InLoc: Indoor Visual Localization with Dense Matching and View Synthesis

Hajime Taira[1] Masatoshi Okutomi[1] Torsten Sattler[2] Mircea Cimpoi[3]
Marc Pollefeys[2,4] Josef Sivic[3,5] Tomas Pajdla[3] Akihiko Torii[1]
[1]Tokyo Institute of Technology [2]Department of Computer Science, ETH Zürich
[3]CIIRC, CTU in Prague* [4]Microsoft, Redmond [5]Inria†

Abstract

We seek to predict the 6 degree-of-freedom (6DoF) pose of a query photograph with respect to a large indoor 3D map. The contributions of this work are three-fold. First, we develop a new large-scale visual localization method targeted for indoor environments. The method proceeds along three steps: (i) efficient retrieval of candidate poses that ensures scalability to large-scale environments, (ii) pose estimation using dense matching rather than local features to deal with textureless indoor scenes, and (iii) pose verification by virtual view synthesis to cope with significant changes in viewpoint, scene layout, and occluders. Second, we collect a new dataset with reference 6DoF poses for large-scale indoor localization. Query photographs are captured by mobile phones at a different time than the reference 3D map, thus presenting a realistic indoor localization scenario. Third, we demonstrate that our method significantly outperforms current state-of-the-art indoor localization approaches on this new challenging data.

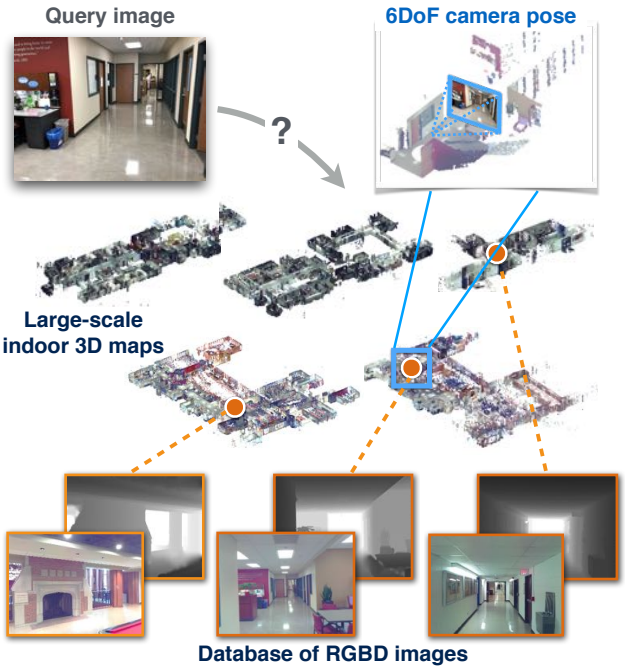


Figure 1. **Large-scale indoor visual localization.** Given a database of geometrically-registered RGBD images, we predict the 6DoF camera pose of a query RGB image by retrieving candidate images, estimating candicate camera poses, and selecting the best matching camera pose. To address inherent difficulties in indoor visual localization, we introduce the "InLoc" approach that performs a sequence of progressively stricter verification steps.

1. Introduction

Autonomous navigation inside buildings is a key ability of robotic intelligent systems [24, 39]. Successful navigation requires both to localize a robot and to determine a path to its goal. One approach to solving the localization problem is to build a 3D map of the building and then use a camera[1] to estimate the current position and orientation of the robot (Figure 1). Imagine also the benefit of an intelligent indoor navigation system that helps you find your way, for example, at Chicago airport, Tokyo Metropolitan station or the CVPR conference center. Besides intelligent systems, the *visual localization* problem is also highly relevant for any type of Mixed Reality application, including Augmented Reality [16, 44, 73].

Due to the availability of datasets, *e.g.*, obtained from Flickr [38] or captured from autonomous vehicles [19, 43], large-scale localization in urban environments has been an active field of research [6, 9, 14, 15, 19, 20, 27, 29, 34, 38, 44, 53–57, 65, 67, 68, 76, 80, 81]. In contrast, indoor localization [11, 12, 39, 58, 59, 64, 70, 75] has received less atten-

*CIIRC - Czech Institute of Informatics, Robotics, and Cybernetics, Czech Technical University in Prague.

†WILLOW project, Departement d'Informatique de l'École Normale Supérieure, ENS/INRIA/CNRS UMR 8548, PSL Research University.

[1]While RGBD sensors could also be used indoors, they are often too energy-consuming for mobile scenarios or have only a short-range to scan close-by objects (faces). Thus, purely RGB-based localization approaches are also relevant in indoor scenes. Obviously, indoor scenes are GPS-denied environments.

tion in the last years. At the same time, indoor localization is, in many ways, a harder problem than urban localization: 1) Due to the short distance to the scene geometry, even small changes in viewpoint lead to large changes in image appearance. For the same reason, ocluders such as humans or chairs often have a stronger impact compared to urban scenes. Thus, indoor localization approaches have to handle significantly larger changes in appearance between a query and reference images. 2) Large parts of indoor scenes are textureless and textured areas are typically rather small. As a result, feature matches are often clustered in small regions of the images, resulting in unstable pose estimates [29]. 3) To make matters worse, buildings are often highly symmetric with many repetitive elements, both on large (similar corridors, rooms, *etc*.) and small (similar chairs, tables, doors *etc*.) scale. While structural ambiguities also cause problems in urban environments, they often only occur in larger scenes [9, 54, 68]. 4) The appearance of indoor scenes changes considerably over the course of a day due to the complex illumination conditions (indirect light through windows and active illumination from lamps). 5) Indoor scenes are often highly dynamic over time as furniture and personal effects are moved through the environment. In contrast, the overall appearance of building facades does not change too much over time.

This paper addresses these difficulties inherent to indoor visual localization by proposing a new localization method. Our approach starts with an image retrieval step, using a compact image representation [6] that scales to large scenes. Given a shortlist of potentially relevant database images, we apply two progressively more discriminative geometric verification steps: (i) We use dense matching of CNN descriptors that capture spatial configurations of higher-level structures (rather than individual local features) to obtain the correspondences required for camera pose estimation. (ii) We then apply a novel pose verification step based on virtual view synthesis that can accurately verify whether the query image depicts the same place by dense pixel-level matching, again not relying on sparse local features.

Historically, the datasets used to evaluate indoor visual localization were restricted to small, often room-scale, scenes. Driven by the interest in semantic scene understanding [10, 23, 79] and enabled by scalable reconstruction techniques [28, 47, 48], large-scale indoor datasets covering multiple rooms or even whole buildings are becoming available [10, 17, 23, 64, 75, 77–79]. However, most of these datasets focus on reconstruction [77, 78] and semantic scene understanding [10, 17, 23, 79] and are not suitable for localization. To address this issue, we create a new dataset for indoor localization that, in contrast to other existing indoor localization datasets [10, 26, 64], has two important properties. First, the dataset is large-scale, capturing two university buildings. Second, the query images are acquired using a smartphone at a time months apart from the date of capture of the reference 3D model. As a result, the query images and the reference 3D model often contain large changes in scene appearance due to the different layout of furniture, occluders (people), and illumination, representing a realistic and challenging indoor localization scenario.

Contributions. Our contributions are three-fold. First, we develop a novel visual localization approach suitable for large-scale indoor environments. The key novelty of our approach lies in carefully introducing dense feature extraction and matching in a sequence of progressively stricter verification steps. To the best of our knowledge, the present work is the first to clearly demonstrate the benefit of dense data association for indoor localization. Second, we create a new dataset suitably designed for large-scale indoor localization that contains large variation in appearance between queries and the 3D database due to large viewpoint changes, moving furniture, occluders or changing illumination. The query images are taken at a different time from the reference database, using a handheld device, and at different moments of the day, to capture enough variability, bridging the gap to realistic usage scenarios. The code and data are publicly available on the project page [1]. Third, the proposed method shows a solid improvement over existing state-of-the-art results, showing an **absolute improvement of 17–20%** in the percent of correctly localized queries within a 0.25 – 0.5 m error, which is of high importance for indoor localization.

2. Related work

We next review previous work on visual localization.

Image retrieval based localization. Visual localization in large-scale urban environments is often approached as an image retrieval problem. The location of a given query image is predicted by transferring the geotag of the most similar image retrieved from a geotagged database [6, 9, 18, 35, 54, 67, 68]. This approach scales to entire cities thanks to compact image descriptors and efficient indexing techniques [7, 8, 22, 31, 33, 49, 63, 71] and can be further improved by spatial re-ranking [51], informative feature selection [21, 22] or feature weighting [27, 32, 54, 68]. Most of the above methods are based on image representations using sparsely sampled local invariant features. While these representations have been very successful, outdoor image-based localization has recently also been approached using *densely sampled* local descriptors [67] or (densely extracted) descriptors based on convolutional neural networks [6, 35, 40, 76]. However, the main shortcoming of all the above methods is that they output only an approximate location of the query, not an exact 6DoF pose.

Visual localization using 3D maps. Another approach is to directly obtain 6DoF camera pose with respect to a pre-

built 3D map. The map is usually composed of a 3D point cloud constructed via Structure-from-Motion (SfM) [2] where each 3D point is associated with one or more local feature descriptors. The query pose is then obtained by feature matching and solving a Perspective-n-Point problem (PnP) [14, 15, 20, 29, 34, 38, 53, 55]. Alternatively, pose estimation can be formulated as a learning problem, where the goal is to train a regressor from the input RGB(D) space to camera pose parameters [11, 34, 59, 74]. While promising, scaling these methods to large-scale datasets is still an open challenge.

Indoor 3D maps. Indoor scene datasets [50, 52, 62, 69] have been introduced for tasks such scene recognition, classification, and object retrieval. With the increased availability of laser range scanners and time-of-flight (ToF) sensors, several datasets include depth data besides RGB images [5, 10, 23, 26, 36, 60, 79] and some of these datasets also provide reference camera poses registered into the 3D point cloud [10, 26, 79], though their focus is not on localization. Datasets focused specifically on indoor localization [59, 64, 70] have so far captured fairly small spaces such as a single room (or a single floor at largest) and have been constructed from densely-captured sequences of RGBD images. More recent datasets [17, 77] provide larger scale (multi-floor) indoor 3D maps containing RGBD images registered to a global floor map. However, they are designed for object retrieval, 3D reconstruction, or training deep-learning architectures. Most importantly, they do not contain query images taken from viewpoints far from database images, which are necessary for evaluating visual localization.

To address the shortcomings of the above datasets for large-scale indoor visual localization, we introduce a new dataset that includes query images captured at a different time from the database, taken from a wide range of viewpoints, with a considerably larger 3D database distributed across multiple floors of multiple buildings. Furthermore, our dataset contains various difficult situations for visual localization, *e.g.*, textureless and highly symmetric office scenes, repetitive tiles, and repetitive objects that confuse the existing visual localization methods designed for outdoor scenes. The newly collected dataset is described next.

3. The InLoc dataset for visual localization

Our dataset is composed of a database of RGBD images geometrically registered to the floor maps augmented with a separate set of RGB query images taken by hand-held devices to make it suitable for the task of indoor localization (Figure 2). The provided query images are annotated with manually verified ground-truth 6DoF camera poses (reference poses) in the global coordinate system of the 3D map.

Database. The base indoor RGBD dataset [77] consists of 277 RGBD panoramic images obtained from scanning two buildings at the Washington University in St. Louis with a Faro 3D scanner. Each RGBD panorama has about 40M 3D points in color. The base images are divided into five scenes: DUC1, DUC2, CSE3, CSE4, and CSE5, representing five floors of the mentioned buildings, and are geometrically registered to a known floor plan [77]. The scenes are scanned sparsely on purpose, to cover a larger area with a small number of scans to reduce the required manual work, as well as due to the long operating times of the high-end scanner used. The area per scan varies between 23.5 and 185.8 m^2. This inherently leads to critical view changes between query and database images when compared with other existing datasets [64, 70, 75][2].

For creating an image database suitable for indoor visual localization evaluation, a set of perspective images is generated by following the best practices from outdoor visual localization [19, 67, 80]. We obtain 36 perspective RGBD images from each panorama by extracting standard perspective views (60° FoV) with a sampling stride of 30° in yaw and ±30° in pitch directions, resulting in 10K perspective images in total (Table 1). Our database contains significant challenges, such as repetitive patterns (stairs, pillars), frequently appearing building structures (doors, windows), furniture changing position, people moving across the scene, and textureless and highly symmetric areas (walls, floors, corridors, classrooms, open spaces).

Query images. We captured 356 photos using a smartphone camera (iPhone 7), distributed only across two floors, DUC1 and DUC2. The other three floors in the database are not represented in the query images, and play the role

	Number	Image size [pixel]	FoV [degree]
Query	356	4,032×3,024	65.57
Database	9,972	1,600×1,200	60

Table 1. Statistics of the **InLoc dataset**.

Figure 2. **Example images from InLoc dataset**. (Top) Database images. (Bottom) Query images. The selected images show the challenges encountered in indoor environments: even small changes in viewpoint lead to large differences in appearance; large textureless surfaces (*e.g.* walls); self-repetitive structures (*e.g.* corridors); significant variation throughout the day due to different illumination sources (*e.g.*, active vs. indirect illumination).

[2] For example, in the database of [64], the scans are distributed on one single floor, and the area per each database image is less than 45 m^2.

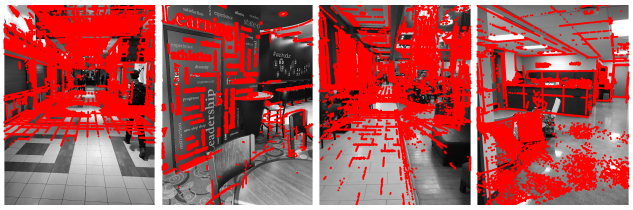


Figure 3. **Examples of verified query poses**. We evaluated the quality of the reference camera poses both visually and quantitatively, as described in section 3. Red dots are the database 3D points projected onto a query image using its estimated pose.

of confusers at search time, contributing to the building-scale localization scenario. Note that these query photos are taken at different times of the day, to capture the variety of occluders and layouts (*e.g.*, people, furniture) as well as illumination changes.

Reference pose generation. For all query photos, we estimate 6DoF reference camera poses w.r.t. the 3D map. Each query camera reference pose is computed as follows:

(i) Selection of the visually most similar database images. For each query, we manually select one panorama location which is visually most similar to the query image using the perspective images generated from the panorama.

(ii) Automatic matching of query images to selected database images. We match the query and perspective images by using affine covariant features [45] and nearest-neighbor search followed by Lowe's ratio test [42].

(iii) Computing the query camera pose and visually verifying the reprojection. All the panoramas (and perspective images) are already registered to the floor plan and have pixel-wise depth information. Therefore, we compute query pose via P3P-RANSAC [25], followed by bundle adjustment [3], using correspondences between query image points and scene 3D points obtained by feature matching. We evaluate the obtained poses visually by inspecting the reprojection of edges detected in the corresponding RGB panorama into the query image (see examples in figure 3).

(iv) Manual matching of difficult queries to selected database images. Pose estimation from automatic matches often gives inaccurate poses for difficult queries which are, *e.g.*, far from any database image. Hence, for queries with significant misalignment in reprojected edges, we manually annotate 5 to 20 correspondences between image pixels and 3D points and apply step (iii) on the manual matches.

(v) Quantitative and visual inspection. For all estimated poses, we measure the median reprojection error, computed as the distance of the reprojected 3D database point to the nearest edge pixel detected in the query image, after removing correspondences with gross errors (with distance over 20 pixels) due to, *e.g.*, occlusions. For query images that have under 5 pixels median reprojection error, we manually inspect the reprojected edges in the query image and finally accept **329 reference poses** out of the 356 query images.

4. Indoor visual localization with dense matching and view synthesis

We propose a new method for large-scale indoor visual localization. We address the three main challenges of indoor environments:

(1) Lack of sparse local features. Indoor environments are full of large textureless areas, *e.g.*, walls, ceilings, floors and windows, where sparse feature extraction methods detect very few features. To overcome this problem, we use *multi-scale dense CNN features* for both image description and feature matching. Our features are generic enough to be pre-trained beforehand on (outdoor) scenes, avoiding costly re-training, *e.g.*, as in [11, 34, 74], of the localization machine for each particular environment.

(2) Large image changes. Indoor environments are cluttered with movable objects, *e.g.*, furniture and people, and 3D structures, *e.g.*, pillars add concave bays, causing severe occlusions when viewed from a close distance. The most similar images obtained by retrieval may therefore be visually very different from a query image. To overcome this problem, we rely on *dense feature matches to collect as much positive evidence as possible*. We employ image descriptors extracted from a convolutional neural network that can match higher-level structures of the scene rather than relying on matching individual local features. In detail, our pose estimation step performs coarse-to-fine dense feature matching, followed by geometric verification and estimation of the camera pose using P3P-RANSAC.

(3) Self-similarity. Indoor environments are often very self-similar, *e.g.*, due to many symmetric and repetitive elements on a large and small scale (corridors, rooms, tiles, windows, chairs, doors, *etc.*). Existing matching strategies count the positive evidence, *i.e.*, how much of the image (or how many inliers) have been matched, to decide whether two images match. This is, however, problematic as large textureless areas can be matched well, hence providing strong (incorrect) positive evidence. To overcome this problem, we propose to count also the *negative evidence*, *i.e.*, what portion of the image does not match, to decide whether two views are taken from the same location. To achieve this, we perform *explicit pose estimate verification based on view synthesis*. In detail, we compare the query image with a virtual view of the 3D model rendered from the estimated camera pose of the query. This novel approach takes advantage of the high quality of the RGBD image database and incorporates both the positive and negative evidence by counting matching and non-matching pixels across the entire query image. As shown by our experiments, this approach is orthogonal to the choice of local

descriptors. The proposed verification by view synthesis is consistently showing a significant improvement regardless of the choice of features used for estimating the pose.

The pipeline of InLoc has the following three steps. Given a query image, (1) we obtain a set of candidate images by finding the N best matching images from the reference image database registered to the map. (2) For these N retrieved candidate images, we compute the query poses using the associated 3D information that is stored together with the database images. (3) Finally, we re-rank the computed camera poses based on verification by view synthesis. The three steps are detailed next.

4.1. Candidate pose retrieval

As demonstrated by existing work [6, 35, 67], aggregating feature descriptors computed densely on a regular grid mitigates issues such as a lack of repeatability of local features detected on textureless scenes, large-illumination changes, and a lack of discriminability of image description, dominated by features from repetitive structures (burstiness). As already mentioned in section 1, these problems are also occurring in large-scale indoor localization, which motivates our choice of using an image descriptor based on dense feature aggregation. Both query and database images are described by NetVLAD [6] (but other variants could also be used), normalized L2 distances of the descriptors are computed, and the poses of the N best matching images from the database are chosen as candidate poses. In section 5, we compare our approach with the state-of-the-art image descriptors based on local feature detection and show benefits of our approach for indoor localization.

4.2. Pose estimation using dense matching

A severe problem in indoor localization is that standard geometric verification based on local feature detection [51,54] does not work on textureless or self-repetitive scenes, such as corridors, where robots (and also humans) often get lost. Motivated by the improvements in candidate pose retrieval with dense feature aggregation (Section 4.1), we use features densely extracted on a regular grid for verifying and re-ranking the candidate images by feature matching and pose estimation. A possible approach would be to match DenseSIFT [41] followed by RANSAC-based verification. Instead of tailoring DenseSIFT description parameters (patch sizes, strides, scales) to match across images with significant viewpoint changes, we use an image representation extracted by a convolutional neural network (VGG-16 [61]) as a set of multi-scale features extracted on a regular grid that describes more higher-level information with a larger receptive field (patch size).

We first find geometrically consistent sets of correspondences using the coarser conv5 layer containing high-level information. Then we refine the correspondence by searching for additional matches on the conv3 layer. Examples in figure 4 demonstrate that our dense CNN matching (4th column) obtains better matches in indoor environments when compared to matching standard local features (3rd column), even for less-textured areas. Notice that dense-feature extraction and description requires no additional computation at query time as the intermediate convolutional layers are already computed when extracting the NetVLAD descriptors as described in section 4.1. As will also be demonstrated in section 5, memory requirements and computational speed of feature matching can be addressed by binarizing the convolutional features without loss in matching performance.

As perspective images in our database have depth values, and hence associated 3D points, the query camera pose can be estimated by finding pixel-to-pixel correspondences between the query and the matching database image followed by P3P-RANSAC [25].

4.3. Pose verification with view synthesis

We propose here to collect both positive and negative evidence to determine what *is* and *is not* matched[3]. This is achieved by harnessing the power of the high-quality RGBD image database that provides a dense and accurate 3D structure of the indoor environment. This structure is used to render a virtual view that shows how the scene would look like from the estimated query pose. The rendered image enables us to count, in a pixel-wise manner, both positive and negative evidence by counting which regions are and are not consistent between the query image and the underlying 3D structure. To gain invariance to illumination changes and small misalignments, we evaluate image similarity by comparing local patch descriptors (DenseRootSIFT [7, 41]) at corresponding pixel locations. The final similarity is computed as the median of descriptor distances across the entire image while ignoring areas with missing 3D structure.

5. Experiments

We first describe the experimental setup for evaluating visual localization performance using our dataset (Section 5.1). The proposed method, termed "InLoc", is compared with state-of-the-art methods (Section 5.2) and we show the benefits of each component in detail (Section 5.3).

5.1. Implementation details

In the candidate pose retrieval step, we retrieve 100 candidate database images using NetVLAD. We use the implementation provided by the authors and the pre-trained Pitts30K [6] VGG-16 [61] model to generate $4,096$-dimensional NetVLAD descriptor vectors.

[3]The impact of negative evidence in feature aggregation is demonstrated in [30].

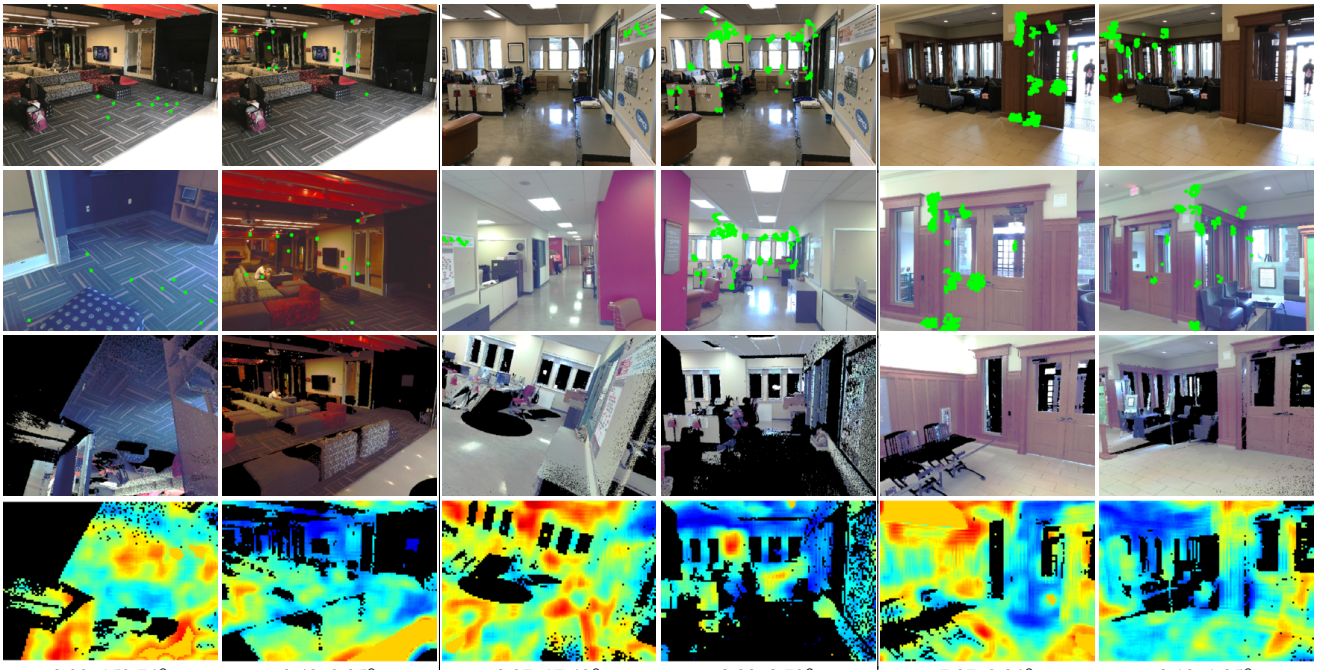


Figure 4. **Qualitative comparison of different localization methods (columns).** From top to bottom: query image, the best matching database image, synthesized view at the estimated pose (without inter/extra-polation), error map between the query image and the synthesized view, localization error (meters, degrees). Green dots are the inlier matches obtained by P3P-LO-RANSAC. Methods using the proposed dense pose estimation (DensePE) and dense pose verification (DensePV) are shown in bold. The query images in the 2nd, 4th and 6th column are well localized within 1.0 meters and 5.0 degrees whereas localization results in the 1st, 3rd and 5th column are incorrect. Additional qualitative results are given in [66].

In the second pose estimation step, we obtain tentative correspondences by matching densely extracted convolutional features in a coarse-to-fine manner: we first find mutually nearest matches among the conv5 features and then find matches in the finer conv3 features restricted by the coarse conv5 correspondences. The tentative matches are geometrically verified by estimating up to two homographies using RANSAC [25]. We re-rank the 100 candidates using the number of RANSAC inliers and keep the top-10 database images. For each of the 10 images, the 6DoF query pose is computed by P3P-LO-RANSAC [37] (referred to as *DensePE*), assuming a known focal length, *e.g.*, from EXIF data, using the inlier matches and depth (*i.e.* the 3D structure) associated to each database image.

In the final pose verification step, we generate synthesized views by rendering colored 3D points while taking care of self-occlusions. For computing the scores that measure the similarities of the query image and the image rendered from the estimated pose, we use the DenseSIFT extractor and its RootSIFT descriptor [7, 41] from VLFeat [72][4]. Finally, we localize the query image by the best pose among its top-10 candidates.

Evaluation metrics. We evaluate the localization accuracy as the consistency of the estimated poses with our reference poses. We measure positional and angular differences in meters and degrees between the estimated poses and the manually verified reference poses.

5.2. Comparison with the state-of-the-art methods

We compare the proposed "InLoc" approach with several state-of-the-art localization methods.

Direct 2D-3D matching [53, 55]. We first compare with a variation[5] of a state-of-the-art 3D structure-based image localization approach [53]. We compute affine covariant RootSIFT features for all the database images and associate them with 3D coordinates via the known scene geometry. Features extracted from a query image are then matched to the database 3D descriptors [46]. We select at most five database images receiving the largest numbers of matches

[4]When computing the descriptors, the blank pixels induced by missing 3D points are filled by linear inter(/extra)-polation using the values of non-blank pixels on the boundary.

[5]Due to the sparse sampling of viewpoints in our indoor dataset, we cannot establish feature tracks between database images. This prevents us from applying algorithms relying on co-visibility [20, 38, 53, 55, 81].

	Direct2D-3D [53]	Disloc [9] +SparsePE	NetVLAD +SparsePE	InLoc (Ours)
0.25m	11.9	20.1	21.3	**38.9**
0.50m	15.8	29.5	30.7	**56.5**
1.00m	22.5	41.0	42.6	**69.9**

Table 2. **Comparison with the state-of-the-art localization methods on the InLoc dataset.** We show the rate (%) of correctly localized queries within a given distance (m) threshold and within a 10° angular error threshold.

and use all these matches together for pose estimation. Similar to [53], we did not apply Lowe's ratio test [42] as it lowered the performance. The 6DoF query pose is finally computed by P3P-LO-RANSAC [37]. As shown in table 2, InLoc outperforms direct 2D-3D matching by a large margin (40.7% at the localization accuracy of 0.5m). We believe that this is because our large-scale indoor dataset involves many distractors and large viewpoint changes that present a major challenge for 3D structure-based methods.

Disloc [9] + sparse pose estimation (SparsePE) [51]. We next compare with the state-of-the-art image retrieval-based localization method. Disloc represents images using bag-of-visual-words with Hamming-Embedding [31] while also taking local descriptor space density into account. We use a publicly available implementation [54] of Disloc with a 200K vocabulary trained on affine covariant features [45], described by RootSIFT [7], extracted from the database images of our indoor dataset. The top-100 candidate images shortlisted by Disloc are re-ranked by spatial verification [51] using (sparse) affine covariant features [45]. The ratio test [42] was not applied here as it was removing too many features that need to be retained in the indoor scenario. Using the inliers, the 6DoF query pose is computed with P3P-LO-RANSAC [37]. To make a fair comparison, we use exactly the same features and P3P-LO-RANSAC for pose estimation as the direct 2D-3D matching method described above. As shown in table 2, Disloc [9]+SparsePE [51] results in a 13.7% performance gain compared to Direct 2D-3D matching [55]. This can be attributed to the image retrieval step that discounts burst of repetitive features. However, the results are still significantly worse compared to our InLoc approach.

NetVLAD [6] + sparse pose estimation (SparsePE) [51]. We also evaluate a variation of the above image retrieval-based localization method. Here the candidate shortlist is obtained by NetVLAD [6], which is then re-ranked using SparsePE [51], followed by pose estimation using P3P-LO-RANSAC [37]. This is a strong baseline building on the state-of-the-art place recognition results obtained by [6]. Interestingly, as shown in table 2, there is no significant difference between NetVLAD+SparsePE and DisLoc+SparsePE, which is in line with results reported in outdoor settings [57]. Yet, NetVLAD outperforms DisLoc (5.8% at the localization accuracy of 0.5m) before re-ranking via SparsePE (*c.f.* figure 5) in this indoor setting (see also figure 4). Overall, both methods, even though they represent the state-of-the-art in outdoor localization, still perform significantly worse than our proposed approach based on dense feature matching and view synthesis.

5.3. Evaluation of each component

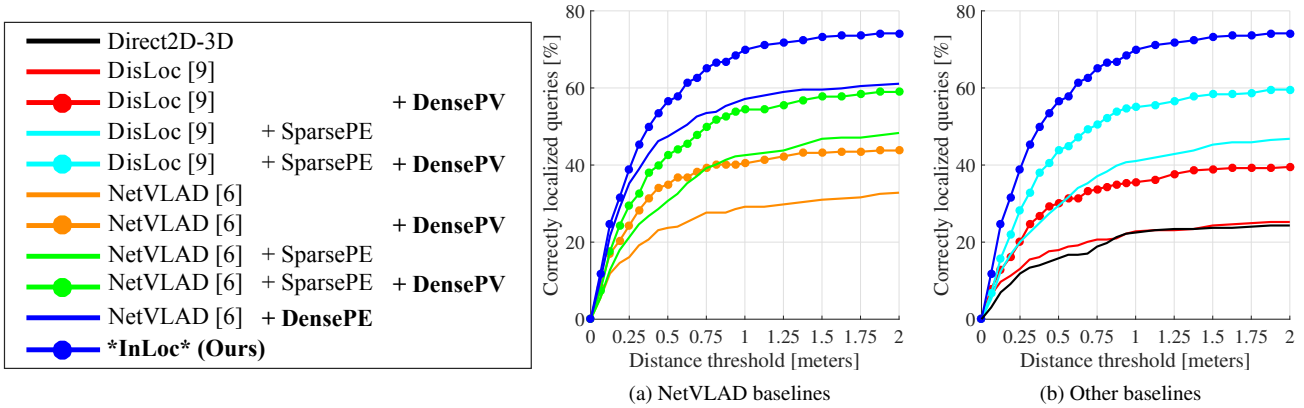
Next, we demonstrate the benefits of the individual components of our approach.

Benefits of pose estimation using dense matching. Using the NetVLAD retrieval as the base retrieval method (Figure 5 (a)), our pose estimation with dense matching (NetVLAD [6]+**DensePE** "–") constantly improves the localization rate by about 15% when compared to the state-of-the-art sparse local feature matching (NetVLAD [6]+SparsePE "–"). This result supports our conclusion that dense feature matching and verification is superior to sparse feature matching for often weakly textured indoor scenes. This effect is also clearly demonstrated in qualitative results in figure 4 (cf. columns 3 and 4).

Benefits of pose verification with view synthesis. We apply our pose verification step (**DensePV**) to the top–10 pose estimates obtained by different spatial re-reranking methods. Results are shown in figure 5 and demonstrate significant and consistent improvements obtained by our pose verification approach (compare "-•-" to "—" in figure 5). Improvements are most pronounced for the position accuracy within 1.5 meters (13% or more).

Binarized representation. A binary representation (instead of floats) of features in the intermediate CNN layers significantly reduces memory requirements. We use feature binarization that follows the standard Hamming embedding approach [31] but without dimensionality reduction. Matching is then performed by computing Hamming distances. This simple binarization scheme results in a negligible performance loss (less than 1% at 0.5 meters) compared to the original descriptors, which is in line with results reported for object recognition [4]. At the same time, binarization reduces the memory requirements by a factor of 32, compressing 428GB of original descriptors to just 13.4GB.

Comparison with learning based localization methods. We have attempted a comparison with DSAC [11], which is a state-of-the-art pose estimator for indoor scenes. Despite our best efforts, training DSAC on our indoor dataset failed to converge. We believe this is because the RGBD scans in our database are sparsely distributed [77] and each scan has only a small overlap with neighboring scans. Training on such a dataset is challenging for methods designed for densely captured RGBD sequences [26]. We believe this would also be the case for PoseNet [34], another method for CNN-based pose regression. We do provide the comparison with DSAC and PoseNet on much smaller datasets next.



(a) NetVLAD baselines (b) Other baselines

Figure 5. **Impact of different components.** The graphs show impact of dense matching (**DensePE**) and dense pose verification (**DensePV**) on pose estimation quality for (a) the pose candidates retrieved by NetVLAD and (b) state-of-the-art baselines. Plots show the fraction of correctly localized queries (y-axis) within a certain distance (x-axis) whose rotation error is at most 10°.

	Disloc [9] +SparsePE	NetVLAD [6] +SparsePE	NetVLAD [6] +DensePE	InLoc (Ours)
90 bldgs.	0.42, 4.58°	0.44, 4.70°	0.23, 2.53°	**0.17, 2.15°**

Table 3. **Comparison on Matterport3D [17].** Numbers show the median positional (m) and angular (degrees) errors.

Scene	PoseNet [34]	ActiveSearch [55]	DSAC [11, 13]	NetVLAD [6] +SparsePE [51]	NetVLAD [6] **+DensePE**
Chess	13, 4.48°	4, 1.96°	**2**, 1.2°	4, 1.83	3, **1.05°**
Fire	27, 11.3°	**3**, 1.53°	4, 1.5°	4, 1.55	**3**, **1.07°**
Heads	17, 13.0°	**2**, 1.45°	3, 2.7°	**2**, 1.65	**2**, **1.16°**
Office	19, 5.55°	9, 3.61°	4, 1.6°	5, 1.49	**3**, **1.05°**
Pumpkin	26, 4.75°	8, 3.10°	**5**, 2.0°	7, 1.87	**5**, **1.55°**
Red kit.	23, 5.35°	7, 3.37°	5, 2.0°	5, 1.61	**4**, **1.31°**
Stairs	35, 12.4°	**3**, **2.22°**	117, 33.1°	12, 3.41	9, 2.47°

Table 4. **Evaluation on the 7 Scenes dataset [26, 59].** Numbers show the median positional (cm) and angular errors (degrees).

5.4. Evaluation on other datasets

We also evaluate InLoc on two existing indoor datasets [17, 59] to confirm the relevance of our results. The Matterport3D [17] dataset consists of RGBD scans of 90 buildings. Each RGBD scan contains 18 images that capture the scene around the scan position with known camera poses. We created a test set by randomly choosing 10% of the scan positions and selected their horizontal views. This resulted in 58,074 database images and a query set of 6,726 images. Results are shown in table 3. Our approach (InLoc) outperforms the baselines, which is in line with results on the InLoc dataset. We also tested PoseNet [34] and DSAC [11] on a single (the largest) building. The test set is created in the same manner as above and contains 1,884 database images and 210 query images. Even in this much easier case, DSAC fails to converge. PoseNet produces large localization errors (24.8 meters and 80.0 degrees) in comparison with InLoc (0.26 meters and 2.78 degrees).

We also report results on the 7 Scenes dataset [26, 59] which is, while relatively small, a standard benchmark for indoor localization. The 7 Scenes dataset [59] consists of geometrically-registered video frames representing seven scenes, together with associated depth images and camera poses. Table 4 shows localization results for our approach (NetVLAD+**DensePE**) compared with state-of-the-art methods [11, 34, 55]. Note that our approach performs comparably to these methods on this relatively small and densely captured data, while it does not need any scene specific training (which is needed by [11, 34]).

6. Conclusion

We have presented InLoc – a new approach for large-scale indoor visual localization that estimates the 6DoF camera pose of a query image with respect to a large indoor 3D map. To overcome the difficulties of indoor camera pose estimation, we have developed new pose estimation and verification methods that use dense feature extraction and matching in a sequence of progressively stricter verification steps. The localization performance is evaluated on a new large indoor dataset with realistic and challenging query images captured by mobile phones. Our results demonstrate significant improvements compared to state-of-the-art localization methods. To encourage further progress on high-accuracy large-scale indoor localization, we make our dataset publicly available [1].

Acknowledgements. This work was partially supported by JSPS KAKENHI Grant Numbers 15H05313, 17H00744, 17J05908, EU-H2020 project LADIO No. 731970, ERC grant LEAP No. 336845, CIFAR Learning in Machines & Brains program and the European Regional Development Fund under the project IMPACT (reg. no. CZ.02.1.01/0.0/0.0/15_003/0000468). The authors would like to express the deepest appreciation to Yasutaka Furukawa for his arrangement to capture query photographs at Washington University in St. Louis.

References

[1] Project webpage. http://www.ok.sc.e.titech.ac.jp/INLOC/. 2, 8

[2] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Comm. ACM*, 54(10):105–112, 2011. 3

[3] S. Agarwal, K. Mierle, and Others. Ceres solver. http://ceres-solver.org. 4

[4] P. Agrawal, R. B. Girshick, and J. Malik. Analyzing the performance of multilayer neural networks for object recognition. In *Proc. ECCV*, 2014. 7

[5] A. Anand, H. S. Koppula, T. Joachims, and A. Saxena. Contextually guided semantic labeling and search for three-dimensional point clouds. *Intl. J. of Robotics Research*, 32(1):19–34, 2013. 3

[6] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proc. CVPR*, 2016. 1, 2, 5, 6, 7, 8

[7] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR*, 2012. 2, 5, 6, 7

[8] R. Arandjelovic and A. Zisserman. All about vlad. In *Proc. CVPR*, 2013. 2

[9] R. Arandjelović and A. Zisserman. Dislocation: Scalable descriptor distinctiveness for location recognition. In *Proc. ACCV*, 2014. 1, 2, 6, 7, 8

[10] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3D Semantic Parsing of Large-Scale Indoor Spaces. In *Proc. CVPR*, 2016. 2, 3

[11] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother. DSAC - Differentiable RANSAC for Camera Localization. In *Proc. CVPR*, 2017. 1, 3, 4, 7, 8

[12] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother. Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image. In *Proc. CVPR*, 2016. 1

[13] E. Brachmann and C. Rother. Learning less is more-6d camera localization via 3d surface regression. In *Proc. CVPR*, 2018. 8

[14] F. Camposeco, T. Sattler, A. Cohen, A. Geiger, and M. Pollefeys. Toroidal constraints for two-point localization under high outlier ratios. In *Proc. CVPR*, 2017. 1, 3

[15] S. Cao and N. Snavely. Minimal scene descriptions from structure from motion models. In *Proc. CVPR*, 2014. 1, 3

[16] R. O. Castle, G. Klein, and D. W. Murray. Video-rate localization in multiple maps for wearable augmented reality. In *ISWC*, 2008. 1

[17] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *Proc. 3DV*, 2017. 2, 3, 8

[18] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, H. Chen, R. Vedantham, R. Grzeszczuk, and B. Girod. Residual enhanced visual vectors for on-device image matching. In *Proc. ASILOMAR*, 2011. 2

[19] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, et al. City-scale landmark identification on mobile devices. In *Proc. CVPR*, 2011. 1, 3

[20] S. Choudhary and P. Narayanan. Visibility probability structure from sfm datasets and applications. In *Proc. ECCV*, 2012. 1, 3, 6

[21] O. Chum, A. Mikulik, M. Perdoch, and J. Matas. Total recall ii: Query expansion revisited. In *Proc. CVPR*, 2011. 2

[22] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV*, 2007. 2

[23] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *Proc. CVPR*, 2017. 2, 3

[24] A. Debski, W. Grajewski, W. Zaborowski, and W. Turek. Open-source localization device for indoor mobile robots. *Procedia Computer Science*, 76:139–146, 2015. 1

[25] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24(6):381–395, 1981. 4, 5, 6

[26] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi. Real-time RGB-D camera relocalization. In *Proc. ISMAR*, 2013. 2, 3, 7, 8

[27] P. Gronat, G. Obozinski, J. Sivic, and T. Pajdla. Learning and calibrating per-location classifiers for visual place recognition. In *Proc. CVPR*, 2013. 1, 2

[28] M. Halber and T. Funkhouser. Fine-To-Coarse Global Registration of RGB-D Scans. In *Proc. CVPR*, 2017. 2

[29] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *Proc. CVPR*, 2009. 1, 2, 3

[30] H. Jégou and O. Chum. Negative evidences and cooccurrences in image retrieval: the benefit of PCA and whitening. In *Proc. ECCV*, 2012. 5

[31] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV*, 2008. 2, 7

[32] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *Proc. CVPR*, 2009. 2

[33] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE PAMI*, 34(9):1704–1716, 2012. 2

[34] A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. In *Proc. CVPR*, 2017. 1, 3, 4, 7, 8

[35] H. J. Kim, E. Dunn, and J.-M. Frahm. Learned contextual feature reweighting for image geo-localization. In *Proc. CVPR*, 2017. 2, 5

[36] K. Lai, L. Bo, and D. Fox. Unsupervised feature learning for 3D scene labeling. In *Proc. Intl. Conf. on Robotics and Automation*, 2014. 3

[37] K. Lebeda, J. Matas, and O. Chum. Fixing the locally optimized ransac–full experimental evaluation. In *Proc. BMVC.*, 2012. 6, 7

[38] Y. Li, N. Snavely, D. P. Huttenlocher, and P. Fua. Worldwide pose estimation using 3d point clouds. In *Proc. ECCV*, 2012. 1, 3, 6

[39] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele. Real-time image-based 6-dof localization in large-scale environments. In *Proc. CVPR*, 2012. 1

[40] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *Proc. ICCV*, 2015. 2

[41] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. Sift flow: Dense correspondence across different scenes. In *Proc. ECCV*, 2008. 5, 6

[42] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 4, 7

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

[44] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt. Scalable 6-dof localization on mobile devices. In *Proc. ECCV*, 2014. 1

[45] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004. 4, 7

[46] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithmic configuration. In *Proc. VISAPP*, 2009. 6

[47] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Proc. ISMAR*, 2011. 2

[48] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3D Reconstruction at Scale using Voxel Hashing. *ACM TOG*, 32(6):169, 2013. 2

[49] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. CVPR*, 2006. 2

[50] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *Proc. ICCV*, 2011. 3

[51] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 2007. 2, 5, 7, 8

[52] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Proc. CVPR*, 2009. 3

[53] T. Sattler, M. Havlena, F. Radenovic, K. Schindler, and M. Pollefeys. Hyperpoints and fine vocabularies for large-scale location recognition. In *Proc. ICCV*, 2015. 1, 3, 6, 7

[54] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys. Large-scale location recognition and the geometric burstiness problem. In *Proc. CVPR*, 2016. 1, 2, 5, 7

[55] T. Sattler, B. Leibe, and L. Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE PAMI*, 39(9):1744–1756, 2017. 1, 3, 6, 7, 8

[56] 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 *Proc. CVPR*, 2018. 1

[57] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, and T. Pajdla. Are large-scale 3D models really necessary for accurate visual localization? In *Proc. CVPR*, 2017. 1, 7

[58] T. Schmidt, R. Newcombe, and D. Fox. Self-Supervised Visual Descriptor Learning for Dense Correspondence. *RAL*, 2(2):420–427, 2017. 1

[59] 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 *Proc. CVPR*, 2013. 1, 3, 8

[60] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *Proc. ECCV*, 2012. 3

[61] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015. 5

[62] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *Proc. ECCV*, 2012. 3

[63] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003. 2

[64] X. Sun, Y. Xie, P. Luo, and L. Wang. A Dataset for Benchmarking Image-based Localization. In *Proc. CVPR*, 2017. 1, 2, 3

[65] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson. City-Scale Localization for Cameras with Known Vertical Direction. *IEEE PAMI*, 39(7):1455–1461, 2017. 1

[66] 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. *arXiv preprint arXiv:1803.10368*, 2018. 6

[67] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *Proc. CVPR*, 2015. 1, 2, 3, 5

[68] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual place recognition with repetitive structures. In *Proc. CVPR*, 2013. 1, 2

[69] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision systems for place and object recognition. In *Proc. ICCV*, 2003. 3

[70] J. Valentin, A. Dai, M. Nießner, P. Kohli, P. Torr, S. Izadi, and C. Keskin. Learning to Navigate the Energy Landscape. In *Proc. 3DV*, 2016. 1, 3

[71] J. C. Van Gemert, C. J. Veenman, A. W. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *IEEE PAMI*, 32(7):1271–1283, 2010. 2

[72] A. Vedaldi and B. Fulkerson. VLFeat - an open and portable library of computer vision algorithms. In *Proc. ACMM*, 2010. 6

[73] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond, and D. Schmalstieg. Real-time detection and tracking for augmented reality on mobile phones. *Visualization and Computer Graphics*, 16(3):355–368, 2010. 1

[74] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-Based Localization Using LSTMs for Structured Feature Correlation. In *Proc. ICCV*, 2017. 3, 4

[75] S. Wang, S. Fidler, and R. Urtasun. Lost shopping! monocular localization in large indoor spaces. In *Proc. ICCV*, 2015. 1, 2, 3

[76] T. Weyand, I. Kostrikov, and J. Philbin. Planet-photo geolocation with convolutional neural networks. In *Proc. ECCV*, 2016. 1, 2

[77] E. Wijmans and Y. Furukawa. Exploiting 2D floorplan for building-scale panorama RGBD alignment. In *Proc. CVPR*, 2017. 2, 3, 7

[78] J. Xiao and Y. Furukawa. Reconstructing the worlds museums. *IJCV*, 110(3):243–258, 2014. 2

[79] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using sfm and object labels. In *Proc. ICCV*, 2013. 2, 3

[80] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. In *Proc. ECCV*, 2010. 1, 3

[81] B. Zeisl, T. Sattler, and M. Pollefeys. Camera Pose Voting for Large-Scale Image-Based Localization. In *Proc. ICCV*, 2015. 1, 6