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 if its ground truth position is less than 25 meters away from 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].

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-
retrieval 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 benchmark 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 methods offer 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 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
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 query 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.

### 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-

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

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.


4.2. 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], Ren Ranking Transformers (RRT) [50] (which uses DELG local features), Patch-Net-VDLAD [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-Net-VDLAD 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-Net-VDLAD applies RANSAC multiple times to match multiple scales (for the multi-scale approach). Patch-Net-VDLAD 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-Net-VDLAD gave poor results, and we found it to be sensible to outliers and reliable only for moderate changes in viewpoint. For Patch-Net-VDLAD 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 present the Recall@100 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.

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.

Table 4. Characteristics of local features extractors. The descriptors size was computed for all methods on the same image of resolution 480x640. For Patch-Net-VDLAD 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.

Table 5. 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], Ren Ranking Transformers (RRT) [50] (which uses DELG local features), Patch-Net-VDLAD [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-Net-VDLAD 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-Net-VDLAD applies RANSAC multiple times to match multiple scales (for the multi-scale approach). Patch-Net-VDLAD 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-Net-VDLAD gave poor results, and we found it to be sensible to outliers and reliable only for moderate changes in viewpoint. For Patch-Net-VDLAD 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 present the Recall@100 (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).

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).
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 \( K \in \{1, 2, 3, \ldots, 100\} \). 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.

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-rank the top-100 candidates for a single query, considering online extraction of local features 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


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


