To Match or Not to Match: Revisiting Image Matching for Reliable Visual Place Recognition

To Match or Not to Match: Revisiting Image Matching for Reliable Visual Place Recognition

Davide Sferrazza[1,2] Gabriele Berton[2] Gabriele Trivigno[2] Carlo Masone[2,3]
[1,2] Politecnico di Torino [3] Focoos AI
[1] {name.surname}@studenti.polito.it [2] {name.surname}@polito.it [3] {name.surname}@focoos.ai

Abstract

Visual Place Recognition (VPR) is a critical task in computer vision, traditionally enhanced by re-ranking retrieval results with image matching. However, recent advancements in VPR methods have significantly improved performance, challenging the necessity of re-ranking. In this work, we show that modern retrieval systems often reach a point where re-ranking can degrade results, as current VPR datasets are largely saturated. We propose using image matching as a verification step to assess retrieval confidence, demonstrating that inlier counts can reliably predict when re-ranking is beneficial. Our findings shift the paradigm of retrieval pipelines, offering insights for more robust and adaptive VPR systems.

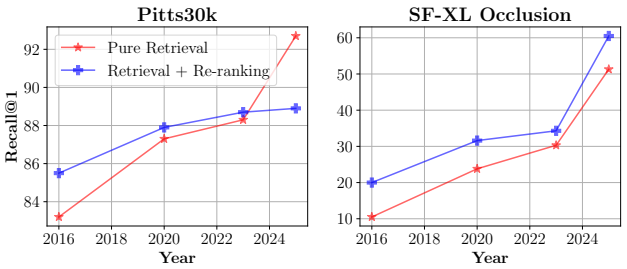


Figure 1. **Re-ranking with SuperGlue with VPR methods from different years** (NetVLAD [5], SFRS [29], EigenPlaces [13], MegaLoc [9]). In the past, re-ranking the top-K VPR results with powerful image matching methods was guaranteed to improve results. With modern VPR models, this is now true only for certain datasets or types of images. This paper explores this phenomenon, aiming to determine whether re-ranking can be adaptively and confidently triggered for individual queries during deployment.

1. Introduction

Visual Place Recognition (VPR) addresses the fundamental question: "Where was this picture taken?". VPR is typically framed as an image retrieval problem, where a query image is localized by comparing it to a database of geotagged images [1, 2, 5, 11, 29, 40, 43, 89, 95]. and it serves as a critical first step in applications such as Structure-from-Motion (SfM) [44, 46, 65], simultaneous localization and mapping (SLAM) [21, 34, 63] and Visual Localization [44, 64, 76, 78]. To address this task in large-scale environments, a comprehensive database is required, which is often composed of daytime Street View images [10, 11, 75, 76]. However, real-world queries may exhibit significant appearance variations due to nighttime conditions, occlusions, or adverse weather. This domain shift between queries and database images remains a major obstacle in VPR research [4, 10, 36, 76, 85, 86, 88, 92]. Hence, a common strategy to improve performance in VPR systems is to adopt a post-processing step to refine retrieval predictions [7, 34, 83]; the underlying idea being that one can apply a more computationally-intensive method on a shortlist of candidate to filter out outliers, which would be too expensive and time-consuming to apply to the entire database. Given the large corpus of literature on re-ranking [19, 34, 41, 55, 83] and image matching [24, 62, 63, 70], this two-step pipeline established itself as the de-facto standard to refine retrieval predictions. As local features are inherently more robust to domain shifts, occlusions and perspective changes, it has been repeatedly shown that this strategy can lead to large improvements in results [7, 34, 83].

Recent advances in VPR literature, such as the introduction of methods based on DINOv2 [56], combined with task-specific aggregations and mining techniques [9, 37, 38] achieved unprecedented results, showing remarkable generalization capabilities. In this work, we propose a *reality check* on the performance of modern VPR and image matching methods, showing that recent advancements have caused a paradigm shift in the typical retrieval+re-ranking pipeline. Specifically, we show in Fig. 1 that (i) modern retrieval methods have reached the point where applying re-ranking can, surprisingly, worsen performance in some cases; and that (ii) current VPR datasets are largely saturated by the current state-of-the-art. The main takeaway from the preliminary experiments in Fig. 1 is that apply-

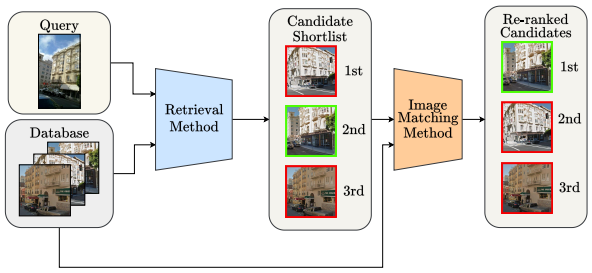


Figure 2. **Re-ranking pipeline.** The standard re-ranking pipeline consists of first retrieving a shortlist of candidates using a retrieval method, followed by sorting these candidates in descending order based on the number of inliers computed using an image matching method.

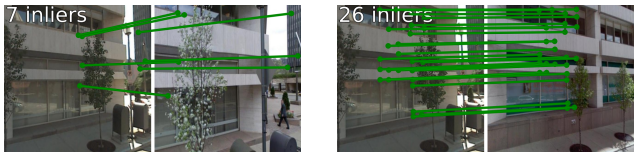
ing a re-ranking step (*cf.* Fig. 2) is not always beneficial. This raises the question of whether an automated approach can discern when retrieval predictions already possess sufficient confidence, preventing potentially detrimental post-processing.

In this work we demonstrate that the number of inliers can serve as a proxy of prediction uncertainty, in turn providing an indication of whether a re-ranking step can improve retrieval performance or not. In essence, we argue that image matching methods should be employed first as a *verification* step, to assess the confidence of the retrieval predictions, and only afterwards, it should be selectively used as a *postprocessing* step for the uncertain estimates. This finding derives from a comprehensive evaluation of image matching methods in both roles, which sets this study apart from prior works that focus solely on re-ranking performance [7, 34, 83].

Our contributions are as follows:

- We conduct an extensive evaluation of state-of-the-art image matching methods for re-ranking in VPR, obtaining the most comprehensive benchmark up-to-date both in terms of methods and datasets;
- Drawing from our comprehensive experimental results, we demonstrate the inadequacy of existing benchmarks in keeping up with the pace of research, showing that most of them are largely saturated, and provide insights on remaining challenges for future works;
- We show that, contrary to common belief, in many cases re-ranking can worsen retrieval performance (see Fig. 3), and propose an approach to quantify prediction uncertainty in VPR using image matching methods, demonstrating that inlier counts provide a reliable measure of confidence for retrieval predictions.

By providing a perspective shift on modern retrieval pipelines, our work advances the state of the art in VPR and provides a foundation for future research in leveraging image matching methods for robust and reliable place recognition.



(a) A query and its positive (b) A query and its negative

Figure 3. **Example of a case when re-ranking through image matching fails.** The top-1 retrieved is shown next to the query on the left, and it's a positive. On the right, the top-2 retrieved image, which is a negative. SuperGlue + RANSAC finds fewer points in common between the pair on the left (only 7 inliers), and more between the wrong pair (26 inliers).

2. Related Work

Visual Place Recognition (VPR) is typically addressed as an image retrieval problem, leveraging a database of geo-tagged images [12, 51, 66]. After the pioneering work of NetVLAD [5], learned representations derived from deep networks became the de-facto standard; initially derived from CNNs [5, 6, 66, 77] and, subsequently, transformer-based architectures [48, 49, 95]. A key challenge addressed by these methods is the generation of compact, yet highly discriminative, global feature descriptors. Techniques for achieving this include various pooling strategies [5, 32, 52, 54, 60, 61, 74], clustering-based feature aggregation [37, 40, 57, 91], MLP-based aggregations [2], and the adoption of a set of learnable tokens [3]. With the growing availability of geo-tagged images, modern research on VPR has moved towards efficient training protocols with stricter supervision, by leveraging curated datasets [1], co-visibility constraints [43] and class-based partitions of the database [11, 13]. A recent breakthrough in VPR [39, 48] has been the adoption of vision foundation models such as DINOv2 [56]. Combining these training techniques and foundation models with optimal transport aggregation [37] and novel mining techniques [9, 38] has led to methods with exceptional generalization capabilities. We show that these recent advances allow to consider many long-standing VPR benchmarks as solved.

Keypoint Detection and Description Finding repeatable keypoints (and associated descriptors) in an image is a long-standing problem of computer vision. Early handcrafted approaches adopted a detect-then-describe approach, typically based on local derivatives of the image [8, 47, 53]. With the advent of deep learning, learning-based approaches gained popularity. Pioneering works [68, 73] employed contrastive learning to learn local descriptors with Convolutional Networks. SuperPoint [23] proposed to generate synthetic shapes to train a neural network via self-supervision. Subsequent works introduced a joint detect-and-describe paradigm, in which keypoints are implicitly defined as local maxima of the extracted features [24, 30, 62, 80, 93]. More

recently, DeDoDe [26] proposes to separately optimize detection and description, in order to improve repeatability by enforcing 3D consistency constraints. Its follow-up Steerers [17] introduces rotation invariant descriptors, enabling several space and medical applications [14, 58, 69].

Image Matching aims at establishing pixel correspondences between different views of a scene. Traditionally, matches were established by finding mutual nearest neighbor on local keypoint descriptors [62]. This strategy can lead to errors as it does not allow reasoning on the global image context. A possible solution is to geometrically verify matches with RANSAC [28], or to employ a learnable matcher such as SuperGlue [63], a graph neural network-based approach. While SuperGlue operates a-posteriori of the matching stage, LoFTR[70] foregoes the detection stage and proposes a detector-free paradigm where image context is incorporated thanks to the global attention mechanism of transformer architectures, thus improving robustness to repetitive patterns and low-texture areas. Following LoFTR, other methods adopted a detector-free paradigm [15, 20, 35, 72, 82, 94]. Alternatively, methods for dense feature matching aim to estimate every matchable pixel pair to obtain a dense warping of the two images [25, 27]. All these methods cast the matching problem in 2D, *i.e.* without explicitly accounting for the geometrical properties of the scene. Recently, Dust3r [84] and its follow-up Mast3r [42], propose to ground matches in 3D, by solving the task of 3D reconstruction from uncalibrated images, and then recovering point correspondences. We conduct a comprehensive benchmark spanning a wide variety of image matching methods applied to re-ranking in VPR, identifying those most suited for the task. We introduce a methodology to automatically assess their potential to enhance retrieval accuracy, and propose a framework to quantify the uncertainty inherent in retrieval through image matching.

Uncertainty Estimation In VPR, naive uncertainty estimation could be obtained directly from the image retrieval model, through the L2-distance in feature space between the query and its nearest neighbors. To improve upon this simple baseline, several techniques have been proposed to explicitly model uncertainty. Examples include STUN [18] and BTL [87], which predict aleatoric uncertainty based only on the query's image content, and SUE [90], which leverages the geographical distribution of the retrieved shortlist of candidates.

3. Datasets

To provide a comprehensive evaluation of the performance of image matching methods for uncertainty estimation and re-ranking in Visual Place Recognition, we use 10 datasets that span a broad spectrum of real-world scenarios, including outdoor and indoor environments, viewpoint variations, seasonal or weather changes, occlusions and day-to-night appearance shifts.

Dataset	# Queries	# Database Images	Scenery	Domain Shift
Baidu	3k	5k	Indoor	Viewpoint Shift/Occlusions
MSLS Val	11k	19k	Urban	Day-Night
Pitts30k	7k	10k	Urban	None
SF-XL Night	466	2.8M	Urban	Day-Night
SF-XL Occlusion	76	2.8M	Urban	Occlusions
SF-XL test V1	1000	2.8M	Urban	Viewpoint / Night
SF-XL test V2	598	2.8M	Urban	Viewpoint
SVOX Night	823	17k	Urban	Day-Night
SVOX Sun	854	17k	Urban	Weather
Tokyo 24/7	315	76k	Urban	Day-Night

Table 1. **Datasets.** For each dataset, the number of queries and database images, scenery and types of domain shift in the test set is provided, except for MSLS, where the validation set is used instead.

For **indoor environments**, we use the Baidu [71] test set, which contains images captured in a mall with varying camera poses. This dataset features challenges such as perceptually aliased structures and distractors (e.g., people), making it ideal for VPR evaluation.

In the domain of **medium-scale** (10k-100k database size) **urban VPR**, we employ three widely used datasets: the validation set of MSLS [86], and the test sets of Pitts30k [5] and Tokyo 24/7 [76]. MSLS consists of over 1 million images from multiple cities and the ones from San Francisco and Copenaghen are used as validation set. Pitts30k, built from Google StreetView images of Pittsburgh, includes 6,816 test queries from different years and is often used as a benchmark in VPR literature. Tokyo 24/7 presents a set of 315 queries from smartphone photos taken in central Tokyo at day, sunset, and night. Thus, it is suitable to assess performance under *varying lighting conditions*.

For **large-scale urban VPR**, we use the San Francisco eXtra Large (SF-XL) [11] dataset, which contains over 41 million images. The SF-XL test set includes 2.8 million images with multiple query sets. The official test sets, V1 and V2, assess *viewpoint changes* and *domain shifts*, with images sourced from Flickr and smartphones, respectively. Additionally, the SF-XL Night and SF-XL Occlusion [7] queries introduce further challenges, with *night-time* imagery and images featuring heavy *occlusions* like cars and pedestrians.

Lastly, to evaluate VPR in **diverse weather conditions**, we use the SVOX [10] dataset, which provides a robust test set for cross-domain VPR. The dataset spans Oxford, UK, using Google StreetView for the database and the Oxford RobotCar [50] dataset for queries. Here, we select the queries from the Sun and Night subsets.

Table 1 provides a summary of all the datasets, showing the number of queries and images in the database, along with the types of scenery and domain shifts.

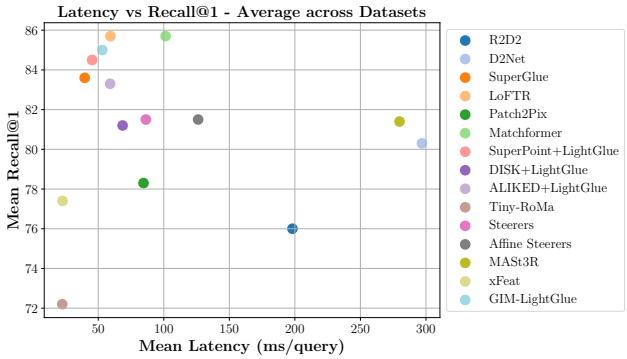


Figure 4. **Plot displaying the mean Recall@1 after re-ranking and mean latency for different methods.** The mean Recall@1 is computed over the datasets, while the mean latency is the average time to process each query over all datasets. The shortlist of candidates for the Recall@1 is obtained with MegaLoc and distance threshold fixed at 25 meters.

4. Experiments

In this work we aim to understand if and when, given the current state of Visual Place Recognition (retrieval) models, image matching methods for re-ranking are still relevant and useful. In the next section we compute and showcase results on such task.

4.1. Re-ranking

Implementation Details: The count of inliers (*i.e.* matches that "survive" the post-processing with RANSAC) can be leveraged to re-rank the candidate shortlist obtained through retrieval methods, thereby enhancing the Recall@K metric.

We use the state-of-the-art MegaLoc [9] as the retrieval model across all datasets. Unless explicitly stated, we follow the standard practices in Visual Place Recognition, considering an image I retrieved from the database a correct match for the query q if and only if their locations are at most 25 meters apart. Formally, the prediction provided by I is correct if $d_g(q, I) \leq 25$, where d_g denotes the geographic distance expressed in meters. Each input image is resized to 322×322 pixels before being processed by MegaLoc. We compare the various image matching methods by using their default hyper-parameters and resizing each image to 512×512 pixels.

To evaluate the re-ranking performance of the image matching methods, the top 100 nearest neighbors for each query are initially retrieved from the database using MegaLoc. The re-ranking process then sorts these 100 candidate images based on the number of inliers $i_q^{(j)}$ between the query q and the j-th nearest neighbor, for $j = 1, 2, \ldots, 100$, in descending order.

Image matching methods: For this analysis, we selected a substantial number of open-source image matching models[1]: R2D2 [62], D2Net [24], SuperGlue [63], LoFTR [70], Patch2Pix [94], Matchformer [82], SuperPoint+LightGlue [23, 45], DISK+LightGlue [45, 80], ALIKED+LightGlue [45, 93], RoMa and Tiny-RoMa [27], Steerers [17], Affine Steerers [16], DUSt3R [84], MASt3R [42], xFeat [59], GIM-DKMv3 [25, 67] and GIM-LightGlue [45, 67].

Baseline: The baseline is represented by the pure retrieval performance of MegaLoc [9]. MegaLoc is trained on a dataset made up of train sets from SF-XL [11], GSV-Cities [1], MSLS [86], MegaScenes [79] and ScanNet [22].

Evaluation Metric: The evaluation is conducted using Recall@K at a fixed distance threshold τ. Recall@K measures the percentage of queries for which at least one of the top-K retrieved images is within τ meters of the query's ground-truth location. Unless otherwise specified, experiments are carried out with $\tau = 25$. A higher value of Recall@K relative to MegaLoc's performance indicates better re-ranking capability of the image matching method.

Results: Table 2 presents the Recall@1 and Recall@10 values after the re-ranking process, along with MegaLoc's performance for each dataset. Since re-ranking is applied to the top 100 retrieved images, the Recall@100 (shown in the table's header) represents the upper bound on performance that can be achieved through re-ranking.

An intriguing observation from our results is that, contrary to previous findings in the literature [7, 31, 34, 48, 81, 83], applying re-ranking does not universally enhance performance: we believe this to be due to recent improvements in the VPR literature (*e.g.* our retrieval baseline MegaLoc), which provide good results that are hard to improve upon by means of re-ranking. Specifically, it is the case of Pitts30k, MSLS, SVOX and SF-XL, where even the best re-ranking methods cause a drop in R@1. At the same time, MegaLoc essentially saturates these long-standing benchmarks through pure retrieval alone. This finding challenges the widely held belief that re-ranking consistently refines initial matches, and motivates us to further investigate when image matching methods can prove beneficial, rather than assuming their unequivocal benefit.

For the datasets that present several occlusions in the query sets, namely Baidu and SF-XL Occlusion, image matching methods–on average–are able to improve the Recall@1 by 3.3% and 0.5%, respectively. However, when considering the average performance across all datasets, only two methods—LoFTR and Matchformer—improve Recall@1, while three methods—SuperPoint+LightGlue, SuperGlue, and MASt3R—improve Recall@10. No single method enhances both metrics. This suggests that image

[1]methods available in the Image Matching Models GitHub repository [14] at https://github.com/alexstoken/image-matching-models

Method	Baidu		MSLS Val		Pitts30k		SF-XL Night		SF-XL Occlusion		SF-XL test V1		SF-XL test V2		SVOX Night		SVOX Sun		Tokyo 24/7		Average	
	R@100 = 99.9		R@100 = 97.6		R@100 = 99.6		R@100 = 85.0		R@100 = 92.1		R@100 = 99.0		R@100 = 99.0		R@100 = 99.6		R@100 = 99.9		R@100 = 99.7		R@100 = 97.1	
	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
-	87.7	98.0	**91.0**	**95.8**	**94.1**	**98.2**	52.8	73.8	51.3	75.0	**95.3**	98.0	94.8	98.5	**95.1**	98.8	**96.5**	99.6	96.5	99.4	85.5	93.5
R2D2 (NeurIPS '19)	90.4	98.5	75.8	88.3	83.8	96.6	34.8	59.2	46.1	67.1	86.6	94.1	92.0	97.8	72.7	84.3	87.5	93.2	89.8	96.8	76.0	87.6
D2Net (CVPR '19)	91.4	98.8	78.1	91.2	86.2	97.2	50.6	71.5	52.6	75.0	90.7	96.8	93.1	98.2	78.4	94.0	88.5	96.7	93.7	97.8	80.3	91.7
SuperGlue (CVPR '20)	91.9	99.1	85.4	94.3	85.5	97.2	60.7	75.5	55.3	81.6	93.4	98.2	90.8	98.3	85.7	97.1	91.3	98.1	95.9	99.4	83.6	93.9
LoFTR (CVPR '21)	**94.8**	98.8	85.4	93.9	87.2	96.3	58.2	75.8	**60.5**	75.0	93.7	97.1	94.1	98.3	91.0	98.2	95.9	99.2	96.5	**99.7**	**85.7**	93.2
Patch2Pix (CVPR '21)	91.1	99.0	79.7	91.1	84.9	96.4	38.0	61.2	52.6	68.4	89.9	96.3	92.5	98.0	67.3	89.1	92.3	98.0	94.9	98.7	78.3	89.6
Matchformer (ACCV '22)	93.7	98.8	83.6	91.9	88.2	97.1	**61.8**	75.8	55.3	71.1	94.0	97.4	**95.5**	98.3	92.6	98.4	95.7	99.2	96.8	99.0	**85.7**	92.7
SuperPoint+LightGlue (ICCV '23)	92.6	**99.3**	83.9	93.5	85.9	97.3	58.8	**76.4**	56.6	**84.2**	94.3	98.2	91.6	98.2	89.4	96.8	93.8	99.3	**98.4**	99.0	84.5	**94.2**
DISK+LightGlue (ICCV '23)	91.4	99.1	84.6	93.6	84.6	96.7	50.2	71.9	53.9	81.6	90.7	97.7	91.5	**98.7**	80.8	89.8	90.7	97.1	94.0	**99.7**	81.2	92.6
ALIKED+LightGlue (ICCV '23)	94.3	99.2	87.5	94.4	85.2	97.4	56.2	73.8	56.6	77.6	92.8	**98.4**	91.5	98.2	83.0	92.0	88.6	96.1	97.5	99.4	83.3	92.7
RoMa (CVPR '24)	88.7	98.3	45.9	88.2	72.8	95.9	44.6	74.9	36.8	77.6	88.0	96.9	88.5	98.0	65.1	98.3	71.4	93.3	84.1	**99.7**	68.6	92.1
Tiny-RoMa (CVPR '24)	88.7	98.3	76.2	91.6	81.8	96.7	41.2	69.1	50.0	71.1	88.0	96.9	88.5	98.0	48.8	85.7	71.4	93.3	87.3	98.4	72.2	89.9
Steerers (CVPR '24)	93.1	98.7	77.0	87.7	85.1	96.8	48.5	67.2	53.9	76.3	92.0	97.6	91.5	98.2	82.7	95.3	93.3	98.1	97.5	**99.7**	81.5	91.6
Affine Steerers (ECCV '24)	91.3	98.0	79.8	90.9	85.1	96.8	50.4	70.2	53.9	76.3	91.3	97.5	91.5	98.2	82.7	95.3	92.5	97.8	96.2	98.4	81.5	91.9
DUSt3R (CVPR '24)	85.0	98.2	63.0	80.4	79.5	93.5	35.6	56.2	42.1	63.2	78.6	92.7	66.1	93.3	60.3	69.9	71.7	82.2	86.0	93.7	66.8	82.3
MASt3R (ECCV '24)	89.8	99.1	71.7	93.0	85.9	98.0	56.2	74.2	56.6	80.3	90.4	98.2	83.4	98.0	92.6	**99.1**	93.4	**99.8**	94.0	**99.7**	81.4	93.9
xFeat (CVPR '24)	86.8	97.8	83.0	92.1	86.6	96.9	45.3	67.6	44.7	75.0	88.7	96.5	91.1	98.3	75.2	90.6	83.5	95.2	88.9	98.4	77.4	90.8
GIM-DKMv3 (ICLR '24)	41.6	94.8	4.4	35.0	40.1	91.4	31.1	71.7	26.3	73.7	33.8	88.2	46.0	94.8	29.6	89.6	21.8	86.1	47.9	97.5	32.3	82.3
GIM-LightGlue (ICLR '24)	92.4	98.8	86.4	94.5	88.8	97.6	59.7	74.0	53.9	77.6	94.4	98.2	92.1	98.3	91.0	96.5	94.7	98.4	96.5	99.0	85.0	93.3

Table 2. **Recalls before and after applying re-ranking.** Recalls are computed by setting the distance threshold to 25 meters. The shortlist of candidates to be re-ranked is obtained with MegaLoc, 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.

matching methods are particularly beneficial when the retrieval model performs poorly and struggles to accurately map retrieved images in its output space. A prime example is SF-XL Night, where Matchformer improves Recall@1 by 9% and Recall@10 by 3%, and seven methods, in total, assist in re-ranking the candidate shortlist for both recall values.

Figure 4 offers a comparison of the analyzed methods, showing the average Recall@1 alongside the time taken to process a single query (*i.e.* re-ranking its top-100 predictions). This includes extracting keypoints for both the query and database image, and matching inliers. Ideally, methods that are both accurate and time-efficient are best suited for real-time re-ranking applications.

4.2. Prediction Uncertainty via Image Matching

Results from the previous section show that re-ranking can prove detrimental for performances in cases where the retrieval R@1 is near 100%. However, in real-world scenarios, there is no such concept as a saturated dataset, as queries are fed to the system individually, and can potentially come from different data distributions. Therefore, it is important to estimate which queries can be solved by retrieval alone, and which ones can benefit from re-ranking. To this end, we posit that, given a reasonable estimation of uncertainty, we can find a correlation between uncertainty and potential improvements attainable via re-ranking. In the rest of this subsection, we aim to validate our hypothesis. Namely, to understand whether a reliable uncertainty value exists (*i.e.* the probability that a given query has been wrongly localized), and whether such value is effectively correlated with the impact that re-ranking has on a given query. Simply put, we aim to verify that, in order to maximize performance, in a real-world application we could apply re-ranking only for high-uncertainty queries, whereas high-confidence predictions can be left untouched in order not to jeopardize positive results.

Baselines: The topic of uncertainty estimation has been studied for image retrieval, either by directly learning to predict aleatoric uncertainty at training time [87], or through post-hoc techniques at inference [90]. Among the latter, a simple technique entails using the L2-distance to the nearest neighbor for each query, $u_q \triangleq d_{(1)}$, and the perceptual aliasing score (PA-score), *i.e.* the ratio of the distances between the first and second nearest neighbors in the database, *i.e.*, $u_q \triangleq \frac{d_{(1)}}{d_{(2)}}$. Additionally, we include SUE [90], the state-of-the-art method for uncertainty estimation in the VPR task, which considers the geographic spread of the shortlist of candidates retrieved by MegaLoc. We further introduce a Random baseline, where uncertainty scores are sampled from a uniform distribution, $u_q \sim \mathcal{U}(0, 1)$.

Besides these methods that focus purely on uncertainty estimation for VPR, image matching models have been shown to provide good results for the task [90]: intuitively, when a retrieved prediction has few matches with the given query, the uncertainty will be high, whereas in the presence of numerous matches between two images we can confidently state that the two represent the same place.

Image matching for uncertainty estimation: To quantify the uncertainty associated with each method, we measure the number of inliers $i_q^{(1)}$ between the query q and the nearest neighbor $I_{(1)}$ (with corresponding L2-distance in the output space of MegaLoc indicated as $d_{(1)}$). The uncertainty is then defined as $u_q \triangleq -i_q^{(1)}$, with fewer inliers indicating greater uncertainty.

Evaluation Metrics: We adopt the evaluation framework of previous uncertainty for VPR papers [90] across all datasets. The evaluation metric is the Area Under the Precision-Recall Curve (AUPRC), where a higher value in-

Method	Baidu	MSLS Val	Pitts30k	SF-XL Night	SF-XL Occlusion	SF-XL test V1	SF-XL test V2	SVOX Night	SVOX Sun	Tokyo 24/7	Average
L2-distance	94.0	97.0	**99.1**	69.8	77.5	99.5	98.0	99.2	99.1	**99.9**	93.3
PA-Score	93.8	96.5	98.9	67.3	71.6	98.6	98.0	99.0	98.9	99.8	92.2
SUE	95.5	97.1	98.6	73.6	73.5	99.1	**98.2**	99.6	99.0	**99.9**	93.4
Random	88.0	90.8	94.3	53.2	45.9	94.7	96.0	94.8	97.6	96.9	85.2
R2D2 (NeurIPS '19)	96.5	96.4	98.4	69.7	80.5	99.5	97.8	99.1	99.6	99.7	93.7
D2Net (CVPR '19)	97.3	96.2	98.4	73.2	78.2	**99.7**	97.7	99.2	99.4	99.8	93.9
SuperGlue (CVPR '20)	**97.4**	97.2	98.8	75.5	84.2	**99.7**	97.7	99.5	99.4	**99.9**	**94.9**
LoFTR (CVPR '21)	97.3	97.0	98.8	73.9	**84.5**	**99.7**	98.0	99.5	99.7	**99.9**	94.8
Patch2Pix (CVPR '21)	96.6	96.7	98.6	73.5	78.0	**99.7**	97.9	99.1	99.4	**99.9**	93.9
Matchformer (ACCV '22)	97.2	96.9	98.9	74.9	83.1	**99.7**	98.1	99.6	99.6	**99.9**	94.8
SuperPoint+LightGlue (ICCV '23)	97.3	97.4	98.9	75.8	82.4	99.6	97.7	99.5	99.5	**99.9**	94.8
DISK+LightGlue (ICCV '23)	95.7	97.0	98.9	70.8	80.9	99.3	96.9	99.2	99.3	99.8	93.8
ALIKED+LightGlue (ICCV '23)	96.9	**97.6**	99.0	70.6	79.6	99.6	97.8	99.5	99.5	**99.9**	94.0
RoMa (CVPR '24)	94.8	96.1	96.7	62.1	72.5	99.3	94.8	99.4	99.7	99.8	91.5
Tiny-RoMa (CVPR '24)	97.1	96.5	99.0	69.4	80.7	99.6	97.0	98.6	99.4	99.8	93.7
Steerers (CVPR '24)	96.8	96.8	99.0	72.5	79.0	99.4	96.8	99.4	99.3	99.8	93.9
Affine Steerers (ECCV '24)	96.5	96.8	98.5	69.7	81.2	99.3	97.1	99.3	99.3	**99.9**	93.8
DUSt3R (CVPR '24)	95.3	97.0	98.7	63.8	59.1	98.0	94.6	98.6	98.7	99.2	90.3
MASt3R (ECCV '24)	95.7	96.8	**99.1**	67.4	79.2	99.6	96.1	**99.8**	**99.8**	**99.9**	93.3
xFeat (CVPR '24)	95.7	96.8	98.6	74.1	79.3	99.6	97.1	98.7	99.3	99.7	93.9
GIM-DKMv3 (ICLR '24)	92.6	92.4	94.5	62.2	63.4	96.6	93.8	97.9	98.7	99.6	89.2
GIM-LightGlue (ICLR '24)	97.1	97.3	98.9	**76.5**	80.4	99.6	98.0	99.6	99.5	**99.9**	94.7

Table 3. **The AUPRC of all the baselines and image matching methods**, split according to group type. The shorlist of candidates is obtained with MegaLoc. Distance threshold is fixed at 25 meters. Best overall results on each dataset are in **bold**, best results for each group are underlined.

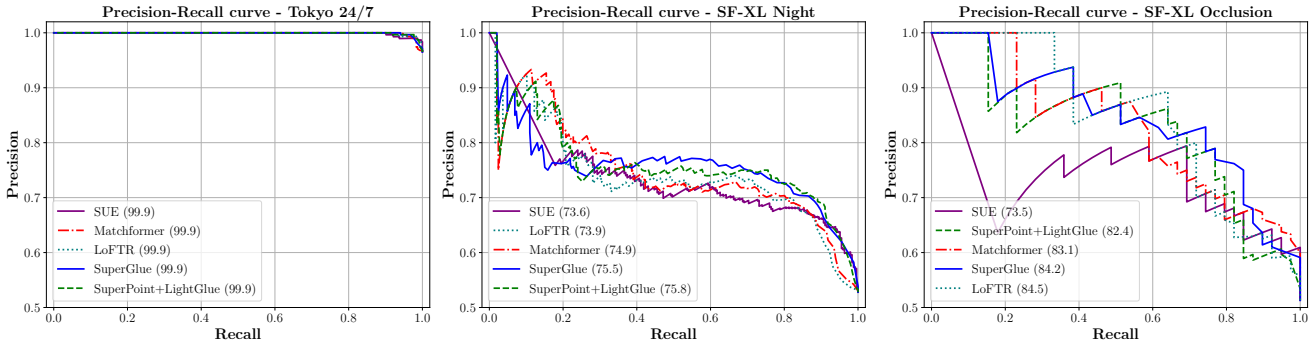


Figure 5. **Precision-Recall curves**, computed for the top-4 image matching methods on Tokyo 24/7, SF-XL Night, and SF-XL Occlusion, together with SUE, which is representative of the baselines when the shortlist of candidates is obtained with MegaLoc. Distance threshold is fixed at 25 meters.

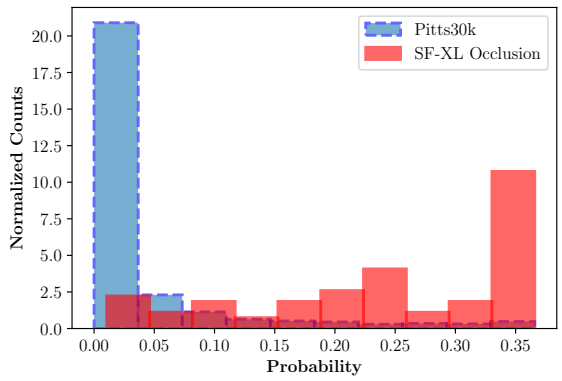


Figure 6. **Histogram of probabilities of being a wrongly localized query**, *i.e.* with a top-1 prediction further than 25 meters. The probabilities are computed on the query sets of Pitts30k and SF-XL Occlusion using a Logistic Regression trained on the uncertainty scores produced by MASt3R on MSLS Val.

dicates better discrimination between correct and incorrect queries based on uncertainty scores.

Results: We articulate our analysis on the relationship between prediction uncertainty, and re-ranking performance through Fig. 6, Tab. 3, and Tab. 2. In Tab. 3 we present the results of uncertainty estimation, across multiple datasets, of existing baselines for uncertainty estimation applied directly on MegaLoc predictions, and several matching methods, for which we use the number of inliers as a confidence score. In Fig. 6, we train a Logistic Regressor to predict the probability of a query being a correct match based on the number of inliers on the top-1 prediction. We train the Logistic Regressor on MASt3R inliers counts on MSLS val, and plot the resulting histogram of probabilities for Pitts30k and SF-XL Occlusion.

From this data, we draw the following conclusions:

Method	Baidu R@100 = 100.0		MSLS Val R@100 = 98.4		Pitts30k R@100 = 100.0		SF-XL Night R@100 = 92.3		SF-XL Occlusion R@100 = 98.7		SF-XL test V1 R@100 = 99.2		SF-XL test V2 R@100 = 99.5		SVOX Night R@100 = 99.8		SVOX Sun R@100 = 99.9		Tokyo 24/7 R@100 = 100.0		Average R@100 = 98.8	
	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
-	94.9	99.8	**95.6**	**97.6**	**98.7**	**99.9**	74.2	84.3	72.4	86.8	**96.4**	98.2	98.7	**99.5**	97.7	99.4	98.7	**99.8**	97.5	99.7	92.5	96.5
R2D2 (NeurIPS '19)	95.9	99.9	80.9	91.8	95.5	99.7	52.1	76.2	64.5	85.5	89.0	95.7	97.5	99.3	76.7	91.9	90.0	97.2	91.1	98.4	83.3	93.6
D2Net (CVPR '19)	96.6	**100.0**	84.2	94.1	95.0	99.7	68.5	82.8	72.4	90.8	92.7	97.3	98.2	**99.5**	84.6	96.8	92.2	98.4	95.6	99.0	88.0	95.8
SuperGlue (CVPR '20)	97.1	**100.0**	91.9	96.6	96.8	99.8	**78.1**	86.5	80.3	**93.4**	95.9	98.6	97.8	**99.5**	91.4	98.5	94.7	99.2	96.8	99.4	92.1	**97.2**
LoFTR (CVPR '21)	97.7	**100.0**	91.0	96.4	96.7	99.8	75.3	85.4	80.3	90.8	95.0	97.8	98.3	**99.5**	95.5	98.9	97.7	99.3	97.5	**100.0**	92.5	96.8
Patch2Pix (CVPR '21)	96.3	99.9	84.8	93.9	96.1	99.7	54.7	76.0	75.0	90.8	92.0	96.9	97.0	99.2	70.8	93.9	94.8	98.9	95.6	98.7	85.7	94.8
Matchformer (ACCV '22)	97.4	**100.0**	88.7	94.9	97.4	99.8	**78.1**	85.6	78.9	88.2	95.9	98.1	**98.8**	**99.5**	94.8	99.3	97.7	99.4	97.1	99.7	92.5	96.5
SuperPoint+LightGlue (ICCV '23)	97.2	**100.0**	90.1	96.1	97.0	99.8	76.6	86.5	80.3	90.8	96.2	98.6	97.7	99.3	93.7	98.1	97.0	99.4	98.7	99.4	92.5	96.8
DISK+LightGlue (ICCV '23)	97.0	**100.0**	91.9	96.2	95.9	99.7	72.3	83.5	77.6	92.1	94.0	98.2	97.8	**99.5**	84.0	92.3	94.6	98.1	96.2	99.7	90.1	95.9
ALIKED+LightGlue (ICCV '23)	**98.1**	**100.0**	93.6	96.6	97.3	99.7	73.2	83.9	80.3	92.1	95.7	98.6	98.5	**99.5**	86.1	94.5	91.2	97.2	98.4	99.7	91.2	96.2
RoMa (CVPR '24)	95.4	99.8	58.4	92.5	96.1	99.8	64.2	**87.8**	67.1	90.8	91.8	97.7	97.8	**99.5**	83.8	99.1	78.0	97.5	96.5	99.7	82.9	96.4
Tiny-RoMa (CVPR '24)	95.4	99.8	83.8	94.7	95.9	99.7	60.5	82.0	65.8	89.5	91.8	97.7	97.8	**99.5**	58.8	93.4	78.0	97.5	90.5	99.0	81.8	95.3
Steerers (CVPR '24)	97.3	99.9	82.7	91.3	97.3	**99.9**	67.8	81.5	72.4	88.2	94.3	98.0	98.5	**99.5**	87.0	96.7	96.1	98.6	98.7	**100.0**	89.2	95.4
Affine Steerers (ECCV '24)	95.7	99.7	85.6	93.9	97.3	**99.9**	69.7	82.4	72.4	88.2	94.2	98.1	98.5	**99.5**	87.0	96.7	95.3	98.5	97.1	99.4	89.3	95.6
DUSt3R (CVPR '24)	94.2	**100.0**	68.8	86.0	94.4	99.5	54.9	71.9	67.1	84.2	88.4	94.9	97.2	99.3	67.7	80.3	84.2	91.3	92.1	96.8	80.9	90.4
MASt3R (ECCV '24)	96.4	**100.0**	81.5	96.0	98.4	**99.9**	75.5	85.6	80.3	**93.4**	95.9	98.6	97.8	**99.5**	**98.2**	**99.6**	**99.2**	**99.8**	**99.7**	**100.0**	92.3	**97.2**
xFeat (CVPR '24)	94.0	99.9	88.3	94.9	95.9	99.7	61.2	80.9	63.2	90.8	90.7	97.0	97.7	**99.5**	80.4	94.2	87.6	97.3	92.4	99.7	85.1	95.4
GIM-DKMv3 (ICLR '24)	75.6	99.7	9.1	48.6	87.0	99.8	58.2	83.3	53.9	89.5	46.6	92.2	73.9	98.8	65.2	97.1	60.5	98.9	76.5	99.7	60.6	90.8
GIM-LightGlue (ICLR '24)	97.9	**100.0**	92.0	96.6	97.3	**99.9**	76.0	84.8	**81.6**	89.5	**96.4**	**98.8**	98.0	**99.5**	93.6	97.8	97.0	99.4	98.4	99.7	**92.8**	96.6

Table 4. **Recalls before and after applying re-ranking, with a threshold of 100 meters.** The shortlist of candidates to be re-ranked is obtained with MegaLoc, 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.

- **Low Uncertainty translates in re-ranking being detrimental**. On datasets where MegaLoc achieves 95+% R@1, applying re-ranking worsens performances. In this scenario, there is little uncertainty on retrieval predictions (*cf.* Fig. 6);
- **High Uncertainty leaves room for improvement via re-ranking**: on Baidu, SF-XL Night and Occlusion, uncertainty is higher (*cf.* Fig. 6), and on these datasets re-ranking generally improves R@1. For instance, MASt3R provides a boost of respectively +2.1%, +5.4%, +5.3%;
- **Image Matching methods are better at estimating uncertainty** . On saturated datasets, even a *Random* uncertainty estimator achieves an AUPRC of over 90% (*cf.* Tab. 3). On the other hand, on the challenging SF-XL Night and Occlusion, using inlier count as a proxy of uncertainty is consistently better than existing baselines in terms of AUC (*cf.* Tab. 2).

Additional insights. In contrast with previous literature [90], we see that L2-distance (in MegaLoc's feature space) can be a fairly good estimator of uncertainty: we hypothesize this discrepancy to be due to MegaLoc being a more robust model w.r.t. VPR models analyzed in previous uncertainty estimation papers [18, 33, 87, 90].

Lastly, Fig. 5 shows the PR curves for Tokyo 24/7 and two of the most challenging datasets for MegaLoc, namely SF-XL Night and Occlusion. These curves illustrate how (i) retrieval alone achieves the ideal curve on saturated benchmarks such as Tokyo; and (ii) image matching methods significantly enhance uncertainty estimation, in scenarios in which the retrieval model actually struggles.

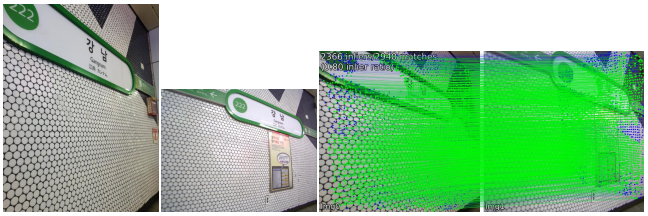
4.3. Additional Experiments

Effect of the distance threshold A potential question is whether our observation that re-ranking can degrade performance is solely due to the 25-meter threshold. It's possible that image matching methods have been trained to recognize broader views of the same location, potentially placing images slightly beyond the 25-meter threshold among the top predictions after re-ranking. To investigate this, we recompute the results using a 100-meter threshold to determine if this is indeed the case. The results on re-ranking for VPR with τ set to 100 meters are presented in Tab. 4. The table shows that our findings are indeed robust to the choice of τ, as it can be seen that re-ranking can have a negative impact on results on 4 datasets.

4.3.1. Failure cases with largest number of inliers

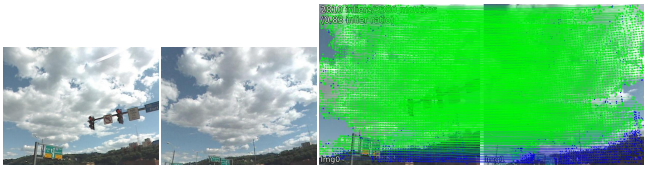
In this section, we provide a deeper analysis of the types of failures encountered by image matching methods in uncertainty estimation. We focus on MASt3R as the representative method, given its consistency across all experiments and datasets and its recent development. Additional examples for the remaining datasets are reported in the Supplementary Material. For each dataset, we identify the query with the highest number of inliers, but where the nearest neighbor retrieved by MegaLoc is located more than 25 meters away from the query's ground-truth location, effectively leading to a wrong (but confident) prediction. These cases are illustrated in Fig. 7, which shows the queries, the top-1 retrieved database images, the matched inliers, and the distance in meters between each pair of images. Based on this, we categorize the failures into two distinct types, as described below.

Noisy GPS Labels / Perceptual Aliasing. The first category of failure arises from pairs of images that look to be from the same place but, according to the GPS labels, are from different places. This is clear in the examples from MSLS, Baidu, SF-XL occlusion. While some of these cases might be due to perceptual aliasing, *i.e.* the phenomenon for which two different places look almost identical (which can happen especially in indoor places, like in the case

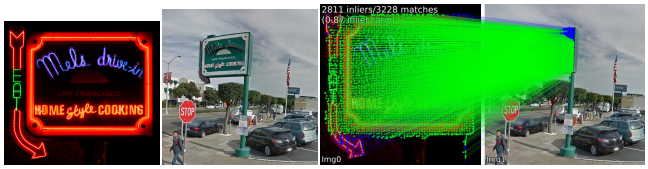


(a) Baidu – inliers: 2366, distance: 119m

(b) MSLS Val – inliers: 2485, distance: 182m



(c) Pitts30k – inliers: 2810, distance: 60m



(d) SF-XL Night – inliers: 2811, distance: 29m

(e) SF-XL Occlusion – inliers: 459, distance: 34m

(f) SF-XL test V1 – inliers: 730, distance: 26m

(g) SVOX Night – inliers: 452, distance: 27m

(h) Tokyo 24/7 – inliers: 345, distance: 35m

Figure 7. **Wrong queries with the largest number of inliers.** For each dataset, the displayed images show the query and a confidently retrieved negative image. The dataset name is accompanied by the number of inliers and the distance (in meters) between the two images, as indicated by their labels.

of Baidu), we believe that in many cases this is due to noisy GPS coordinates: it should be noted that GPS labels, even when post-processed (*e.g*. Mapillary famously post-processes images' GPS with SfM [86]), can be wrong. As examples, it can be noted the pairs of images from MSLS, which, although the distance according to the GPS is over 50 meters, it is likely that the two photos have been taken from a much smaller distance.

Large Distance and Viewpoint Variation. The second category consists of failure cases where the retrieved image is just above the 25 meters threshold, as in the cases of Tokyo 24/7 and SF-XL V1, whose predictions from Fig. 7 are 35 and 26 meters away.

4.3.2. Limitations

This paper presents, among other insights, the most comprehensive benchmark for re-ranking in VPR, both in terms of models and datasets. Although we aimed to obtain results that are as fair and comparable to each other, it must be noted that some of the chosen hyperparameters could benefit one method over another: for example, the image resolution was set to 512×512, which is a common resolution in VPR [1, 11], and some image matching methods might benefit from this more than others (*e.g*. RoMa [27] is known to prefer higher resolutions).

5. Conclusions

In this work, we revisit the conventional retrieval-and-re-ranking pipeline in the context of recent advances in the field. Our findings reveal that current state-of-the-art retrieval methods have effectively saturated historically challenging benchmarks, uncovering a counter-intuitive side effect: re-ranking can degrade performance when applied to near-perfect predictions. The key insight of this work is that image matching methods remain valuable; not as a default mechanism for trading computational cost for performance, but as a strategic tool to assess the confidence of retrieval predictions. When uncertainty is detected, image matching can then be employed selectively to refine and improve results. Finally, we present a comprehensive benchmark of re-ranking techniques for visual place recognition, encompassing a diverse range of methods and datasets to guide future research in the field.

Acknowledgements Gabriele Trivigno and Carlo Masone were supported by FAIR - Future Artificial Intelligence Research which received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.
We acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources.

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, 8

[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

[3] 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 (CVPR)*, pages 17794–17803, 2024. 2

[4] 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

[5] 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

[6] Artem Babenko, Anton Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. *ArXiv*, abs/1404.1777, 2014. 2

[7] Giovanni Barbarani, Mohamad Mostafa, Hajali Bayramov, Gabriele Trivigno, Gabriele Berton, Carlo Masone, and Barbara Caputo. Are local features all you need for cross-domain visual place recognition? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6155–6165, 2023. 1, 2, 3, 4

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

[9] Gabriele Berton and Carlo Masone. Megaloc: One retrieval to place them all. *arXiv preprint arXiv:2502.17237*, 2025. 1, 2, 4

[10] 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, 2021. 1, 3

[11] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *CVPR*, 2022. 1, 2, 3, 4, 8

[12] 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. 2

[13] Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. EigenPlaces: Training Viewpoint Robust Models for Visual Place Recognition . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11046–11056, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 1, 2

[14] Gabriele Berton, Gabriele Goletto, Gabriele Trivigno, Alex Stoken, Barbara Caputo, and Carlo Masone. Earthmatch: Iterative coregistration for fine-grained localization of astronaut photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3, 4

[15] Georg Bökman and Fredrik Kahl. A case for using rotation invariant features in state of the art feature matchers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5110–5119, 2022. 3

[16] Georg Bökman, Johan Edstedt, Michael Felsberg, and Fredrik Kahl. Affine steerers for structured keypoint description. In *European Conference on Computer Vision*, pages 449–468. Springer, 2024. 4

[17] Georg Bökman, Johan Edstedt, Michael Felsberg, and Fredrik Kahl. Steerers: A framework for rotation equivariant keypoint descriptors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4885–4895, 2024. 3, 4

[18] Kaiwen Cai, Chris Xiaoxuan Lu, and Xiaowei Huang. Stun: Self-teaching uncertainty estimation for place recognition. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6614–6621. IEEE, 2022. 3, 7

[19] 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

[20] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David Mckinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022. 3

[21] Mark Cummins and Paul Newman. Fab-map: Probabilistic localization and mapping in the space of appearance. *The International journal of robotics research*, 27(6):647–665, 2008. 1

[22] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 4

[23] Tomasz Malisiewicz Daniel DeTone and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Computer Vision and Pattern Recognition Workshop*, 2018. 2, 4

[24] 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, 4

[25] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3, 4

[26] Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Dedode: Detect, don't describe—describe, don't detect for local feature matching. In *2024 International Conference on 3D Vision (3DV)*, pages 148–157. IEEE, 2024. 3

[27] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 3, 4, 8

[28] 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

[29] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *Computer Vision – ECCV 2020*, pages 369–386, Cham, 2020. Springer International Publishing. 1

[30] Pierre Gleize, Weiyao Wang, and Matt Feiszli. Silk: Simple learned keypoints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22499–22508, 2023. 2

[31] Stephen Hausler and Peyman Moghadam. Pair-vpr: Place-aware pre-training and contrastive pair classification for visual place recognition with vision transformers. In *Arxiv*, 2024. 4

[32] S. Hausler, A. Jacobson, and M. Milford. Multi-process fusion: Visual place recognition using multiple image processing methods. *IEEE Robotics and Automation Letters*, 4(2): 1924–1931, 2019. 2

[33] Stephen Hausler, Tobias Fischer, and Michael Milford. Unsupervised complementary-aware multi-process fusion for visual place recognition. *arXiv preprint arXiv:2112.04701*, 2021. 7

[34] 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, 4

[35] Dihe Huang, Ying Chen, Yong Liu, Jianlin Liu, Shang Xu, Wenlong Wu, Yikang Ding, Fan Tang, and Chengjie Wang. Adaptive assignment for geometry aware local feature matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5425–5434, 2023. 3

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

[37] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17658–17668, 2024. 1, 2

[38] Sergio Izquierdo and Javier Civera. Close, but not there: Boosting geographic distance sensitivity in visual place recognition. In *Computer Vision – ECCV 2024*, pages 240–257, Cham, 2025. Springer Nature Switzerland. 1, 2

[39] 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, 2024. 2

[40] 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

[41] 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, 2022. 1

[42] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 3, 4

[43] María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. Data-efficient large scale place recognition with graded similarity supervision. *CVPR*, 2023. 1, 2

[44] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[45] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 4

[46] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2372–2381, 2017. 1

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

[48] Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 4

[49] 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16772–16782, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 2

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

[51] Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547, 2021. 2

[52] Riccardo Mereu, Gabriele Trivigno, Gabriele Berton, Carlo Masone, and Barbara Caputo. Learning sequential descriptors for sequence-based visual place recognition. *IEEE Robotics and Automation Letters*, 7(4):10383–10390, 2022. 2

[53] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 2

[54] Peer Neubert and Stefan Schubert. Hyperdimensional computing as a framework for systematic aggregation of image descriptors. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16933–16942, 2021. 2

[55] 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

[56] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V 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*, 2023. 1, 2

[57] 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, 2021. 2

[58] Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Natasa Sladoje. Comir: Contrastive multimodal image representation for registration. *Advances in neural information processing systems*, 33:18433–18444, 2020. 3

[59] Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, and Erickson R Nascimento. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2691, 2024. 4

[60] 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. 2

[61] A. Razavian, J. Sullivan, A. Maki, and S. Carlsson. Visual Instance Retrieval with Deep Convolutional Networks. *CoRR*, abs/1412.6574, 2015. 2

[62] Jerome Revaud, Philippe Weinzaepfel, César Roberto de Souza, and Martin Humenberger. R2D2: repeatable and reliable detector and descriptor. In *NeurIPS*, 2019. 1, 2, 3, 4

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

[64] 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, 2018. 1

[65] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[66] Stefan Schubert, Peer Neubert, Sourav Garg, Michael Milford, and Tobias Fischer. Visual place recognition: A tutorial [tutorial]. *IEEE Robotics & Automation Magazine*, 31 (3):139–153, 2024. 2

[67] Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. *arXiv preprint arXiv:2402.11095*, 2024. 4

[68] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE international conference on computer vision*, pages 118–126, 2015. 2

[69] Alex Stoken and Kenton Fisher. Find my astronaut photo: Automated localization and georectification of astronaut photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 6196–6205, 2023. 3

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

[71] Xun Sun, Yuanfan Xie, Pei Luo, and Liang Wang. A dataset for benchmarking image-based localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7436–7444, 2017. 3

[72] Shitao Tang, Jiahui Zhang, Siyu Zhu, and Ping Tan. Quadtree attention for vision transformers. *arXiv preprint arXiv:2201.02767*, 2022. 3

[73] 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. 2

[74] Giorgos Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. *CoRR*, abs/1511.05879, 2016. 2

[75] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla. Visual place recognition with repetitive structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2346–2359, 2015. 1

[76] 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, 3

[77] Gabriele Trivigno, Gabriele Berton, Juan Aragon, Barbara Caputo, and Carlo Masone. Divide&Classify: Fine-grained

classification for city-wide visual geo-localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11142–11152, 2023. 2

[78] Gabriele Trivigno, Carlo Masone, Barbara Caputo, and Torsten Sattler. The unreasonable effectiveness of pre-trained features for camera pose refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12798, 2024. 1

[79] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *ECCV*, 2024. 4

[80] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 2, 4

[81] Issar Tzachor, Boaz Lerner, Matan Levy, Michael Green, Tal Shalev, Gavriel Habib, Dvir Samuel, Noam Zailer, Or Shimshi, Nir Darshan, and Rami Ben-Ari. Effovpr: Effective foundation model utilization for visual place recognition, 2024. 4

[82] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Proceedings of the Asian Conference on Computer Vision*, pages 2746–2762, 2022. 3, 4

[83] 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, 2022. 1, 2, 4

[84] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3, 4

[85] Ziqi Wang, Jiahui Li, Seyran Khademi, and Jan van Gemert. Attention-aware age-agnostic visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[86] 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*, 2020. 1, 3, 4, 8

[87] Frederik Warburg, Martin Jørgensen, Javier Civera, and Søren Hauberg. Bayesian triplet loss: Uncertainty quantification in image retrieval. In *Proceedings of the IEEE/CVF International conference on Computer Vision*, pages 12158–12168, 2021. 3, 5, 7

[88] 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, 2022. IEEE Computer Society. 1

[89] 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

[90] Mubariz Zaffar, Liangliang Nan, and Julian FP Kooij. On the estimation of image-matching uncertainty in visual place recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 5, 7

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

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

[93] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter CY Chen, Qingsong Xu, and Zhengguo Li. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–16, 2023. 2, 4

[94] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4669–4678, 2021. 3, 4

[95] 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. 1, 2