

Yes, we CANN: Constrained Approximate Nearest Neighbors for local feature-based visual localization

Dror Aiger André Araujo Simon Lynen

Google Research

{aigerd, andrearaujo, slynen}@google.com

Abstract

Large-scale visual localization systems continue to rely on 3D point clouds built from image collections using structure-from-motion. While the 3D points in these models are represented using local image features, directly matching a query image’s local features against the point cloud is challenging due to the scale of the nearest-neighbor search problem. Many recent approaches to visual localization have thus proposed a hybrid method, where first a global (per image) embedding is used to retrieve a small subset of database images, and local features of the query are matched only against those. It seems to have become common belief that global embeddings are critical for said image-retrieval in visual localization, despite the significant downside of having to compute two feature types for each query image. In this paper, we take a step back from this assumption and propose Constrained Approximate Nearest Neighbors (CANN), a joint solution of k -nearest-neighbors across both the geometry and appearance space using only local features. We first derive the theoretical foundation for k -nearest-neighbor retrieval across multiple metrics and then showcase how CANN improves visual localization. Our experiments on public localization benchmarks demonstrate that our method significantly outperforms both state-of-the-art global feature-based retrieval and approaches using local feature aggregation schemes. Moreover, it is an order of magnitude faster in both index and query time than feature aggregation schemes for these datasets. Code will be released.

1. Introduction

In this paper we focus on the problem of image retrieval for visual localization. Modern visual localization approaches are predominantly based on 3D point clouds that represent the geometry and appearance of large scale scenes [30, 19, 43, 31]. These 3D points are estimated from image collections using Structure-from-Motion (SfM), where each 3D point has an associated descriptor derived from pixels.

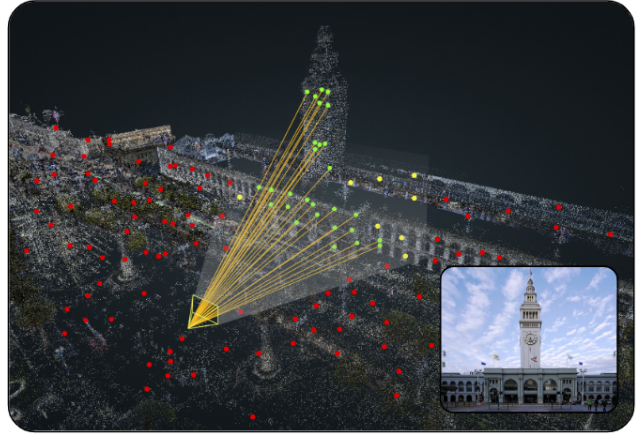


Figure 1: The proposed Constrained Approximate Nearest Neighbor algorithm allows to find the best subset of 3D points that are both close to query features in appearance space and that are consistently seen by the same camera, leading to high overlap with the initially unknown query camera pose (shaded area). Jointly solving for these two metrics in a single search algorithm is a long-known open question in the community and CANN provides to the best of our knowledge the first practical solution. Red points in the figure show neighbors retrieved by an unconstrained search using the features from the query image (bottom right). Using CANN it’s more likely to retrieve points that are inliers to geometric verification (green) and less likely to fetch unrelated outlier points (yellow).

To localize a query image against such 3D models, a set of local features is extracted from it and 2D-3D correspondences are estimated based on descriptor similarity. In practice, this data association problem suffers from various challenges: visual aliasing, scene change, noise, etc. Because the final localization solution is computed using geometric inference from these 2D-3D correspondences, not finding enough correct matches can lead the entire localization process to fail.

Simply establishing many more matches per query key-point (red points in Fig. 1) however causes long runtime in geometric verification [3]. It is thus important to find

a small 2D-3D set which has high probability to contain “good” matches (yellow/green points in Fig. 1): In fact we know that the 3D points of “good” matches should all lie inside one (unknown) camera frustum which is the one of the query image (shaded area in Fig. 1).

There exist several approximations to this problem, ranging from clustering nearest-neighbor matches in the 3D model’s covisibility graph [42] to using image retrieval methods to obtain a small set of candidate images for which local features are matched subsequently [41]. The latter approach, leveraging recent advances in global (per image) embeddings, has gained substantial traction recently [18, 42, 41, 8], to a degree that it appears the community has abandoned the idea of finding a solution that jointly solves for appearance and geometry using local features only. For example, the benchmark we evaluate on [18] didn’t even consider local feature based retrieval approach at publication time.

We don’t consider the case of using local features closed and therefore propose an approach to obtain matches that are close in appearance space while obtaining geometric consistency at the same time – which is a long-known open question in the community.

Contributions. In this paper we make three contributions:

(1) Our first and main contribution is a new method, referred to as Constrained Approximate Nearest Neighbors (CANN), that efficiently obtains a high quality, small set of 2D-3D correspondences. CANN performs nearest neighbor search in descriptor space in a constrained manner, so that matches are compact in 3D space. We provide both a brute-force solution as well as an efficient implementation and associated complexity analysis of this *colored* nearest neighbor search algorithm.

(2) Our second contribution is to make the connection of colored nearest neighbor search to the problem space of image retrieval and localization, proposing a metric to rank cameras, which can serve as a way to evaluate future work in this area.

(3) Lastly we provide an extensive evaluation of both global and local feature based methods on four large scale datasets from [18]: “Baidu-Mall”, “Gangnam Station”, “RobotCar Seasons” and “Aachen Day-Night v1.1”. We demonstrate that local feature based methods are not only competitive, but in fact strongly outperform global embedding based approaches; which goes contrary to the trend in the community. We hope to provide new impulse to techniques that aim for jointly searching in appearance and geometry space, which is more efficient and elegant than previously proposed two-step approaches.

2. Related Work

Visual Localization using local features without image retrieval: A large body of work in visual localization [44, 46, 61, 45, 29, 28, 47, 31, 6, 50] is based on sparse 3D

point clouds built from image collections using Structure-from-Motion (SfM). These methods directly establish 2D-3D matches between local features from the query image and the descriptors associated with 3D points in the model. As mentioned before, these matches often contain many outliers and thus directly feeding them to geometric verification is typically impractical[3]. Therefore several post-filtering techniques have been proposed, such as clustering in the SfM covisibility graph [45, 46] or applying voting in the space of camera poses [61, 31]. Remaining 2D-3D matches typically have a sufficiently low fraction of outliers, so that they can be efficiently processed by geometric verification, using minimal pose solvers [26, 51] in a RANSAC [14] scheme.

Visual Localization using local features for retrieval and 2D-3D matching: Image retrieval approaches promise to

both reduce the cost of matching against features in the SfM model and achieving high quality matches by limiting the search to only a subset of the model[37]. Such approaches either remove features that don’t belong to top-ranking images or perform an additional matching step to top-ranking images before running geometry verification using the obtained local feature matches. Our proposed algorithm provides an alternative to these two-step filtering approaches, by directly optimizing for compactness of nearest neighbor matches in the covisibility graph or 3D space *during* the search.

Visual Localization using global features for retrieval and local features for 2D-3D matching: Image retrieval

using local features however has most recently lost attention from the community and instead global features (*e.g.*, DELG-GLDV2 [9] and AP-GeM [38]) have dominated benchmarks [18]. While using global features offers significant speedups due to the much smaller database size, the full-image embeddings are not appropriate for high quality localization due to their global nature [18]. In order to obtain an accurate localization result, some approaches [41, 42] compute additionally local features, which are matched only between the query image and top-ranking images from the database. While there are attempts to concurrently compute local and global features to reduce cost/latency [9], the accuracy of the local feature keypoints remain inferior to approaches that compute dedicated local features [43].

Local feature-based image retrieval techniques: Despite the image retrieval community’s recent focus on global features, local feature-based retrieval has a long history, with well-established methods [49, 36, 24, 55, 34]. Among these, the most relevant method today is the Aggregated Selective Match Kernels (ASMK), which continues to be explored recently in conjunction with deep-learned local features [52, 56, 58]. ASMK (like VLAD [5]) performs local descriptor aggregation and essentially produces high-dimensional global image representations, which are however sparse and can be searched efficiently. In contrast, our method operates directly on local descriptor space and avoids aggregation,

which makes it more suitable to match against partial views and unique details that do not get lost in aggregation.

Approximate nearest neighbor methods: Another related field is the area of proximity problems in high dimensional spaces with its many applications in computer vision [23, 7, 11, 2] (to name a few). The most common of this kind is nearest neighbor search, where given a set P of n points in a high-dimensional space R^d we wish to find the point(s) in P closest to a query point q . Extensive research on this problem has led to a variety of interesting solutions, both exact and approximate [21]. In many use cases, indexed points in the “database” are equipped with additional attributes, such as vector-valued attributes or simple scalars, such as an ID (“color”) that indicates a grouping of points.

The term Constrained Approximate Nearest Neighbors that we propose in this paper refers to a way to apply nearest neighbors in one space given constraints in the space of these attributes. The simplest such case is “colored nearest neighbor search”: each point in P is assigned with an ID and for a given query point q (with or without colors), we want to use the IDs of points in P as constraints during the search. A simple example, which is the use case in this work, is to return nearest neighbors for all query points, provided that all of the neighbors have the same ID. The optimal result are those points in P that all have the same ID and optimize some metric, such as the sum of distances to the query points.

Colored range searching and nearest neighbors (also known as “categorical range searching”, or “generalized range searching”) have been extensively studied in computational geometry since the 1990s [22, 16, 17]. The colored versions of nearest neighbor (or range searching) problems tend to be harder than their uncolored counterparts and several different problems and solutions were proposed, see e.g. [35]. To the best of our knowledge, no previous problem and solution fits into the requirement that we need in this work and the Constrained Approximate Nearest Neighbor problem we address here is new.

3. Method

3.1. Ranking Images for Visual Localization using Constrained Approximate Nearest Neighbors

We first propose a natural metric to rank cameras and then show that this ranking can be efficiently computed during the feature-matching stage instead of requiring post processing. For simplicity of presentation we consider the case of a single optimal camera/image from the index. This is without loss of generality, since in practice, we may use k -best cameras or simply weight matches by the rank of each image.

The metric: We are given a large d -dimensional space containing local feature descriptors extracted from all images in a large collection. Denote $I = \{0, 1, 2, \dots\}$ the set of

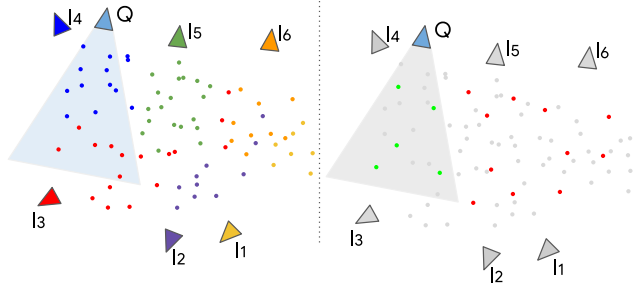


Figure 2: A visual depiction of CANN: the image on the left shows 3D points colored by the camera from which they were reconstructed. CANN leverages this information to retrieve feature matches that are consistently seen in the same camera. This contrasts with prior art (on the right), where unconstrained feature matching returns many unrelated outlier matches (red), which then need to be filtered out subsequently by geometric verification to obtain inlier matches (green).

image IDs in that collection. We assign each local descriptor the ID of the image, $i \in I$, it was computed from, so we obtain the set P of “ID-colored” points (see colors in Fig. 2 on the left). Then, at query time, for a query image with a set of features $Q = \{q_j\}$ extracted from it, let $d_{ij} = d(q_j, NN_i(q_j))/R$ be the (normalized) Euclidean distance in descriptor space between the feature q_j to its nearest neighbor descriptor in image i . R is some fixed maximum distance that we use for normalization such that $d_{ij} \in [0, 1]$. We then compute a score for each image i in the dataset

$$s_i = \sum_j (1.0 - d_{ij}^{\frac{p}{1-p}})^{\frac{1-p}{p}} \quad (1)$$

and use it to rank all images with respect to the query image features $q_j \in Q$. To obtain this per-image compact set of descriptors from the set of all indexed descriptors P (with their “ID-color”), we have to develop an efficient colored version of nearest neighbors. Such algorithm obtains the nearest neighbor of each q_j for all colors at once, provided that its distance is at most R . We observe that depending on a tuned parameter p , we can crop the distances at R such that all distances larger than R have score at most some very small value (say 10^{-6}). This allows to get good bound on the runtime of the search for NN . Figure 3 shows our metric.

3.2. Preliminaries

To explain the proposed Constrained Approximate Nearest Neighbors algorithm we refer to standard tools like Approximate Nearest Neighbors (ANN) and Approximate Range searching (RS) and make the (common) assumption that there is a maximum distance, R , known at local descriptor indexing time. We also assume that randomization is allowed, i.e. all results are correct with (arbitrary) high

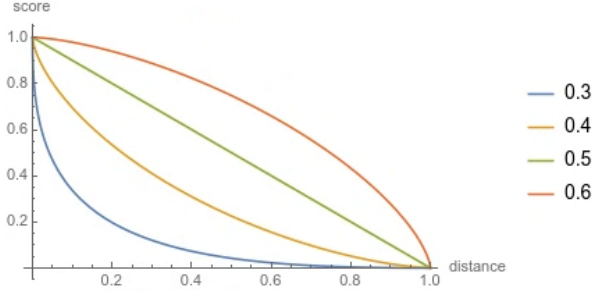


Figure 3: Our score for $R = 1$ and various p different values in Equation 2. p is a parameter of our metric that we tune upfront and is used to compute s_i for all $d_{i,j}$.

probability. Details on the exact definitions of ANN and RS for the case of bounded distance can be found in [12]. We can assume (for simplicity of presentation) that ANN and RS data structures can be created in $O(C_I(d, c) * n)$ and a point query takes $O(C_q(d, c) + k)$ time, $C_q(d, c)$ is a constant depending on the dimension d and the approximation factor, c and k is the maximum number of items it returns. For image retrieval, this runtime is multiplied by the number of features in the image, $|Q|$.

Colored Nearest Neighbors vs Colored Range Searching

As can be seen from Equation 1, we need a colored NN data structure to compute the scores for all relevant images given one query point q_j . Such algorithm returns for each q_j the set of 1-NN in all cameras within radius R . We see from the metric that cameras without such neighbor don't contribute to the sum, so we want as many neighbors with as low Euclidean distance from the query as possible. We are not aware of any efficient algorithm to perform this with a better time complexity than a brute force method using $|I|$ separate NN structures (See Section 3.4). Fortunately, we can reduce this colored NN problem to a fixed R colored range searching which can be implemented efficiently. A reduction from the fixed radius decision problem: "is there a point within distance R from the query" to the approximate NN is well known from LSH [20] using a form of binary search over several R 's. While this approach isn't directly applicable for colored searches, we can use similar ideas as outlined in the following section.

3.3. Colored Range Searching

In this section we explain the colored nearest neighbor search for computing the scores in Eq. (1). While there are multiple versions of this problem, we're specifically interested in *colored range reporting*: For a set of colored points in R^d , report all the distinct colors in the query range. Even with approximations, this problem is computationally hard with $O(C_q(d, c) + |I|)$ [35, 53] as lower bound on the runtime. For a large set of images this bound can be

very high, yet in practice it can be solved quite efficiently by introducing the threshold distance R . The most recent work [53] on this exact problem shows that the problem is already hard for low dimensional spaces, even with integer coordinates and considering only orthogonal queries (an axis-aligned box vs. a sphere). For a set of n colored points in three dimensions, the authors of [53] describe a randomized data structure with $O(n * \text{polylog}(n))$ space that can report the distinct colors within an axis-aligned box using $O(k * \text{polyloglog}(n))$ time, with k as number of distinct colors in the range, assuming that coordinates are in $\{1, \dots, n\}$. In this paper we show that with R known at index time and allowing for approximation, we can develop a more efficient data structure for the colored range search that allows us to efficiently compute the 1-NN across all images at once. Besides being essential for solving the Constrained Nearest Neighbors problem we believe that this data-structure is interesting on its own and beyond the problem of image localization.

3.4. A brute force method using ANN

There exist two straightforward algorithms for colored range searching: First build $|I|$ separate regular nearest neighbor structures, one for each color in $O(C_q(d, c) * |P_I| * |I|)$ indexing time, with $|P_I|$ as the color I in P . Then call them sequentially for each query point q_j with cost $O(C_q(d, c) * |I|)$. This is independent of R and thus much worse than the above lower bound. The advantage is that the runtime of this version is asymptotically independent of the threshold radius R .

The second simple algorithm, that we call CANN-RS, is applicable for small thresholds R using Randomized Range Searching [12]: We index points into a RS data-structure for radius R and then for each query feature, enumerate all neighbors within R , keeping a tally of neighbors per image I . Because we only retrieve a small subset of points within radius R we only obtain a few colors (cameras or images) with number of features, much less than $|I|$. This approach has runtime $O(C_q(d, c) + k)$, here, k is the expected number of neighbors in each such range query over all images. The drawback is that for each query feature we must enumerate *all indexed points in the range R* where most of them do not realize any nearest neighbor for any color. This number (k above) can still be quite large.

3.5. Efficient algorithm using Random Grids

To implement an efficient variant of the algorithm (CANN-RG), we leverage Random Grids [2, 1], an efficient c -approximate nearest neighbor algorithm based on applying randomized rotations and shifts to high-dimensional vectors prior to hashing them to a set of keys. We extend the Random Grids to support colored range searching. We show that our algorithm avoids the enumeration of all points in

the range R (as in CANN-RS) and doesn't require distance computation in descriptor space which can take most of the time in practice due to high dimensional feature descriptors.

Our algorithm works as follows: For each query point q_j CANN-RG should report all colors in the range R from q_i approximately by factor $c > 1$, i.e. any color that has a feature at distance at most R is reported with high probability and any reported color (image) has a feature at distance at most cR . The points are indexed using Algorithm 1, where we store a set of distinct integers using hash-sets and use hashing to create a key for each non-empty grid cell in a high dimensional space following [2, 1]. At query time we retrieve points from the grid indices using Algorithm 2.

Note that since we're only interested in the color of points within range R , the index only holds point colors not point coordinates and the query results similarly only comprise colors without exact distances.

Algorithm 1: Efficient colored Range Searching indexing (CANN-RG-INDEX-R)

Data: P : d -points, $\{p_i\}$ with colors $color(p_i)$ for each $p_i \in P$, $R > 0$: Range, $c > 1$: Approximation factor

Result: Index RG for the colors $color(p)$ such that for a given query point q , all colors at distance at most R from q are reported quickly.

Function IndexColors(P, R, c):

```

    Impose a grid of cell size  $w = R * c / \sqrt{d}$  on  $P$ 
    Create  $L$  such grids for  $L$  randomly rotated and translated versions of  $P$ 
    /*  $L$  is determined from the analysis below */
    for all  $p_i \in P$  do
        Add  $color(p_i)$  to a distinct set of colors hashed in each of  $L$  cells the transformed  $p_i$  falls into
    end
    /* Each cell now contains a set of distinct colors of all images that have point inside it */
    /* NOTE: We do not store the coordinates of the points, just their color (16 bits per color) */
    return The set of hashed cells with their lists of colors

```

Analysis In this section we analyze indexing and query algorithms of CANN-RG. First we make concrete the constants $C_I(d, c)$ and $C_q(d, c)$ which appear in all Random Grids implementations: For the grid cell of size $l * c / \sqrt{d}$, a random vector of length l in R^d will be captured in a given cell with probability at least $e^{-\sqrt{d}/w} = e^{-d/c}$ [1]. We thus need $L = e^{d/c}$ random grids in Algorithm 1 to ensure that, if there is a point in P at distance at most R from q , its color will be found in at least one of the grid cells with constant

Algorithm 2: Query an arbitrary point in the index built for a given R (CANN-RG-QUERY-R)

Data: RG : A Random Grid index for colors (Algorithm 1), q : query d -point

Result: A set of colors (images) that have a point at distance at most R from q

Function QueryPoint(RG, q):

```

    Create an empty set distinct colors,  $S$ 
    for all grids  $g \in RG$  do
        Rotate and translate  $q$  according to  $g$  to obtain  $q_t$ 
        Retrieve the set of colors,  $colors(g)$ , from the grid cell in  $g$  that  $q_t$  falls into
        Insert all colors in  $colors(g)$  into  $S$ 
    end
    /*  $S$  now contains a set of distinct colors that have a point at distance at most  $cR$  from  $q$  */
    return  $S$ 

```

probability. On the other hand, any color of a point found in a grid cell that also contains q_t (the rotated and translated version of q for that grid) is at distance at most cR from q due to the size of the grid cells.

Because we do not care about the coordinates of indexed points, we only store each color at most once per grid cell. Therefore the data structure build time ($C_I(d, c) * |P|$) = $O(e^{d/c} * |P|)$ and storage $O(e^{d/c} * |P|)$ are linear in $|P|$. For each query point q , we retrieve the points in the grid cells where all rotated and shifted versions of q fall into. The runtime is then $O(e^{d/c} + k_c) = O(C_q(d, c) + k_c)$ ignoring the constant for matrix rotation that depends on d . Note that for Random Grids implementation we have $C_I(d, c) = C_q(d, c)$. In contrast to k in CANN-RS, k_c here refers to the number of *distinct colors* found in the enumerated cells. As in [2], the probability of success can be amplified to $1 - \gamma$ by repeating the randomized indexing $\ln(1/\gamma)$ times, which increases the data structure, space and query time accordingly. The number of grids that we need in practice is much smaller than the above worst case depending on the intrinsic dimension of the data [2].

Constructing and querying CANN-RG The above algorithms allow indexing colors of P for a given R such that for any query point q , the colors that have points at distance at most R from q are reported quickly. Given that we omitted the computation of point distances to enable efficient queries, we're still missing a way to compute the scores in Equation 1. We now show how we move from fixed radius Range Search to 1-NN.

To fill this gap, let r be a constant denoting the minimum distance between points in P that we aim to distinguish. For each $l \in \{rc^0, rc^1, \dots, R\}$, we generate a sequence of random grid indexes $B^l = \{B_i^l, \dots, B_n^l\}$ of radius l . Then,

given query point q , we query q in all B_i in order and keep only the closest (first observed) color. This maps the list of colors to the B_i^l they came from and thus to a c -approximate distance from the query. Given these minimum distances, Equation 1 provides a score per point and thus a ranking of all index images by summing over all $q_j \in Q$. This scoring operation increases the runtime of the query by logarithmic factor of R/r . Note that CANN-RG is output sensitive on k_c , the number of actual neighbor colors we find for each query.

4. Experiments

4.1. Experimental setup

Datasets: We evaluated our method on four public datasets from [18], “Baidu-Mall”, “Gangnam Station”, “RobotCar Seasons” and “Aachen Day-Night v1.1”. These datasets demonstrate performance in “regular” outdoor scenarios as well as repetitive indoor environments. “RobotCar Seasons” and “Aachen Day-Night v1.1” have day and night subsets.

Metrics: We evaluated two metrics: (1) The image retrieval performance using the same equal weighted barycenter (EWB) interpolation as in [18] which is based solely on the retrieved images and their known poses. (2) The effect on final localization quality using the existing localization pipeline from [18] where camera localization is computed using only features from the top- k ranking images.

Local and global feature baselines: Following [18, 33], we compared our method against state-of-the-art global features AP-GeM [38], DELG [9], DenseVLAD [57], NetVLAD [4]. For local-features we compare performance and cost for both indexing and query to ASMK [54] with HOW and FIRE local features. Results for the latter were not previously published and only recently made available on the codebase for image retrieval methods. R2D2 features were computed using code from the same codebase. Storage cost for the baselines is discussed analytically.

Local feature types: We experiment with three state-of-the-art local image features: HOW [56], FIRE [59] and R2D2 [40]. These three approaches have different operation characteristics and thus show the power of CANN in being adaptable to different local features. HOW and FIRE are designed for image retrieval, and are not suitable to the local feature matching part of the visual localization pipeline. R2D2, on the other hand, is designed for image matching tasks and a common choice in structure-from-motion and visual localization evaluations [25, 18]. We use a recent and lighter R2D2 version (referred to as “Feather2d2 20k”) described in [18]’s codebase, where we can download the

local features (the model is not publicly available). When using HOW and FIRE, our visual localization system requires indexing two different feature types: HOW for retrieval and R2D2 for matching. When using R2D2, we only need to index one feature type – which is appealing since it simplifies the overall system. For our experiments we used 1000 per image for all indexed and query images and all methods.

Implementation details: We implemented CANN-RS and CANN-RG (Section 3) in C++, given that it performs well for low intrinsic dimensions of the features: 32D for R2D2 and 128D for HOW. Even though CANN-RS can be implemented with any of-the-shelf range search data structures, we used Random Grids also here as it has the ability to exploit the fact that we know the range in advance. The Random Grids were adjusted to different intrinsic dimensions by tuning its parameters, which is also required to trade off performance vs runtime using the c -approximation. Both our algorithms are very simple, trivially parallelized and are very fast (down to 20ms per query image).

Tuning: The parameters of our metric are p and R and we tune them for each feature type separately. Note that in contrast to ASMK which creates a codebook that depends on the distribution of the data, CANN-RG and CANN-RS only tune for the metric itself. One can therefore provide theoretic bounds of the (approximate) algorithmic result quality for a given metric. This may make CANN more resilient to different datasets which is not the case for codebook methods, even though the latter can perform better if the distribution of features between query and training set matches. For CANN-RS, we set the grid cell size to slightly above $1/\sqrt{d}$ and the number of grids accordingly to balance result quality and runtime (see Section 3.4). For CANN-RG we set $c = 1.1$ in all datasets and the metric parameters (p , R) were tuned using a subset of 500 queries from “Baidu-Mall” separately per local feature type. To the best of our knowledge, the datasets of [18] provide no tune/eval/test split and only the “Baidu-Mall” has ground-truth available to enable tuning. For ASMK we only evaluated R2D2 features, taking results for other features from [18] or used previously unpublished results provided by the authors. We train the ASMK codebook on “GangnamStyle” as it is the largest set among the four. To validate generalization, we used the same set of parameters for evaluation on all other datasets.

4.2. Results

As mentioned above, we evaluate the CANN-RG and CANN-RS algorithms on four large-scale datasets, in an outdoor, urban setting and covering an indoor scenario. Following [18, 33] we evaluate across two regimes/metrics (“EWB” and “SFM”) discussed above. Figure 3 shows our results of all methods and datasets with one figure per each metric.

In general, we can observe local features outperforming global features almost everywhere and by a large margin. Datasets that are more appropriate for global features are those that have many approximately similar viewpoints in the index so there is almost always one close neighbor for a given query image. Local features are naturally better where the query contains only partial overlap with the indexed images. Qualitative results are available in the appendix.

Runtime One of the main advantages of CANN-RG (and CANN-RS as well) comparing to ASMK for image retrieval using local features is its simplicity and its runtime in both indexing and query. Table 1 shows numbers across datasets using HOW features. Since our implementation of CANN-RG and CANN-RS does not use GPU, we compared runtime on CPU using 48 cores. The table does not contain the codebook creation for ASMK and tuning for CANN-RG. CANN-RG has a nice run-time/quality trade-off: In its upper bound quality, we have the results of CANN-RS and with CANN-RG can pay in quality for much better runtime. The significance of this is that CANN-RG can achieve runtimes of a few milliseconds for query image, which is otherwise only possible with global features. Table 1 provides results demonstrating the trade-off of runtime and quality. To obtain a cheaper, yet representative quality measure, we compute the EWB using the top-1 retrieved image. The indexing time for CANN-RG is larger due to the fact that we have factor $O(\log R)$ more data structures.

Dataset	Index		
	CANN-RS	CANN-RG	ASMK
Baidu	0.88	9.08	253.34
Gangnam	6.41	168.49	1467.17
Aachen	11.19	244.09	2782.43
Robotcar	33.02	852.12	8104.98
Dataset	Query		
	CANN-RS	CANN-RG	ASMK
Baidu	0.37(12.47)	0.02(12.12)	0.47(12.52)
Gangnam	1.6(12.66)	0.05(11.35)	0.41(11.03)
Aachen	1.38(29.1)	0.06(28.8)	0.48(28.0)
Robotcar	5.29(94.2)	0.04(93.6)	0.53(91.0)

Table 1: Indexing and average runtime per query image (seconds) for CANN-RS, CANN-RG and ASMK using HOW features. An indication of quality/runtime trade-off can be taken from the simplified EWB metric, computed using the top-1 retrieved image and provided in parentheses.

Preliminary results on general image retrieval To re-emphasize the generalization of the algorithm and its scalability (20-50ms per query image), we also evaluated it for general image retrieval on the ROxford dataset. Global retrieval benchmarks evaluate the full rank of all indexed images, which requires also scoring the tail of the retrieved images. Since ranking the tail of the index is not typically

meaningful for local features, we evaluated a combination of CANN with global features by computing a weighted average of DELG and CANN-RG+HOW or CANN-RG+FIRE, for all image scores. We compare CANN and this combined approach to the SOTA for global/local features. Very recently, a new method called Correlation Verification [27] was published which is, to our knowledge the best performing method on the ROxford dataset. Correlation Verification however includes (significantly expensive) spatial verification of local features and is thus not comparable to CANN-RG which doesn't use geometry or spatial reasoning of features (out of the cameras). Like for localization, spatial reasoning is an additional step that can be applied on top of CANN-RG. Table 2 shows comparisons of SOTA approaches including [27] with our proposed approach (bold).

(A) Local feature aggregation	Medium	Hard
DELG-D2R-R-ASMK* (GLDv1) [52]	73.3	47.6
R50-HOW-ASMK,n=2000 [15]	79.4	56.9
(B) Global feature		
R101-GeM [13, 48]	65.3	39.6
R101-GeM-AP (GLDv1) [39]	66.3	42.5
R101-GeM+SOLAR (GLDv1) [32]	69.9	47.9
R50-DELG (Global-only, GLDv2-clean) [10]	73.6	51.0
R101-DELG (Global-only, GLDv2-clean) [10]	76.3	55.6
R50-DOLG (GLDv2-clean) [60]	80.5	58.8
R101-DOLG (GLDv2-clean) [60]	81.5	61.1
R101-CVNet-Global (GLDv2-clean) [27]	80.2	63.1
DELG+CANN-FIRE (weighted)	82.4	62.3
DELG+CANN-HOW (weighted)	83.3	64.2

Table 2: Results of DELG+CANN compared to state-of-the-art reranking (local aggregation and global features) on ROxford (numbers of related work from [27])

Limitations. Using local features throughout the stack requires that the entire map fit in memory. Approaches that use global features can be more easily scaled, in that the local features per spatial region are kept out-of-memory and are only loaded after image retrieval.

5. Conclusions

In this paper, we proposed CANN, a novel nearest neighbor searching approach that finds the best matches in both appearance and geometry space to improve visual localization using only local features. Unlike the state-of-the-art in the field, which uses global features for image retrieval and local features for 2D-3D matching, our approach uses only local features, while providing significantly better performance than the state-of-the-art at very competitive runtime cost. By providing the relevant metric and theoretical foundation of the algorithm, as well as two efficient algorithmic solutions, we hope to inspire a revived interest in solving visual localization with local features only.

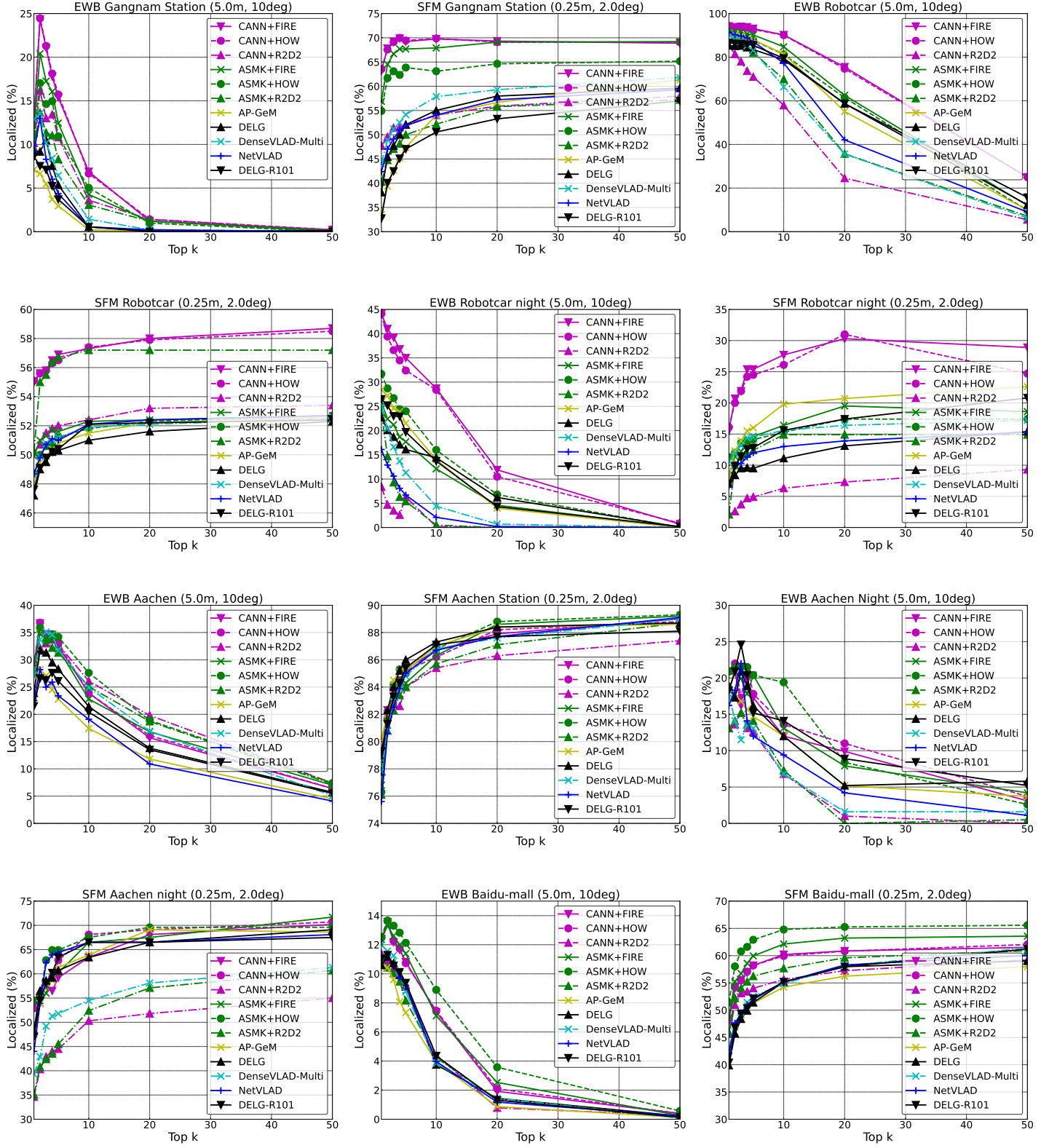


Table 3: Results from four public benchmarks. Following [18, 33] we evaluate across image retrieval using equal weighted barycenter (EWB) interpolation and final localization quality using the existing localization pipeline from [18] [SFM].

A. Additional Qualitative Results

We include additional qualitative results in Figures 4,5,6,7,8,9 taken from all datasets, showing that CANN retrieves good results also in images with heavy occlusions. Cases like these, where there is only partial overlap between the query image and database images are very difficult for global features. We use HOW [56] for local features with both CANN-RG (ours) and ASMK [55]. The query image is on the left and the top 5 retrieved images are on the right. Our method retrieves all correct images, while other methods retrieve occasionally incorrect images ranked high among the top 5. We see that some global methods retrieve incorrect images due to scene clutter or high-frequency textures, while CANN provides diverse set of correct results. In several cases, we see that CANN+HOW outperforms ASMK+HOW. Retrieved images are marked red (bad) or green (good).

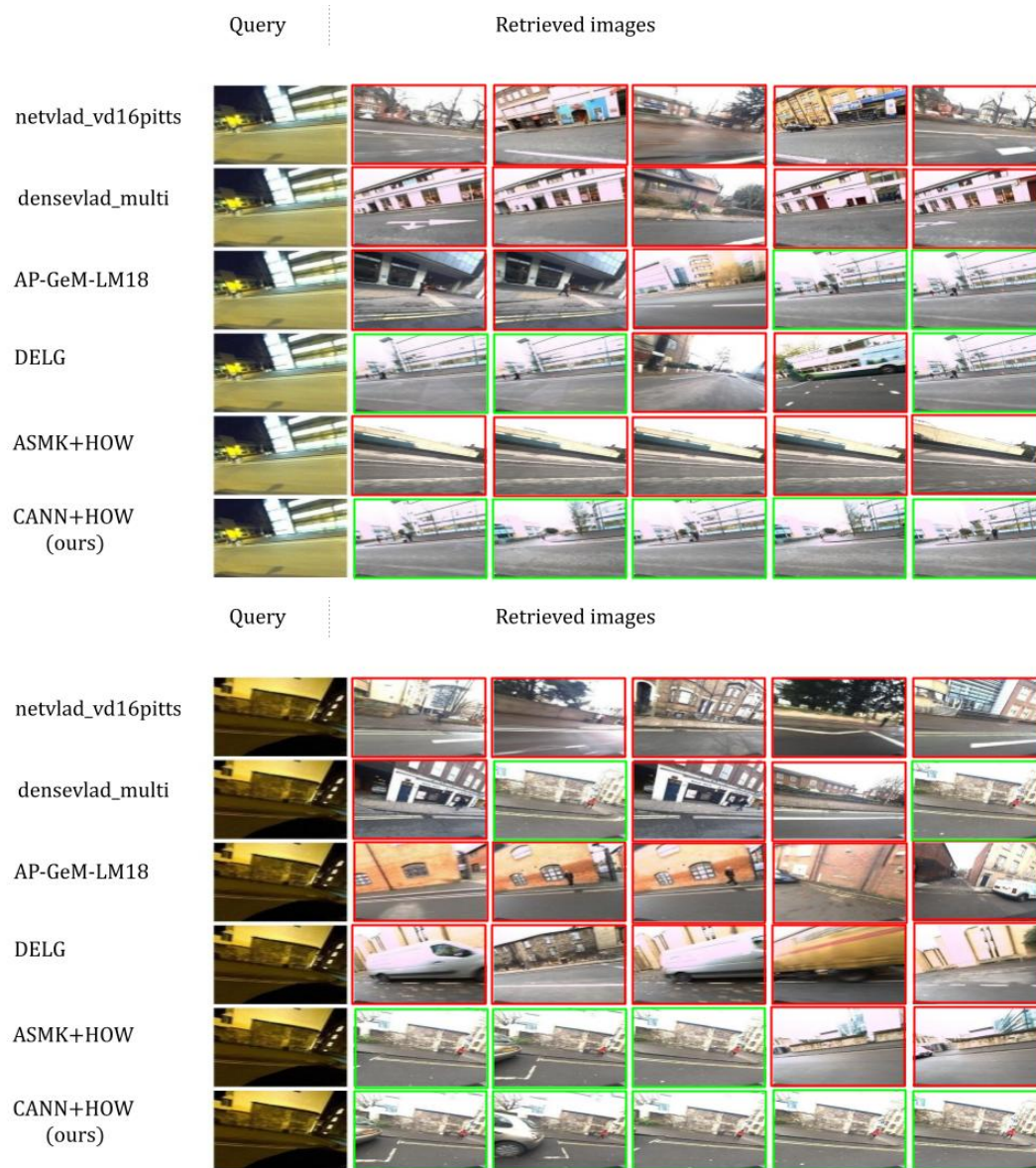


Figure 4: Robotcar

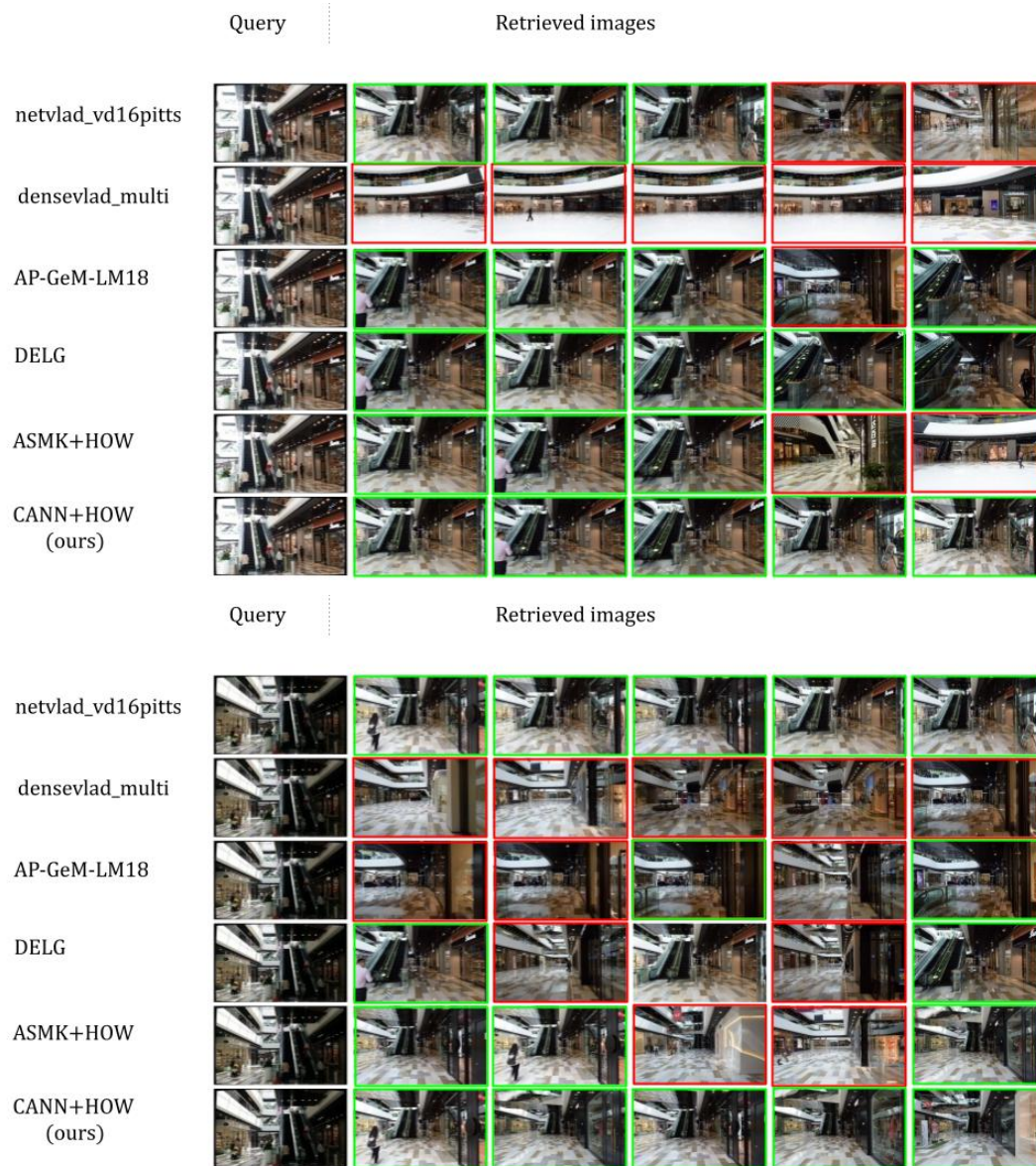


Figure 5: Gangnam

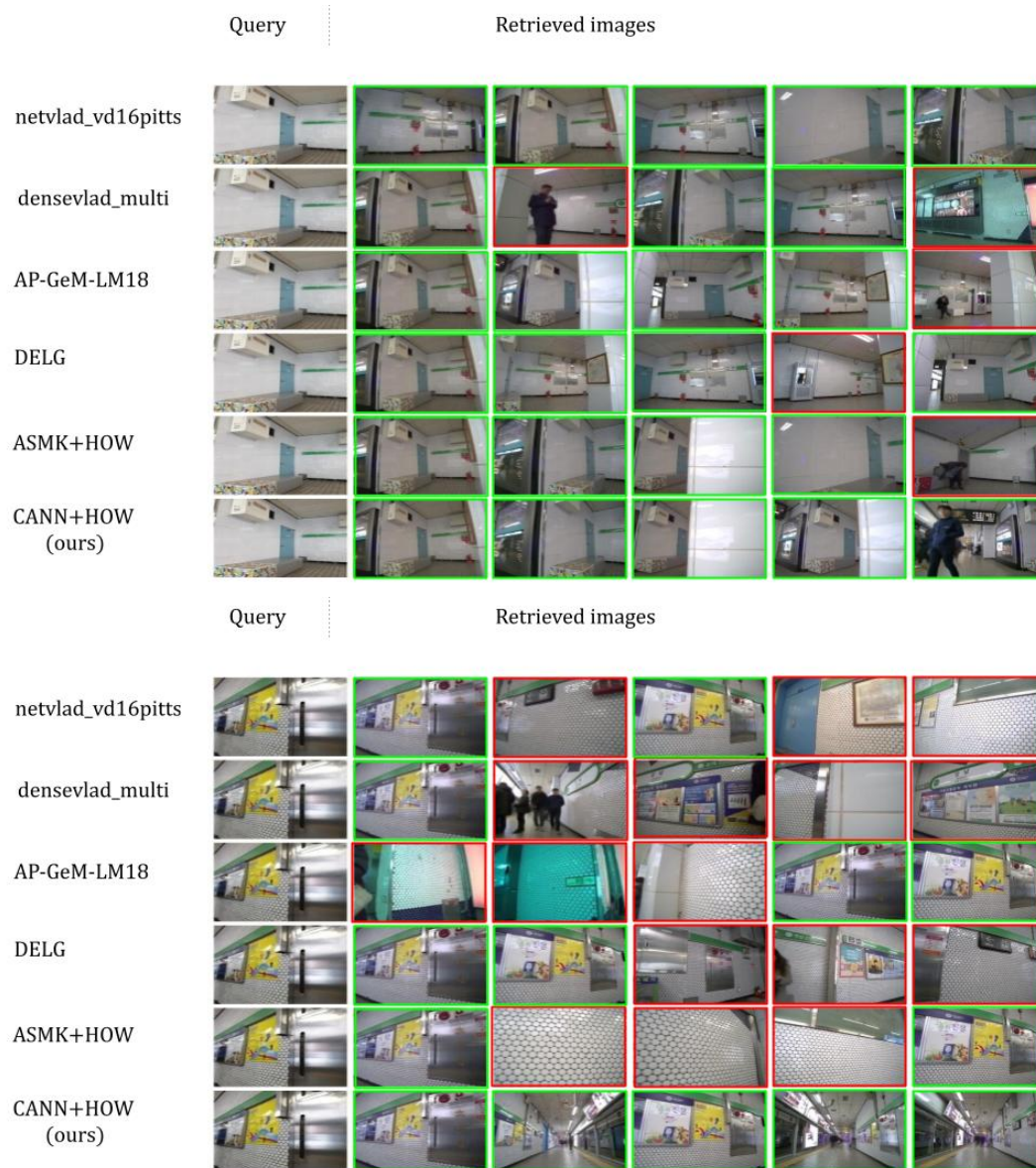


Figure 6: Baidu

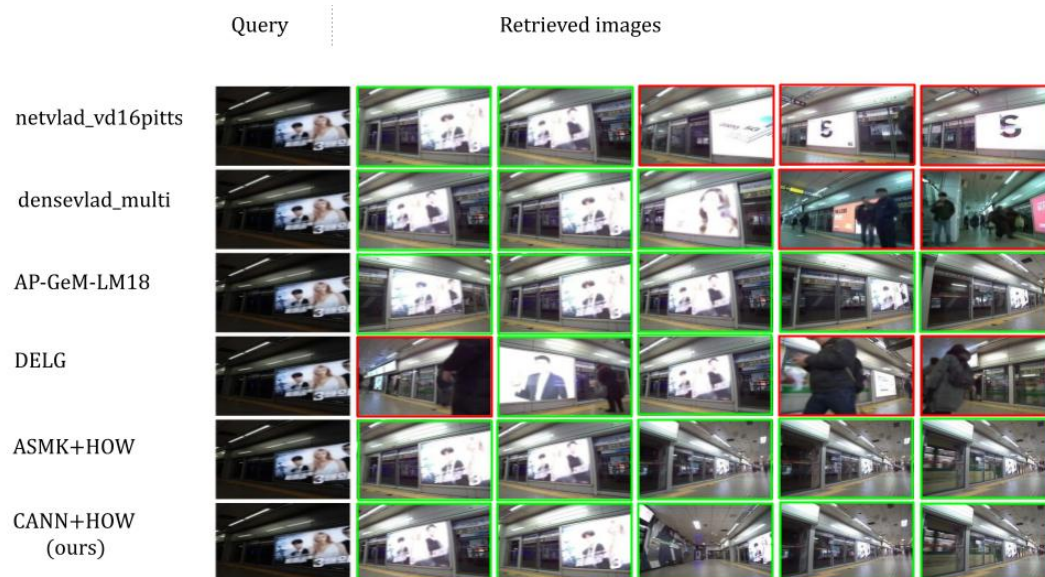


Figure 7: Baidu



Figure 8: Aachen



Figure 9: Aachen

References

- [1] Dror Aiger, Haim Kaplan, and Micha Sharir. Reporting neighbors in high-dimensional euclidean space. *SIAM Journal on Computing*, 2014.
- [2] Dror Aiger, Efi Kokiopoulou, and Ehud Rivlin. Random grids: Fast approximate nearest neighbors and range searching for image search. In *ICCV*, 2013.
- [3] Dror Aiger, Simon Lynen, Jan Hosang, and Bernhard Zeisl. Efficient large scale inlier voting for geometric vision problems. In *ICCV*, 2021.
- [4] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In *CVPR*, 2016.
- [5] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *CVPR*, 2013.
- [6] Clemens Arth, Daniel Wagner, Manfred Klopschitz, Arnold Irschara, and Dieter Schmalstieg. Wide area localization on mobile phones. In *ISMAR*, 2009.
- [7] Artem Babenko and Victor Lempitsky. The inverted multi-index. *PAMI*, 2014.
- [8] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *CVPR*, 2022.
- [9] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *ECCV*, 2020.
- [10] Bingyi Cao, André Araujo, and Jack Sim. Unifying deep local and global features for image search. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV, Lecture Notes in Computer Science*, 2020.
- [11] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, 2004.
- [12] Ehud Rivlin Dror Aiger, Efi Kokiopoulou. Random grids: Fast approximate nearest neighbors and range searching for image search. In *ICCV*, 2013.
- [13] Giorgos Tolias Filip Radenovic and Ondrej Chum. Finetuning cnn image retrieval with no human annotation. *PAMI*, 2018.
- [14] M. Fischler and R. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 1981.
- [15] Tomas Jenicek Giorgos Tolias and Ondrej Chum. Learning and aggregating deep local descriptors for instance-level recognition. *ECCV*, 2019.
- [16] Prosenjit Gupta, Ravi Janardan, and Michiel Smid. Algorithms for generalized halfspace range searching and other intersection searching problems. *Computational Geometry*, 1996.
- [17] Prosenjit Gupta, Ravi Janardan, and Michiel Smid. A technique for adding range restrictions to generalized searching problems. *Information Processing Letters*, 1997.
- [18] Martin Humenberger, Yohann Cabon, Noé Pion, Philippe Weinzaepfel, Donghwan Lee, Nicolas Guérin, Torsten Sattler, and Gabriela Csurka. Investigating the role of image retrieval for visual localization. *IJCV*, 2022.
- [19] Apple Inc. Apple, arkit geotracking, 2022.
- [20] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In Jeffrey Scott Vitter, editor, *ACM*, 1998.
- [21] Joseph O’Rourke Jacob E. Goodman. Handbook of discrete and computational geometry, second edition. In *Chapman and Hall/CRC*, 2004.
- [22] Ravi Janardan and Mario Lopez. Generalized intersection searching problems. *International Journal of Computational Geometry & Applications*, 1993.
- [23] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *PAMI*, 2010.
- [24] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating Local Image Descriptors into Compact Codes. *PAMI*, 2012.
- [25] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *IJCV*, 2021.
- [26] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *CVPR 2011*, 2011.
- [27] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *CVPR*, 2022.
- [28] Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua. Worldwide pose estimation using 3d point clouds. In *ECCV*, 2012.
- [29] Yunpeng Li, Noah Snavely, and Daniel P Huttenlocher. Location recognition using prioritized feature matching. In *ECCV*, 2010.
- [30] Google LLC. Google, arcore geospatial api, 2022.
- [31] Simon Lynen, Bernhard Zeisl, Dror Aiger, Michael Bosse, Joel Hesch, Marc Pollefeys, Roland Siegwart, and Torsten Sattler. Large-scale, real-time visual-inertial localization revisited. *IJRR*, 2020.
- [32] Tony Ng, Vassileios Balntas, Yurun Tian, and Krystian Mikołajczyk. SOLAR: second-order loss and attention for image retrieval. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV, Lecture Notes in Computer Science*, 2020.
- [33] Gabriela Csurka Yohann Cabon Torsten Sattler Noé Pion, Martin Humenberger. Benchmarking image retrieval for visual localization. In *3DV*, 2020.
- [34] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-Scale Image Retrieval with Attentive Deep Local Features. In *ICCV*, 2017.
- [35] S. Rahul P. Gupta, R. Janardan and M. H. M. Smid. Computational geometry: Generalized (or colored) intersection searching. In *Handbook of Data Structures and Applications, CRC Press, chapter 67*, 2018.
- [36] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object Retrieval with Large Vocabularies and Fast Spatial Matching. In *CVPR*, 2007.
- [37] Noé Pion, Martin Humenberger, Gabriela Csurka, Yohann Cabon, and Torsten Sattler. Benchmarking image retrieval for visual localization. In *3DV*, 2020.

- [38] J. Revaud, J. Almazan, R. Sampaio de Rezende, and C. Roberto de Souza. Learning with Average Precision: Training Image Retrieval with a Listwise Loss. In *ICCV*, 2019.
- [39] Jérôme Revaud, Jon Almazán, Rafael S. Rezende, and César Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *ICCV*, 2019.
- [40] Jerome Revaud, Cesar De Souze, Philippe Weinzaepfel, and Martin Humenberger. R2D2: Repeatable and Reliable Detector and Descriptor. In *NeurIPS*, 2019.
- [41] P-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *CVPR*, 2019.
- [42] Paul-Edouard Sarlin, Frédéric Debraine, Marcin Dymczyk, Roland Siegwart, and Cesar Cadena. Leveraging deep visual descriptors for hierarchical efficient localization. In *CRL*, 2018.
- [43] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *CVPR*, 2021.
- [44] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *2011 International Conference on Computer Vision*, 2011.
- [45] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving image-based localization by active correspondence search. In *ECCV*, 2012.
- [46] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *PAMI*, 2016.
- [47] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? In *CVPR*, 2017.
- [48] Oriane Siméoni, Yannis Avrithis, and Ondrej Chum. Local features and visual words emerge in activations. In *CVPR*, 2019.
- [49] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, 2003.
- [50] Linus Svärm, Olof Enqvist, Fredrik Kahl, and Magnus Oskarsson. City-scale localization for cameras with known vertical direction. *PAMI*, 2016.
- [51] Chris Sweeney, John Flynn, Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. Efficient computation of absolute pose for gravity-aware augmented reality. In *ISMAR*, 2015.
- [52] M. Teichmann, A. Araujo, M. Zhu, and J. Sim. Detect-to-Retrieve: Efficient Regional Aggregation for Image Search. In *CVPR*, 2019.
- [53] Timothy M. Chan, Qizheng He, Yakov Nekrich. Further Results on Colored Range Searching. In *SoCG*, 2020.
- [54] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *ICCV*, 2013.
- [55] G. Tolias, Y. Avrithis, and H. Jegou. Image Search with Selective Match Kernels: Aggregation Across Single and Multiple Images. *IJCV*, 2015.
- [56] G. Tolias, T. Jenicek, and O. Chum. Learning and Aggregating Deep Local Descriptors for Instance-Level Recognition. In *ECCV*, 2020.
- [57] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, 2015.
- [58] P. Weinzaepfel, T. Lucas, D. Larlus, and Y. Kalantidis. Learning Super-Features for Image Retrieval. In *ICLR*, 2022.
- [59] Weinzaepfel, Philippe and Lucas, Thomas and Larlus, Diane and Kalantidis, Yannis. Learning Super-Features for Image Retrieval. In *ICLR*, 2022.
- [60] Min Yang, Dongliang He, Miao Fan, Baorong Shi, Xuotong Xue, Fu Li, Errui Ding, and Jizhou Huang. DOLG: single-stage image retrieval with deep orthogonal fusion of local and global features. In *ICCV*, 2021.
- [61] Bernhard Zeisl, Torsten Sattler, and Marc Pollefeys. Camera pose voting for large-scale image-based localization. In *ICCV*, 2015.