The Overlooked Value of Test-time Reference Sets in Visual Place Recognition

Anonymous ICCV submission

Paper ID *****

Abstract

Given a query image, Visual Place Recognition (VPR) is the task of retrieving an image of the same place from a reference database with robustness to viewpoint and appearance changes. Recent works show that some VPR benchmarks are solved by methods using Vision-Foundation-Model backbones and trained on large-scale and diverse VPR-specific datasets. Several benchmarks remain challenging, particularly when the test environments differ significantly from the usual VPR training datasets. We propose a complementary, unexplored source of information to bridge the train-test domain gap, which can further improve the performance of State-of-the-Art (SOTA) VPR methods on such challenging benchmarks. Concretely, we identify that the test-time reference set, the "map", contains images and poses of the target domain, and must be available before the test-time query is received in several VPR applications. Therefore, we propose to perform simple Reference-Set-Finetuning (RSF) of VPR models on the map, boosting the SOTA (≈ 2.3% increase on-average for Recall@1) on these challenging datasets. Finetuned models retain generalization, and RSF works across diverse test datasets.

1. Introduction

Given a query image and a database of geo-tagged reference images, the task of a Visual Place Recognition (VPR) method is to retrieve from the database a correct matching reference image for this qfrom the database uery. What is considered as a correct match is ill-defined, but most VPR benchmarks consider any reference image within a fixed (e.g., 25-meter) circular radius of the query location as a correct match [9]. VPR has many applications, such as in landmark retrieval [43], 3D modeling [1], image search [37] and map-based localization [35, 49]. *These applications of VPR require that the test time reference set (the map) is available offline, i.e., before a test-time query is received.*[1]

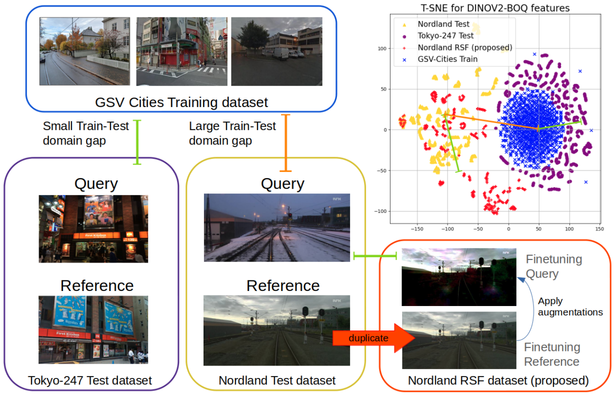

Figure 1. Large-scale VPR training datasets are created usually from Google Street View [2], e.g., the GSV-cities dataset. Thus models trained in these environments perform well (SOTA Recall@5 $\sim 98-99\%$) for similar test datasets, e.g., the Tokyo-247 dataset [5] but suffer in unseen environments, e.g., the railway-tracks of the Nordland dataset [34]. A train-test domain gap exists, as evident in the T-SNE projection of descriptors computed using BoQ-DinoV2 [4] for randomly sampled images of these datasets. Descriptors from the Tokyo-247 dataset form a single cluster with the GSV-cities dataset, while the Nordland dataset is further away. Creating a finetuning dataset by using the freely available test-time reference images could help bridge the train-test domain gap.

Traditionally, the most investigated challenges in VPR have been viewpoint and appearance changes between the matching query and reference images, the so-called query-ref domain gap [23]. Thus, the objective of VPR methods is to extract representations robust to these variations. Given this objective, VPR benefited significantly through neural networks trained on large-scale VPR-specific datasets [9]. More recently, this has been complemented by adapting strong general-purpose *Vision-Foundation-Model backbones* (VFM) to the task of VPR, e.g., the DinoV2 vision transformer [4, 17, 24, 29]. As a result, test datasets with large query-ref domain gap (e.g., Tokyo-247 [5] and SVOX-night/snow [10]) that were previously challenging for VPR methods now seem solved ($\sim 98-99\%$ Recall@5) by the State-of-the-Art (SOTA) [4, 17, 24].

[1] We acknowledge that there are other applications of VPR where the reference map may not be available offline, such as in SLAM. These applications are not the focus in this work.

However, another important but less investigated challenge in VPR is the train-test domain gap, i.e., when the test dataset is from a different environment and/or device than the training dataset. It could be hypothesized that SOTA VPR methods would be already robust to this gap, since VFM backbones are known to generalize across datasets and tasks [6], and more so, when finetuned on diverse VPR-specific training data [2]. We examine this hypothesis, revealing that the current SOTA VPR methods still suffer from the train-test domain gap. Details of this will follow later in the section 3.2.

To address the challenge posed by the large train-test domain gap in VPR, we propose a strategy complementary to the typical curation of larger training datasets and/or using stronger VFM backbones. A case is made for using the unexplored reference set in test datasets to finetune the SOTA in VPR. We argue that since this reference set with labeled (poses) images is freely available beforehand in various VPR applications and/or could even be obtained online, it is permissible to use it to bridge the train-test domain gap. Thus, outlining the two assumptions made in our work: a) the test-time reference set is available offline, b) there are resources available at test-time to finetune a VPR model.

Given this argument, we illustrate in Fig. 1 the train-test domain gap in VPR. A T-SNE [38] projection of two VPR test datasets, Tokyo-247 [5] and Nordland [34], is shown along with the diverse GSV-Cities training dataset. The Tokyo-247 dataset contains urban scenes similar to the GSV-Cities and hence both form a single cluster, while the Nordland dataset contains railway-tracks unlike GSV-Cities and forms a separate cluster. Our proposal is simply that the reference set in test datasets (e.g., Nordland) could be combined with image augmentations to create a new finetuning dataset that has a smaller train-test domain gap than the original GSV-Cities dataset. Domain knowledge can then be injected into the model using this proposed finetuning dataset, akin to domain adaptation in other computer vision tasks such as classification [19].

However, this raises several questions: a) Is finetuning of VFM-based VPR methods on small test datasets useful? b) Do the finetuned models still generalize to other test datasets? c) Can a single finetuning strategy work across diverse test datasets? We will present a simple self-supervised strategy, namely, Reference-Set-Finetuning (RSF), to answer these questions.

2. Related Work

Visual place recognition was first surveyed in the seminal work of Lowry *et al.* [23], which coincided well with the rise of deep learning for computer vision. The three most fundamental challenges identified by Lowry *et al.* in VPR are matching images given viewpoint changes, appearance changes due to illumination, seasons, dynamic objects, etc., and perceptual-aliasing [47]. For handling viewpoint and appearance changes, VPR requires robust image representations, and thus this formulation of VPR as a (deep) representation learning problem led to many works that achieved state-of-the-art VPR performance under challenging conditions [3, 5, 8, 20–22, 31, 32, 40, 46].

Deep-learning-based VPR methods can be broadly categorized based on their underlying novelty, such as the use of a novel loss function [22, 32, 36], better training data [2, 8], new architectures [40, 45, 48], different data augmentations [12, 18, 28], and new methods for feature aggregation [3, 5, 16, 31]. The work of Berton *et al.* [9] recently created a model zoo based on different combinations of the aforementioned key modules of a VPR system, which is freely accessible online. Since deep-learning mainly benefits from larger training datasets, a number of training datasets have been proposed and used in VPR, e.g., the Pitts-250k dataset [5], Mapillary Street Level Sequences dataset [42], San-francisco-XL [8] dataset, or the GSV-Cities dataset [2].

The use of vision-transformers in VPR was first studied in TransVPR [40], where image features are first extracted using a CNN and then a transformer encoder is used to aggregate these features into a global descriptor. This work was followed up by R2former [50], where a vision transformer is used for both retrieval and re-ranking, and operates directly on image patches.

VPR has benefited from advances in related fields that also require learning robust image representations. Thus, after the release of DinoV2 [29] Vision-Foundation Model (VFM), it was quickly adopted for VPR, where Anyloc [20] investigated using DinoV2 as an off-the-shelf feature extractor. Many concurrent works subsequently showed that the performance benefits are significantly larger when DinoV2 is finetuned on VPR-specific data and training objectives [4, 17, 24, 25]. CricaVPR [24] proposes to use correlation between images in the batch with feature aggregation at multiple scales to produce robust global features. SALAD [17] uses the Sinkhorn algorithm to aggregate the global and local DinoV2 tokens for VPR. Authors of SelaVPR [25] add serial and parallel adapters to the DinoV2 architecture. Finally, BoQ [4] proposes to learn queries from scratch that are useful for VPR using the attention mechanism of transformers, and demonstrates that these learnable queries work with both older (ResNet) and newer (DinoV2) feature extraction backbones.

These methods collectively show that VFMs (e.g., DinoV2) have directly benefited the VPR community and that stronger backbones, i.e., larger models trained on larger datasets, can directly improve VPR. However, we report that some VPR benchmarks, with a large train-test domain gap still remain unsolved. In this context, the contributions of our work are as follows:

- Our comparison of concurrent VFM-based SOTA VPR methods reveals that these methods suffer from a train-test domain gap. It is demonstrated that the freely available test-time reference set can be used to extract useful domain knowledge for VPR applications where the reference map is available offline.
- A simple Reference-Set-Finetuning (RSF) strategy is proposed to address the train-test domain gap for such VPR applications. The proposed finetuning improves the SOTA in VPR, and the RSF models retain generalization to other test datasets. RSF works across diverse datasets and is compatible with different VPR methods.

3. Methodology

We first formalize VPR, then formulate the use of deep learning in VPR, and finally describe the RSF strategy proposed in this work.

3.1. Formalizing VPR

The goal of VPR is to find one or multiple reference images $I_i \in \mathcal{I}_\mathcal{R}$ that match the place of a query image $I_q \in \mathcal{I}_\mathcal{Q}$ given a set of reference images $\mathcal{I}_\mathcal{R}$ with known poses $\mathcal{P}_\mathcal{R}$. The pose of I_q is then approximated by the pose of its nearest neighbour references in $\mathcal{I}_\mathcal{R}$. In its standard formulation, VPR consists of an offline map preparation stage and an online retrieval stage. The unknown pose p_q for the query I_q can then be approximated from the poses of the matched references $p_i \in \mathcal{P}_\mathcal{R}$ [30].

In the offline phase, a VPR method G is applied to every reference image $I_i \in \mathcal{I}_\mathcal{R}$ to obtain D-dimensional reference feature descriptors $f_i = G(I_i)$. The method G is usually a trained neural network [26] or a handcrafted feature descriptor [14]. The resulting VPR map $\mathcal{M} = (\mathcal{I}_\mathcal{R}, \mathcal{R}, \mathcal{P}_\mathcal{R})$ contains the reference feature descriptors set $\mathcal{R} = \{f_1, \cdots f_N\}$, where each descriptor f_i is associated with a corresponding pose $p_i \in \mathcal{P}_\mathcal{R}$.

In the online retrieval stage, the same method G is applied to the query image I_q, and its descriptor $f_q = G(I_q)$ is compared to the reference descriptors in the map $\mathcal{M}$. This can be achieved through an efficient K-nearest neighbor lookup, considering the L2-distances $d_i = ||f_i - f_q||_2$ between each reference i and the query q.

3.2. Relating the current SOTA in VPR to train-test domain gap

VPR in deep-learning is generally formulated either as a representation learning task [5] or a classification [8] task. We use the former formulation in this paper. A deep-learning-based VPR method G consists of four major choices: a feature extraction backbone B, a feature aggregator P, a training dataset D, and a metric-learning loss function $\mathcal{L}$. The backbone B and aggregator P are compositional and together form the method G, such that $f_i = G(I_i) = P(B(I_i))$. This VPR method G is then trained on the training dataset D by minimizing the loss $\mathcal{L}$. The training dataset D is itself composed of four sets, such that $D = (\mathcal{I}_\mathcal{Q}{}^{train}, \mathcal{P}_\mathcal{Q}{}^{train}, \mathcal{I}_\mathcal{R}{}^{train}, \mathcal{P}_\mathcal{R}{}^{train})$, where for every $I_q \in \mathcal{I}_\mathcal{Q}$, the true and false matching reference images I_i are defined based on the spatial proximity of their corresponding poses in $\mathcal{P}_\mathcal{Q}{}^{train}$ and $\mathcal{P}_\mathcal{R}{}^{train}$, respectively.

The choice of backbone in VPR is primarily motivated by advances in other vision tasks, and we have thus seen a change from using VGG [5] and ResNet-based backbones [3, 7] to domain-agnostic Vision-Foundation-Model (VFM) backbones [4, 17, 20, 24]. For a fixed backbone B, different types of aggregators could be used as P, for example, a NetVLAD layer [5], GeM layer [31], or the recently proposed Bag-of-learnable-Queries (BoQ) [4], etc. BoQ has been shown to outperform other aggregators trained on the same dataset with the same backbone [4].

Once the architecture $G = P(B(I_i))$ is fixed, the training loss $\mathcal{L}$ could be the distance-based loss [36], relative-pose-based loss [27], triplet loss [39], or the multi-similarity loss [41], etc. These losses could be minimized on different training datasets, for example, the Pitts-250k dataset [5], Mapillary Street Level Sequences dataset [42], San-francisco-XL [8] dataset, or the GSV-Cities dataset [2]. The purpose of these training datasets is to learn a generalizable feature extractor G that works well in different domains, and thus the training datasets must be as diverse as possible. From existing literature, GSV-cities dataset [2] is the most diverse training dataset in VPR.

Provided this formulation, would a VPR method G, employing a VFM backbone (e.g., DinoV2) trained on a large-scale diverse VPR dataset (e.g., GSV-Cities) with SOTA aggregation (e.g., BoQ), resolve the train-test domain gap? We examine this by benchmarking the performance (Recall@5) in Table 1 of three DinoV2-based SOTA VPR methods that were published almost simultaneously [4, 17, 24]. All methods are trained on the GSV-cities dataset [2]: the most diverse training dataset in VPR, containing viewpoint and appearance changes from many streets across the world. The reported performance suggests that the test datasets with small train-test domain gap are almost solved by these SOTA VPR methods, despite their large query-ref domain gap. But some other test datasets, such as Nordland [34] and AmsterTime [44] with archival reference images, where the test environments differ significantly from the training dataset, still present a challenge.[2]

[2] Please note that we do *not* refer to the presence/absence of train-test domain gap in the various VPR test datasets in binary terms, but in a proportional manner. That is, while there is still a train-test domain gap between the GSV-cities dataset and the solved test datasets, this gap is larger for the unsolved datasets.

	Backbone	SVOX-Snow	SVOX-Night	Pitts-250k	Tokyo-247	Nord.	Eyn.	Ams-AR	Avg.
Query-Ref gap		✓✓✓	✓✓✓	✓✓	✓✓✓	✓✓✓	✓	✓✓✓	
Train-Test gap		✓	✓	✓	✓	✓✓✓	✓✓	✓✓✓	
MixVPR [3] ('23)	ResNet50	98.4	79.5	98.2	91.7	86.8	93.2	60.4	88.5
BoQ [4] ('24)	ResNet50	99.5	94.7	98.5	95.9	91.1	94.9	75.4	93.8
Crica [24] ('24)	DinoV2	99.0	95.0	99.0	97.1	96.2	94.9	83.9	95.6
SALAD [17] ('24)	DinoV2	99.7	99.3	99.1	96.8	93.5	95.0	79.7	95.4
BoQ [4] ('24)	DinoV2	99.7	99.4	99.1	97.8	95.9	95.5	83.5	**96.4**

Table 1. *Recall*@5 of some of the SOTA foundation-model-based VPR methods on various test datasets. All methods are trained on the most diverse VPR training dataset: the GSV-Cities dataset. The second row represents the domain gap of the respective test dataset from the GSV-Cities training dataset. ✓ indicates a small gap and ✓✓✓ indicates a large gap. On average, BoQ-DinoV2 is the SOTA in VPR, outlined in Bold, and thus our primary baseline. To indicate the margin of improvement left for BoQ, the datasets are ranked from left-to-right and colored. Datasets with small train-test gap are almost solved, but a large train-test domain gap presents a challenge even for the SOTA VPR methods.

3.3. Our proposed Reference-Set-Finetuning (RSF)

The preceding discussion suggests that although the training dataset D could be carefully curated to maximize diversity, it might still lack the domain knowledge needed for G to perform well on the test-time queries $\mathcal{I_Q}$. Here we make our key observation: $\mathcal{I_R}$ is already available at the map preparation stage as well as its corresponding set of poses $\mathcal{P_R}$. Therefore, we propose *Reference-set-finetuning (RSF)*, an unexplored but straightforward and effective procedure to adapt a trained model G to the target domain. Concretely, RSF (1) creates a **finetuning dataset** $D_{ft} = (\mathcal{I_Q}^{ft}, \mathcal{P_Q}^{ft}, \mathcal{I_R}^{ft}, \mathcal{P_R}^{ft})$, and (2) updates G on D_{ft} with pose-aware triplet mining, as illustrated in Fig. 2, and described in the following.

For D_{ft}, the finetuning query set $\mathcal{I_Q}^{ft}$ should represent a combination of viewpoint and appearance changes typically seen between the matching queries and references. Thus, a query $I_q^{ft} \in \mathcal{I_Q}^{ft}$ is formulated as $I_q^{ft} = A(I_i^{ft})$, where $A(.)$ represents an **augmentation operation**. Ideally, $A(.)$ approximates the viewpoint and appearance changes expected between the queries and references. An M number of different augmentations could be chosen as $A(.)$. In conclusion, the choices follow:

$$\mathcal{I_R}^{ft} = \mathcal{I_R}, \tag{1}$$

$$\mathcal{P_R}^{ft} = \mathcal{P_Q}^{ft} = \mathcal{P_R}, \tag{2}$$

$$\text{and} \qquad |\mathcal{I_Q}^{ft}| = M \times |\mathcal{I_R}^{ft}|. \tag{3}$$

The finetuning queries $\mathcal{I_Q}^{ft}$ and references $\mathcal{I_R}^{ft}$ are encoded as feature vectors with G, positives and hard negatives [5] are **mined given the poses** $\mathcal{P_Q}^{ft}$ and $\mathcal{P_R}^{ft}$, and the network G is **finetuned** using a standard triplet loss [15]: $L_{triplet} = max\{d(f_q^{ft}, f_p^{ft}) - d(f_q^{ft}, f_n^{ft}) + m, 0\}$, with a Euclidean distance function $d(f_1, f_2) = ||f_1 - f_2||_2$ and a margin m. A hard-negative for a given query is the wrong reference image further than some fixed physical distance threshold that is the closest in the feature space.

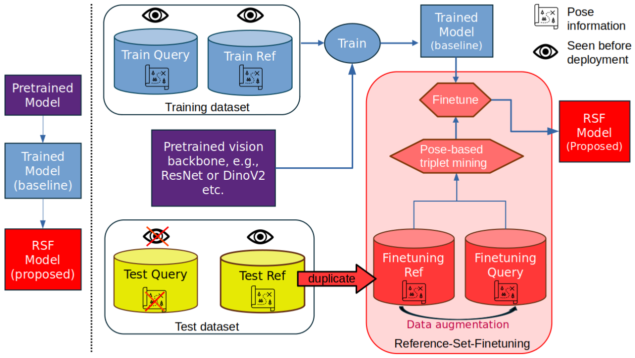

Figure 2. Deep learning for VPR usually utilizes a pretrained neural network that is further trained on a VPR dataset in a supervised manner with ground-truth poses. This usual pipeline assumes that we do not have any access to the test environment and that the training dataset is diverse enough to cover features of the test domain. However, there is always a train-test domain gap. We propose that the reference images in the test set are freely available offline in VPR and could be used to finetune VPR methods using simple data augmentations. This novel take on the problem setting of VPR, results in reference-set-finetuned (RSF) models that are more robust than the original trained model.

4. Experiments

First, we present the experimental setup of our work, then report the qualitative and quantitative performance of RSF models compared to baselines, and finally evaluate the various aspects of RSF.

4.1. Datasets and evaluation metric

To evaluate RSF, we use three public VPR datasets which have large train-test domain gap and hence pose challenges to SOTA VPR methods, and one dataset with a small train-test domain gap. Our ground-truth usage is similar to the standard formats in VPR [9], All of these datasets are summarized in Table 2.

	Queries	Refs.	Q-R gap	Train-test gap
Nord.	27.6k	27.6k	✓✓	✓✓✓
Amst-AR	1231	1231	✓✓✓	✓✓✓
Eyns.	24k	24k	✓	✓ ✓
SVOX-Ni	823	17.2k	✓✓	✓

Table 2. The datasets used in this work. We report the total number of query images, the total number of reference images, the presence of a domain gap between the queries and references, and the presence of a domain gap between the respective test dataset and the GSV-Cities training dataset. ✓ indicates a small gap and ✓✓✓ indicates a large gap.

The **Nordland dataset** [34] consists of a railway-track traversal through Norway during two different seasons: summer and winter. The summer traversal acts as reference images while the winter images are queries. This dataset is challenging due to the unstructured environment depicted in different seasons. We also use the challenging AmsterTime dataset [44] that contains archival imagery of Amsterdam and their corresponding Google Street View images. We use the archival images as references and street view images as queries, which depicts the task of retrieving an archival image of a place given a query image. We refer to this version as **AmsterTime-AR dataset**, outlining that the Archival images acts as References. We use the **Eynsham dataset** [13] that contains only grayscale images presenting a lack of color information for VPR. Finally, we use the **SVOX-Night dataset** [10] that contains night-time images as queries and day-time images as references collected through Google Street View in Oxford.

Following the existing literature, Recall@N is used as the evaluation metric. Ground-truths are as-is used by others [4, 9, 17, 24]. A retrieval is successful if the Top-N retrieved reference images were within a 25-meter radius of the query image.

4.2. Implementation details

Given the standards and SOTA described earlier in section 3.2, Dino-V2 [29] backbone with BoQ [4] aggregation trained on the GSV-cities dataset is used as the primary baseline VPR method G, since it is the current SOTA in VPR. Nevertheless, we also report performance of SALAD [17] when used with the proposed RSF. We use the complete reference set of each respective test dataset for performing RSF as described in section 3.3. A small learning rate of 1e-7 is used for all datasets for both the VPR techniques. Simple image-level augmentations from the Kornia library [33] are used as A; examples are shown in Fig. 3. More sophisticated augmentations such as domain translations using image-to-image vision foundation models could also be considered [11]. The Kornia augmentations are applied on the fly and randomly chosen during training. To avoid overfitting the test set, we validate our model on the Pitts30k validation set [9]. RSF is done on a single NVIDIA A100 80GB GPU and on-average takes only a few hours ($\approx 3-5$) depending on the size of the reference set.

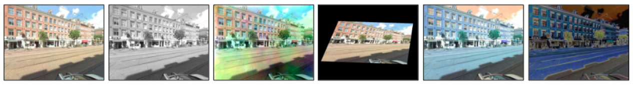

Figure 3. Examples of the augmentations applied to create finetuning queries using Kornia augmentations [33]. Left-most is the original reference image.

4.3. Results

Baseline comparison: Table 3 contains the performance of RSF models in comparison to baselines. Models finetuned using our proposed RSF outperform existing methods by a large margin for both the metrics. Please note that this performance improvement is *without* the use of new training data or a stronger backbone. The performance benefits are more significant for the challenging Nordland and AmsterTime-AR datasets, which are the primary focus due to their large train-test domain gap. We also note that the proposed RSF is beneficial for the datasets without a large train-test domain gap, e.g., the SVOX-Night and Eynsham datasets. However, the performance improvement is less significant than on other datasets. More importantly, we show that both the SOTA VPR methods, BoQ and SALAD, benefit from RSF.

We further show in Fig. 4 examples of queries that are correctly matched after the proposed RSF, and also some failure cases. Since BoQ with RSF is the best-performing method in our baseline comparison, we focus on this method in the remainder of the experiments.

Model generalization: A key component of this study is the desire for the RSF models to retain generalization to the other test datasets. For this, we report in Table 4 the performance of an RSF model finetuned on a given reference dataset and evaluated on the other test datasets. Interestingly, we note that not only do the finetuned models retain generalization to other test datasets, but also that the RSF finetuned models consistently outperform the original model, agnostic to the reference set used for finetuning. This is attributed to the additional finetuning of SOTA on VPR-specific data; however, quite expectedly, we see a diagonal trend in the bold numbers, such that the best-performing RSF model for each test dataset is always the model that was finetuned on the same test dataset's reference map.

Attention masks: We visualize the attention masks for a learned BoQ query in Fig. 5 for the original model and the RSF model. Note that the RSF model strongly attends to the unique facades of windows in the building on the right, while the original BoQ only attends to edges.

	Nordland		Amster-AR		SVOX-Night		Eynsham		Average	
	$R@1$	$R@5$	$R@1$	$R@5$	$R@1$	$R@5$	$R@1$	$R@5$	$R@1$	$R@5$
MixVPR [3]	76.1	86.8	38.3	60.4	63.1	79.5	89.4	93.2	66.7	80.0
BoQ-Res [3]	83.3	91.1	52.1	75.4	85.7	94.7	91.2	94.9	78.1	89.0
CricaVPR [24]	91.2	96.2	64.7	83.9	86.9	95.0	91.6	94.9	83.6	92.5
SALAD [17]	85.9	93.5	58.7	79.7	95.0	99.3	91.5	95.0	82.8	91.9
BoQ [4]	90.4	95.9	61.9	83.5	97.1	99.4	92.1	**95.5**	85.4	93.6
SALAD-RSF	91.4	96.2	59.9	80.6	96.1	98.8	91.8	95.2	84.8	92.7
BoQ-RSF	**94.2**	**97.7**	**65.6**	**86.3**	**98.8**	**99.6**	**92.2**	95.4	**87.7**	**94.8**

Table 3. The recalls of SOTA VPR methods tested on various challenging test datasets. The first two rows: MixVPR and BoQ-Res use ResNet-50 backbone, while the remainder use DinoV2 backbone. All methods are trained on the GSV-Cities dataset. Best is in Bold.

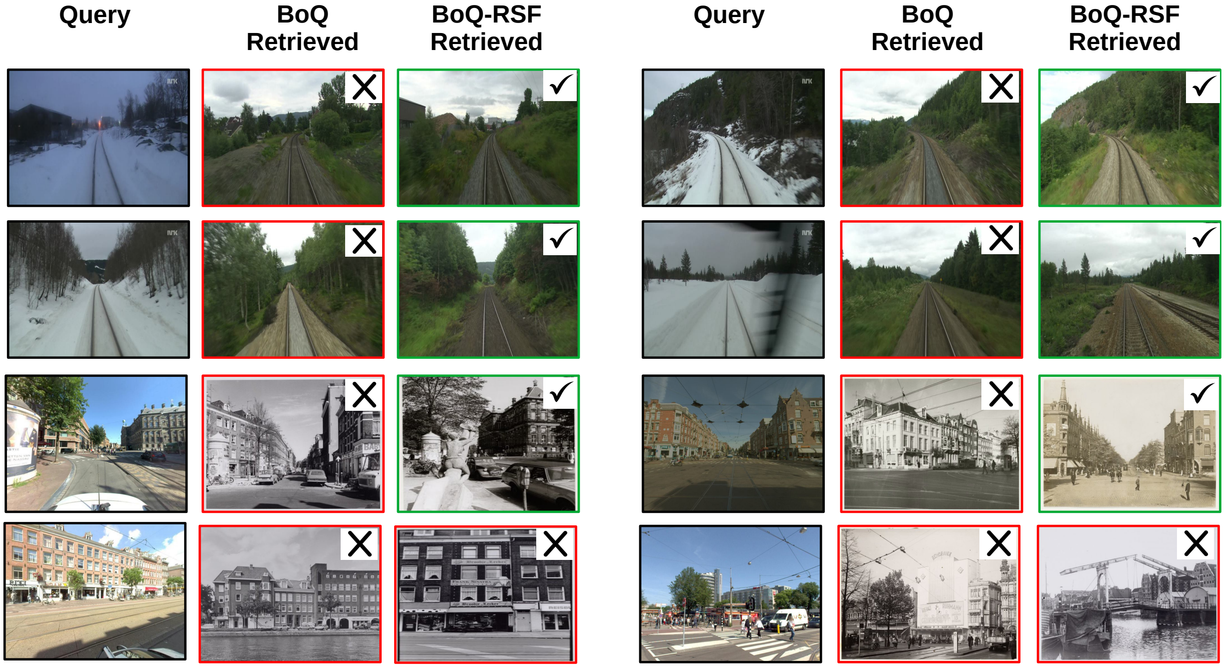

Figure 4. Examples of queries that are mismatched by the original BoQ-DinoV2 model but correctly matched by our reference-set-finetuned BoQ-RSF model, except for the last row which demonstrates two BoQ-RSF failure cases.

4.4. Ablations

We have argued in this work that the reference poses are freely available offline in VPR and are thus used in pose-based triplet mining for RSF. However, it is possible to have image-retrieval use-cases where reference images are available without pose information, e.g., image cataloging, landmark identification, etc. Table 5 thus reports the performance of our baseline in comparison to RSF models trained with and without access to pose information in the reference set. It is observed that although the reference pose information is helpful for RSF and such models are consistently the best-performing, but even without access to reference pose information, RSF models are still better than the baseline.

We further report in Table 6 the effect of Kornia augmentations on our proposed RSF for BoQ. These results show that augmentations are required to benefit from fine-tuning on the reference set, and that appearance augmentations are more useful than viewpoint augmentations for the chosen datasets. Only having viewpoint augmentations and no appearance augmentations is hurtful for RSF. We hypothesize that using viewpoint augmentations as A is distractful for the model finetuned on the Nordland dataset, since there is

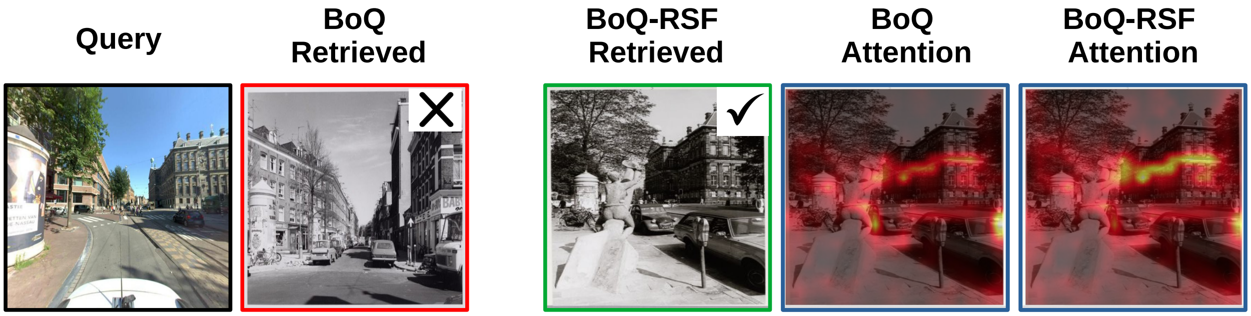

Figure 5. Learned attention for the original BoQ and the BoQ-RSF model on a ground-truth reference image is shown. The RSF model attends more to facades in the building while BoQ attends to edges. These attention masks are for the *same* BoQ query of the original and the BoQ-RSF model.

	Test dataset		
	Nord.	Amst-AR	SVOX-Ni.
Baseline BoQ	90.4	61.9	97.1
BoQ-RSF (Nord.)	**94.2**	64.4	**98.9**
BoQ-RSF (Amst-AR)	92.3	**65.6**	**98.9**
BoQ-RSF (SVOX-Ni.)	93.4	64.7	**98.9**

Table 4. The Recall@1 of RSF models on various test datasets. The first column reports the reference set used for BoQ-RSF. RSF models retains generalization. Bold numbers in the diagonal indicate that the best-performing method for each dataset is the model finetuned on that dataset's reference set.

	Nordland	Amst-AR
Baseline BoQ	95.9	83.5
BoQ-RSF (without poses)	97.1	85.3
BoQ-RSF (with poses)	**97.7**	**86.3**

Table 5. The Recall@5 performance of a baseline BoQ method is compared with RSF two test datasets with and without access to the test-time reference poses. The availability of test-time reference poses allows for hard-negative mining and gives SOTA performance compared to random negative mining when pose information is not accessible. However, even without access to the reference poses, RSF model performs better than the baseline BoQ.

almost no viewpoint change between the queries and the references in this dataset. The choice of augmentations in practice should follow from the expected query-reference domain gap, and in case of no prior knowledge about the expected Q-R gap, we recommend that the viewpoint augmentations be used together with appearance augmentations as a thumb rule.

5. Conclusions

In this work, we demonstrate that even the strong vision-foundation models-based VPR methods trained on large-scale Google Street View data struggle on test datasets which represent a domain different from the training data. We thus proposed that the reference set in test datasets is a free and valuable source of information that can be used to bridge this train-test domain gap. A simple Reference-Set-Finetuning (RSF) strategy is proposed that boosts the performance of SOTA VPR methods by large margins. The proposed RSF is shown to work for multiple datasets. The resulting finetuned models retain generalization to other test datasets. We also show that the same RSF strategy could be applied to other VPR methods, albeit the performance benefits vary. Future works could investigate further how different formulations of RSF, particularly the augmentations, could benefit different VPR methods.

Chosen A	Amster-AR	Nordland
No augmentations	83.51	95.92
No viewpoint augmentations	86.31	**97.80**
No appearance augmentations	76.20	91.13
All augmentations	**86.32**	97.70

Table 6. The Recall@5 performance of BoQ-RSF with different types of augmentations chosen as A.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54 (10):105–112, 2011. 1

[2] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022. 1, 2, 3

[3] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2998–3007, 2023. 2, 3, 4, 6

[4] Amar Ali-Bey, Brahim Chaib-draa, and Philippe Giguère.

Boq: A place is worth a bag of learnable queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17794–17803, 2024. 1, 2, 3, 4, 5, 6

[5] 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*, pages 5297–5307, 2016. 1, 2, 3, 4

[6] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[7] Gabriele Berton, Carlo Masone, Valerio Paolicelli, and Barbara Caputo. Viewpoint invariant dense matching for visual geolocalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12169–12178, 2021. 3

[8] 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*, pages 4878–4888, 2022. 2, 3

[9] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5407, 2022. 1, 2, 4, 5

[10] 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*, pages 2918–2927, 2021. 1, 5

[11] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 5

[12] Chao Chen, Zegang Cheng, Xinhao Liu, Yiming Li, Li Ding, Ruoyu Wang, and Chen Feng. Self-supervised place recognition by refining temporal and featural pseudo labels from panoramic data. *IEEE Robotics and Automation Letters*, 2024. 2

[13] Mark Cummins. Highly scalable appearance-only slam-fabmap 2.0. *In Proceedings of the Robotics: Sciences and Systems (RSS) Conference*, 2009. 5

[14] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 886–893. IEEE, 2005. 3

[15] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017. 4

[16] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-CNN: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021. 2

[17] 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*, pages 17658–17668, 2024. 1, 2, 3, 4, 5, 6

[18] Suji Jang and Ue-Hwan Kim. On the study of data augmentation for visual place recognition. *IEEE Robotics and Automation Letters*, 8(9):6052–6059, 2023. 2

[19] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4893–4902, 2019. 2

[20] 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. 2, 3

[21] Ahmad Khaliq, Shoaib Ehsan, Zetao Chen, Michael Milford, and Klaus McDonald-Maier. A holistic visual place recognition approach using lightweight CNNs for significant viewpoint and appearance changes. *IEEE Transactions on Robotics*, 2019.

[22] María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. Generalized contrastive optimization of siamese networks for place recognition. *arXiv preprint arXiv:2103.06638*, 2021. 2

[23] 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. 1, 2

[24] 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*, pages 16772–16782, 2024. 1, 2, 3, 4, 5, 6

[25] Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. *arXiv preprint arXiv:2402.14505*, 2024. 2

[26] Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547, 2021. 3

[27] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Relative camera pose estimation using convolutional neural networks. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 675–687. Springer, 2017. 3

[28] Mohamed Adel Musallam, Vincent Gaudillière, and Djamila Aouada. Self-supervised learning for place representation generalization across appearance changes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7448–7458, 2024. 2

[29] 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. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024. 1, 2, 5

[30] Noé Pion, Martin Humenberger, Gabriela Csurka, Yohann Cabon, and Torsten Sattler. Benchmarking image retrieval for visual localization. In *International Conference on 3D Vision (3DV)*, pages 483–494. IEEE, 2020. 3

[31] 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. 2, 3

[32] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5107–5116, 2019. 2

[33] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020. 5

[34] Sindre Skrede. Nordland dataset. `https://bit.ly/2QVBOym`, 2013. 1, 2, 3, 5

[35] Janine Thoma, Danda Pani Paudel, Ajad Chhatkuli, Thomas Probst, and Luc Van Gool. Mapping, localization and path planning for image-based navigation using visual features and map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7383–7391, 2019. 1

[36] Janine Thoma, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Geometrically mappable image features. *IEEE Robotics and Automation Letters*, 5(2):2062–2069, 2020. 2, 3

[37] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. Image search with selective match kernels: aggregation across single and multiple images. *International Journal of Computer Vision*, 116(3):247–261, 2016. 1

[38] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 2

[39] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393, 2014. 3

[40] Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. TransVPR: Transformer-based place recognition with multi-level attention aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13648–13657, 2022. 2

[41] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5022–5030, 2019. 3

[42] Frederik Warburg, Soren Hauberg, Manuel López-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*, pages 2626–2635, 2020. 2, 3

[43] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2575–2584, 2020. 1

[44] Burak Yildiz, Seyran Khademi, Ronald Maria Siebes, and Jan Van Gemert. Amstertime: A visual place recognition benchmark dataset for severe domain shift. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2749–2755. IEEE, 2022. 3, 5

[45] Jun Yu, Chaoyang Zhu, Jian Zhang, Qingming Huang, and Dacheng Tao. Spatial pyramid-enhanced CNN with weighted triplet loss for place recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 31(2):661–674, 2019. 2

[46] Mubariz Zaffar, Sourav Garg, Michael Milford, Julian Kooij, David Flynn, Klaus McDonald-Maier, and Shoaib Ehsan. VPR-Bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *International Journal of Computer Vision*, 129(7): 2136–2174, 2021. 2

[47] Mubariz Zaffar, Liangliang Nan, and Julian FP Kooij. On the estimation of image-matching uncertainty in visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17743–17753, 2024. 2

[48] Jian Zhang, Yunyin Cao, and Qun Wu. Vector of locally and adaptively aggregated descriptors for image feature representation. *Pattern Recognition*, 116:107952, 2021. 2

[49] Jianliang Zhu, Yunfeng Ai, Bin Tian, Dongpu Cao, and Sebastian Scherer. Visual place recognition in long-term and large-scale environment based on CNN feature. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1679–1685. IEEE, 2018. 1

[50] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19370–19380, 2023. 2