Are Local Features All You Need for Cross-Domain Visual Place Recognition?

Are Local Features All You Need for Cross-Domain Visual Place Recognition?

Giovanni Barbarani, Mohamad Mostafa, Hajali Bayramov,
Gabriele Trivigno, Gabriele Berton, Carlo Masone and Barbara Caputo
Politecnico di Torino
[giovanni.barbarani, mohamad.mostafa, hajali.bayramov]@studenti.polito.it
[gabriele.trivigno, gabriele.berton, carlo.masone, barbara.caputo]@polito.it

Abstract

Visual Place Recognition is a task that aims to predict the coordinates of an image (called query) based solely on visual clues. Most commonly, a retrieval approach is adopted, where the query is matched to the most similar images from a large database of geotagged photos, using learned global descriptors. Despite recent advances, recognizing the same place when the query comes from a significantly different distribution is still a major hurdle for state of the art retrieval methods. Examples are heavy illumination changes (e.g. night-time images) or substantial occlusions (e.g. transient objects). In this work we explore whether re-ranking methods based on spatial verification can tackle these challenges, following the intuition that local descriptors are inherently more robust than global features to domain shifts. To this end, we provide a new, comprehensive benchmark on current state of the art models. We also introduce two new demanding datasets with night and occluded queries, to be matched against a city-wide database. Code and datasets are available at https://github.com/gbarbarani/re-ranking-for-VPR.

1. Introduction

The task of Visual Place Recognition (VPR) aims to answer the question "*Where was this picture taken?*". In the literature the most popular approach is to cast the task as an image retrieval problem, where a given query is localized via comparison to a previously collected database of geotagged images [1, 2, 4, 6, 7, 17, 23, 25, 29, 34, 63, 65], and the query is considered correctly localized of its ground truth position is less then 25 meters away form the prediction. VPR can be used as a first step before more precise visual localization, and can find multiple applications in fields like autonomous driving, SLAM and augmented reality. Given these applications, the task is usually performed in large-scale outdoor scenarios, for which the database is collected in an automated fashion, typically via Street View data [6, 8, 53, 54], which ends up being made up of mostly day-time images. On the other hand, the queries that a real-world VPR system receives once deployed may be subject to high appearances changes, due to night-time images, occlusions, critical meteorological conditions. This domain shift between queries and database still represents a major challenge in the literature [3, 8, 20, 32, 39, 53, 56, 57, 62, 64].

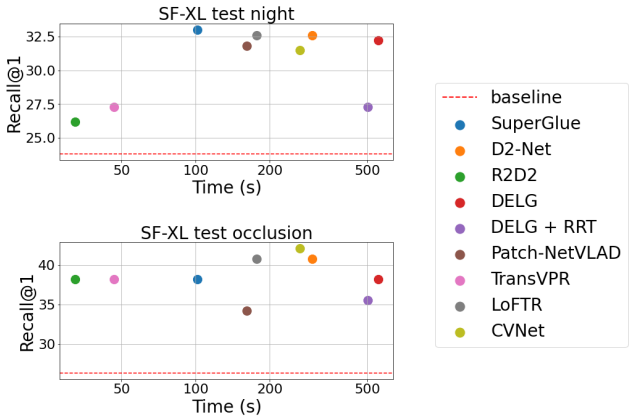


Figure 1. **Plot showing the Recall@1 and latency for different methods on multiple datasets.** Latency is to re-rank 100 candidates of a single query, considering local features extraction to be performed online. We can see that there is no single method that outperforms all others on all scenarios, and the ideal choice of a re-ranking method for a VPR system depends on multiple factors, such as time requirements and expected domain shifts.

To improve results and address these issues, a number of previous works noted that local features [9, 24, 33, 55] and image matching [15, 42, 43, 49] methods are inherently more robust to domain shifts , and that these can be used to re-rank a set of candidates (usually through spatial verification) provided through image retrieval methods, leading to large improvements in results [18, 55].

Our work aims to quantify the effectiveness of these methods that provide a matching score between two images, when applied to re-rank the top-N candidates of a re-

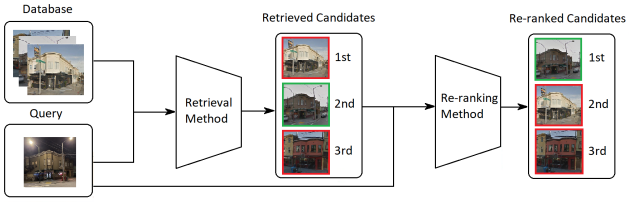


Figure 2. **Overview of the re-ranking pipeline.** First, a retrieval method performs a similarity search on the global descriptors extracted from query and database to output a set of top-k candidates. Then re-ranking is applied to refine the retrieved candidates.

trieval module for VPR. Despite the recent interest in this category of methods, they are yet to be studied in the VPR setting, and previous work only compare a small number of these methods [18, 55]. Furthermore, previous comparisons of such methods in a VPR setting are hindered by the use of different underlying retrieval methods: for example, in [55] the authors use DELG, Patch-NetVLAD, TransVPR and SuperGlue to re-rank a shortlist of candidates provided by different retrieval methods, making it difficult to understand whether better results are granted by re-ranking or the retrieval module.

To establish which methods are most suited for re-ranking in real-world VPR, we perform an extensive benchmarks of existing image matching pipelines with a focus on the domain shift problem. In particular, we focus on creating fair benchmarking conditions, by providing all the re-ranking methods with the same pool of candidates to score. Where possible, we use the same backbone for local feature extraction, and use the same hardware to carry out extensive efficiency evaluation. Our benchmark reveals that even highly challenging datasets can be nearly solved by combining SOTA retrieval and re-ranking methods (*e.g.* on Tokyo night, which uses only the night queries of Tokyo 24/7, we achieved a Recall@$1 > 95\%$).

We therefore propose two new challenging datasets, in order to provide a stimulating and challenging benchmark to foster future research. In the creation of these two datasets we focused on the two most challenging domains for VPR: the first has night-time queries, whereas the second has queries with heavy occlusions due to dynamic objects (*e.g.* vehicles and pedestrians). For both datasets, we collected (and manually verified) queries from Flickr, whereas as a database we use the San Francisco eXtra Large (SF-XL) dataset.

Our contributions can be summarized in the following points:

- We propose two new query sets to allow to evaluate the performance on night-time and occluded images against a city-wide database. Both query sets have been collected from Flickr and manually curated.
- We construct a benchmark to explore the applicability of spatial verification techniques for re-ranking in VPR. We create comparable setups to isolate the performances of the tested methods, quantify their gains with respect to the state of the art in VPR.
- We find that re-ranking methods are able to greatly improve the results over commonly used retrieval methods, and we observe that there is no clear winning solution, as different scenarios require different methods.

2. Related Work

Visual Place Recognition through Image Retrieval

Image Retrieval is the most common way to approach the task of Visual Place Recognition. A neural network is used to extract global descriptors from the query and database images, and then a kNN is performed to find the matches to the query. Among the proposed global extractors, NetVLAD [4], a deep learning successor of VLAD [22], established a milestone in the field of VPR. NetVLAD led to the birth of a large number of work that proposed improvements to it, commonly trained with variants of the weakly supervised triplet loss: among them we note ApaNet [66], CRN [23], SARE [29], SFRS [17], AppSVR [36] and SralNet [35]. Such methods have been recently outperformed by models that do not rely on NetVLAD, use smaller descriptors, and propose new training techniques to scale to large training sets: the most notable examples of this trend are CosPlace [6], Conv-AP [1] and MixVPR [2].

A separate line of works [3, 8, 56] propose to explicitly tackle the domain shift issue through domain adaptation, although such methods are by nature focusing on a single domain, lacking generalization capabilities.

While retrieval methods have been covered by a number of benchmarks throughout the years [7, 40, 44, 63], no benchmark has focused on the possibilities that re-ranking method offers and the computational trade-offs they entail.

Local features for spatial verification

Local features-based spatial verification represents an established paradigm that has been applied to several computer vision tasks, ranging from structure from motion (SfM) [26, 28, 46], simultaneous localization and mapping (SLAM) [11, 18, 43] and visual localization [26, 44, 49, 53]. For many years hand-crafted feature extractors have represented a remarkably strong baseline [5, 30], whereas pioneering learning-based methods showed large margin for improvements under perspective and lightning changes [13, 61]. In recent years, we have witnessed a flourishing literature on learnable detectors and descriptors exploiting local features for pose and homography estimation [12, 14, 15, 21, 42, 43, 49]. Although these methods are not specifically designed for retrieval, they can be naturally used to re-rank retrieval candidates by assigning a higher similarity to pairs of images that share more matches across

Database	Database source	Database # images	Query set	Queries source	# queries
Tokyo 24/7 [53]	Google StreetView	75k	Tokyo night (night queries from Tokyo 24/7)	Collected with a smartphone by [53]	105
SVOX [8]	Google StreetView	17k	SVOX Night	Oxford RobotCar	823
SF-XL [6]	Google StreetView	2.8M	SF-XL test v1	Flickr	1000
			SF-XL test v2	Collected with a smartphone by [10]	598
			SF-XL test night (**ours**)	Flickr	466
			SF-XL test occlusion (**ours**)	Flickr	76

Table 1. **Summary of the datasets considered in our experiments.** The table reports the raw data sources, every author has cleaned and processed them in customized way, refer to their papers for the details. In general streetview panoramas have been cropped in patches and turned in multiples suitable references for the databases. Flickr queries have been filtered and manually checked for positive references. Oxford RobotCar [31] data have been collected and processed with a modality analogues to streetview panoramas, although Tokyo 24/7 [53] queries have been collected with smartphone devices they are scenes compatible with a moving vehicle point of view. While Flickr and San Francisco Landmark [10] data contains a broader range of point of views and camera types.

local features.

However, methods trained for outdoor image matching can be less robust to dynamic objects (*e.g.* they may match the cars between two images instead of the buildings), than other methods that were specifically trained for image retrieval and re-ranking [9, 18].

Generally, local features are represented as pairs of keypoint (*i.e.* the pixel coordinate of the feature) and descriptor (a fixed size vector). Given a pair of images, the local features are then cross-matched to find pairs of keypoints across the two images, in a procedure known as spatial verification. While this step is usually performed with heuristics like RANSAC [16], data-driven approaches like SuperGlue [43] have been proposed for the task.

SuperGlue uses graph neural networks to learn data-dependent priors on matches given two sets of keypoints as an input. This approach has been subsequently generalized by LoFTR [49], which removes the dependence from an underlying detector exploiting cross-attention transformers for directly selecting keypoints matches among an image pair.

Some of these methods have been evaluated on outdoor datasets that are not commonly used in the VPR literature. Some examples are Oxford5k [37] and Paris6k [38], not suited for VPR since they do not provide a dense database (densely covering a given area), and others, like Aachen [44, 45], cover only a small area of a city, and are mainly used for pose estimation. The same consideration holds for Madrid Metropolis, Gendarmenmarkt and Tower of London, proposed in [59].

Local features for re-ranking

Local features have been explored also for image retrieval, mainly with purpose of re-ranking the shortlist of top-N candidates proposed by the retrieval module [9, 18, 50, 55]. To this end, the matching algorithm needs to be turned in a scoring algorithm, commonly using the number of matches found in an image pair as a proxy of confidence. In other cases, instead of an explicit re-ranking step, local features are used either to enhance global descriptors or to match directly reference images [33, 52, 58, 60]. DELG [9] uses a global extractor trained with a large margin cosine loss, coupled with local features refined following unsupervised criteria for discriminativeness and reliability. Patch-NetVLAD [18] obtains a set of dense local features performing the VLAD aggregation on local patches, while TransVPR [55] is based on a transformer architecture that selects a subset of patches through its multi-scale attention maps. These last two methods were designed for VPR. The majority of re-ranking methods exploit RANSAC, using the number of inliers to assign a score to the candidates [9, 18, 33, 55].

Recently, researchers have been explored end-to-end learnable architectures able to estimate a similarity score between pairs of images, as alternative to RANSAC. In [50], the authors feed the local and global features from DELG to a transformers architecture and then cast the problem as a binary classification task. Similarly, CVNet [24] builds a pyramid of 4D correlation maps from the feature maps of a CNN. The 4D maps are then reduced to a similarity score through 4D convolutions. Unlike many methods for homography estimation, the algorithms mentioned in this section are trained without patch-level supervision.

3. Dataset

Previous datasets

Several datasets with query splits that explicitly aim to measure the domain shift adaptation capability have been proposed in the VPR literature [6, 8, 31, 53].

Among these, Tokyo 24/7 [53] presents three queries splits taken from different times of the day, respectively daytime, sunset, night, providing a widely used dataset for cross-domain VPR. While Tokyo 24/7 provides a well curated dataset and is widely used in literature [4, 17, 27, 29], we argue that it is quite limited in size: there are 105 queries per domain and a database of 75k images, covering just a small part of the city.

In [8] presented SVOX, a cross-domain dataset with database from Google StreetView. The queries come from RobotCar [31] and belong to a number of domains, namely snow, rain, sun, night and overcast. SVOX uses a database

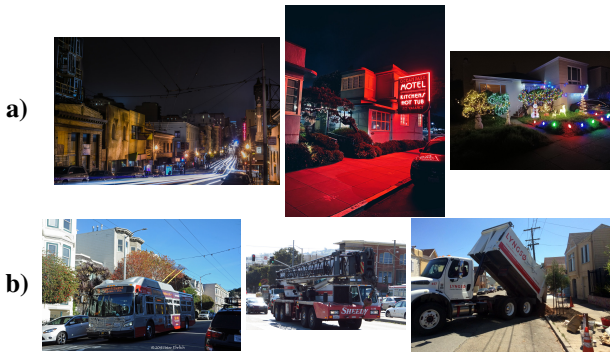


Figure 3. **Examples of queries from a) SF-Night and b) Sf-occlusion.**

smaller than Tokyo 24/7's, and it only provides frontal-view images, *i.e*. with the facing straight along the road.

Another cross-domain dataset is San Francisco eXtra Large (SF-XL) [6], with a large-scale database that covers the whole city of San Francisco with 2.8M StreetView images. SF-XL provides two query sets named *test v1* and *test v2*: (1) *test v1* are 1000 images from Flickr, uniformly distributed across the whole city, providing some domain shifts (mostly viewpoint and a few night images); (2) *test v2* is a set of 598 images taken with a smartphone in the city center, with images providing mild to moderate viewpoint shift w.r.t. to the database. Although SF-XL large scale makes it a relevant option for research in VPR, its lack of well defined query splits from multiple domains makes it difficult to understand which methods can perform well in certain domain shifts.

To overcome the aforementioned limitations of previous datasets, we built two new sets of night and occluded images respectively, to be used against the database of SF-XL.

Our new datasets To obtain realistic and diverse queries, we downloaded hundreds of thousands images from Flickr for the area of San Francisco, similarly to [6, 37, 38, 41].

With the help of trained classifiers, we then removed indoor images, and proceeded with the creation of two challenging sets of queries:

- *SF-XL test night* is a set of night images, which we automatically selected with a trained classifier.
- *SF-XL test occlusion* is a set of images that present heavy occlusions: the images were automatically selected using an object detection model, keeping those with a dynamic object (*e.g*. car, truck, person) with width $> 50\%$ and height $> 30\%$ the size of the image.

Due to the inaccuracy of Flickr geotag information we manually verified the positions of each image, which resulted in 466 and 76 images for the *SF test night* and *SF test Occlusion* respectively. A sample of queries is shown in Fig. 3.

Given the availability in literature of an open-source large-scale dataset that covers the city of San Francisco, namely SF-XL [6], we match our proposed sets of queries against the SF-XL dataset, in practice using it as a database.

A summary of the datasets that we use in our benchmarks is shown in Tab. 1.

Retrieval Method	Descriptors Dimension	SF-XL test v1 R@1	R@5	R@10
NetVLAD	4096	33.1	45.0	50.4
TransVPR	256	9.7	16.6	20.3
CVNet	2048	70.1	81.2	84.6
DELG	2048	64.3	73.0	76.1
CosPlace	512	**76.7**	**82.5**	**85.6**
Conv-AP	4096	49.1	60.6	65.6
MixVPR	4096	72.3	79.5	81.4

Table 2. **Recalls with different retrieval methods.** We used only global descriptors for this table (*i.e*. no re-ranking is applied to DELG and CVNet). NetVLAD uses a VGG-16 [47] (and PCA), TransVPR a custom transformer model, while for all other methods we used the author's ResNet-50 [19] implementation.

4. Experiments

4.1. Benchmark Methodology

To further motivate our benchmark, we point out that the application of spatial verification methods to the VPR task is not straightforward in light of the fact that many of them [43, 49] are trained on 3D models from SfM [26], thus having access to accurate matching labels. On the other hand, in the place recognition settings, matches are more loosely defined (within 25 m [4, 7, 40]) and thus positive matches may share only a small portion of a scene. These differences raise doubts on the performances of these methods against complex perspective shifts and transient objects. In our proposed benchmark we shed a light on these previously unexplored research questions.

In our set of experiments, our aim is to maximize the results given the following two-step pipeline (see Fig. 2):

1. first we obtain a shortlist of K candidates using global descriptors methods (*i.e*. the K nearest neighbor to the query in features space);
2. sort the K candidates with a re-ranking algorithm.

Given that a large body of literature on image retrieval through global descriptors already exists in the specific task of VPR [1, 2, 4, 6, 7, 63], our benchmark focuses on the second step, *i.e*. the re-ranking algorithms. To this end, we obtain a shortlist of candidates using CosPlace [6] (using a ResNet-50 backbone), which outperforms all other methods on SF-XL test v1 (see Tab. 2). Then, we perform the re-ranking step with a number of methods from the literature, namely SuperGlue, D2-Net, R2D2, DELG, PatchNetVLAD, TransVPR, LoFTR and CVNet. By providing these algorithms with the same set of candidates to re-rank, it is possible to disentangle the effect of the local features from the global extractor performance. In this way we ob-

Features Extractor	Features Matching	Tokyo night R@100 = 96.2			SVOX night R@100 = 90.3			SF-XL test v1 R@100 = 92.5			SF-XL test v2 R@100 = 97.7			SF-XL test night R@100 = 41.6			SF-XL test occlusion R@100 = 60.5		
		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
-	-	80.0	88.6	91.4	51.6	68.8	76.1	76.7	82.5	85.6	89.0	95.3	96.3	23.8	29.0	31.5	26.3	38.2	46.1
SuperPoint	SuperGlue	**95.2**	95.2	95.2	77.9	**85.2**	**86.5**	**88.6**	**91.6**	**91.9**	92.8	96.7	**97.7**	**33.0**	38.0	39.1	38.2	44.7	50.0
D2-net	RANSAC	92.4	**96.2**	**96.2**	78.9	85.1	86.4	87.5	90.3	90.8	94.0	96.3	97.0	32.6	38.2	39.5	40.8	48.7	51.3
R2D2	RANSAC	86.7	90.5	92.4	72.5	80.7	82.9	85.1	88.2	89.6	**94.1**	**96.8**	96.8	26.2	32.2	33.9	38.2	47.4	50.0
DELG	RANSAC	94.3	95.2	**96.2**	**80.1**	84.1	86.0	88.5	91.2	91.5	93.8	96.2	97.0	32.2	37.6	39.2	38.2	50.0	53.9
DELG	RRT	84.8	94.3	95.2	66.3	81.7	85.7	85.3	89.6	90.4	88.6	96.0	97.2	27.3	35.6	38.6	35.5	48.7	52.6
Patch-NetVLAD	RANSAC	90.5	94.3	94.3	67.2	80.6	83.6	77.0	84.7	87.0	91.0	95.2	96.2	31.8	37.3	38.4	34.2	47.4	52.6
Patch-NetVLAD	Rapid Scoring	73.3	87.6	92.4	42.2	66.3	73.1	69.3	80.3	84.1	90.0	94.6	95.8	21.7	31.3	35.4	25.0	38.2	42.1
TransVPR	RANSAC	88.6	95.2	95.2	63.8	79.2	83.2	84.0	87.6	89.1	92.5	96.2	96.7	27.3	34.3	36.7	38.2	46.1	52.6
LoFTR		93.3	95.2	95.2	80.0	84.0	85.3	87.9	89.8	90.7	93.3	96.3	97.2	32.6	37.6	38.2	40.8	48.7	51.3
CVNet		94.3	**96.2**	**96.2**	74.6	**85.2**	**86.5**	84.8	91.0	91.6	88.0	95.8	97.0	31.5	**39.3**	**39.9**	**42.1**	**52.6**	**56.6**

Table 3. **Recalls before and after applying re-ranking.** The shortlist of candidates to be re-ranked is obtained with CosPlace, and the results with such shortlist are shown in the first row. Re-ranking has been applied to the first 100 candidates (*i.e.* $K = 100$). Next to each dataset's name, we show the R@100, which in practice sets the upper bound of the maximum recalls achievable after re-ranking. Best results are in **bold**, second best are underlined.

Model	Descriptors size (num. × dim.)	Backbone	Designed for re-ranking	Sparse Keypoints
DELG	1000 x 128	ResNet-50	✓	✓
Patch-NetVLAD	2826 x 4096	VGG-16	✓	✗
TransVPR	522 x 256	Custom CNN+transformer	✓	✓
R2D2	4126 x 128	custom L2-Net [51]	✗	✓
D2Net	2775 x 512	VGG-16	✗	✓
SuperPoint	1034 x 256	custom VGG	✗	✓

Table 4. **Characteristics of local features extractors.** The descriptors size was computed for all methods on the same image of resolution 480x640. For Patch-NetVLAD descriptors size depends only on the resolution, because it uses dense keypoints/features, whereas for all other methods the number of descriptors depends on the visual content of the image.

tain an indication of the benefit one can expect from the spatial verification step on this challenging task, quantifying the expected gains when modularly integrating these models a pre-existing VPR pipeline.

Given that re-ranking is performed on K candidates, the value of K is of great importance: higher values of K reduce the speed of the search, but (might) also lead to higher results (we investigate this effect in Sec. 4.4). For our main experiments we set $K = 100$, following [9, 18, 50, 55]; in Sec. 4.4 we investigate how different values of K affect speed and results.

Following the VPR literature [4, 6, 7, 17, 29, 65], we use the Recall@N (R@N) as metric, which indicates the proportion of queries for which at least one of the first N predictions is correct, *i.e.* within a given threshold distance from the query. The threshold distance is set to 25 meters, although in Fig. 4 we investigate how results change with a distance of 50 and 100 meters. Note that a positive image might not share any visual content with the query (e.g. distance ¡ 25 meters but opposite viewpoint direction with the query), although the chance that one of these positives is matched to a query by pure chance decreases as the database increases, and that it would otherwise be unfeasible to obtain unbiased ground-truth covisibility (as it would rely on one of the methods used for testing).

4.2. Implementation details

To provide a relevant benchmark, we use a large number of methods, some of which were specifically designed for re-ranking and some for tasks like spatial verification and image matching. Specifically, we use SuperGlue [43] (which uses SuperPoint [12] local features), D2-Net [15], R2D2 [42], DELG [9], Reranking Transformers (RRT) [50] (which uses DELG local features), Patch-NetVLAD [18] with both its RANSAC and Rapid Scoring implementation, TransVPR [55], LoFTR [49] and CVNet [24].

For all methods, we use the official implementations and weights released by the authors, without fine-tuning. When more options were available we chose the configuration with best performance on *SF-XL test v2*, with the only exception that we preferred models with ResNet50 backbone to ResNet100 counterpart.

A number of methods use some kind of multi-scale approach, namely DELG, R2D2, D2-Net, Patch-NetVLAD and CVNet. Most spatial verification methods rely on the standard RANSAC with 8 parameters to describe a homography, with the exceptions of DELG, which uses an affine version of RANSAC (*i.e.* with 6 parameters). Patch-NetVLAD applies RANSAC multiple times to match multiple scales (for the multi-scale approach). Patch-NetVLAD also provides a fast version, which uses a novel alternative to RANSAC called Rapid Scoring, which is a non-iterative heuristics that has been proposed as a faster option to RANSAC as it does not require an iterative algorithm to be computed. Preliminary experiments with Rapid Scoring on methods other than Patch-NetVLAD gave poor results, and we found it to be sensible to outliers and reliable only for moderate changes in viewpoint. For Patch-NetVLAD and TransVPR we resized the images to have a resolution of 480x640, following the original implementations.

4.3. Quantitative evaluations of results

Results from our experiments are shown in Tab. 3. We point out how the majority of the methods effectively pro-

vide a boost with respect to the baseline performance. For example, LoFTR and SuperGlue grant on average a 21% and 13% boost, respectively on night and occlusion benchmarks. This supports one of the motivation of our paper, that is showing the potential of methods based on local features to overcome the limitations of global descriptors in these challenging scenarios. We summarize other detailed findings from the experiment table in the following points:

- Interestingly, methods designed for Image Matching turn out to be highly competitive even when extended to VPR systems. In particular, SuperGlue achieves the best recalls for *Tokyo night*, *SF-XL test v1* and *SF-XL test night*. This disproves the intuitive hypothesis that these models would suffer the absence of a training protocol that explicitly encodes a prior on ignoring transient objects. The same considerations hold as well for D2-Net and LoFTR, which perform very closely to SuperGlue across the board.
- Among the native re-ranking methods, DELG paired with RANSAC is the more versatile option. It reaches the best performance on *SVOX night*, and on every other dataset its R@1 is comparable with the highest score. Regarding CVNet, it grants the highest R@5 and R@10 in almost every benchmark, despite some drops in R@1 for *SVOX night* and *SF-XL test v1*. TransVPR and Patch-NetVLAD end up being by far the less robust to the night domain.
- CVNet is the single best model on *SF-XL test occlusion*. This confirm the effectiveness of its training procedure that involves Hide-and-Seek [48] augmentations for robustness against occlusions.
- Rapid scoring reaches very modest results, failing to provide a faster alternative for matching. Likewise, scoring with RRT improves the baseline but it is significantly worse than RANSAC variant on night datasets.
- Finally, it emerges how our newly proposed datasets are far from solved; we believe that this new and challenging benchmarks will inspire the community, paving the way for future research.

4.4. Ablation on K and different positive threshold distances

Given that re-ranking can be orders of magnitude more expensive than standard retrieval (through a nearest neighbor search), it is important to understand the ideal number of candidates to be re-ranked for an efficient VPR system. To this end, we use the best-performing methods from Tab. 3 and run a new set of experiments by using different values of K, precisely any $K \in \{1, 2, 3, ...100\}$. Furthermore, we investigate how this affects the Recall@1 not only when using a threshold for positives of 25 meters, but also when increasing it to 50 and 100 meters. Given the large number of combinations, we report the results on the two datasets that we believe would best represent a real-world scenario: *SF-XL test v1* and *SF-XL test night*.

Interestingly, we find that no single method achieves best results across the board:

- Firstly, it can be noted that simply increasing the threshold up to 100m allows to count more matches as correct. This is relevant as many applications that do not require high localization precision can exploit this effect. In particular, increasing the threshold grants higher gains on the more challenging datasets.
- DELG and CVNet exhibit superior performance when the threshold is increased to 100 meters: this is probably due to them being trained on the Google Landmark Dataset, which provides photos of buildings from far apart. In particular DELG scores higher on *SF-XL test v1* whereas CVNet is superior in handling night images and transient occlusions.
- D2-Net, SuperGlue and DELG provide more precise matches under domain shift. They achieve the top scores with a threshold of 25 meters on *SF-XL test night*.
- As a rule of thumb, the higher the K the better the results. However, it should be noted that the more challenging the dataset, the earlier this curve plateaus. This effect is especially visible with lower thresholds; in particular on *SF-XL test occlusion* in many cases increasing K leads to higher false positives ratio. Considering that the cost of re-ranking scales linearly with K, the choice of this parameter must be devoted the utmost attention.
- Lastly, we can see that the upper bound (*i.e.* the Recall@K with CosPlace) is still much higher than any of the re-ranking methods, proving that there is still a large margin for improvements.

4.5. Qualitative evaluations of results

To give an intuition to the reader over the strengths and weaknesses of three re-ranking methods (*i.e.* SuperGlue, DELG and CVNet), we report in Fig. 5 a number of queries and the first prediction with each method.

4.6. Is the night domain a real challenge?

In this section we disentangle if the errors in night time datasets are due to the difficult illumination or other factors. We considered the 101 queries of *Tokyo night* for which CosPlace provides at least one positive within the first 100 candidates. Of these 101 queries, we found that CVNet is able to solve every single query, while DELG and SuperGlue fail in one case (which is shown in Fig. 5 (d)). Given these results, we argue that the difficulty of the *SF-XL test night* dataset is not solely due to the night domain: for example, a factor could be that the night photos from Flickr often contain several other challenges, repre-

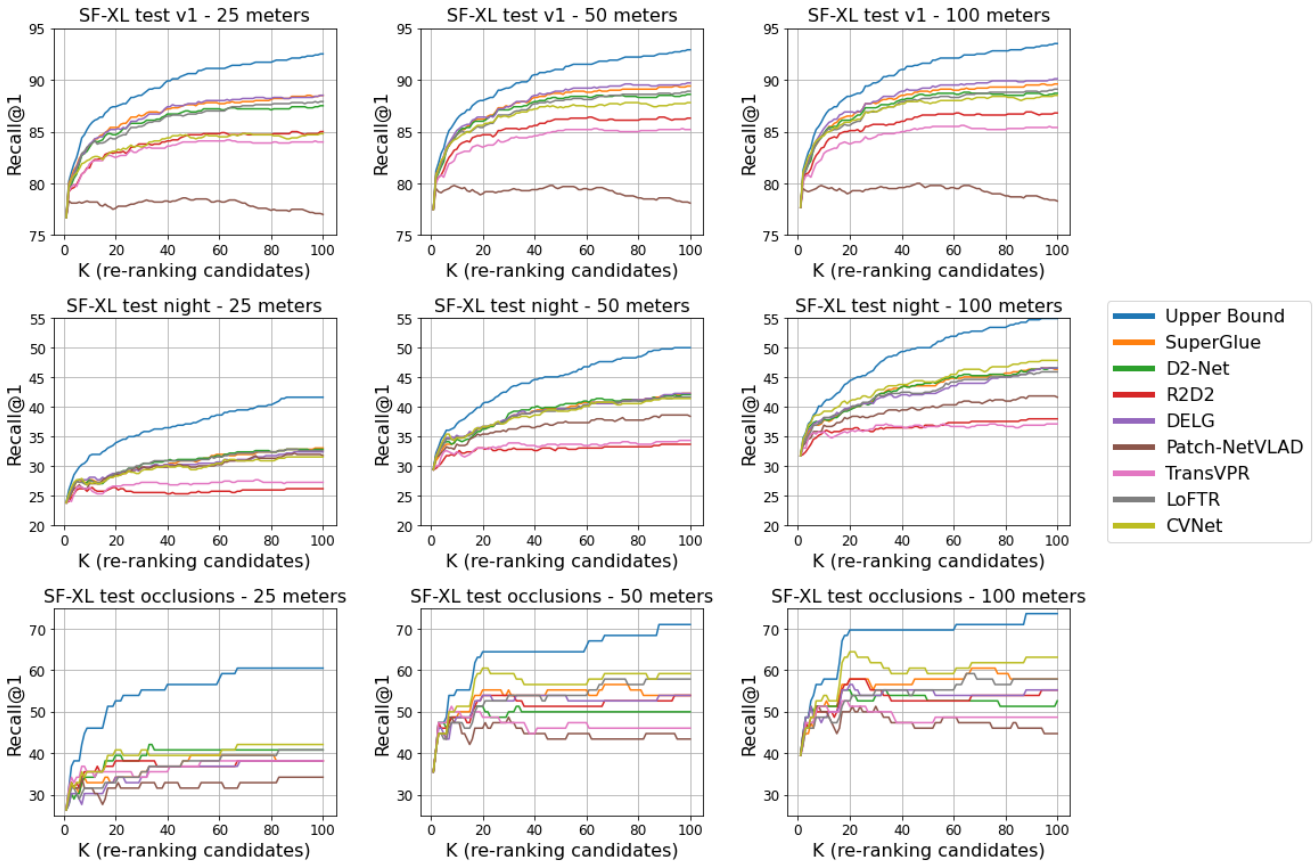


Figure 4. **Re-ranking with different values of K, from 1 to 100.** The "Upper Bound" is the Recall@K without applying re-ranking (*i.e.* with CosPlace). For DELG and Patch-NetVLAD we used the version with RANSAC.

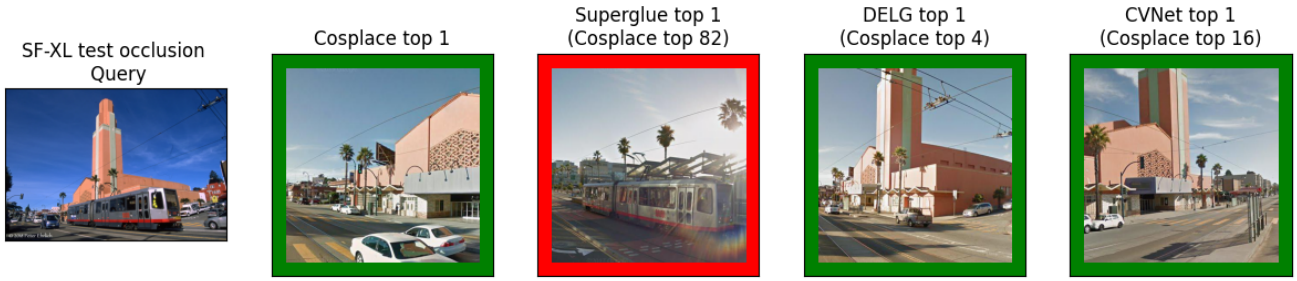
sentative of realistic use-cases. For instance heavy viewpoint shifts, and artificial lights such as signboards, decorative lights can highly affect the visual appearance of a place (*e.g.* see queries in Fig. 3). These results prove that state-of-the-art local features are indeed very robust to illumination changes, and that our newly proposed *SF-XL test night* highlights the real challenges that photos in the wild can present.

4.7. Computational cost

In Fig. 1 we study the computational requirements for the considered re-ranking methods in relationship to their performances on *SF-XL test night* and *SF-XL test occlusion*. We plot the time required to re-rerank the top-100 candidates for a single query, considering online extraction of local features for the candidates from the database. Although many works consider this step to be computed offline, without any quantization techniques the storage cost would quickly explode on realistic large scale databases for VPR. For instance, storing SuperPoint local descriptors for the SF-XL database would require roughly 1 Tb. Since quantization techniques must be studied accurately for each case [7, 33], we kept the most generally applicable implementation and considered online local features extraction. In general, methods with lighter backbones for feature extraction are the fastest, namely R2D2 and TransVPR. Whereas DELG, either with RANSAC or RRT, is the costlier approach. Including performances into the equation, SuperGlue, LoFTR and CVNet attain the best trade-off overall. Nevertheless, these delays of the order of hundreds of seconds may not be acceptable in many practical applications, and this trade-off should be carefully evaluated together with the ablation on the number of candidates to re-rank presented in Fig. 4. It shows that, despite it is common practice in the re-ranking literature to adopt $K = 100$ or more [9, 18, 33, 50, 55], in many cases it is possible to cut down inference time substantially reducing K without suffering big performance hits.

5. Conclusion

In this paper we investigate how re-ranking techniques can be used to improved results in visual place recognition. Specifically, we experiment on the relevant setting when the



(a) A failure of SuperGlue due to a dynamic object (a tram), which SuperGlue (unlike DELG and CVNet) has not been trained to ignore. We can also see that CVNet finds a positive with very different viewpoint than the query, even though candidates closer to the query are available.

(b) DELG and CVNet failures for this case are most likely due to those methods using a combination of local and global scoring system. The global features see trees and a red line (which for the query is on the bus, and for the predictions is an awning).

(c) This example shows the robustness of CVNet to strong occlusions, which is learned thanks to its use of Hide-and-Seek data augmentation [48].

(d) The only example from *Tokyo night* where DELG and SuperGlue fail to find the correct prediction.

Figure 5. **Qualitative examples of 3 queries and the first prediction with 4 relevant methods**, namely CosPlace (retrieval baseline) SuperGlue, DELG and CVNet. Predictions are in green if they are less than 100 meters away from the query's ground truth.

queries come from a different domain than the database, with a focus on night and occluded queries. We propose two challenging query sets, on which even the best combination of methods achieve a Recall@1 $< 50\%$.

We provide a large set of experiments to show which methods perform best for the task of cross-domain re-ranking for VPR, finding that many methods achieve pareto-optimal solutions when time and recalls are considered. We also find that different domain shifts require different approaches, and that there is no clear winner across all datasets, even when latency is not an issue.

We believe that our work can shed light on how to design a highly performant VPR system on multiple conditions, and that our proposed datasets will foster further research to continue to improve the state of the art.

Acknowledgements. This work was supported by CINI.

References

[1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022. 1, 2, 4

[2] 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. 1, 2, 4

[3] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE, 2019. 1, 2

[4] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1437–1451, 2018. 1, 2, 3, 4, 5

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110:346–359, 06 2008. 2

[6] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *CVPR*, June 2022. 1, 2, 3, 4, 5

[7] 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 Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 4, 5, 7

[8] Gabriele Berton, Valerio Paolicelli, Carlo Masone, and Barbara Caputo. Adaptive-attentive geolocalization from few queries: A hybrid approach. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2918–2927, January 2021. 1, 2, 3

[9] B. Cao, A. Araujo, and J. Sim. Unifying deep local and global features for image search. In *European Conference on Computer Vision*, pages 726–743. Springer Int. Publishing, 2020. 1, 3, 5, 7

[10] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 737–744, 2011. 3

[11] Mark Cummins and Paul Newman. Appearance-only slam at large scale with fab-map 2.0. *The International Journal of Robotics Research*, 30(9):1100–1123, 2011. 2

[12] Tomasz Malisiewicz Daniel DeTone and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Computer Vision and Pattern Recognition Workshop*, 2018. 2, 5

[13] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Toward geometric deep slam. *arXiv preprint arXiv:1707.07410*, 2017. 2

[14] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Toward geometric deep slam. *arXiv preprint arXiv:1707.07410*, 2017. 2

[15] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 5

[16] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3

[17] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 369–386, Cham, 2020. Springer International Publishing. 1, 2, 3, 5

[18] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021. 1, 2, 3, 5, 7

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[20] Sarah Ibrahimi, Nanne van Noord, Tim Alpherts, and Marcel Worring. Inside out visual place recognition. In *British Machine Vision Conference*, 2021. 1

[21] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021. 2

[22] Hervé Jégou, Matthijs Douze, Jorge Sánchez, Patrick Perez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34, 12 2011. 2

[23] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Learned contextual feature reweighting for image geo-localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3251–3260, 2017. 1, 2

[24] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5374–5384, June 2022. 1, 3, 5

[25] María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. Data-efficient large scale place recognition with graded similarity supervision. *CVPR*, 2023. 1

[26] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 4

[27] Dongfang Liu, Yiming Cui, Liqi Yan, Christos Mousas, Baijian Yang, and Yingjie Chen. Densernet: Weakly supervised visual localization using multi-scale feature aggregation. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6101–6109, May 2021. 3

[28] Liu Liu, Hongdong li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. pages 2391–2400, 10 2017. 2

[29] Liu Liu, Hongdong Li, and Yuchao Dai. Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization. In *IEEE International Conference on Computer Vision*, 2019. 1, 2, 3, 5

[30] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. 2

[31] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research*, 2017. 3

[32] Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547, 2021. 1

[33] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3476–3485, 2017. 1, 3, 7

[34] Valerio Paolicelli, Antonio Tavera, Carlo Masone, Gabriele Moreno Berton, and Barbara Caputo. Learning semantics for visual place recognition through multi-scale attention. In *Image Analysis and Processing – ICIAP 2022*, 2022. 1

[35] Guohao Peng, Yufeng Yue, Jun Zhang, Zhenyu Wu, Xiaoyu Tang, and Danwei Wang. Semantic reinforced attention learning for visual place recognition. In *IEEE International Conference on Robotics and Automation*, pages 13415–13422. IEEE, 2021. 2

[36] Guohao Peng, Jun Zhang, Heshan Li, and Danwei Wang. Attentional pyramid pooling of salient visual residuals for place recognition. In *IEEE International Conference on Computer Vision*, pages 885–894, October 2021. 2

[37] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2007. 3, 4

[38] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2008. 3, 4

[39] Nathan Piasco, Desire Sidibé, Cédric Demonceaux, and Valérie Gouet-Brunet. A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognit.*, 74:90–109, 2018. 1

[40] Noé Pion, Martin Humenberger, Gabriela Csurka, Yohann Cabon, and Torsten Sattler. Benchmarking image retrieval for visual localization. In *2020 International Conference on 3D Vision (3DV)*, pages 483–494, 2020. 2, 4

[41] F. Radenović, G. Tolias, and O. Chum. Fine-tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 4

[42] Jerome Revaud, Philippe Weinzaepfel, César Roberto de Souza, and Martin Humenberger. R2D2: repeatable and reliable detector and descriptor. In *NeurIPS*, 2019. 1, 2, 5

[43] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1, 2, 3, 4, 5

[44] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla. Benchmarking 6DOF outdoor visual localization in changing conditions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8601–8610, June 2018. 2, 3

[45] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 3

[46] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 2

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 4

[48] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3544–3553, 2017. 6, 8

[49] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 1, 2, 3, 4, 5

[50] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *IEEE International Conference on Computer Vision*, 2021. 3, 5, 7

[51] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 661–669, 2017. 5

[52] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 460–477. Springer, 2020. 3

[53] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):257–271, 2018. 1, 2, 3

[54] Akihiko Torii, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. Visual place recognition with repetitive structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2346–2359, 2015. 1

[55] 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 (CVPR)*, pages 13648–13657, June 2022. 1, 2, 3, 5, 7

[56] Ziqi Wang, Jiahui Li, Seyran Khademi, and Jan van Gemert. Attention-aware age-agnostic visual place recognition. In

2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pages 1437–1446, 2019. 1, 2

[57] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2020. 1

[58] Weinzaepfel, Philippe and Lucas, Thomas and Larlus, Diane and Kalantidis, Yannis. Learning Super-Features for Image Retrieval. In *ICLR*, 2022. 3

[59] Kyle Wilson and Noah Snavely. Robust global translations with 1dsfm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 3

[60] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuetong Xue, Fu Li, Errui Ding, and Jizhou Huang. Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, pages 11772–11781, 2021. 3

[61] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 467–483. Springer, 2016. 2

[62] B. Yildiz, S. Khademi, R. Siebes, and J. 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, Los Alamitos, CA, USA, aug 2022. IEEE Computer Society. 1

[63] 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. 1, 2, 4

[64] Xiwu Zhang, Lei Wang, and Yan Su. Visual place recognition: A survey from deep learning perspective. *Pattern Recognition*, 113:107760, 2021. 1

[65] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1, 5

[66] Yingying Zhu, Jiong Wang, Lingxi Xie, and Liang Zheng. Attention-based pyramid aggregation network for visual place recognition. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 99–107. ACM, 2018. 2