdo and Civera 2024) redefines NetVLAD (Arandjelovic et al. 2016)'s soft-assignment as the optimal transmission problem aggregating local features from DINOv2. However, one-stage methods are prone to perceptual aliasing (Lu et al. 2024d) due to the fact that they only utilize global information for the purpose and ignore spatial information that can be used for position calibration. In contrast, our proposed FoL models and exploits image spatial discriminative regions so that the network pays more attention to valuable regions to alleviate the perceptual aliasing of global retrieval and imposes spatial information validation through local matching between discriminative regions to further refine the global retrieval results efficiently.

Two-Stage VPR: A recent trend in VPR is the use of a two-stage retrieval strategy (Hausler et al. 2021; Lu et al. 2024a), which first performs an initial ranking based on the global representation similarity, and then re-ranks the top K retrieved candidates in the second stage using local features containing spatial information. The two-stage VPR methods achieve a more competitive performance due to the use of spatial information. Patch-NetVLAD (Hausler et al. 2021) is a two-stage method that first retrieves candidate images using global NetVLAD features and subsequently reorders these candidates by leveraging multi-scale patch-level features derived from the NetVLAD framework. TransVPR (Wang et al. 2022b) leverages the transformer architecture's multi-level attention mechanism to derive global features for candidate retrieval, and uses an attention mask to filter the feature map to generate key patch descriptors for reordering candidates. R2Former(Zhu et al. 2023) proposes a unified framework for place recognition that integrates retrieval and re-ranking, where the re-ranking module considers feature correlation, attention values, and coordi-

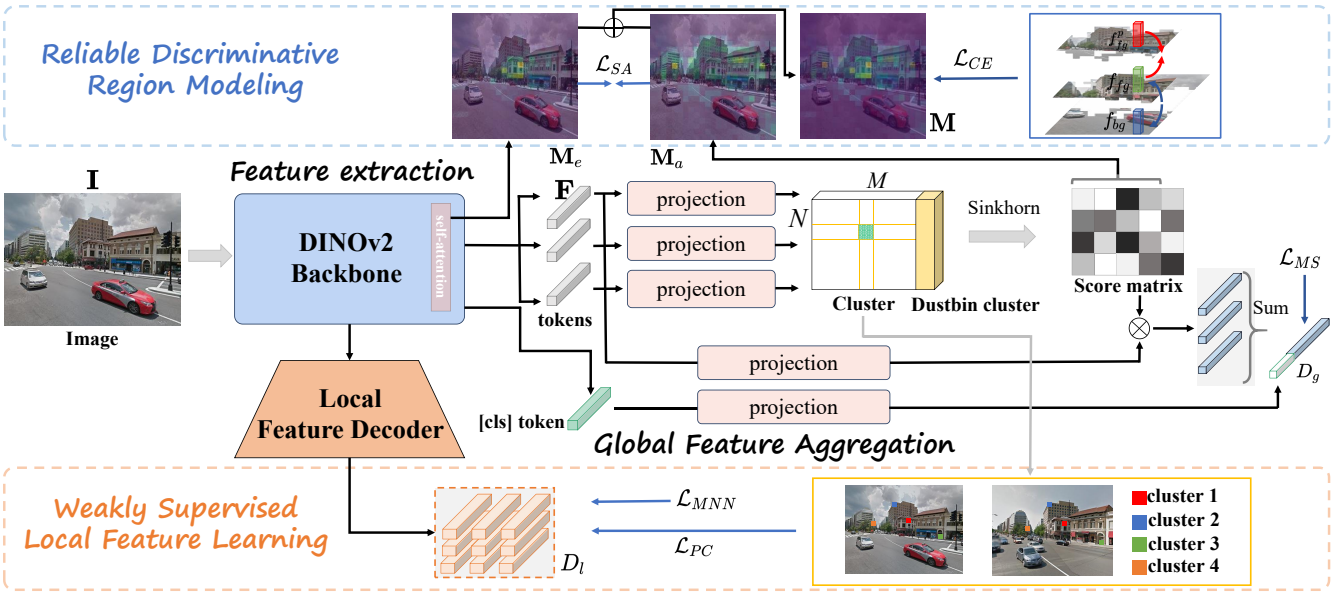


Figure 2: Illustration of our FoL's pipeline during training. In addition to including common feature extraction and global feature aggregation steps, our FoL also includes Reliable discriminative region modeling and weakly supervised local feature learning to fully introduce spatially localized information to simultaneously improve the performance of the two-stage VPR method for image retrieval and re-ranking.

nates. Recently, SelaVPR(Lu et al. 2024c) integrated trainable adapters into the DINOv2 architecture then used GeM pooling techniques for feature aggregation and performed re-ranking by introducing an adapter that produces local features. Different from existing two-stage methods, our proposed FoL explicitly models reliable discriminative regions by leveraging extraction-aggregation spatial alignment and foreground-background contrast enhancement losses. These mined regions can be used to facilitate efficient and accurate local feature matching in the re-ranking stage. Additionally, our FoL also introduce pseudo-correspondence-based weakly-supervised local feature learning techniques to enhance the accuracy of the re-ranking stage.

Methodology

Network Architecture

As shown in Figure 2, the image is first fed into the backbone network to extract the features, and then goes through two branches to derive the global feature and local features respectively. Specifically, global feature for image retrieval are obtained after feature aggregation, while local features are obtained by a lightweight resolution recovery decoder.

Feature Extraction. Following recent works (Keetha et al. 2023), we use the transformer-based vision foundation model DINOv2 (Oquab et al. 2023) as a feature extractor, for input image $\mathbf{I} \in \mathbb{R}^{h \times w \times 3}$ after feature extraction to obtain feature $\mathbf{F}$ containing $N = h \times w/14^2$ tokens $\mathbf{t}_i, i \in \mathbb{R}^N$.

Global Feature Aggregation. We adopt the optimal assignment based on Sinkhorn algorithm (Cuturi 2013) proposed by SALAD (Izquierdo and Civera 2024) to aggregate the features $\mathbf{F}$ to construct the global feature $\mathbf{D}_g$. Specifically, tokens from $\mathbf{F}$ are assigned to M clusters according to the score matrix computed by the Sinkhorn algorithm, and an additional dustbin cluster can is set up to discard useless background region tokens. In particular, we refer to the tokens regions assigned to the M clusters outside the dustbin cluster can as the feature aggregation attention mask $\mathbf{M}_a$.

Local Feature Decoder. The decoder is lightweight consisting of only two Deconvolution layers and a ReLU layer.

Reliable Discriminative Region Modeling

In FoL, we design two losses, Extraction-Aggregation Spatial Alignment Loss (SAL) and Foreground-Background Contrast Enhancement Loss (CEL), to model reliable discriminative regions. We seek to construct discriminative regions by utilizing the spatial information capturing capabilities readily available in feature extraction and global feature aggregation of VPR pipeline.

For feature extraction, the self-attention score computation of the transformer architecture incorporates the level of attention paid to each token representing an image patch, which can be viewed as modeling spatial information. Specifically, $\mathbf{M}_e$ is obtained by averaging the attention scores of the last multi-head attention layer [cls] token over patch tokens. For feature aggregation, the dustbin cluster is dedicated to hold background features, while other foreground features can be obtained by assigning to the set of M clusters, thus we can obtain discriminative foreground region mask $\mathbf{M}_a$. Finally, we obtain the discriminative region mask $\mathbf{M}$ by fusing $\mathbf{M}_e$ and $\mathbf{M}_a$, *i.e.*, $\mathbf{M} = (\mathbf{M}_e + \mathbf{M}_a)/2$.

Extraction-Aggregation Spatial Alignment Loss. We propose Extraction-Aggregation Spatial Alignment Loss (SAL)

to achieve mutually reinforcing self-supervised learning by aligning the regions $\mathbf{M}_e$ of interest of the feature extractor with the regions $\mathbf{M}_a$ that are not discarded by the feature aggregator. Specifically, we align $\mathbf{M}_a$ and $\mathbf{M}_e$ by Kullback-Leibler divergence (Van Erven and Harremos 2014):

$$\mathcal{L}_{SA} = \sum_{i}^{h\times w} \mathbf{M}_a^i log\frac{\mathbf{M}_a^i}{\mathbf{M}_a^i} + \sum_{i}^{h\times w} \mathbf{M}_e^i log\frac{\mathbf{M}_a^i}{\mathbf{M}_e^i}, \quad (1)$$

where $i \in \mathbb{R}^{h\times w}$. In practice, we found that direct alignment leads to network learning difficulties due to the significant difference in distribution between $\mathbf{M}_a$ and $\mathbf{M}_e$. Therefore, we smoothed $\mathbf{M}_a$ and $\mathbf{M}_e$. Taking $\mathbf{M}_a$ as an example, it can be defined as:

$$\mathbf{M}_a^{(i)} = \begin{cases} \mathbf{M}_a^{(i)} & \text{if } \mathbf{M}_a^{(i)} < \text{perc}_{10}(\mathbf{M}_a^{(i)}) \\ \text{perc}_{10}(\mathbf{M}_a^{(i)}) & \text{otherwise,} \end{cases} \quad (2)$$

where $\text{perc}_{10}(\mathbf{M}_a^{(i)})$ denotes the 10-th percentile value of the i-th row in matrix $\mathbf{M}_a$. This formula is used to truncate the top 10% of values in each row to the 10-th percentile value, thereby smoothing the distribution of $\mathbf{M}_a$ and $\mathbf{M}_e$.

Foreground-Background Contrast Enhancement Loss. We further introduce a Foreground-Background Contrast Enhancement Loss (CEL) to encourage the network to generate regions that are both reliable (*i.e.*, produce consistent attention across images) and discriminative (ensure that they are well distinguished from the background). We first reshape $\mathbf{F}$ containing the patch tokens into $\mathcal{F} \in \mathbb{R}^{h\times w\times d}$ that is consistent with the shape of M. From this, we can obtain the foreground comparison prototype f_{fg}:

$$f_{fg} = \frac{\sum_i^{h\times w} \mathbf{M}(i)\cdot\mathcal{F}(i)}{||\sum_i^{H\times W} \mathbf{M}(i)\cdot\mathcal{F}(i)||}, \quad (3)$$

where $||\cdot||$ means L2 norm. Similarly, we can get the background prototype f_{bg}:

$$f_{bg} = \frac{\sum_i^{h\times w}(1-\mathbf{M}(i))\cdot\mathcal{F}(i)}{||\sum_i^{H\times W}(1-\mathbf{M}(i))\cdot\mathcal{F}(i)||}. \quad (4)$$

Next, Based on triplet loss (Hermans, Beyer, and Leibe 2017) $\mathcal{L}_{CE}$ can be defined as:

$$\mathcal{L}_{CE} = \max(0, 1-(f_{fg}\cdot f_{fg}^p - f_{fg}\cdot f_{bg})), \quad (5)$$

where f_{fg}^p is the foreground prototype of the positive sample (different images of the same scene), as shown in Figure 2. Notice that in order to force the network to optimize only $\mathbf{M}$ without affecting feature extraction, we truncate the gradient of $\mathcal{F}$ when computing the $\mathcal{L}_{CE}$ loss. Fueled by $\mathcal{L}_{CE}$, $\mathbf{M}$ tends to produce consistent and reliable foreground discriminative region activation, and foreground and background regions are encouraged to be clearly distinguished. With these two losses proposed above, FoL can model reliable discriminative regions and benefit both image retrieval and re-ranking stages. On the one hand, the network can be induced to pay more attention to critical regions when extracting features and aggregating global features by constraining the attention of the transformer and the score matrix of the aggregation phase. On the other hand, discriminative regions can be used as a priori information in the re-ranking stage to accelerate the efficiency and accuracy of local matching.

Algorithm 1: Pseudo-correspondence Ground Truth Construction

1: **Input:** Discriminative Region Map M, Assignment Score Matrix S
2: **Parameters:** $thr1 = 0.8$, $thr2 = 0.5$, $N = 8$
3: We sort the patches using the activation values in M. Then, we select the patch feature f_p with the highest activation value and exclude it from the sequence.
4: The corresponding cluster order number is found through the index assignment matrix S. Then all the patches in the positive sample image that are assigned to that cluster are found based on the cluster order number as candidate features.
5: If candidate features $\{f_{p'_{1th}}, f_{p'_{2th}}\}$ corresponding to f_p satisfy $sim(f_p, f_{p'_{1th}}) > thr1$ and $sim(f_p, f_{p'^{2th}})/sim(f_p, f_{p'_{1th}}) < thr2$ with the highest and second-highest similarity to the corresponding feature f_p, perform step 6 else perform step 3.
6: p and $p^{'}_{1th}$ as pseudo labels $(p, p^{'})$ and then perform step 3 until N labels are obtained or there are no candidates.
7: **Return:** N pseudo-correspondence ground truth.

Weakly Supervised Local Feature Learning

Due to the lack of pixel-level correspondence labels in the VPR dataset, existing two-stage methods can usually only supervise the final matching results, such as R2Former (Zhu et al. 2023)'s classification loss and CricaVPR (Lu et al. 2024b)'s mutual nearest neighbor loss. In FoL, we explore the use of the clustering results of local features in the global feature aggregation process to construct pseudo-correspondence truths thereby enabling pixel-level supervision of local feature matching.

Pseudo-correspondence Ground Truth Construction. As shown in Figure 2, in the global feature aggregation phase, each local patch is assigned to a specific cluster, which in fact indirectly includes the correspondence between patches, *i.e.*, the corresponding pair of patches will be assigned to the same cluster. We develop the pseudo correspondence truth construction algorithm based on the above observation as shown in Algorithm 1.

Weakly Supervised Pseudo correspondence Loss. We obtain $(D_l^p, D_l^{p^{'}})$ by sampling the network output dense local features D_l using the pseudo correspondence $(p, p^{'})$. From this, $\mathcal{L}_{PC}$ can be defined as:

$$\mathcal{L}_{PC} = \frac{\sum exp(sim(f_{p_i}, f_{p'_i}))\cdot(1-sim(D_l^{p_i}, D_l^{p_i^{'}}))}{\sum exp(sim(f_{p_i}, f_{p'_i}))}. \quad (6)$$

Since not true ground truth is used, we use similarity as a confidence for evaluating the quality of labels to mitigate the negative impact of inaccurate labels on model training, hence $\mathcal{L}_{PC}$ is a weakly supervised loss.

Efficient Re-ranking with Discriminative Region Guidance

We propose to accelerate and enhance the re-ranking process using the local regions mask $\mathbf{M}$ modeled in previous section as a priori information for local matching.

First, we convert $\mathbf{M}$ into a binary matrix, where the top k positions in $\mathbf{M}$ are set to 1, and others to 0. This can be expressed as:

$$\mathbf{M}_{\text{bin}} = \begin{cases} 1 & \text{if index} \in \text{top } k \text{ of } \mathbf{M} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where k is set to the top 40%.

Then, we interpolate $\mathbf{M}_{\text{bin}}$ and select the parts of the dense local features D_l where the mask is 1:

$$D_l^{\mathbf{M}} = D_l[\mathbf{M}_{\text{bin}} = 1] \tag{8}$$

Finally, we only perform nearest-neighbor local feature matching on $D_l^{\mathbf{M}}$, which reduces the probability of wrong matching and improves the efficiency of matching due to the narrowed matching range.

Total Loss Function

Following previous works (Izquierdo and Civera 2024), we also use the multi-similarity loss (Wang et al. 2019a) $\mathcal{L}_{MS}$ with an online hard mining strategy to optimize the global features. To optimize the local features matching, we also adopt the mutual nearest neighbor local feature loss $\mathcal{L}_{MNN}$ proposed in SelaVPR(Lu et al. 2024c). Combined with the techniques presented in the previous sections, the total loss can be defined as:

$$\mathcal{L} = \mathcal{L}_{MS} + \mathcal{L}_{MNN} + \mathcal{L}_{CE} + \alpha\mathcal{L}_{SA} + \beta\mathcal{L}_{PC}, \tag{9}$$

where $\{\alpha, \beta\}$ are hyperparameters to adjust the magnitude of the different losses.

Experiments

Benchmarks and Performance Evaluation

Our experiments utilize several VPR benchmark datasets to assess the performance of our models, focusing primarily on Tokyo24/7, Pitts250k, and MSLS, with additional evaluations on Nordland, AmsterTime, SVOX, SPED, and SF-XL. **Tokyo24/7** (Torii et al. 2015) includes approximately 76,000 database images and 315 query images from urban environments, each image involving three different viewpoints and three different times of the day, showcasing significant changes in lighting conditions. **Pitts250k** (Torii et al. 2013) consists of images from Google Street View panoramas, exhibiting significant viewpoint changes, moderate condition variations, and a small number of dynamic objects. **MSLS** (Mapillary Street-Level Sequences) (Warburg et al. 2020) is a large-scale dataset with over 1.6 million images collected from urban, suburban, and natural scenes, testing models under varying conditions such as illumination, weather, and dynamic objects. **Nordland** (Sünderhauf, Neubert, and Protzel 2013) captures images from a fixed viewpoint in the front of a train on the same route over four seasons, providing images from suburban and natural environments with significant seasonal and lighting changes. **AmsterTime** (Yildiz et al. 2022) presents historical grayscale queries and contemporary RGB references from Amsterdam, challenging models with temporal and modality variations. **SVOX** (Berton et al. 2021) evaluates performance under different weather and lighting conditions, with queries extracted from the Oxford RobotCar dataset. The **SPED** (Zaffar et al. 2021) and **SF-XL** (Berton, Masone, and Caputo 2022b) datasets add further diversity with varying degrees of viewpoint and condition changes. We follow the common evaluation metrics used in previous works (Ali-bey, Chaib-draa, and Giguère 2024). We assess performance using Recall@N, with a 25-meter threshold for Tokyo24/7, Pitts30k, and MSLS, and ± 10 frames for Nordland, effectively measuring retrieval accuracy under various conditions.

Implementation Details

We initialize the ViT-L backbone with pre-trained DINOv2 weights and fine-tune only the last four layers of the backbone. The remaining modules are set to learnable. In the feature extraction stage, the number of clusters M is set to 64. In the re-ranking stage, the number of channels up-conv are 256 and 128, and the convolution kernel size was 3×3, stride=2, padding=1. The ViT architecture supports variable input sizes, provided images can be partitioned into 14 × 14 patches. To speed up training, we used 322 × 322 images but evaluated on 504 × 504 resolution. We trained with one A100 GPU on GSV-Cities (Ali-bey et al. 2022), a large city location dataset collected by Google Street View. Batch size is 60 and each batch is described by 4 images. The AdamW optimizer with a linear learning rate schedule was used, with a learning rate of $6e-5$ and a weight decay of $9.5e-9$. The training converged after 5 epochs. To ensure the validity of our experiments and optimize hyperparameter selection, we continuously monitored recall performance on the MSLS validation set. We follow mainstream works to use 25 meters as the threshold for correct scene and report recall@k (k=1,5,10) as evaluation metrics.

Comparisons with state-of-the-art Methods

In this section, we compare our one-stage (FoL-global) and two-stage (FoL-re-ranking) results with previous state-of-the-art VPR methods, including eight one-stage methods that utilize global feature retrieval: NetVLAD (Arandjelovic et al. 2016), SFRS (Ge et al. 2020), CosPlace (Berton, Masone, and Caputo 2022b), MixVPR (Ali-bey, Chaib-draa, and Giguère 2023), EigenPlaces (Berton et al. 2023), CricaVPR (Lu et al. 2024b), SALAD (Izquierdo and Civera 2024), and BoQ (Ali-bey, Chaib-draa, and Giguère 2024). Additionally, we compare our method with three two-stage methods that incorporate re-ranking: Patch-NetVLAD (Hausler et al. 2021), R2Former (Zhu et al. 2023), and SelaVPR (Lu et al. 2024d). Our method, similar to CricaVPR and SALAD, is trained on the GSV-Cities (Ali-bey et al. 2022). Quantitative results are presented in Table 1. Our FoL-global method achieves the best

Method	Pitts250k-test			MSLS-val			MSLS-challenge			Tokyo24/7		
	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 CVPR' 2016 (Arandjelovic et al. 2016)	90.5	96.2	97.4	53.1	66.5	71.1	35.1	47.4	51.7	60.6	68.9	74.6
SFRS ECCV' 2020 (Ge et al. 2020)	90.7	96.4	97.6	69.2	80.3	83.1	41.6	52.0	56.3	81.0	88.3	92.4
Patch-NetVLAD CVPR' 2021 (Hausler et al. 2021) †	-	-	-	79.5	86.2	87.7	48.1	57.6	60.5	86.0	88.6	90.5
CosPlace CVPR' 2022 (Berton, Masone, and Caputo 2022b)	92.4	97.2	98.1	82.8	89.7	92.0	61.4	72.0	76.6	81.9	90.2	92.7
MixVPR WACV' 2023 (Ali-bey, Chaib-draa, and Giguère 2023)	94.6	98.3	99.0	88.2	93.1	94.3	64.0	75.9	80.6	86.7	92.1	94.0
R2Former CVPR' 2023 (Zhu et al. 2023) †	93.2	97.5	98.3	89.7	95.0	96.2	73.0	85.9	88.8	88.6	91.4	91.7
EigenPlaces ICCV' 2023 (Berton et al. 2023)	94.1	98.0	98.7	89.1	93.8	95.0	67.4	77.1	81.7	93.0	96.2	97.5
SelaVPR ICLR' 2024 (Lu et al. 2024d) †	95.7	98.8	99.2	90.8	96.4	97.2	73.5	87.5	90.6	94.0	96.8	97.5
CricaVPR CVPR' 2024 (Lu et al. 2024b)	95.6	98.9	**99.5**	90.0	95.4	96.4	69.0	82.1	85.7	93.0	97.5	98.1
SALAD CVPR' 2024 (Izquierdo and Civera 2024)	95.1	98.5	99.1	92.2	96.2	97.0	75.0	88.8	91.3	94.6	97.5	97.8
BoQ CVPR' 2024 (Ali-bey, Chaib-draa, and Giguère 2024)	95.0	98.4	99.1	91.4	94.5	96.1	-	-	-	-	-	-
FoL-global	96.5	99.1	**99.5**	93.1	**96.9**	97.4	78.7	90.8	**93.0**	96.2	98.7	98.7
FoL-re-ranking†	**97.0**	**99.2**	**99.5**	**93.5**	**96.9**	**97.6**	**80.0**	**90.9**	**93.0**	**98.4**	**99.1**	**99.4**

Table 1: Comparison to state-of-the-art methods on four benchmarks. The best results are highlighted in **bold** and the second best are underlined. Two-stage methods are marked with †.

Method	Nordland	Amster Time	SF-XL Occlusion	SF-XL Night	SVOX Night	SVOX Sun	SVOX Snow	SVOX Overcast	SVOX SPED	
SFRS ECCV' 2020 (Ge et al. 2020)	16.0	29.7	-	-	28.6	54.8	-	-	-	-
CosPlace CVPR' 2022 (Berton, Masone, and Caputo 2022b)	43.8	38.7	-	-	44.8	67.3	-	-	-	-
MixVPR WACV' 2023 (Ali-bey, Chaib-draa, and Giguère 2023)	58.4	40.2	-	-	64.4	84.8	-	-	-	-
EigenPlaces ICCV' 2023 (Berton et al. 2023)	54.4	48.9	32.9	23.6	58.9	86.4	93.1	93.1	98.0	70.2
SelaVPR ICLR' 2024 (Lu et al. 2024d)	85.2	55.2	35.5	38.4	89.4	90.2	97.0	97.0	97.2	88.6
CricaVPR CVPR' 2024 (Lu et al. 2024b)	90.7	64.7	42.1	35.4	85.1	93.8	96.0	96.7	97.8	91.3
SALAD CVPR' 2024 (Izquierdo and Civera 2024)	76.0	58.8	51.3	46.6	95.4	97.2	98.9	**98.3**	98.2	**92.1**
BoQ CVPR' 2024 (Ali-bey, Chaib-draa, and Giguère 2024)	74.4	53.0	-	-	85.2	96.5	98.4	**98.3**	-	86.2
FoL-global	87.8	64.6	51.3	53.4	98.3	98.1	99.1	97.9	98.4	**92.1**
FoL-re-ranking	**92.6**	**70.1**	**61.8**	**60.5**	**98.8**	**98.8**	**99.3**	**98.3**	**98.9**	91.8

Table 2: Comparison (R@1) to state-of-the-art methods on more challenging datasets.

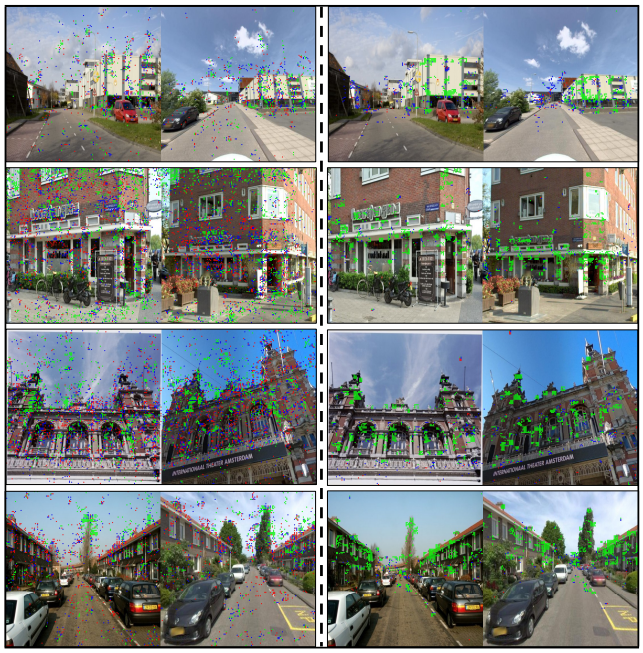
R@1/R@5/R@10 across all datasets using only global features for direct retrieval, significantly outperforming other one-stage and two-stage methods. Specifically, FoL-global improves the absolute R@1 by 0.8%, 0.9%, 3.7%, and 1.6% on Pitts250k-test, MSLS-val, MSLS-challenge, and Tokyo24/7, respectively. This highlights the effectiveness of the single global representation learned through our approach. After re-ranking with local features, the complete FoL method outperforms all other methods by a substantial margin. Specifically, FoL-re-ranking improves the absolute R@1 by 1.3%, 1.3%, 5.0%, and 3.8% on Pitts250k-test, MSLS-val, MSLS-challenge, and Tokyo24/7, respectively. These results demonstrate that our model can overcome various visual changes induced by illumination (Tokyo24/7), viewpoint (Pitts250k), weather, season, and dynamic objects (MSLS).

Subsequently, we demonstrate the effectiveness of our method in the most challenging VPR benchmark scenarios through experiments on ten demanding datasets. The detailed results in Table 2 emphasize the significant advantages of our method over previous methods on these datasets. FoL-re-ranking shows improvements of +1.9%, +5.4%, +10.5%, +13.9%, +3.4%, and +1.6% on Nordland, Amster Time, SF-XL Occlusion, SF-XL Night, SVOX Night, and SVOX Sun, respectively. Notably, we observe an improvement of over 10% on the large-scale SF-XL dataset, demonstrating that our model can generalize to highly challenging scenarios involving viewpoint changes and occlusions.

Method	Extraction Time (s)	Matching Time (s)	Total Time (s)
TransVPR	0.007	2.632	2.639
R2Former	0.009	0.347	0.356
SelaVPR	**0.023**	0.071	0.094
FoL-re-ranking	0.025	**0.032**	**0.057**

Table 3: The single query runtime comparison of two-stage methods on Pitts250k-test.

Efficiency is another critical metric for evaluating VPR methods. In Table 3, we compare the runtime of our method with other two-stage methods on Pitts250k-test, including both feature extraction time and matching/retrieval time. Due to the absence of geometric verification in our re-ranking stage and the involvement of only discriminative regions in matching, our method demonstrates a significant speed advantage. Figure 3 and Figure 4 show the visualization results of re-ranking local matching and VPR retrieval.



(a) (b)

Figure 3: Visualization of local feature matching in the re-ranking stage. (a) shows the *w/o* Discriminative Region Guidance, while (b) displays *w/* Discriminative Region Guidance. Red •, Blue •, and Green • represent low, medium, and high similarity matching points, respectively (• → • → • indicates increasing similarity).

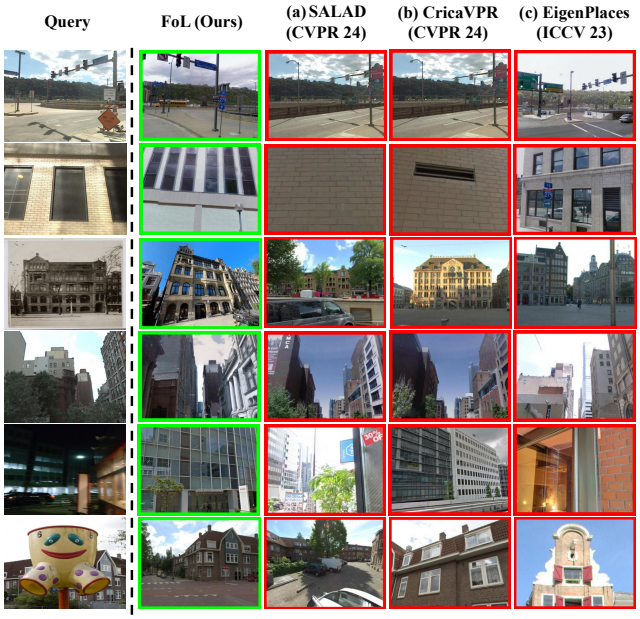


Figure 4: Qualitative VPR comparison results. Our FoL accurately matching the query images, while other methods such as SALAD, CricaVPR, and EigenPlaces cause erroneous matches under complex lighting and viewpoint variations.

Ablation Study

Table 4 reports the results of our proposed FoL ablation experiments on the MSLS Challenge benchmark. Here, the Baseline denotes using only the SALAD model trained with $\mathcal{L}_{MS}$ and $\mathcal{L}_{MNN}$ losses, without any of the proposed techniques. We first apply a simple re-ranking strategy to the Baseline, which aims to refine the initial retrieval results by re-examining candidate lists. This step slightly improves performance from (74.8, 88.4, 91.1) to (75.7, 89.1, 91.8). Next, we introduce $\mathcal{L}_{SA}$ to explicitly model discriminative regions for the VPR task. After applying $\mathcal{L}_{SA}$, we observe increments of ($+1.5\%$, $+0.3\%$, $+0.9\%$) on R@1, R@5, and R@10, respectively. Subsequently, adding $\mathcal{L}_{CE}$ further leverages spatial information, yielding increments of ($+0.9\%$, $+1.2\%$, $+0.1\%$). Then, we impose $\mathcal{L}_{PC}$ loss to enhance local feature representation in the re-ranking phase. As illustrated in Table 4, this leads to another improvement of ($+0.6\%$, $+0.2\%$, $+0.2\%$), highlighting the crucial role of improving local feature quality for re-ranking. Finally, we adopt our proposed Efficient Re-ranking with Discriminative Region Guidance *i.e.*, DRG. The performance reaches (80.0, 90.9, 93.0), and notably, the proposed DRG significantly reduces the matching time compared to dense local feature matching, resulting in a substantial inference speedup.

Configurations	MSLS Challenge		
	R@1	R@5	R@10
Baseline	74.8	88.4	91.1
+re-ranking	75.7	89.1	91.8
$+\mathcal{L}_{SA}$	77.2	89.4	92.7
$+\mathcal{L}_{CE}$	78.1	90.6	<u>92.8</u>
$+\mathcal{L}_{PC}$	<u>78.7</u>	<u>90.8</u>	**93.0**
+ DRG	**80.0**	**90.9**	**93.0**

Table 4: Ablation Study on MSLS Challenge benchmark. Here DRG means Efficient Re-ranking with Discriminative Region Guidance.

Conclusions

In this paper, we propose a two-stage VPR method called FoL and focus our research vision on the role of mining spatial local information for VPR. We explicitly model reliable and discriminative local regions with two well-designed losses and demonstrate the importance of discriminative region modeling for image retrieval and local matching in VPR tasks. We also introduce a weakly supervised local feature training strategy based on pseudo-correspondence to improve re-ranking performance. In addition, we propose discriminative region-guided efficient re-ranking strategy to further improve the accuracy and efficiency. Experimental results show that our FoL achieves the latest state-of-the-art performance on mainstream benchmarks for VPR tasks with high efficiency.

Acknowledgements

This work was supported by National Science and Technology Major Project (No.2022ZD0116800), and in part by the Taishan Scholars Program (Nos. TSQN202211214 and TSQN202408245), Shandong Excellent Young Scientists Fund Program (Overseas) No.2023HWYQ-113, Beijing Natural Science Foundation (No.JQ23014), National Natural Science Foundation of China (Nos.62271074, 32271983, 62171321, 62162044), Shandong Provincial Natural Science Foundation for Young Scholars (Nos.ZR2024QF110 and ZR2024QF052) and the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2023B01).

References

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. 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.

Ali-bey, A.; et al. 2022. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513: 194–203.

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.; Tuytelaars, T.; and Van Gool, L. 2006. Surf: Speeded up robust features. *Lecture notes in computer science*, 3951: 404–417.

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

Berton, G.; Masone, C.; and Caputo, B. 2022b. Rethinking visual geo-localization for large-scale applications. In *CVPR*, 4878–4888.

Berton, G.; Paolicelli, V.; Masone, C.; and Caputo, B. 2021. Adaptive-Attentive Geolocalization From Few Queries: A Hybrid Approach. In *IEEE Winter Conference on Applications of Computer Vision (WACV)n*, 2918–2927.

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.; Maffra, F.; Sa, I.; and Chli, M. 2017. Only look once, mining distinctive landmarks from convnet for visual place recognition. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9–16. IEEE.

Cuturi, M. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2): 1–60.

Ge, Y.; Wang, H.; Zhu, F.; Zhao, R.; and Li, H. 2020. Self-supervising fine-grained region similarities for large-scale image localization. In *ECCV*, 369–386. Springer.

Guo, J.; Wang, H.; Cheng, Z.; Zhang, X.; and Yan, D.-M. 2020. Learning local shape descriptors for computing non-rigid dense correspondence. *Computational Visual Media*, 6: 95–112.

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.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Izquierdo, S.; and Civera, J. 2024. Optimal Transport Aggregation for Visual Place Recognition. In *Accepted to CVPR*.

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*.

Li, J.; Xie, Q.; Chang, X.; Xu, J.; and Liu, Y. 2024. Mutually-Guided Hierarchical Multi-Modal Feature Learning for Referring Image Segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(12): 1–18.

Lu, F.; Dong, S.; Zhang, L.; Liu, B.; Lan, X.; Jiang, D.; and Yuan, C. 2024a. Deep Homography Estimation for Visual Place Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10341–10349.

Lu, F.; Lan, X.; Zhang, L.; Jiang, D.; Wang, Y.; and Yuan, C. 2024b. CricaVPR: Cross-image Correlation-aware Representation Learning for Visual Place Recognition. In *Accepted to CVPR*.

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

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

Middelberg, S.; Sattler, T.; Untzelmann, O.; and Kobbelt, L. 2014. Scalable 6-dof localization on mobile devices. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, 268–283. Springer.

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*.

Sünderhauf, N.; Neubert, P.; and Protzel, P. 2013. Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons. In *Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA)*.

Torii, A.; Arandjelovic, R.; Sivic, J.; Okutomi, M.; and Pajdla, T. 2015. 24/7 place recognition by view synthesis. In *CVPR*, 1808–1817.

Torii, A.; Sivic, J.; Pajdla, T.; and Okutomi, M. 2013. Visual place recognition with repetitive structures. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 883–890.

Van Erven, T.; and Harremos, P. 2014. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7): 3797–3820.

Wang, C.; Xu, R.; Lu, K.; Xu, S.; Meng, W.; Zhang, Y.; Fan, B.; and Zhang, X. 2023. Attention weighted local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10632–10649.

Wang, C.; Xu, R.; Xu, S.; Meng, W.; and Zhang, X. 2022a. CNDesc: Cross normalization for local descriptors learning. *IEEE Transactions on Multimedia*, 25: 3989–4001.

Wang, R.; Shen, Y.; Zuo, W.; Zhou, S.; and Zheng, N. 2022b. TransVPR: Transformer-based place recognition with multi-level attention aggregation. In *CVPR*, 13648–13657.

Wang, X.; Han, X.; Huang, W.; Dong, D.; and Scott, M. R. 2019a. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5022–5030.

Wang, Y.; Guo, J.; Yan, D.-M.; Wang, K.; and Zhang, X. 2019b. A robust local spectral descriptor for matching non-rigid shapes with incompatible shape structures. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 6231–6240.

Warburg, F.; Hauberg, S.; López-Antequera, M.; Gargallo, P.; Kuang, Y.; and Civera, J. 2020. Mapillary street-level sequences: A dataset for lifelong place recognition. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2626–2635.

Wu, J.; Xu, R.; Wood-Doughty, Z.; Wang, C.; Xu, S.; and Lam, E. Y. 2023. Segment anything model is a good teacher for local feature learning. *arXiv preprint arXiv:2309.16992*.

Xu, S.; Chen, S.; Xu, R.; Wang, C.; Lu, P.; and Guo, L. 2024. Local feature matching using deep learning: A survey. *Information Fusion*, 107: 102344.

Yildiz, B.; Khademi, S.; Siebes, R. M.; and van Gemert, J. 2022. AmsterTime: A Visual Place Recognition Benchmark Dataset for Severe Domain Shift. *arXiv preprint arXiv:2203.16291*.

Yuan, D.; Geng, G.; Shu, X.; Liu, Q.; Chang, X.; He, Z.; and Shi, G. 2024. Self-supervised discriminative model prediction for visual tracking. *Neural Computing and Applications*, 36(10): 5153–5164.

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 (IJCV)*, 1–39.

Zhu, S.; Yang, L.; Chen, C.; Shah, M.; Shen, X.; and Wang, H. 2023. R^2Former: Unified retrieval and reranking transformer for place recognition. In *CVPR*, 19370–19380.