Benchmarking Image Retrieval for Visual Localization

Noé Pion[1] Martin Humenberger[1] Gabriela Csurka[1] Yohann Cabon[1] Torsten Sattler[2]
[1]NAVER LABS Europe [2]Czech Technical University in Prague
noe.pion@gmail.com, {firstname.lastname}@naverlabs.com, torsten.sattler@cvut.cz

Abstract

Visual localization, i.e.*, camera pose estimation in a known scene, is a core component of technologies such as autonomous driving and augmented reality. State-of-the-art localization approaches often rely on image retrieval techniques for one of two tasks: (1) provide an approximate pose estimate or (2) determine which parts of the scene are potentially visible in a given query image. It is common practice to use state-of-the-art image retrieval algorithms for these tasks. These algorithms are often trained for the goal of retrieving the same landmark under a large range of viewpoint changes. However, robustness to viewpoint changes is not necessarily desirable in the context of visual localization. This paper focuses on understanding the role of image retrieval for multiple visual localization tasks. We introduce a benchmark setup and compare state-of-the-art retrieval representations on multiple datasets. We show that retrieval performance on classical landmark retrieval/recognition tasks correlates only for some but not all tasks to localization performance. This indicates a need for retrieval approaches specifically designed for localization tasks. Our benchmark and evaluation protocols are available at https://github.com/naver/kapture-localization.*

1. Introduction

Visual localization is the problem of estimating the exact camera pose for a given image in a known scene, *i.e.*, the exact position and orientation from which the image was taken. Localization algorithms are core components of systems such as self-driving cars [34], autonomous robots [54], and mixed reality applications [5, 16, 59, 62, 101].

Traditionally, visual localization algorithms rely on a 3D scene representation of the target area [36, 51, 52, 77, 88], constructed from reference/database images with known poses. They use 2D-3D matches between a query image and the 3D representation for pose estimation. This representation can be an explicit 3D model, often obtained via Structure-from-Motion (SFM) [33, 83, 87] using local features for 2D-3D matching, or an implicit representation through a machine learning algorithm [10, 61, 85]. In the latter case, the learning algorithm is trained to regress 2D-3D matches. These structure-based methods can be scaled to large scenes through an intermediate image retrieval step [11, 24, 30, 74, 75, 80, 91, 92]. The intuition is that the top retrieved images provide hypotheses about which parts of the scene are likely visible in a query image. 2D-3D matching can then be restricted to these parts.

The pre-processing step of building a 3D scene representation is not strictly necessary. Instead, the camera pose of a query image can be computed using the known poses of the top database images found, again using image retrieval. This can be achieved via relative pose estimation between query and retrieved images [109, 113], by estimating the absolute pose from 2D-2D matches [111], via relative pose regression [9, 26] or by building local 3D models on demand [100]. If high pose accuracy is not required, the query pose can be approximated very efficiently via a combination of the poses of the top retrieved database images [99, 100, 108].

As illustrated in Fig. 1, there are various roles that image retrieval can play in visual localization systems. **Efficient pose approximation** by representing the pose of a query image by a (linear) combination of the poses of retrieved database images [99, 100, 108] **(Task 1)**. **Accurate pose estimation without a global 3D map** by computing the pose of the query image relative to the known poses of retrieved database images [9, 26, 48, 100, 109, 113] **(Task 2a)**. **Accurate pose estimation with a global 3D map** by estimating 2D-3D matches between features in a query image and the 3D points visible in the retrieved images [15, 30, 36, 74, 80, 91] **(Task 2b)**.

These three tasks have differing requirements on the results of the retrieval stage: Task 1 requires the retrieval step to find images taken from poses as similar as possible to the query, *i.e.*, the image representation should not be too robust or invariant to changes in viewpoint. Tasks 2a and 2b require the retrieval stage to find images depicting the same part of the scene as the query image. However, the retrieved images do not need to be taken from a similar

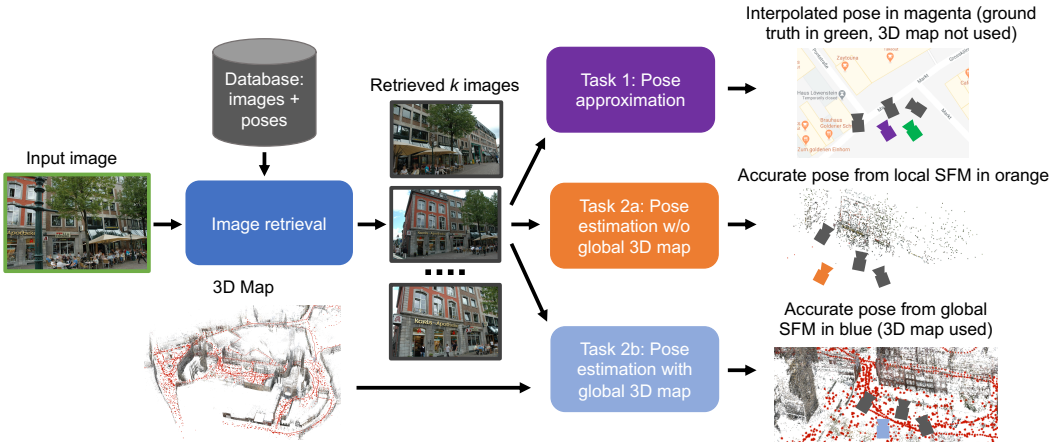

Figure 1. This paper analyzes the role of image retrieval in three visual localization tasks through extensive experiments.

pose as the query as long as local feature matching succeeds. In fact, Task 2a usually requires retrieving multiple images from a diverse set of viewpoints that differ from the query pose [48, 113]. Task 2b benefits from retrieving images of high visual overlap with the query image and (in theory) requires only one relevant database image.

Despite differing requirements, modern localization methods [27, 30, 74, 91, 100, 113] indiscriminately use the same representations based on compact image-level descriptors [1, 97]. These descriptors are typically trained for landmark retrieval/place recognition tasks with the goal to produce similar descriptors for images showing the same building or place independently of the pose or other viewing conditions [55, 69]. Interestingly, to the best of our knowledge, there is no work analyzing the suitability of such descriptors on the three visual localization tasks.

In order to close this gap in the literature, this paper investigates the role of image retrieval for visual localization. We design a benchmark to measure the correlation between localization and retrieval/recognition performance for each task. Our benchmark enables a fair comparison of different retrieval approaches by fixing the remaining parts of the localization pipeline. Our main contributions are a set of extensive experiments and the conclusions we draw from them: (**1**) there is no correlation between landmark retrieval/place recognition performance and Task 1. (**2**) Similarly, retrieval/recognition performance is not a good indicator for performance on Task 2a. (**3**) Task 2b correlates with the classical place recognition task. (**4**) Our results clearly show that there is a need to design image retrieval approaches specifically tailored to the requirements of some of the localization tasks. To foster such research, our benchmark and evaluation protocols are publicly available.

2. Related Work

Landmark retrieval. Landmark retrieval is the task of identifying all relevant database images depicting the same landmark as a query image. Early methods relying on global image statistics were significantly outperformed by methods based on aggregating local features, most notably the bag of visual words representations for images [22, 86] and its extensions such as Fisher Vectors [64] and the Vector of Locally Aggregated Descriptors (VLAD) [38]. More recently, deep representation learning has led to further improvements. They apply various pooling mechanisms [1, 7, 39, 69, 70, 95, 95] on activations in the last convolutional feature map of CNNs in order to construct a global image descriptor. They learn the similarity metric by using ranking losses such as contrastive, triplet, or average precision (AP) [8, 31, 69, 71]. Several benchmark papers compare such image representations on the task of instance-level landmark retrieval [4, 63, 66, 68, 105, 112]. In contrast, this paper explores how state-of-the-art landmark retrieval approaches perform in the context of visual localization.

Visual localization. Traditionally, structure-based methods establish 2D-3D correspondences between a query image and a 3D map, typically via matching local feature descriptors [23, 84] and use them to compute the camera pose by solving a perspective-n-point (PNP) problem [44, 46, 47] robustly inside a RANSAC [20, 28, 49] loop. More recently, scene coordinate regression techniques determine these correspondences using random forests [61, 85] or CNNs [10, 11, 61]. Earlier methods trained a regressor specifically for each scene while recent models are able to adapt the trained model on-the-fly to new scenes [17, 18, 106]. Even if scene coordinate regression methods achieve high pose accuracy on small datasets, they currently do not scale up well to larger and more complex scenes [11, 53, 79, 91, 92, 104]. This is why we focus on feature-based localization methods that use image retrieval to cope with this problem [27, 72–74].

Absolute pose regression methods forego 2D-3D matching and train a CNN to directly predict the full camera pose from an image for a given scene [12, 41, 42, 103]. However, they are significantly less accurate than structure-

based methods [81] and currently not (significantly) more accurate than simple retrieval baselines [81] but significantly less scalable [79]. This is why we focus on image retrieval for efficient and scalable pose approximation instead.

Accurate real-world visual localization needs to be robust to a variety of conditions, including day-night, weather and seasonal variations. [79] introduces several benchmark datasets specifically designed for analyzing the impact of such factors on visual localization using query and training images taken under varying conditions. For our benchmark, we use the Aachen Day-Night-v1.1 [79, 110], the RobotCar Seasons [60], and the Baidu shopping mall dataset [89].

For more details see recent survey papers that cover different aspects of visual localization [13, 29, 58, 67, 107] and benchmark these methods [79, 81, 100].

Place recognition. Place recognition, also referred to as visual geo-localization [107], lies between landmark retrieval and visual localization. While, similarly to the latter, its goal is to estimate the camera location, a coarse geographic position of the image is considered sufficient [32, 82, 102, 108]. It is often important to explicitly handle confusing [45, 82] and repetitive scene elements [2, 76, 98], especially in large urban scenes. To improve scalability, a popular strategy is to perform visual and geo-clustering [6, 15, 21, 40, 50].

As image matching and retrieval are key ingredients of place recognition methods, several papers proposed improved image representations using GPS and geometric information as a form of weak supervision [1, 43, 69, 102]. In this paper, among others, we use NetVLAD [1] which is probably the most popular representation trained this way, as well as DenseVLAD [97], its handcrafted counterpart.

3. The proposed benchmark

Modern localization algorithms tend to only use state-of-the-art landmark retrieval and place recognition representations. However, different localization tasks have different requirements on the retrieved images and thus on the used retrieval representations. In this paper, we are interested in understanding how landmark retrieval/place recognition performance relates to visual localization performance. In particular, we are interested in determining whether current state-of-the-art retrieval/recognition representations are sufficient or whether specialized (task-dependent) representations for localization are needed.

This section presents an evaluation framework designed to answer this question. Our framework enables a fair comparison of different retrieval approaches for each of the three localization tasks by fixing the remaining parts of the localization pipeline. It consists of two parts, one measuring localization performance for the three tasks identified above (Sec. 3.1) and the other measuring landmark retrieval/place recognition performance (Sec. 3.2), both on the same datasets. Relating the performance of state-of-the-art retrieval representations on all these tasks thus enables us to understand the relation between image retrieval in visual localization and landmark retrieval/recognition tasks.

3.1. Visual localization tasks

As outlined in Sec. 1, we consider two roles for image retrieval in the context of visual localization: identifying reference images taken from a similar pose as the query image (Task 1) and retrieving database images depicting the same part of the scene as the query image but not necessarily from similar poses (Tasks 2a and 2b).

Task 1: Pose approximation. Methods falling into the first category are inspired by place recognition [81, 99, 100, 108] and aim to efficiently approximate the query pose from the poses of the top k retrieved database images.

We represent a camera pose as a tuple $\mathbf{P} = (\mathbf{c}, \mathbf{q})$. Here, $\mathbf{c} \in \mathbb{R}^3$ is the position of the camera in the global coordinate system of the scene and $\mathbf{q} \in \mathbb{R}^4$ is the rotation of the camera encoded as a unit quaternion. We compute the pose of the query image as a weighted linear combination $\mathbf{P}_q = \sum_{i=1}^{k} w_i \mathbf{P}_i$, where $\mathbf{P}_i$ is the pose of the top i retrieved image and w_i is a corresponding weight[1]. As a consequence, for $k = 1$ we directly use the pose of the top retrieved image.

We consider three variants: **equal weighted barycenter** (**EWB**) assigns the same weight to all of the top k retrieved images with $w_i = 1/k$. **Barycentric descriptor interpolation** (**BDI**) [81, 99] estimates w_i as the best barycentric approximation of the query descriptor via the database descriptors with

$$\left\| \mathbf{d}_q - \sum_{i=1}^{k} w_i \mathbf{d}_i \right\|_2 \quad \text{subject to} \quad \sum_{i=1}^{k} w_i = 1 \ . \tag{1}$$

Here, $\mathbf{d}_q$ and $\mathbf{d}_i$ are the global image-level descriptors of the query image and the top i retrieved database image, respectively.

In the third approach, w_i is based on the **cosine similarity** (**CSI**) between L2 normalized descriptors:

$$w_i = \frac{1}{z_i} \left(\mathbf{d}_q^T \mathbf{d}_i \right)^{\alpha} \ , \text{with} \ z_i = \sum_{j=1}^{k} \left(\mathbf{d}_q^T \mathbf{d}_j \right)^{\alpha} \ . \tag{2}$$

Setting $\alpha = 0$ reduces this method to **EWB**. At the opposite, as $\alpha \to \infty$, the pose obtained is the one of the image giving the highest similarity. We fix $\alpha = 8$ based on preliminary results on the Cambridge Landmarks [42] dataset.

[1] Note that, $\mathbf{q}_q = \sum_i w_i \mathbf{q}_i$ is re-normalized to be a quaternion.

Task 2a: Pose estimation without a global map. In theory, using the top 1 retrieved image would be sufficient for this task if the relative pose between the query and this image could be estimated accurately, including the scale of the translation [9, 26]. In practice, retrieving $k > 1$ images improves the accuracy because k relative poses can be considered [48, 100, 109, 113]. Once the relative poses between query and database images are estimated, triangulation can be used to compute the absolute pose [48, 109, 113]. However, pose triangulation fails if the query pose is colinear with the poses of the database images, which is often the case in autonomous driving scenarios.

Therefore, we follow [100] where the retrieved database images with known poses are used to build the 3D map of the scene on-the-fly[2] and then register the query image within this map using PNP. Similar to pose triangulation, this **local SFM** approach fails if (i) less than two images among the top k database images depict the same place as the query image, (ii) the viewpoint change among the retrieved images and/or between the query and the retrieved images is too large (to be handled by local feature matching) or (iii) the baseline between the retrieved database images is not large enough to allow stable triangulation of enough 3D points. Thus, this approach requires retrieving a diverse set of images depicting the same scene as the query image from a variety of viewpoints. As such, methods for Task 2a benefit from image representations that are robust but not invariant to viewpoint changes.

Task 2b: Pose estimation with a global map. In contrast to Task 2a, this task uses a pre-built global 3D model of the scene rather than reconstructing it locally on-the-fly. We follow a standard local feature-based approach from the literature [30, 36, 74, 80]: an SFM model of the scene provides the correspondences between local features in the database images and 3D points in the map. Establishing 2D-2D matches between the query image and top ranked database images yields a set of 2D-3D matches which are then used for pose estimation via PNP and RANSAC.

In theory, retrieving a single relevant image among the top k is sufficient as long as the viewpoint change between the query and this image can be handled by the local features. Retrieving more relevant images increases the chance for accurate pose estimation. For efficiency, k should be as small as possible[3] as local feature matching is often the bottleneck in terms of processing time.

Overall, we expect this task to benefit from retrieval representations that are moderately robust to viewpoint changes while still allowing reliable local feature matching.

[2]Note that compared to the query pose, the 3D points are very seldomly colinear with the reference poses and can thus be accurately triangulated.

[3]In scenes with ambiguous structures, *e.g.*, the InLoc dataset [92], retrieving more images can lead to a decrease in accuracy due to wrong, but geom. consistent, matches. We did not observe this in our experiments.

Visual localization metrics. We follow common practices [79, 85] to measure localization performance by computing the position and rotation errors between an estimated and a reference pose. For evaluation, we use the percentage of query images localized within a given pair of (position, rotation) error thresholds $(X\text{m}, Y^\circ)$ [79, 85].

3.2. Landmark retrieval and place recognition tasks

In order to correlate visual localization with landmark retrieval/place recognition performance, we evaluate the latter two tasks on the same datasets used for localization.

Landmark retrieval. This is an instance retrieval task where all images containing the main object of interest shown in the query image are to be retrieved from a large database of images. Thus, image representations should be as robust as possible to viewpoint and viewing condition changes in order to identify all relevant images.

In order to determine whether a retrieved image is relevant for a query, we follow the 3D model-based definition from [69]: the similarity of two images is computed as the intersection over union (IoU) for the sets of 3D scene points observed by both images in an SFM model. We consider a database image relevant for a given query image if this IoU score is strictly positive, *i.e.* they have shared 3D points. This is in contrast to classical landmark retrieval where relevant images might depict unrelated parts of the same landmark, *e.g.*, opposite sides of the same building.

In order to compute the visible 3D points of the query images, we compute an SFM model containing the database and the query images using R2D2 features [72] and COLMAP [83]. To accelerate the image matching, we only match image pairs if their viewing frusta (limited by a far plane) overlap [9]. The strategy can fail in two ways: the query images can either be too far away for the frusta to overlap or there are not enough good local feature correspondences in the resulting image pair. Hence some query images do not have reconstructed 3D points and are ignored during evaluation.

Landmark retrieval metric. The classical mean Average Precision (mAP) metric, most commonly used in the literature [65, 66, 68, 94] to measure landmark retrieval performance, reports a single number integrating over different numbers of retrieved images. We use the related *mean Precision@k* (P@k) measure to determine the link between number of retrieved images and localization performance.

Place recognition. This task aims to approximately determine the place from which a given query image was taken. Since the place is defined by the location of the retrieved images, this requires at least one relevant reference image amongst the top k retrieved ones. A database image is typically considered relevant if it was taken within a neighborhood of the query image [2, 76, 97, 98]. When the camera

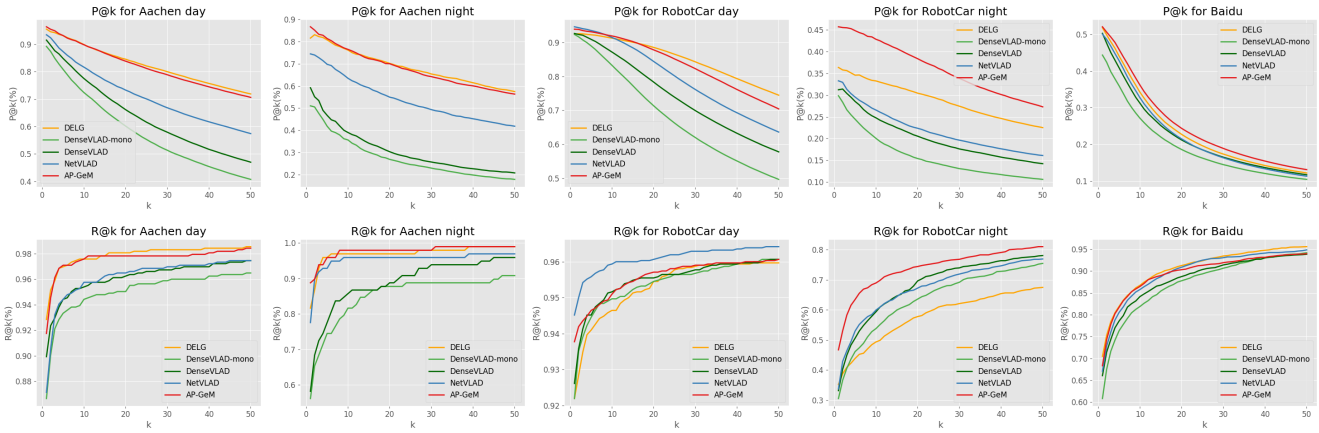

Figure 2. **Landmark retrieval/place recognition**. Image retrieval (top row) and place recognition (bottom row) performance as measured with Precision@k respectively Recall@k. Results per dataset are shown per column. There are clear differences between the representations with learned descriptors typically outperforming the handcrafted DenseVLAD descriptors. Best performance on both measures are obtained with AP-GeM and DELG, except on RobotCar night where DELG performs significantly worse than AP-GeM.

orientation is available, the angle between the cameras' orientations can also be taken into account. Alternatively, the above mentioned IoU similarity can be considered as well to determine whether or not a database image represents the same place [19]. We consider both measures, discussing the latter in the main paper and showing results/correlations with the pose distance in the supplementary material.

Place recognition metric. We follow the standard protocol and measure place recognition performance via *Recall@k* (R@k) [1,2,97,97]. R@k measures the percentage of query images with *at least* one relevant database image amongst the top k retrieved ones.

4. Experimental evaluation

After first describing our experimental setup, Sec. 4.1 evaluates the representations on retrieval and place recognition tasks. Sec. 4.2 and 4.3 present results for pose approximation (Task 1) and accurate visual localization (Task 2).

Experimental setup. We use **DenseVLAD** [97] and three popular deep image representations, **NetVLAD** [1], **AP-GeM** [71], and **DELG** [14], for image retrieval. DenseVLAD pools densely extracted SIFT [56] descriptors through the VLAD representation [38], resulting in a compact image-level descriptor. We use two variants: DenseVLAD extracts descriptors at multiple scales, while DenseVLAD-mono uses only a single scale. NetVLAD uses CNN features instead of SIFT features and was trained on the Pitts30k [1, 97] dataset. Both DenseVLAD and NetVLAD have been used for visual localization [30, 74, 79, 81, 100] and place recognition [1] before. AP-GeM and DELG represent state-of-the-art representations for landmark retrieval [68], while AP-GeM was recently used for visual localization as well [35]. Both models were trained on the Google Landmarks dataset (GLD) [63], where each training image has a class label based on the landmark visible in the image. Relevance between images is established based on these labels. Hence, two images can be relevant to each other without showing the same part of a landmark. We use the best pre-trained models[4] released by the authors for all experiments (*c.f.* suppl. mat.).

For Tasks 2a and 2b, *i.e.*, pose estimation without and with a global map, we use R2D2 [72] to extract local image features and COLMAP [83] for SFM. We observed similar behavior for D2-Net [27] and SIFT [56] (*c.f.* suppl. mat.).

We use three public datasets for our experiments: **Aachen Day-Night-v1.1** [79, 80, 110], **RobotCar Seasons** [60, 79], and **Baidu Mall** [89]. RobotCar represents an autonomous driving scenario with little viewpoint change between query and reference images. In contrast, the outdoor dataset Aachen exhibits stronger viewpoint changes as the camera can freely move through the scene. Both datasets come with **day-** and **nighttime** queries. Nighttime queries introduce the challenge of handling strong illumination changes as all reference images were taken during the day. The Baidu dataset contains medium viewpoint and limited illumination changes between the query and database images but exhibits strong differences in image quality, occlusion from people, and other distractions such as reflections on storefronts.

Following [79], for all three datasets we use three threshold pairs for evaluating localization for **low** (5m, 10°), **medium** (0.5m, 5°), and **high** (0.25m, 2°) accuracy. We only show here results with thresholds *low* and *high* as these are the most relevant for pose approximation and accurate pose estimation (for *medium*, see suppl. mat.).

[4] It is common practice that localization methods rely on pre-trained models, *e.g.* [27, 30, 78] use off-the-shelf DenseVLAD or NetVLAD for their results and [35] uses AP-GeM.

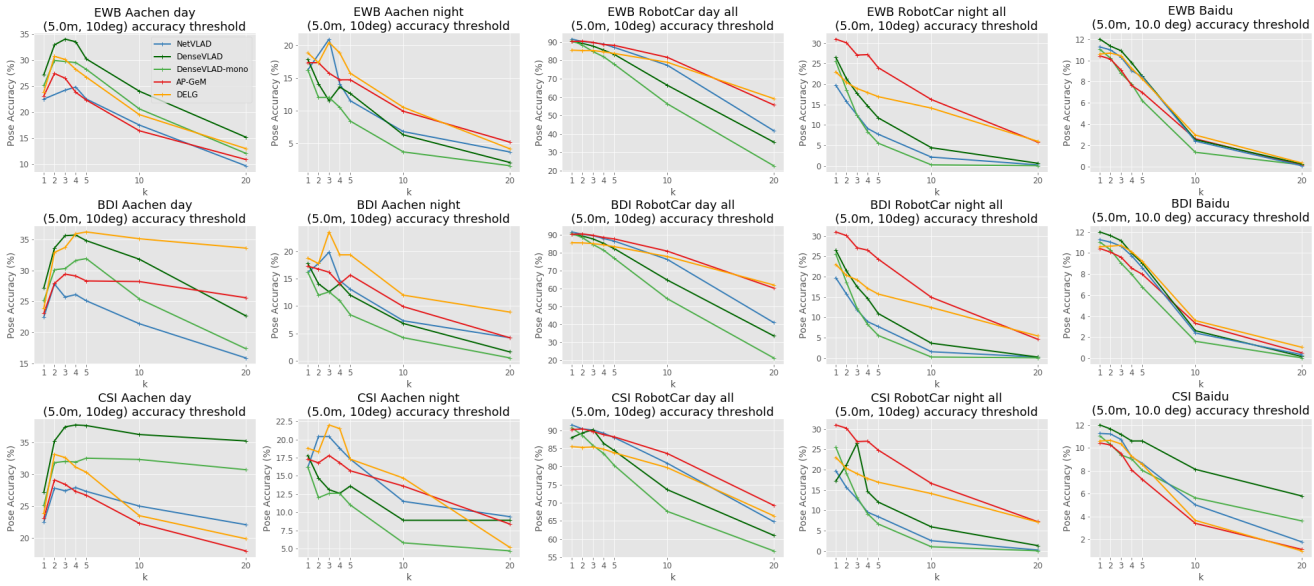

Figure 3. **Task 1 (pose approximation)**. The rows show results obtained via equal weighted barycenter (EWB), barycentric descriptor interpolation (BDI), and cosine similarity (CSI), respectively. The best results are obtained with CSI, however simply using the top-retrieved pose works best for RobotCar and Baidu due to their limited variation between query and reference poses. Interestingly, the results on Baidu and the daytime queries for Aachen and RobotCar show that the rather low-level DenseVLAD descriptors perform as good or better than the learned descriptors.

4.1. Landmark retrieval and place recognition

Figure 2 shows the results for landmark retrieval (top) and place recognition (bottom) tasks from Sec. 3.2. As expected, the learned descriptors (NetVLAD, AP-GeM, and DELG) typically outperform the SIFT-based DenseVLAD. There are two interesting observations: (1) NetVLAD outperforms both AP-GeM and DELG under the R@k measure for small k on the daytime RobotCar queries. This can be attributed to the fact that NetVLAD was trained on street-view images captured at daytime from a vehicle while AP-GeM and DELG were trained with a large variety of landmark images taken from very different viewpoints. (2) On R@k for the RobotCar nighttime queries, DELG performs significantly worse than the others. We attribute this to the low-quality nighttime images, which often exhibit strong motion blur and color artifacts which are not reflected in the training set of DELG. AP-GeM, trained on the same data, avoids this problem through adequate data augmentation.

Discussion. Our experiments confirm that the state-of-the-art DELG and AP-GeM descriptors remain the best choices for place retrieval/recognition tasks. To complement this, Table 1 shows the performance obtained on $\mathcal{R}$Oxford ($\mathcal{R}$O) and $\mathcal{R}$Paris ($\mathcal{R}$P) landmark retrieval benchmarks using the Medium (m) and Hard (h) protocols [68]. Here, DELG and AP-GeM descriptors also outperform DenseVLAD and NetVLAD by a large margin. However, if the aim is place recognition, *i.e.*, if the requirement is to find at least one relevant image, the gap is significantly lower and in some cases the ranking might even change (*e.g.*, the Robotcar dataset).

	$\mathcal{R}$O(m)	$\mathcal{R}$O (h)	$\mathcal{R}$P(m)	$\mathcal{R}$P(h)
DenseVLAD	36.8	13.0	42.5	13.7
NetVLAD	37.1	13.8	59.8	35.0
AP-GeM	67.4	42.8	80.4	61.0
DELG	69.7	45.1	81.6	63.4

Table 1. Performance evaluation (mAP) on $\mathcal{R}$Oxford ($\mathcal{R}$O) and $\mathcal{R}$Paris ($\mathcal{R}$P) using the Medium (m) and Hard (h) protocols.

4.2. Task 1: Pose approximation

Figure 3 shows pose approximation results for the three approaches discussed in Sec. 3.1. We show the percentage of query images localized within a given error threshold w.r.t. the ground truth poses as a function of the number k of retrieved images used for pose approximation. We only report results for the low (5m, 10°) thresholds as even fewer images are localized for stricter thresholds (*c.f.* suppl. mat.). Equal weighted barycenter (EWB) uses the same weight for each of the top k retrieved images. In contrast, both barycentric descriptor interpolation (BDI) and cosine similarity (CSI) give a higher weight to higher ranked images. They thus assume correlation between descriptor and pose similarity. As can be seen in Fig. 3, retrieving $k > 1$ images only improves performance on the Aachen dataset because there is a larger pose difference between query and reference images than for the other datasets. This allows better poses to be estimated by interpolating between the poses of the retrieved database images. Here, CSI performs best as the exponent α in Eq. 2 effectively downweigths unrelated images, whereas unrelated images among the top k receive a larger weight for BDI and EWB.

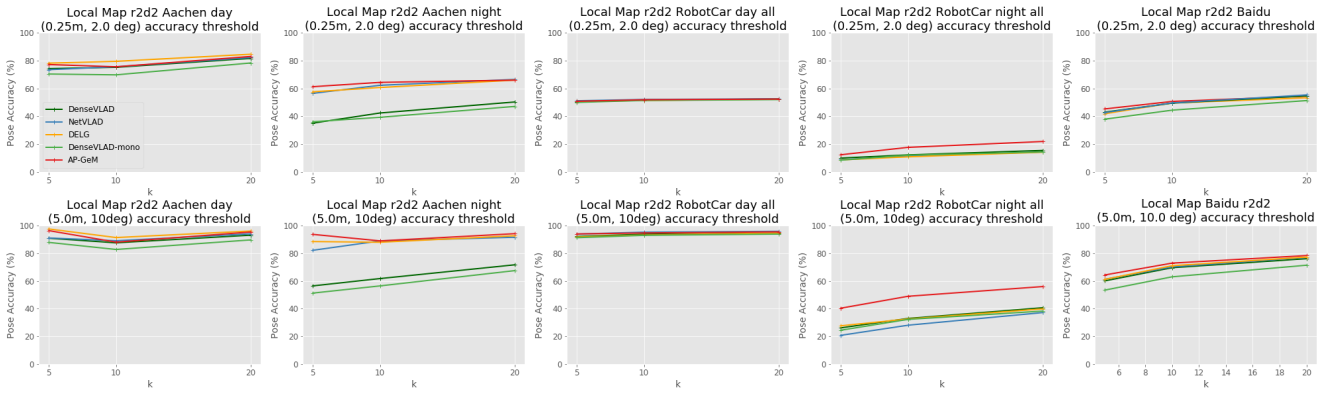

Figure 4. **Task 2a (pose estimation without a global map)**. Top/bottom row: percentage of images localized within the high/low accuracy threshold as a function of the number k of retrieved images. All representations perform similarly on outdoor day images. AP-GeM slightly outperforms the other descriptors on night images as well as on Baidu, while DenseVLAD performs worst in these cases.

Comparing Fig. 2 and 3, we observe a correlation between these curves and P@k but none with R@k, except for Aachen day[5]. In the latter case, we observe that while DenseVLAD performs poorly on the landmark retrieval/place recognition tasks it offers good performance for pose approximation. The reason might be that DenseVLAD descriptors are less robust to viewpoint changes. As discussed in Sec. 3.1, this is desirable for pose approximation as more similar poses will lead to more similar descriptors. However, DenseVLAD is less robust to illumination changes than AP-GeM and DELG. This explains why the latter perform better on the nighttime queries of RobotCar and Aachen.

Finally, while AP-GeM performs similar for retrieval/recognition on Aachen (see Fig. 2), interpolating top retrieved poses with DELG outperforms AP-GeM. This suggests that DELG retrieves reference images that are better spread through the scene, which is beneficial for pose interpolation.

Discussion. Overall, our results show that while pose approximation metrics is somewhat correlated to the top precision of image retrieval, there is no correlation with the place recognition task (R@k). These suggest that learning scene representations tailored to pose approximation, instead of using off-the-shelf methods trained for landmark retrieval, is an interesting direction for future work.

The best results for RobotCar and Baidu are obtained for $k = 1$, in which case all three methods perform the same. For RobotCar, this result is surprising as interpolating between two consecutive images should give finer approximation. This indicates that there is potential for improvement by designing image retrieval representations specifically suited for pose interpolation.

[5]This is confirmed by the scatter plots in the supplementary material.

4.3. Task 2: Accurate pose estimation

We show results obtained with high and low accuracy thresholds for Task 2a (local SFM) in Fig. 4 and for Task 2b (global map) in Fig. 5.

Task 2a. We observe that all representations, even the handcrafted DenseVLAD descriptors, perform similarly well on the outdoor daytime and the indoor images, with DenseVLAD-mono being the exception in the latter case. We again observe that learned descriptors can perform better than DenseVLAD under day-night changes, with AP-GeM yielding the best results overall.

Comparing Fig. 2 and Fig. 4, we do not see a clear correlation between retrieval/recognition and Task 2a: despite significant differences in P@k and R@k, all methods perform similarly well for Task 2a on Aachen and RobotCar daytime images. In contrast, the better retrieval/recognition performance of AP-GeM on RobotCar night translates to better performance on Task 2a. However, the better P@k performance of DELG and AP-GeM on Aachen night does not translate to a better Task 2a performance. In fact, there is no correlation between P@k, which decreases with increasing k (*c.f.* Fig. 2), and Task 2a performance, which remains the same or increases with increasing k.

Discussion. To achieve good pose accuracy using a local SFM model for a given set of retrieved images, a high R@k score is not sufficient. This is due to the fact that more than one relevant image is needed to build the local map. At the same time, not all of the top k retrieved images need to be relevant, *i.e.*, a high P@k score is not needed.

Overall, retrieval/recognition performance is not a good indicator for Task 2a performance[6]. This indicates that better retrieval representations can be learned that are tailored to the task of pose estimation without a global map, *e.g.*, by designing a loss that optimizes pose accuracy.

[6]Correlation plots between Task 2 performance and P@k respectively R@k are shown in the suppl. mat.

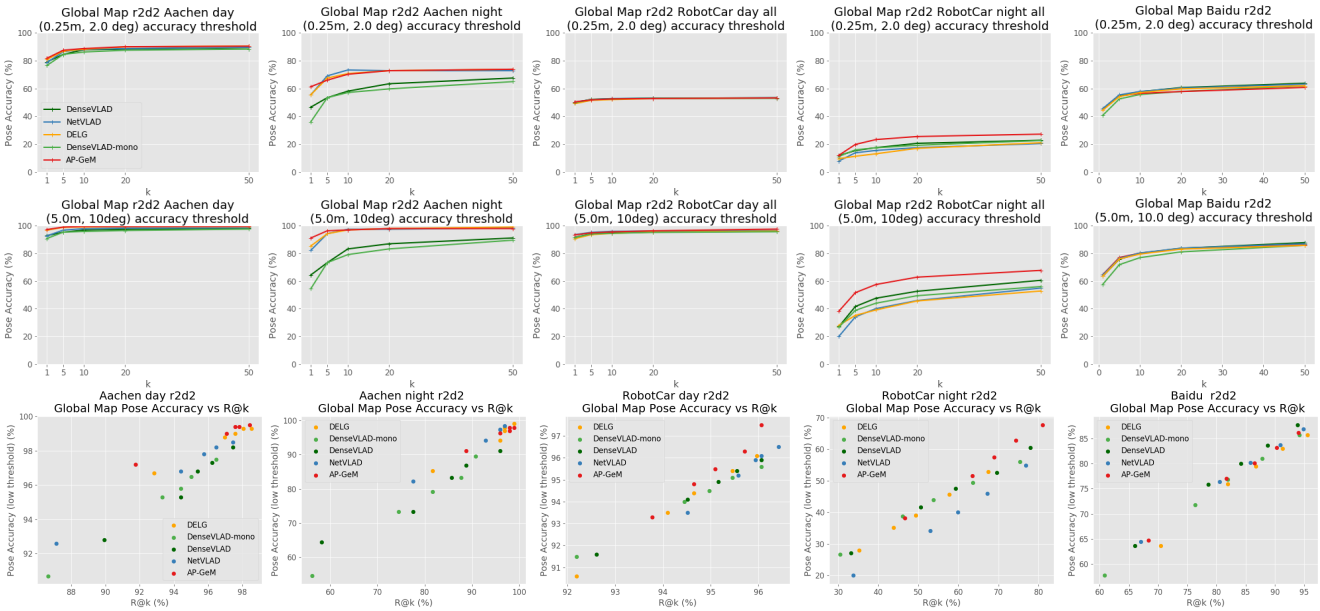

Figure 5. **Task 2b (pose estimation with a global map)**. Top/middle row: percentage of images localized within the high/low accuracy threshold as a function of the number k of retrieved images. The representations perform similarly for daytime scenarios and the indoor case, but AP-GeM best handles nighttime illumination changes. Bottom row: scatter plots showing low accuracy localization results versus the retrieval metric R@k for $k = 1, 5, 10, 20, 50$. We observe a clear correlation between Task 2b and the landmark recognition task.

Task 2b. A similar behavior of the different representations as in Task 2a can be seen in Fig. 5, *i.e.*, all methods perform well on daytime and indoor images while learning-based methods perform better on the Aachen nighttime queries.

The lower performance on the RobotCar nighttime and Baidu images can be explained by the comparably low R@k for both datasets (*c.f.* Fig. 2), especially for $k \leq 10$. Without retrieving at least one relevant image amongst the top k, pose estimation is bound to fail.

Discussion. As shown in the bottom row of Fig. 5, there is a clear correlation between a high R@k and good performance for the coarse pose threshold. A higher R@k increases the chance that an image can be localized at all. This validates the common practice to use state-of-the-art representations trained for retrieval/recognition for localization. However, the row also shows that a high R@k does not necessarily imply a high pose accuracy. One explanation is that the retrieved images are relevant but share little visual overlap with the query image. In this case, all matches will be found in a small area of the query image resulting in an unstable (and thus likely inaccurate) pose estimate.

5. Conclusion

Image retrieval plays an important role in modern visual localization systems. Retrieval techniques are often used to efficiently approximate the pose of the query image or as an intermediate step towards obtaining a more accurate pose estimate. Most localization systems simply use state-of-the-art image representations trained for landmark retrieval or place recognition. In this paper, we analyzed the correlation between the tasks of visual localization and retrieval/recognition through detailed experiments.

Our results show that state-of-the-art image-level descriptors for place recognition are a good choice when localizing an image against a pre-built map as performance on both tasks is correlated. We can see that on the night images as well as on Baidu, AP-GeM often outperforms the other features. One of the reason might be that AP-GeM is the only feature that was trained not only with geometric data augmentation but also with color jittering. This might explain why it better handles day-night variations.

We also show that the tasks of pose approximation and localization without a pre-built map (local SFM) are not directly correlated with landmark retrieval/place recognition. In the case of pose approximation, representations that reflect pose similarity in their descriptor similarities, *i.e.*, exhibit robustness only to illumination changes, are preferable as they tend to retrieve closer images. For local SFM, there is a complex relationship between the retrieved images that is not captured by the classical Precision@k and Recall@k measures used for retrieval and recognition. Our results suggest that developing suitable representations tailored to these tasks are interesting directions for future work. Our code and evaluation protocols are publicly available to support such research.

Acknowledgments. This work received funding through the EU Horizon 2020 project RICAIP (grant agreement No 857306) and the European Regional Development Fund under IMPACT No. CZ.02.1.01/0.0/0.0/15 003/0000468.

A. Introduction

This supplementary material complements the main paper with further details on the experimental setup (Section B) and additional results (Section C) to strengthen the findings.

Section B is structured as follows. First, in Section B.1 we give more information on the localization datasets used in our experiments. In Section B.2 we describe the global image representations that were compared in the benchmark and we provide links to the codes we used to extract them. In Section B.3, we briefly recall the SFM pipeline and point to the codes on which we relied upon for running the experiments. In Section B.4 we recall the evaluation metrics to make the plots in the supplementary easily understandable.

In Section C, we provide additional figures to Sections **4.1**, **4.2** and **4.3** of the main paper, corresponding here to Sections C.1, C.2, and C.3. In Section C.4, we show additional correlation analyses between localization and retrieval measures.

B. Experimental setup

B.1. Datasets

To evaluate the role of image retrieval in visual localization, we selected three public datasets aimed at benchmarking visual localization: **Aachen Day-Night-v1.1** [79, 80, 110], **RobotCar Seasons** [60, 79] and **Baidu Mall** [89]. Altogether, the selected datasets cover a variety of application scenarios: large-scale outdoor handheld localization under varying conditions (Aachen Day-Night), small-scale indoor handheld localization with occlusions (Baidu), and large-scale autonomous driving (RobotCar Seasons).

The Aachen Day-Night-v1.1 [79, 80, 110] dataset. contains 6,697 high-quality training/database and 1015 test/query images from the old inner city of Aachen, Germany. The database images are taken under daytime conditions using handheld cameras. The query images are captured with three mobile phones at day and at night. This dataset represents a handheld scenario similar to augmented or mixed reality applications in city-scale environments.

The RobotCar Seasons [60, 79] dataset. is based on a subset of the RobotCar dataset [60], captured in the city of Oxford, UK. The training sequences (26,121 images) are captured during daytime, the query images (11,934) are captured during different traversals and under changing weather, time of the day, and seasonal conditions. In contrast to the other two datasets used, the RobotCar dataset contains multiple synchronized cameras and the images are provided in sequences. However, in our benchmark we did not use this additional information. Note that integrating this information in the benchmark can be an interesting follow up of this paper.

The Baidu Mall [89] dataset. was captured in a modern indoor shopping mall in China. It contains 689 training images captured with high resolution cameras in the empty mall and about 2,300 mobile phone query images taken a few months later while the mall was open. The images were semi-manually registered into a LIDAR scan in order to obtain the ground truth poses. The query images are of much poorer quality compared to the database images. In contrary to the latter, where all images were taken in parallel or perpendicular with respect to the main corridor of the mall, query images were taken from more varying viewpoints. Furthermore, the images contain reflective and transparent surfaces, moving people, and repetitive structures which are all important challenges for visual localization and image retrieval.

B.2. Global image representations for retrieval

Our benchmark compares the following 4 popular image representations using the models provided by the authors.

DenseVLAD[7] **[96].** To obtain the DenseVLAD image representation for an image, first RootSIFT [3, 57] descriptors are extracted on a multi-scale (we used 4 different scales corresponding to region widths of 16, 24, 32 and 40 pixels), regular, densely sampled grid, and then aggregated into an intra-normalized VLAD [38] descriptor followed by PCA (principle component analysis) compression, whitening, and L2 normalization [37]. We also used DenseVLAD-mono, which is a variant where the local features are extracted only on a single scale (with the region width equal to 24).

NetVLAD[8] **[1].** The main component of the NetVLAD architecture is a generalized VLAD layer that aggregates mid-level convolutional features extracted from the entire image into a compact single vector representation for efficient indexing similarly to VLAD [38]. The resulting aggregated representation is then compressed using PCA to obtain a final compact descriptor of the image. NetVLAD is trained with geo-tagged image sets consisting of groups of images taken from the same locations at different times and seasons, allowing the network to discover which features are useful or distracting and what changes should the image representation be robust to. This makes NetVLAD very interesting for the visual localization pipeline and motivated our choice to integrate it in our benchmark. Furthermore, NetVLAD has been used in state-of-the-art localization pipelines [30, 74] and in combination with D2-Net [27].

AP-GeM[9] **[71].** This model, similarly to [69], uses a generalized-mean pooling layer (GeM) to aggregate CNN-based descriptors of several image regions at different scales but instead of a contrastive loss, it directly optimizes the Average Precision (AP) approximated by histogram binning to make it differentiable. It is one of the state-of-the art image representation on popular landmark retrieval benchmarks (*e.g.*, $\mathcal{R}$Oxford and $\mathcal{R}$Paris [68]). The model we used was trained on the Google Landmarks v1 dataset (GLD) [63], where each training image has a class label based on the landmark contained in the image. AP-GeM has been used successfully used for visual localization in [35].

DELG[10] **[14].** DELG is designed to extract local and global

[7] Code available at http://www.ok.ctrl.titech.ac.jp/~torii/project/247/.

[8] Matlab code and pretrained models are available at https://github.com/Relja/netvlad. We used the VGG-16-based NetVLAD model trained on Pitts30k [1].

[9] Pytorch implementation and models are available at https://europe.naverlabs.com/Research/Computer-Vision/Learning-Visual-Representations/Deep-Image-Retrieval/. We used the Resnet101-AP-GeM model trained on Google Landmarks [63] to extract image representations.

[10] We used the TensorFlow code publicly available at https://github.com/tensorflow/models/tree/master/

features using one CNN. After a common backbone, the model is split into two parts (heads) one to detect relevant local features and one which describes the global content of the image as a compact descriptor. The two networks are jointly trained in an end-to-end manner using the ArcFace [25] loss for the compact descriptor and leveraging the Google Landmark v1 [63] dataset, which provides image-level labels. The method was originally designed for image search, where the local features enable geometric verification and re-ranking.

Our choice of VLAD with densely extracted features (DenseVLAD, NetVLAD) is based on [90, 97]. It is shown that DenseVLAD and DenseFV (Fisher Vectors) significantly outperform SparseVLAD and SparseFV (based on local features) under strong illumination changes as using densely extracted features eliminates potential repeatability problems of feature detectors. Furthermore, [97] reports that DenseVLAD performs on par with advanced sparse bag of visual words representations. At the same time, both DenseVLAD and NetVLAD are used in state-of-the-art localization pipelines [74, 90, 93]. Investigating aggregation of modern local features (e.g. D2-Net, R2D2) via ASMK [94], VLAD [38] or FV [64] is an interesting research direction with practical benefits.

B.3. SFM pipeline with and without a global model

For our global SFM experiments, we created an SFM model for each local feature type considered (SIFT, D2Net and R2D2). To not introduce a bias towards one specific global feature type and because matching all possible training image pairs would potentially introduce noise and would require large computational resources, we selected the image pairs to match by their frustum overlaps. In detail, we fitted a sphere into the overlapping space of two frusta (we used 50m as maximum distance) and used the radius of this sphere as our measure for image overlap. For Aachen Day-Night and Baidu, we used all image pairs with an overlapping-sphere-radius of 10m or more. For RobotCar, this threshold resulted in too many image pairs to process, thus we only used the 50 most overlapping pairs. To generate the 3D model we triangulated the 3D points from the local feature matches (obtained from the image pairs) using the provided camera poses of the training images and applied bundle adjustment for global optimization. In order to localize an image within this map, we match the query images with the top k retrieved database images and use PNP [44, 46] to register them within the map.

Our local SFM experiments are inspired by the SFM-on-the-fly approach from [100] where the retrieved database images are used to create a small SFM map on-the-fly and the query images are registered within this map using PNP. The SFM pipeline is the same as described above with the difference that the database image pairs to match are generated using all possible pairwise combinations of the retrieved images.

In both cases, for SFM and query image registration, we used COLMAP[11] [83].

`research/delf/delf/python/delg`, and the model that similarly to AP-GeM has a ResNet101 backbone architecture and was trained on GLD [63]

[11] Code available at https://colmap.github.io/.

B.4. Evaluation metrics

To increase the readability of the following experimental results, we first recall the evaluation metrics used in the main paper.

Visual localization metrics.. To measure localization performance, we follow common practice from the literature [42, 79, 85]. Let $\mathtt{R} \in \mathbb{R}^{3\times3}$ be the camera rotation and $\mathbf{c} \in \mathbb{R}^3$ be the camera position, *i.e.*, a 3D point $\mathbf{X}_g$ in world coordinates is mapped to local camera coordinates as $\mathbf{X}_l = \mathtt{R}(\mathbf{X}_w - \mathbf{c})$. Following [79], the position and rotation errors between an estimated and the reference pose are defined as

$$c_{\text{error}} = \|\mathbf{c}_{\text{estimated}} - \mathbf{c}_{\text{reference}}\|_2 \ , \quad (3)$$

$$R_{\text{error}} = \arccos\left(\frac{\text{trace}\left(\mathtt{R}_{\text{estimated}}^{-1} \cdot \mathtt{R}_{\text{reference}}\right) - 1}{2}\right) \ , \quad (4)$$

where R_{error} is the angle of the smallest rotation aligning $\mathtt{R}_{\text{estimated}}$ and $\mathtt{R}_{\text{reference}}$.

For evaluation, we use the percentage of images localized within a given error threshold $(X\text{m}, Y^\circ)$ [79, 85], *i.e.*, the percentage of query images for which $c_{\text{error}} < X$ and $R_{\text{error}} < Y$. Following [79], for all three datasets we use three different threshold pairs for evaluating **low** (5m, 10°), **medium** (0.5m, 5°), and **high** (0.25m, 2°) accuracy localization.

Retrieval metrics.. As indicated in the main paper, (index, retrieved image) pairs are considered positive based on either the IoU similarity or the distance between the camera poses. We use them to compute both *Precision@k* (P@k) and *Recall@k* (R@k). In the case of IoU similarity, we use the presence of overlapping triangulated 3D points between query and database images to classify a pair as positive. In the case of pose similarity, we consider a database image as relevant if it was taken within a neighborhood of the query image [2, 76, 96, 98]. When the camera orientation is available (which is true in our case), the angle between the cameras' orientations can also be taken into account. Following [98], we consider two images relevant if they were taken within 25 meters from each other and if the camera orientations are less than 45 degrees apart. While in the main paper we mainly focused on IoU measure-based relevance, here we provide P@k / R@k results obtained with both relevance measures.

C. Additional experimental results

This section presents results that complement the main paper. In Section C.1, we present the retrieval and place recognition results (*c.f.* Sec. 3.2, Place recognition, in the main paper). We compare the results obtained with ground truth image relevance defined based on visual overlap with the ones defined based on pose proximity. In Section C.2, we present further details on our experiments on pose approximation. In Section C.3, we provide results for various localization accuracy settings with SFM maps built with different local feature types. Finally, in Section C.4, we study the correlation between visual localization and place recognition/landmark retrieval metrics as well as the correlation between the interpolation-based localization results and *Recall@k*.

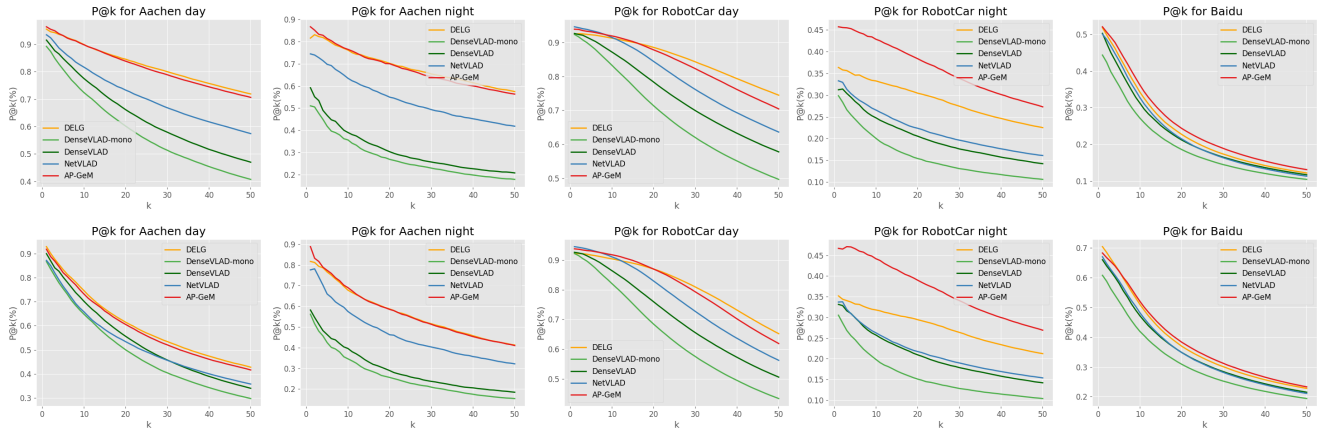

Figure 6. **P@k curves**. Landmark retrieval result where the ground truth relevance between two images was defined by visual overlap (top row) respectively pose similarity (bottom row).

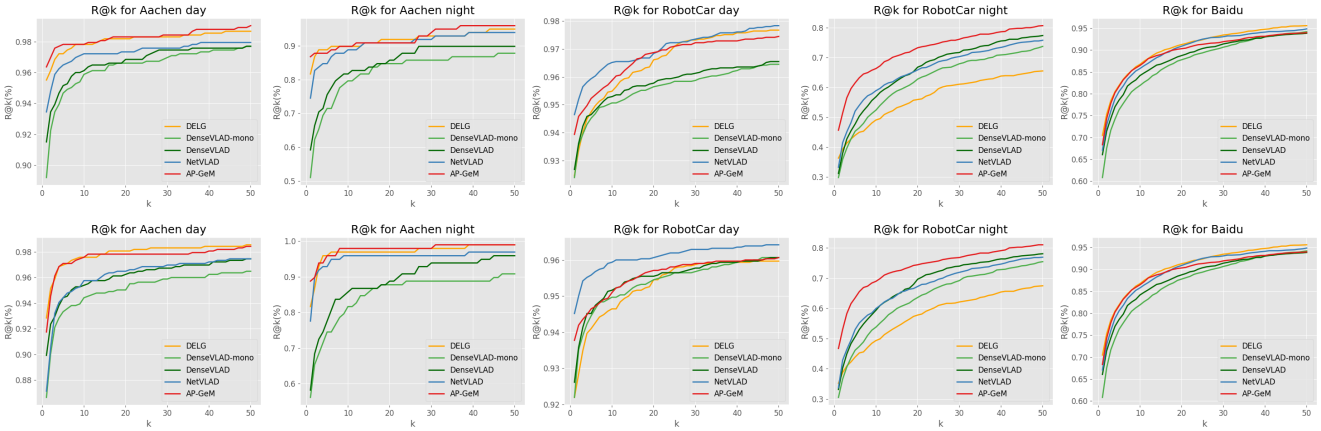

Figure 7. **R@k curves**. Place recognition results where the ground truth relevance between two images is defined by visual overlap (top row) respectively pose proximity (bottom row).

C.1. Landmark retrieval and place recognition

As discussed in the main paper, we use two retrieval metrics from the literature: *Precision@*k (P@k) and *Recall@*k (R@k) where we considered two different ways to define whether two images are related / relevant to each other. One is based on visual overlap (measured using jointly visible 3D points) and the other is based on pose proximity (whether the two images were roughly taken from the same position and, if available, orientation).

In Fig. 6, we show results for the landmark retrieval scenario evaluated with the P@k metric considering ground truth relevance scores obtained either with visual overlap or with pose proximity. Similarly, Fig. 7 shows results for the place recognition scenario measured by R@k using the two relevance scores. Overall, we can see that the results with both relevance scores are very similar yielding to the same rankings between image representations. This shows that the exact definition of the relevance score between two images does not play a major role when evaluating which retrieval representation works well for landmark retrieval and place recognition tasks.

C.2. Task 1: Pose approximation

In the main paper, we show pose approximation results for the three approaches (EWB, BDI, and CSI) considering the low localization accuracy threshold (5m, 10°). In Fig. 8, here we show pose approximation results for RobotCar day for the medium accuracy threshold (0.5m, 5°). We only show results for this dataset as for the other datasets less than 1-2% of the queries were successfully localized with medium accuracy. The better results obtained for this dataset are due to the small variation in camera pose between query and reference images. Note that the lower performance for the RobotCar nighttime queries is due to the lower quality of the nighttime images (which exhibit strong motion blur) and the challenge of bridging the appearance gap between day- and nighttime images. Interestingly, our results show that for the medium accuracy threshold, NetVLAD and DenseVLAD perform better than DELG and AP-GeM. This is in contrast to the low accuracy threshold where DELG and AP-GeM perform better (see main paper) and means that more images were successfully localized with DELG and AP-GeM but not very accurately.

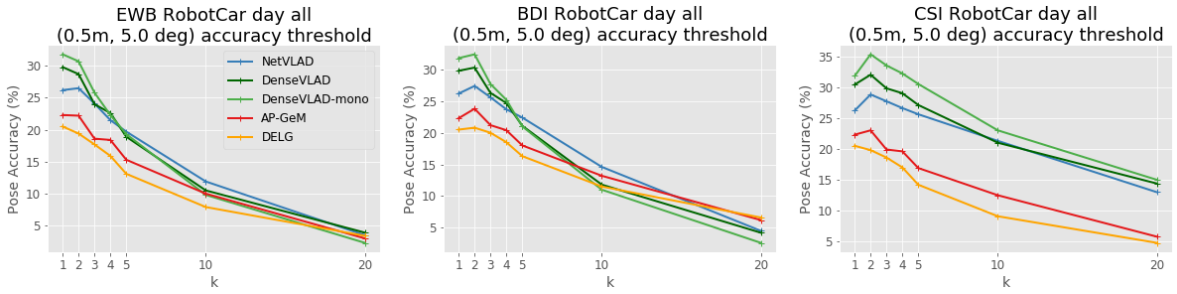

Figure 8. **Task 1: Pose approximation with medium accuracy threshold**. We only show results for RobotCar day as less than 1-2% of the queries were localized for the others. On these plots, we observe that NetVLAD and DenseVLAD outperform DELG and AP-GeM.

C.3. Task 2: Accurate pose estimation

As discussed in Section 4 of the main paper, in the following we present the full set of results for **Task 2a** and **Task 2b** considering all three accuracy thresholds and using three different local feature types to build the SFM maps, namely R2D2 [72], D2-Net [27], and SIFT [57].

Figures 9 to 11 show results for Task 2a (pose estimation without global map) and Figures 12 to 14 show results for Task 2b (pose estimation with global map).

We first observe that the choice of local features does not change the ranking of image representations for any of the three datasets considered. Furthermore, as already observed in [27, 72], SIFT is outperformed by R2D2 and D2-Net. As a consequence, we observe that the performance gap between different image representations shrinks when using SIFT: retrieving a relevant image does not help if the local feature type used is unable to establish good correspondences. We can therefore confirm the statement of the main paper that our conclusions are not tied to a specific type of local features.

Similarly, while increasing the accuracy threshold decreases the number of localized images, the rank according to the global representations does not change. Nevertheless, we observe that the gap between deep representations and DenseVLAD is smaller for higher accuracy, which suggest that the images retrieved with DenseVLAD, when relevant, are well suited for localization.

C.4. Visual localization versus retrieval metrics

This section analyses the correlation between typical visual localization (measured via the percentage of the images retrieved at a given accuracy threshold) and retrieval metrics (measured by precision and recall). We only consider the lowest accuracy threshold for the following results. On the one hand we observed that the ranking is similar for various localization accuracy thresholds. On the other hand, high pose accuracy is not necessary for both landmark retrieval and place recognition. To compute the retrieval metrics, we only use the visual overlap based ground truth relevance because, as we see in Section C.1, the two strategies yield similar results. Finally, concerning local features we only show correlations with the results obtained with the R2D2 SFM maps for the accurate localization scenarios.

To analyse the correlation, we generate scatter plots where we select pairs of (Pose Accuracy, Retrieval Metric) for corresponding top k retrieved images. Figure 15 shows scatter plots for localization based on a global SFM map. As we see in the main paper, there appears to be a clear correlation between the R@k and pose accuracy for this task. On the other hand, the Precision at top k (P@k) does not seem to correlate with localization performance confirming that it is not necessary that all top retrieved images are relevant for a query as long as a few relevant images are found.

Figure 16 shows scatter plots for localization based on the local SFM approach. We can observe a weaker correlation between this method and the Recall at k (R@k) than when using a global map. Again, we do not observe an obvious correlation with the Precision at k (P@k).

Finally, Figure 17 shows scatter plots for pose approximation via the EWB pose interpolation scheme. We observe a weak correlation between the pose estimation and the P@k measures, but no correlation, or even in several cases an inverse correlation, with R@k. This seems to be opposite to the trends observed for **Task 2a/b**. An interesting thing to notice is that the points associated to DenseVLAD in the P@k plots are consistently on the left of the other methods. For equal P@k, DenseVLAD achieves better pose accuracy, *i.e.*, retrieves images closer to the query. This confirms that DenseVLAD is less robust to viewpoint changes and thus retrieves closer images.

Overall, we notice some correlation between P@k and pose accuracy for **Task 1**, and R@k and pose accuracy for **Task 2a/b**. The correlation is clearest between R@k and pose accuracy for **Task 2b**, where it appears in every dataset. These correlations are understandable: to have a good pose approximation, all of the retrieved images need to be close enough to the query. A single database image taken far away from the query, even with low weight in the interpolation scheme, can significantly affect the accuracy of the pose approximation. For the geometry-based approaches, the correlation with R@k can be explained by the fact that the localization pipeline does filtering based on the local feature matches. Thus, the system is less sensitive to wrong retrievals which will be filtered out if good local feature correspondences can be estimated. Still, at least one correctly retrieved image is necessary to facilitate pose estimation.

References

[1] R. Arandjelović, P. Gronát, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. Dislocation: Scalable Descriptor Distinctiveness for Location

Recognition. In *Asian Conference on Computer Vision (ECCV)*, 2014.

[3] R. Arandjelović and A. Zisserman. Three Things Everyone Should Know to Improve Object Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[4] R. Arandjelović and A. Zisserman. All about VLAD. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[5] C. Arth, D. Wagner, M. Klopschitz, A. Irschara, and D. Schmalstieg. Wide Area Localization on Mobile Phones. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* , 2009.

[6] Y. Avrithis, Y. Kalantidis, G. Tolias, and E. Spyrou. Retrieving Landmark and Non-Landmark Images from Community Photo Collections. In *ACM Multimedia*, 2010.

[7] A. Babenko and V. Lempitsky. Aggregating Deep Convolutional Features for Image Retrieval. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[8] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural Codes for Image Retrieval. In *European Conference on Computer Vision (ECCV)*, 2014.

[9] V. Balntas, S. Li, and V. Prisacariu. RelocNet: Continuous Metric Learning Relocalisation Using Neural Nets. In *European Conference on Computer Vision (ECCV)*, 2018.

[10] E. Brachmann and C. Rother. Learning Less Is More - 6D Camera Localization via 3D Surface Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] E. Brachmann and C. Rother. Expert sample consensus applied to camera re-localization. In *ICCV*, 2019.

[12] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. Kautz. Geometry-Aware Learning of Maps for Camera Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] J. Brejcha and M. Čadík. State-of-the-art in Visual Geolocalization. *Pattern Analysis and Applications (PAA)*, 20(3):613–637, 2017.

[14] B. Cao, A. Araujo, and J. Sim. Unifying Deep Local and Global Features for Efficient Image Search. *arXiv*, 2001.05027, 2020.

[15] S. Cao and N. Snavely. Graph-Based Discriminative Learning for Location Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[16] R. O. Castle, G. Klein, and D. W. Murray. Video-rate localization in multiple maps for wearable augmented reality. In *ISWC*, 2008.

[17] T. Cavallari, L. Bertinetto, J. Mukhoti, P. Torr, and S. Golodetz. Let's take this online: Adapting scene coordinate regression network predictions for online rgb-d camera relocalisation. In *3DV*, 2019.

[18] T. Cavallari, S. Golodetz, N. Lord, J. Valentin, V. Prisacariu, L. Di Stefano, and P. H. S. Torr. Real-Time RGB-D Camera Pose Estimation in Novel Scenes using a Relocalisation Cascade. *TPAMI*, 2019.

[19] D. Chen, G. Baatz, K. Köser, S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale Landmark Identification on Mobile Devices. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[20] O. Chum and J. Matas. Optimal Randomized RANSAC. *PAMI*, 30(8):1472 –1482, 2008.

[21] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the World's Photos. In *International Conference on World Wide Web (WWW)*, 2009.

[22] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual Categorization with Bags of Keypoints. In *ECCV Workshops* , 2004.

[23] G. Csurka, C. Dance, and M. Humenberger. From Handcrafted to Deep Local Invariant Features. *arXiv*, 1807.10254, 2018.

[24] Q. Cui, V. Fragoso, C. Sweeney, and P. Sen. GraphMatch: Efficient Large-Scale Graph Construction for Structure from Motion. In *International Conference on 3D Vision*, 2017.

[25] J. Deng, J. Guo, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. 01 2018.

[26] M. Ding, Z. Wang, J. Sun, J. Shi, and P. Luo. CamNet: Coarse-to-Fine Retrieval for Camera Re-Localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[27] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler. D2-Net: a Trainable CNN for Joint Description and Detection of Local Features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[28] M. Fischler and R. Bolles. Random Sampling Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography. *CACM*, 24:381–395, 1981.

[29] E. Garcia-Fidalgo and A. Ortiz. Vision-based Topological Mapping and Localization Methods: A Survey. *ACM Conference on Recommender Systems (CRS)*, 64(2):1—-20, 2015.

[30] H. Germain, G. Bourmaud, and V. Lepetit. Sparse-to-Dense Hypercolumn Matching for Long-Term Visual Localization. In *International Conference on 3D Vision (3DV)*, 2019.

[31] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. End-to-End Learning of Deep Visual Representations for Image Retrieval. *International Journal of Computer Vision (IJCV)*, 124:237—-254, 2017.

[32] J. Hays and A. Efros. IM2GPS: Estimating Geographic Information from a Single Image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[33] J. Heinly, J. L. Schönberger, E. Dunn, and J. M. Frahm. Reconstructing the world* in six days. In *CVPR*, 2015.

[34] L. Heng, B. Choi, Z. Cui, M. Geppert, S. Hu, B. Kuan, P. Liu, R. Nguyen, Y. C. Yeo, A. Geiger, G. H. Lee, M. Pollefeys, and T. Sattler. Project AutoVision: Localization and 3D Scene Perception for an Autonomous Vehicle with a Multi-Camera System. In *International Conference on Robotics and Automation (ICRA)*, 2019.

[35] M. Humenberger, Y. Cabon, N. Guerin, J. Morat, J. Revaud, P. Rerole, N. Pion, C. de Souza, V. Leroy, and G. Csurka.

Robust image retrieval-based visual localization using kapture. *arXiv 2007.13867*, 2020.

[36] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From Structure-from-Motion Point Clouds to Fast Location Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[37] H. Jégou and O. Chum. Negative Evidences and Cooccurrences in Image Retrieval: the Benefit of PCA and Whitening. In *European Conference on Computer Vision (ECCV)*, 2012.

[38] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating Local Descriptors Into a Compact Image Representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[39] Y. Kalantidis, C. Mellina, and S. Osindero. Cross-dimensional Weighting for Aggregated Deep Convolutional Features. In *ECCV Workshops* , 2016.

[40] Y. Kalantidis, G. Tolias, Y. Avrithis, M. Phinikettos, E. Spyrou, P. Mylonas, and S. Kollias. VIRaL: Visual Image Retrieval and Localization. *Multimedia Tools and Applications (MTA)*, 74(9):3121–3135, 2011.

[41] A. Kendall and R. Cipolla. Geometric Loss Functions for Camera Pose Regression with Deep Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[42] A. Kendall, M. Grimes, and R. Cipolla. PoseNet: a Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[43] H. Kim, E. Dunn, and J.-M. Frahm. Learned Contextual Feature Reweighting for Image Geo-Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[44] L. Kneip, D. Scaramuzza, and R. Siegwart. A Novel Parametrization of the Perspective-three-point Problem for a Direct Computation of Absolute Camera Position and Orientation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[45] J. Knopp, J. Sivic, and T. Pajdla. Avoding Confusing Features in Place Recognition. In *ECCV*, 2010.

[46] Z. Kukelova, M. Bujnak, and T. Pajdla. Real-Time Solution to the Absolute Pose Problem with Unknown Radial Distortion and Focal Length. In *ICCV*, 2013.

[47] V. Larsson, Z. Kukelova, and Y. Zheng. Making Minimal Solvers for Absolute Pose Estimation Compact and Robust. In *ICCV*, 2017.

[48] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala. Camera Relocalization by Computing Pairwise Relative Poses Using Convolutional Neural Network. In *ICCV Workshops* , 2017.

[49] K. Lebeda, J. E. S. Matas, and O. Chum. Fixing the Locally Optimized RANSAC. In *British Machine Vision Conference (BMVC)*, 2012.

[50] Y. Li, D. Crandall, and D. Huttenlocher. Landmark Classification in Large-Scale Image Collections. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[51] Y. Li, N. Snavely, and D. Huttenlocher. Location Recognition Using Prioritized Feature Matching. In *European Conference on Computer Vision (ECCV)*, 2010.

[52] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3d point clouds. In *ECCV*, pages 15–29. Springer, 2012.

[53] Y. Li, N. Snavely, and D. P. Huttenlocher. Location Recognition using Prioritized Feature Matching. In *ECCV*, 2010.

[54] H. Lim, S. Sinha, M. Cohen, M. Uyttendaele, and H. Kim. Real-time Monocular Image-based 6-DoF Localization. *International Journal of Robotics Research*, 34(4–5):476–492, 2015.

[55] L. Liu, H. Li, and Y. Dai. Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[56] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2), 2004.

[57] D. Lowe. Distinctive Image Features from Scale-invariant Keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[58] S. Lowry, N. Sünderhauf, P. Newman, J. Leonard, D. Cox, P. Corke, and M. Milford. Visual Place Recognition: A Survey. *Transactions on Robotics*, 32(1):1–19, 2016.

[59] S. Lynen, T. Sattler, M. Bosse, J. Hesch, M. Pollefeys, and R. Siegwart. Get out of my Lab: Large-scale, Real-Time Visual-Inertial Localization. In *Robotics: Science and Systems Conference (RSS)*, 2015.

[60] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *IJRR*, 36(1):3–15, 2017.

[61] D. Massiceti, A. Krull, E. Brachmann, C. Rother, and P. Torr. Random Forests versus Neural Networks - What's Best for Camera Localization? In *International Conference on Robotics and Automation (ICRA)*, 2017.

[62] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt. Scalable 6-DoF Localization on Mobile Devices. In *European Conference on Computer Vision (ECCV)*, 2014.

[63] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-Scale Image Retrieval with Attentive Deep Local Features. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[64] F. Perronnin and C. Dance. Fisher Kernels on Visual Vocabularies for Image Categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[65] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[66] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[67] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet. A Survey on Visual-Based Localization: On the Benefit of Heterogeneous Data. *Pattern Recognition*, 74(2):90–109, 2018.

[68] F. Radenović, A. Iscen, G. Tolias, and O. Avrithis, Yannis Chum. Revisiting Oxford and Paris: Large-scale Image Retrieval Benchmarking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[69] F. Radenović, G. Tolias, and O. Chum. Fine-Tuning CNN Image Retrieval with no Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 41(7):1655–1668, 2019.

[70] A. Razavian, J. Sullivan, S. Carlsson, and A. Maki. Visual Instance Retrieval with Deep Convolutional Networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2015.

[71] J. Revaud, J. Almazan, R. S. de Rezende, and C. R. de Souza. Learning with Average Precision: Training Image Retrieval with a Listwise Loss. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[72] J. Revaud, P. Weinzaepfel, C. De Souza, and M. Humenberger. R2D2: Reliable and Repeatable Detectors and Descriptors. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[73] J. Revaud, P. Weinzaepfel, C. R. de Souza, N. Pion, G. Csurka, Y. Cabon, and M. Humenberger. R2D2: repeatable and reliable detector and descriptor. *CoRR*, abs/1906.06195, 2019.

[74] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[75] T. Sattler, M. Havlena, F. Radenovic, K. Schindler, and M. Pollefeys. Hyperpoints and fine vocabularies for large-scale location recognition. In *ICCV*, 2015.

[76] T. Sattler, M. Havlena, K. Schindler, and M. Pollefey. Large-Scale Location Recognition and the Geometric Burstiness Problem. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[77] T. Sattler, B. Leibe, and L. Kobbelt. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(9):1744–1756, 2017.

[78] T. Sattler, B. Leibe, and L. Kobbelt. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *PAMI*, 39(9):1744–1756, 2017.

[79] 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 (CVPR)*, 2018.

[80] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image Retrieval for Image-Based Localization Revisited. In *BMVA British Machine Vision Conference (BMVC)*, 2012.

[81] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixé. Understanding the Limitations of CNN-based Absolute Camera Pose Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[82] G. Schindler, M. Brown, and R. Szeliski. City-Scale Location Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[83] J. Schönberger and J.-M. Frahm. Structure-from-motion Revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[84] J. L. Schönberger, H. Hardmeier, T. Sattler, and M. Pollefeys. Comparative Evaluation of Hand-Crafted and Learned Local Features. In *CVPR*, 2017.

[85] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[86] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2003.

[87] N. Snavely, S. Seitz, and R. Szeliski. Modeling the World from Internet Photo Collections. *IJCV*, 80(2):189–210, 2008.

[88] Stephen Se, D. Lowe, and J. Little. Global localization using distinctive visual features. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2002.

[89] X. Sun, Y. Xie, P. Luo, and L. Wang. A Dataset for Benchmarking Image-based Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[90] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and T. Akihiko. InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[91] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii. InLoc: Indoor visual localization with dense matching and view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[92] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii. InLoc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, 2018.

[93] H. Taira, I. Rocco, J. Sedlar, M. Okutomi, J. Sivic, T. Pajdla, T. Sattler, and A. Torii. Is This the Right Place? Geometric-Semantic Pose Verification for Indoor Visual Localization. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[94] G. Tolias and H. Jégou. Visual Query Expansion with or Without Geometry: Refining Local Descriptors by Feature Aggregation. *Computer Vision and Image Understanding (CVIU)*, 47(10):3466–3476, 2014.

[95] G. Tolias, R. Sicre, and H. Jégou. Particular Object Retrieval with Integral Maxpooling of CNN Activations. In *International Conference on Learning Representations (ICLR)*, 2016.

[96] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 Place Recognition by View Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[97] 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.

[98] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla. Visual Place Recognition with Repetitive Structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 37(11):2346–2359, 2015.

[99] A. Torii, J. Sivic, and T. Pajdla. Visual Localization by Linear Combination of Image Descriptors. In *ICCV Workshops* , 2011.

[100] A. Torii, H. Taira, J. Sivic, M. Pollefeys, M. Okutomi, T. Pajdla, and T. Sattler. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019.

[101] J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg. Global Localization from Monocular SLAM on a Mobile Phone. *IEEE Transactions on Visualization and Computer Graphics*, 20(4):531–539, 2014.

[102] N. Vo, N. Jacobs, and J. Hays. Revisiting IM2GPS in the Deep Learning Era. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[103] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-based Localization Using LSTMs for Structured Feature Correlation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[104] P. Weinzaepfel, G. Csurka, Y. Cabon, and M. Humenberger. Visual Localization by Learning Objects-Of-Interest Dense Match Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[105] T. Weyand, A. Araujo, B. Cao, and J. Sim. Google Landmarks Dataset v2 – A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. *CoRR*, (arXiv:2004.01804), 2020.

[106] L. Yang, Z. Bai, C. Tang, H. Li, Y. Furukawa, and P. Tan. SANet: Scene Agnostic Network for Camera Localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[107] A. Zamir, A. Hakeem, L. Gool, M. Shah, and S. Richard. *Large-Scale Visual Geo-localization*. Advances in Computer Vision and Pattern Recognition. Springer, 2016.

[108] A. R. Zamir and M. Shah. Accurate Image Localization Based on Google Maps Street View. In *European Conference on Computer Vision (ECCV)*, 2010.

[109] W. Zhang and J. Kosecka. Image Based Localization in Urban Environments. In *International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT)*, 2006.

[110] Z. Zhang, T. Sattler, and D. Scaramuzza. Reference Pose Generation for Visual Localization via Learned Features and View Synthesis. *arXiv*, 2005.05179, 2020.

[111] E. Zheng and C. Wu. Structure From Motion Using Structure-Less Resection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[112] L. Zheng, Y. Zhao, S. Wang, J. Wang, and Q. Tian. Good Practice in CNN Feature Transfer. *arXiv*, 1605.06636, 2016.

[113] Q. Zhou, T. Sattler, M. Pollefeys, and L. Leal-Taixé. To Learn or not to Learn: Visual Localization from Essential Matrices. In *International Conference on Robotics and Automation (ICRA)*, 2020.

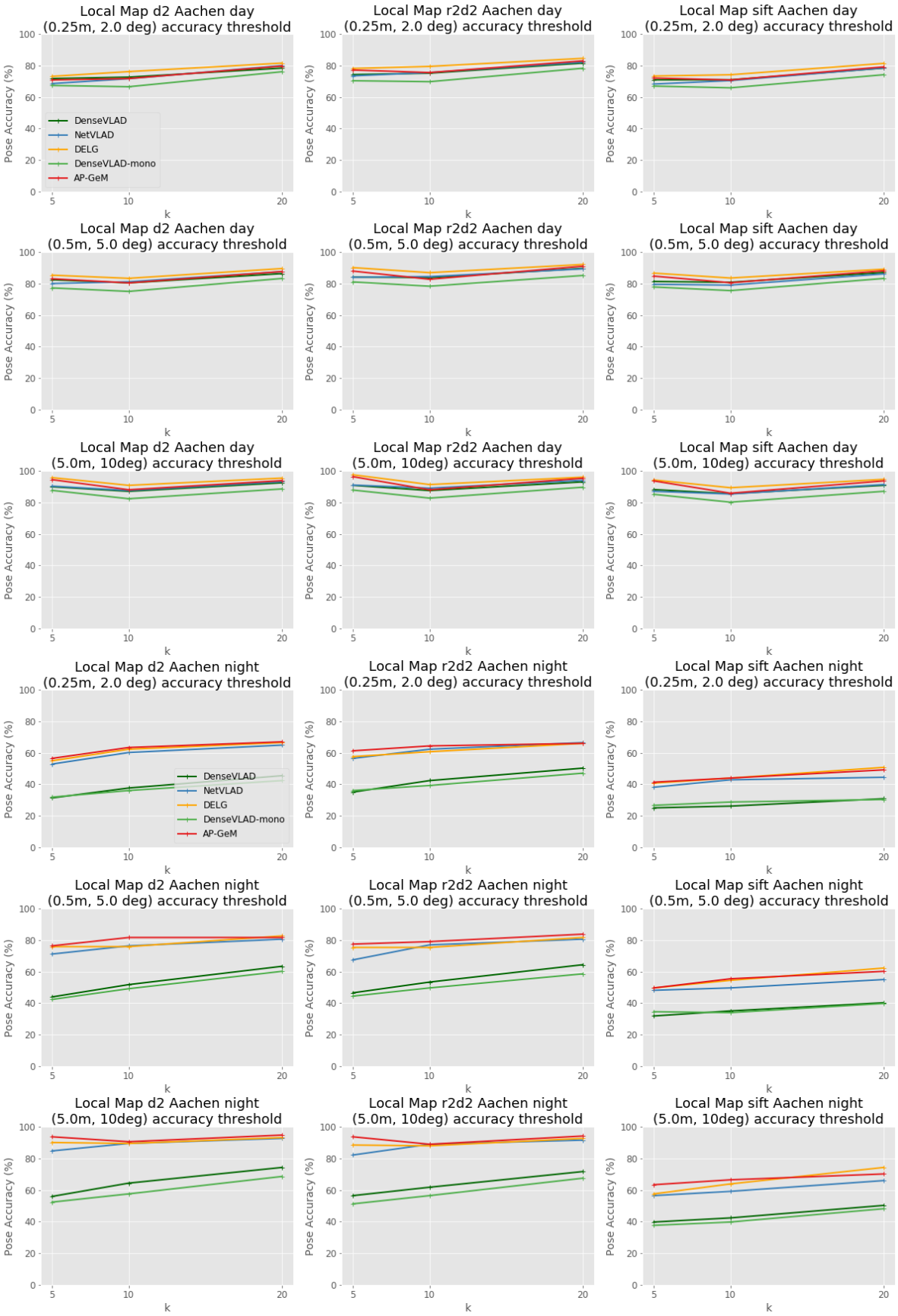

Figure 9. **Task 2a (pose estimation without a global map) results for Aachen Day-Night**. We show results for D2-Net (left column), R2D2 (middle column), and SIFT (right column).

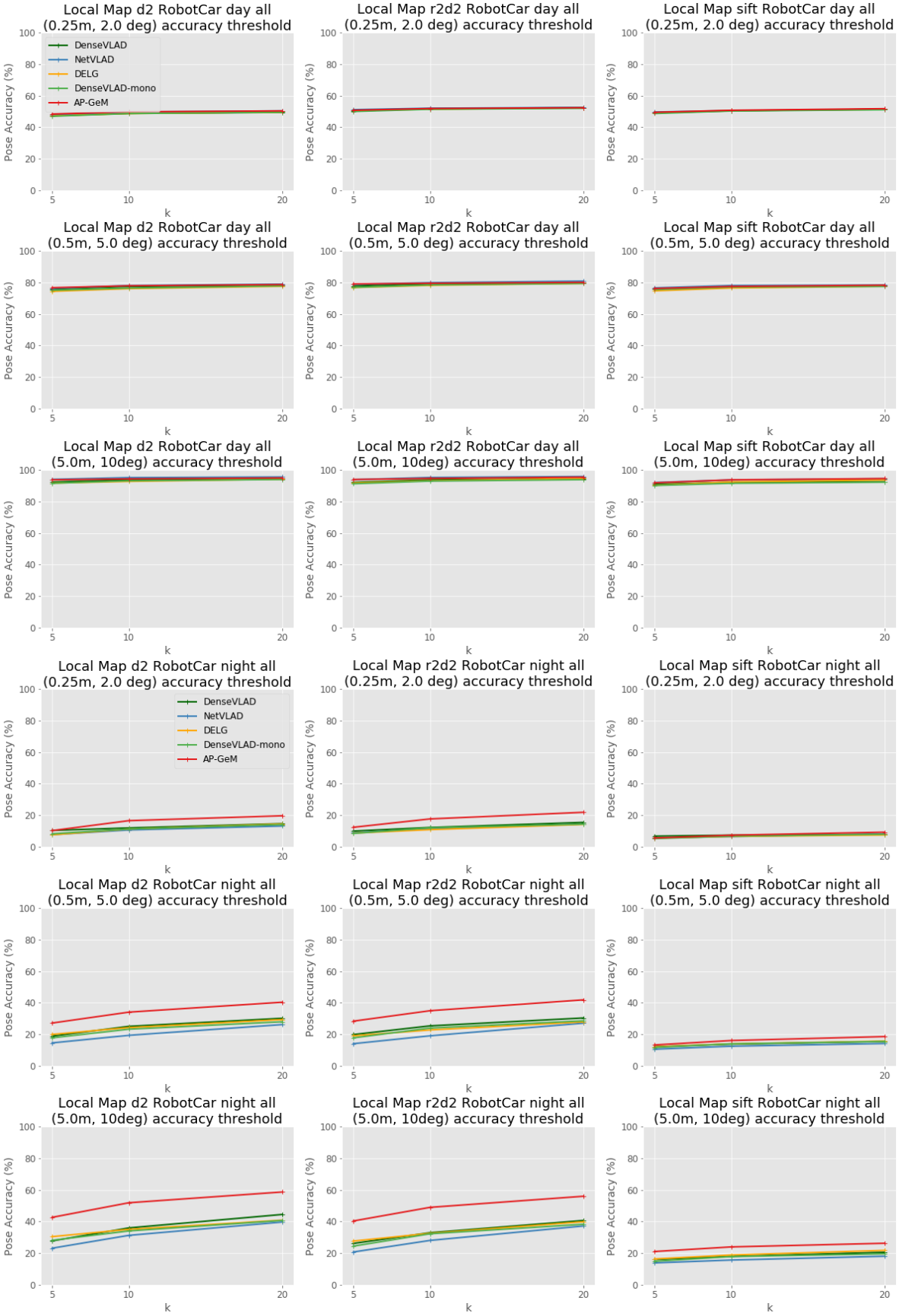

Figure 10. **Task 2a (pose estimation without a global map) results for RobotCar**. We show results for D2-Net (left column), R2D2 (middle column), and SIFT (right column).

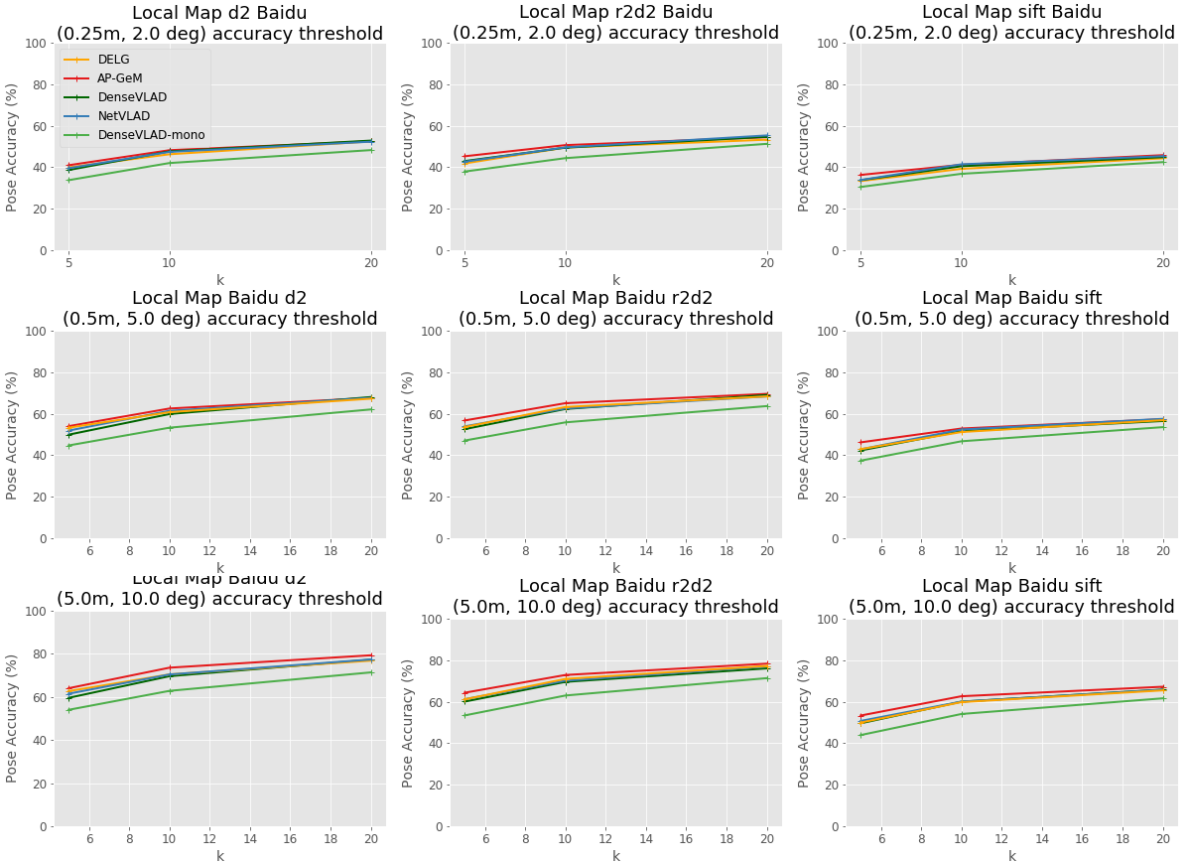

Figure 11. **Task 2a (pose estimation without a global map) results for Baidu**. We show results for D2-Net (left column), R2D2 (middle column), and SIFT (right column).

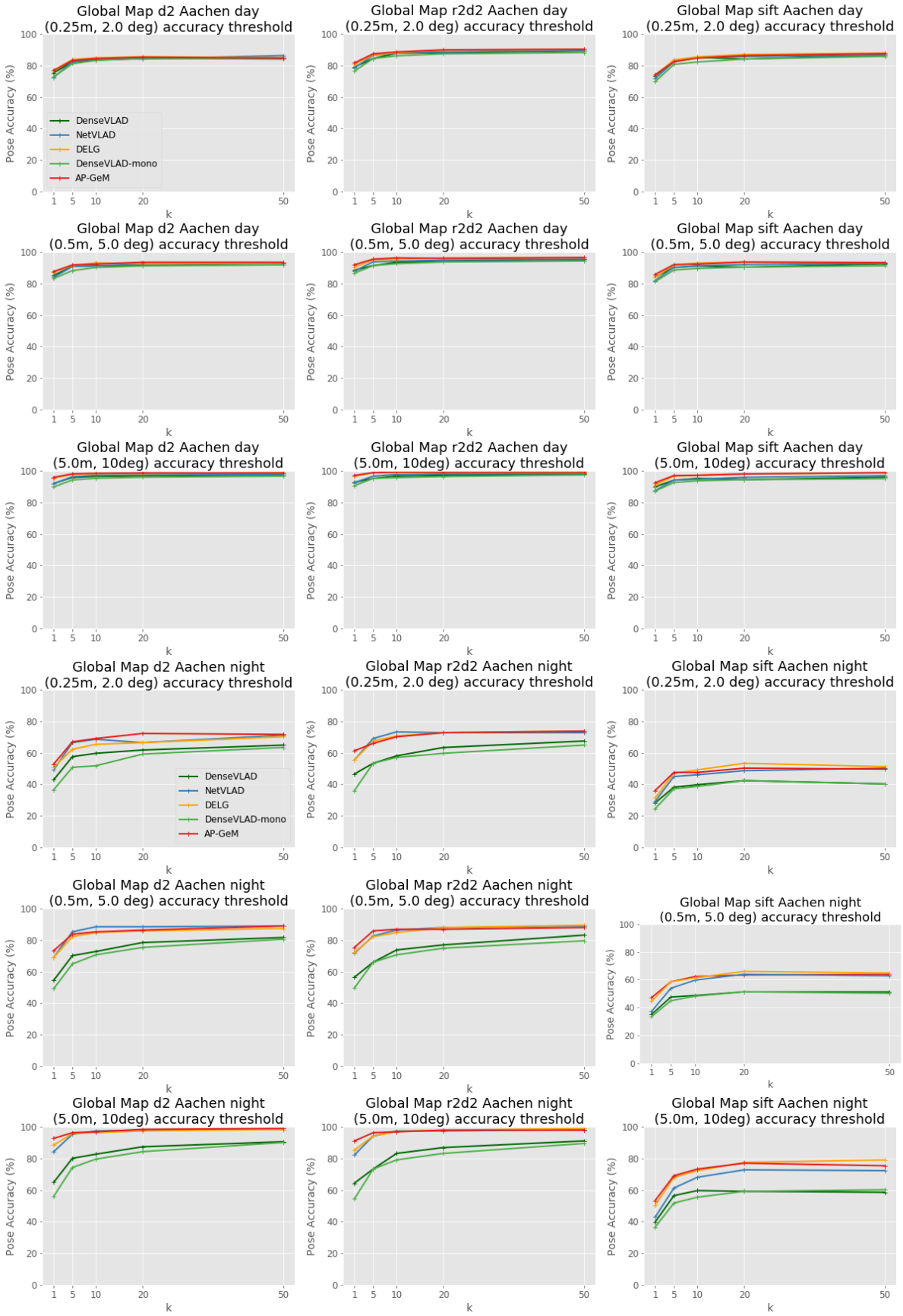

Figure 12. **Task 2b (pose estimation with a global map) results for Aachen Day-Night**. We show results for D2-Net (left column), R2D2 (middle column), and SIFT (right column).

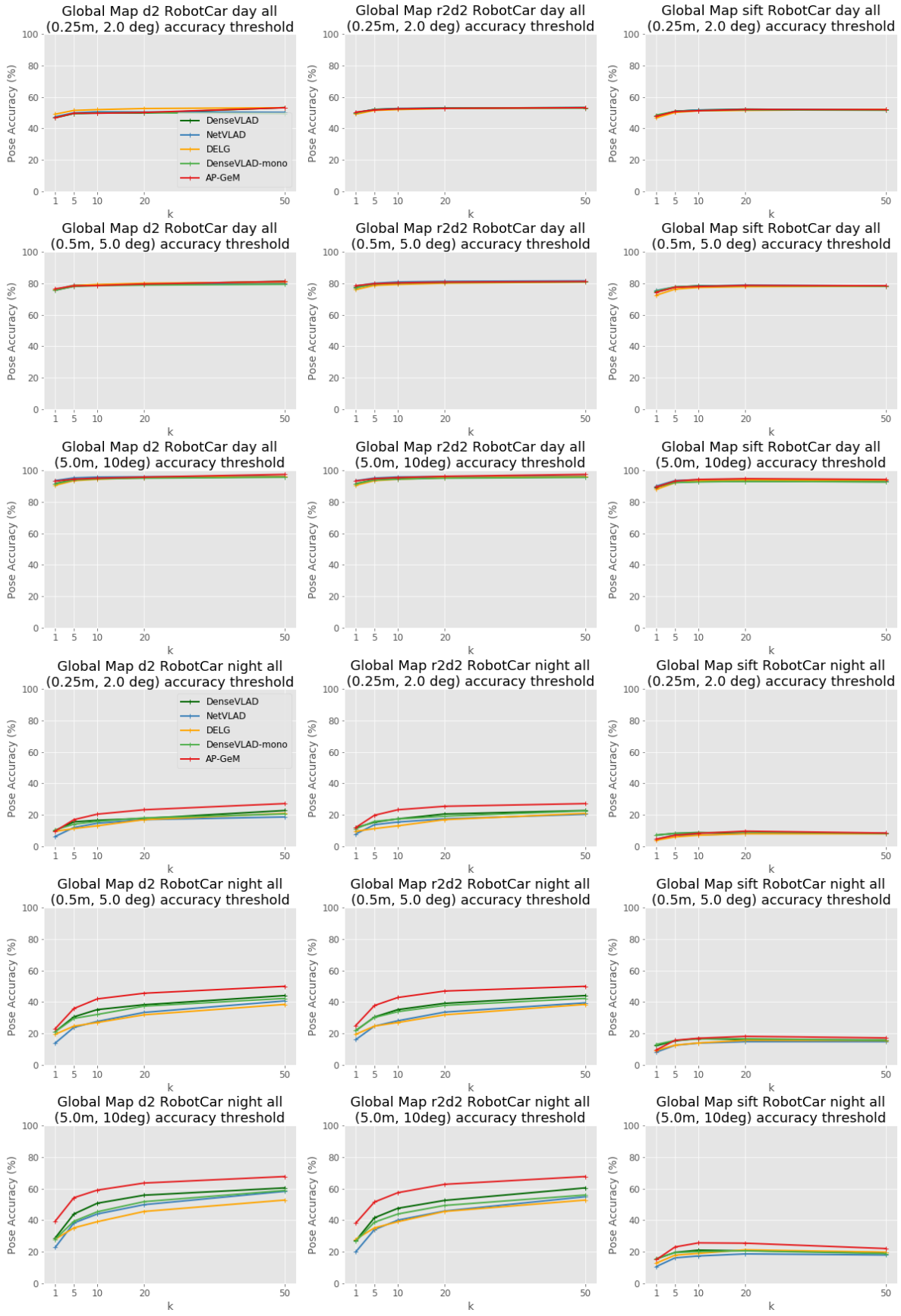

Figure 13. **Task 2b (pose estimation with a global map) results for RobotCar**. We show results for D2-Net (left column), R2D2 (middle column), and SIFT (right column).

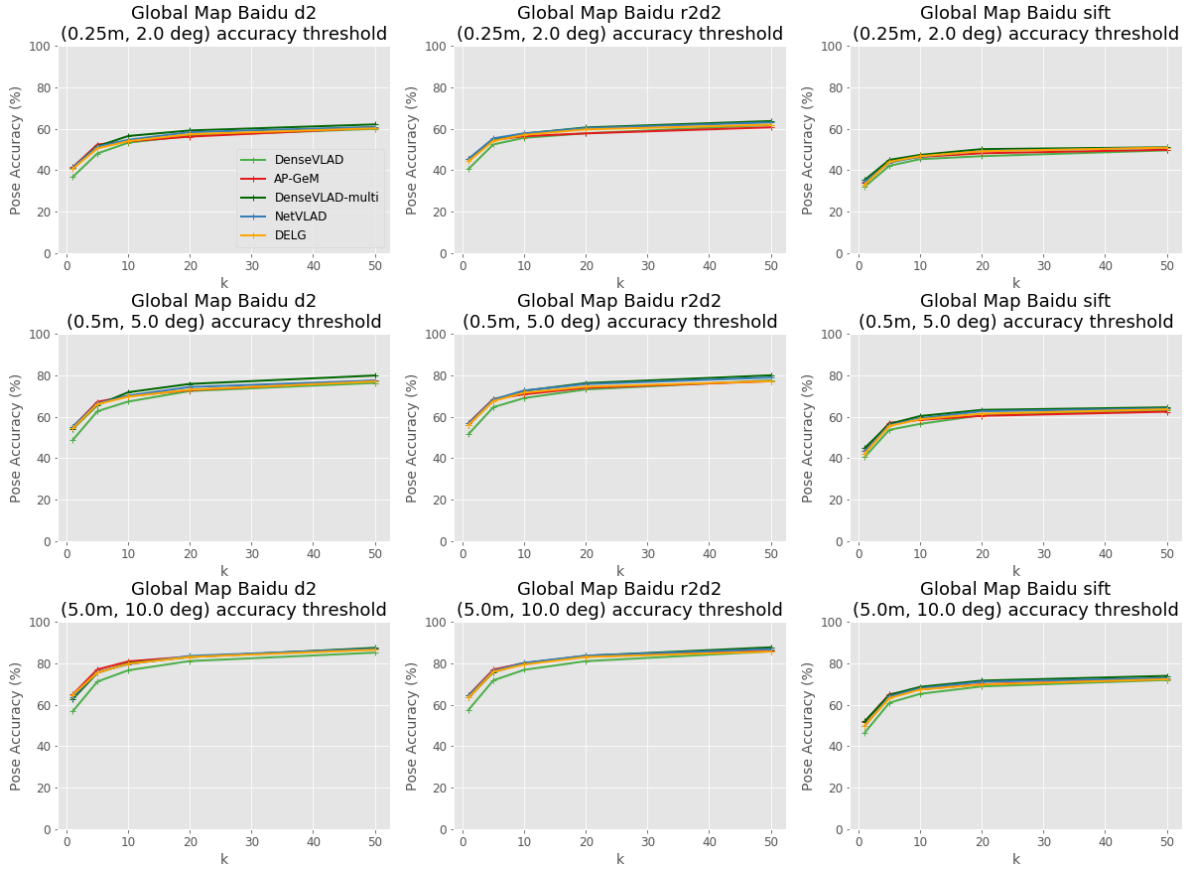

Figure 14. **Task 2b (pose estimation with a global map) results for Baidu**. We show results for D2-Net (left column), R2D2 (middle column), and SIFT (right column).

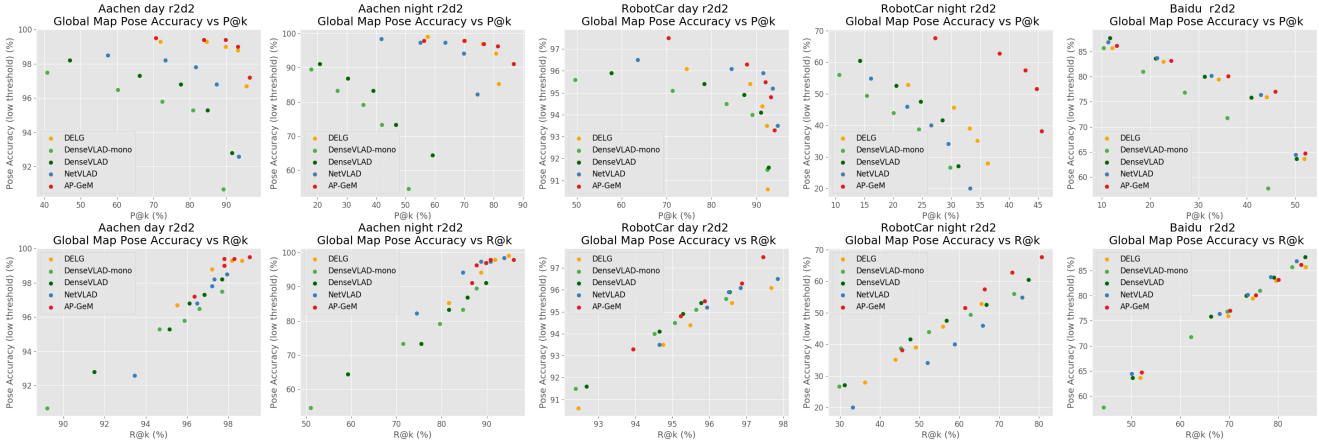

Figure 15. **Task 2b (pose estimation with a global map) vs retrieval metrics**. We see that while there appears to be a clear correlation between R@k and pose accuracy for this task, the Precision at top k (P@k) does not seem to correlate with localization performance.

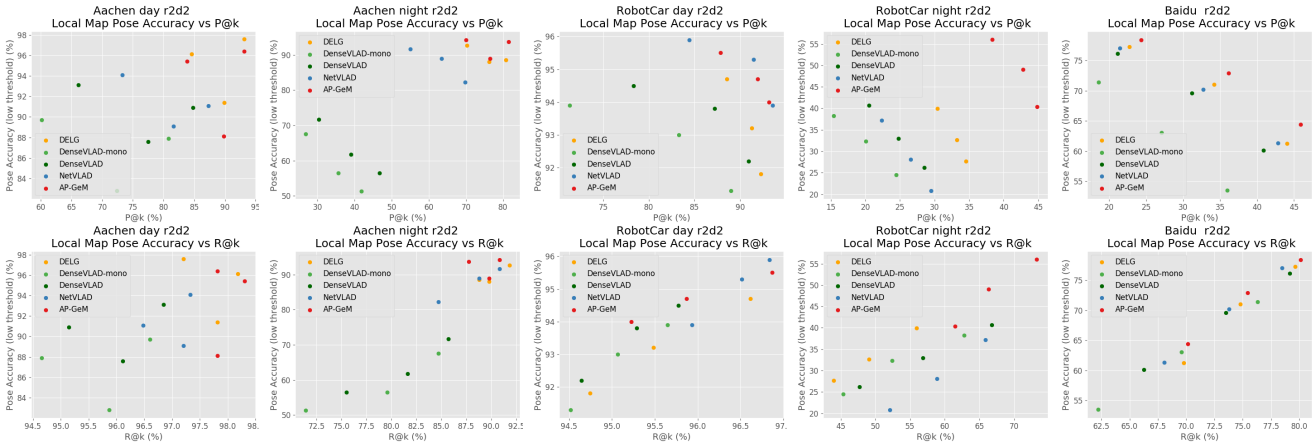

Figure 16. **Task 2a (pose estimation with local map) vs retrieval metrics**. We can observe some weak correlation between pose estimation with local SFM and the Recall at k (R@k), but there is no obvious correlation with the Precision at k (P@k).

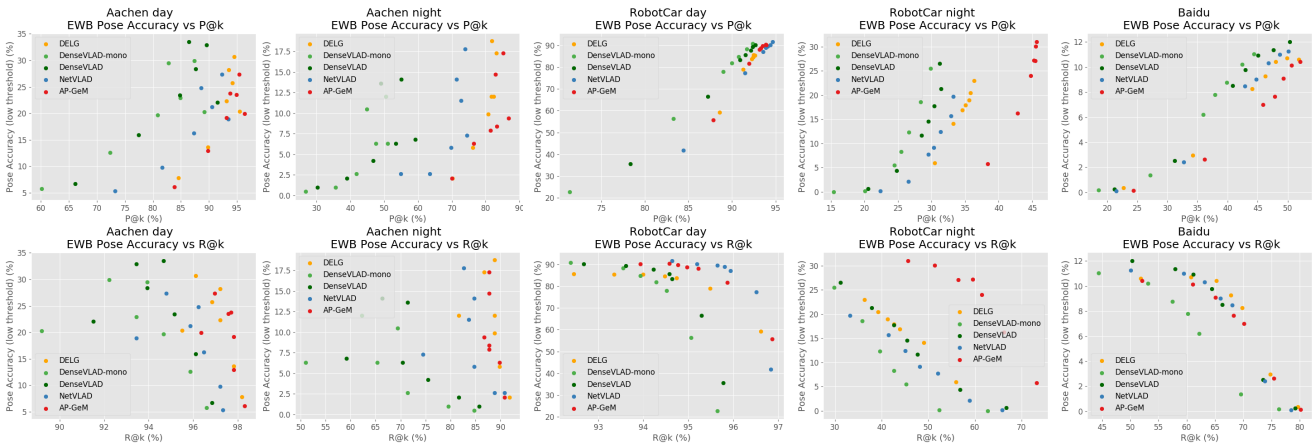

Figure 17. **Task 1a (Approximate localization) vs retrieval metrics**. Here, correlation appears to be between pose estimation based on interpolation and P@k, but not with R@k.