JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY

ReLoc: Indoor Visual Localization with Hierarchical Sitemap and View Synthesis

Abstract Nowadays indoor visual localization, i.e. 6 DoF camera pose estimation for a query image with respect to a known scene, is gaining much more attentions driven by rapid progress of applications such as robotics and augmented reality. However, drastic visual discrepancies between an onsite query image and prerecorded indoor images cast a big challenge for visual localization. In this paper, based on the key observation of the constant existence of planar surfaces such as floors or walls in indoor scenes, we propose a novel system incorporating geometric information to address issues only using pixelated images. In system implementation, we contribute a hierarchical structure consisting of pre-scanned images and point cloud as well as a distilled representation of planar elements layout extracted from the original dataset. A view synthesis procedure is designed for generating synthetic images as complementary to that of sparse sampled dataset. Moreover, a global image descriptor based on the image statistic modality, called BMVC, is employed to speed up the candidate poses identification incorporated with traditional CNN descriptor. Experimental results on a popular benchmark demonstrate that the proposed method outperforms the state-of-the-art approaches in visual localization validity and accuracy.

Keywords visual localization, planar surface, statistic information, view synthesis

1 Introduction

Visual localization is the task of 6 degree-of-freedom (DoF) pose estimation for a query image with respect to a known scene. It is a key problem in robotics and computer vision, highly relevant to Structure-from-Motion (SfM) [1], Simultaneous Localization and Mapping (SLAM) [2, 3], and applications such as autonomous driving and Augmented Reality (AR).

State-of-the-art approaches of precise visual localization are generally divided into three categories, namely structure-based methods [4]-[7], image retrieval-based methods [8]-[13] and learning-based methods [14]-[16]. Traditional structure-based and image retrieval-based methods can be boiled down to a descriptor-matching problem. Recently learning-based approaches have become popular, some of which aims to regress the camera pose in the end-to-end fashion without the need for a 3D model and others tend to make one or more modules in the traditional pipeline learnable. However, those methods without using 3D models sometimes return poor accuracy of pose estimation [17], while the image retrieval-based manner is promising as it can be easily generalized to novel scenes.

The coarse-to-fine image retrieval-based localization [12, 13, 18, 19] paradigm combines the strengths of structure-based and learning-based approaches, and is gaining popularity with recent advances of machine learning. It first leverages a learning-based global image descriptor to retrieve location hypotheses and then performs local feature matching to estimate the agent pose from those candidates. Usually, image descriptors are extracted from Convolutional Neural Networks (CNNs), VGG [18] or ResNet [19], for approaches using the hierarchical localization paradigm [20, 21]. NetVLAD [22], one of the state-of-the-art CNN-based descriptors, is widely utilized in the task of visual localization for both outdoor [9], [23]-[26] and indoor [12, 13] environments. It performs well even under large variations in image appearance such as day-night and seasonal change which commonly happened in urban environment. However, it still needs more work especially for indoor localization problem.

The task of visual localization for indoor environments has received considerably less attention compared with that for outdoors. InLoc and its variants [12, 13] are state-of-the-art indoor localization ap-

Regular Paper

proaches using NetVLAD for retrieval as the first step in the pipeline, however, the top-ranked candidates may not include location hypotheses. Furthermore, we observe that these approaches sometimes fail to regress the correct pose due to sparse discretization of the database. Indoor localization is a harder problem than urban localization [12] in many ways, e.g., 1) drastic change in scene appearance over time as furniture move and people walk, 2) large variation in viewpoints, and 3) common occurrence of symmetrical layout and repetitive elements, etc. Under such situations, the problem of perceptual aliasing easily arises which significantly decreases the accuracy of localization.

To handle these difficulties, we propose a two-stage ReLoc system which includes off-line structured sitemap construction and on-line visual localization. In detail, our contributions are mainly three-fold. 1) Propose hierarchical sitemap construction in the off-line stage. We generate synthetic views from different viewpoints on the extracted scene layout to enrich the database, and establish a hierarchical sitemap for the extended database to offer convenient query/track of the image/geometry data. The constructed sitemap, thereby, is not only a structured description of original database, but also a distilled representation of extended database. 2) Propose Blocked Mean, Variance and Colors (BMVC), a novel global image descriptor based on the statistic information of an image. We use BMVC to re-rank the shortlist of image candidates that have been retrieved via CNN-based global descriptor. 3) Propose a novel similarity function to determine the final pose from multiple candidate poses in the pose verification stage. It focuses on both the image appearance and the geometry layout, which is invariant to drastic variation in scene appearance over time as furniture move and people walk.

2 Related Work

2.1 Visual Localization

Structure-based approaches rely on 2D-3D matches between 2D pixels and 3D scene points for pose estimation. Matches are established by descriptor matching [4, 5] or by regressing 3D coordinates from pixel patches [27, 28]. 3D coordinate regression methods currently achieve a higher pose accuracy at small scale, but have not yet been shown to scale to larger scenes [27].

Recently, learning-based approaches have become popular. Some approaches make certain modules in the traditional localization pipeline learnable, e.g., learning-based detectors, learning-based descriptor [16], learning-based matchers [17], or scene coordinate regression [29], and others [14, 15] aim to learn in a single network to regress the camera pose from a test image without using a 3D model.

Image retrieval-based localization [8]-[13] is promising as it can be easily generalized to novel scenes. However, image retrieval-based localization baseline has so far paid less attention to re-ranking. Re-ranking and the pose-estimation step are tightly coupled for some approaches [12, 13]. They re-rank the retrieved database images by the count of match inliers, which is computationally expensive. Hierarchical MNV [21] re-rank the image candidates by clustering the locations by co-visibility in terms of the feature point, but can not deal with large parts of textureless scenes for indoor environments. In contrast, we bring to attention re-ranking by integrating additional statistic information.

2.2 Map Representation and Visual Localization Benchmarks

Scene map is mainly generated by two scanning approaches, namely Panorama RGB-D scanning and

RGB-D streaming, and the corresponding core technologies are SfM [1] and SLAM [2, 3], respectively. In the industry, the way of panorama RGB-D scanning has been extensively used as it directly provides visible panorama images as well as 3D point cloud. For large scenes such as airport terminals and museums, high-end panorama camera scanning are almost the only choice. After the success of Kinect Fusion [30], RGB-D streaming has become popular in the field of Computer Vision.

With the maturity of the SLAM framework and the rapid development of sensors, research about fusion of hybrid primitive features such as points, lines, planes and semantics has received much attention [31, 32]. Scene maps that have been constructed so far from densely-captured RGB-D sequences fairly focus on relatively small spaces ranging from room-scale to floor-scale at largest. For the task of localization, scene maps are further enhanced by, e.g., assigning feature descriptors or semantic clues to individual points in the point cloud [4]-[6], [35], or generating perspective RGB-D images from panoramic scans with a certain sampling stride [12].

Challenging visual localization benchmarks, e.g., RobotCar Seasons dataset [24] which contains lots of challenging conditions including illumination changes, day-night, season changes as well as weather variations, Aachen Day-Night dataset [25] which includes large viewpoint changes and query images taken at night time; Extended CMU Seasons dataset [26] which features the large variations in appearance of the scene, especially those in the suburban and park regions; and InLoc benchmark [12] which contains various difficult situations to make it suitable for the task of indoor localization such as weakly textured scenes, repetitive elements and symmetrical layout, drastic changes in viewpoint and large variations in appearance of the scene, etc.

We focus on the visual localization for indoor environments, thereby we test the proposed approach using the InLoc benchmark.

2.3 Pose Verification

The classical approach for pose verification is to select the pose with the largest number of inliers [10, 39, 40]. However, inlier count is not suitable in the scene with repetitive elements, which is usually appeared in indoor scenes. Recently the state-of-the-art approaches focus on the geometric [12, 13] or semantic [13], [33]-[35] consistency of feature matches. InLoc [12] proposes to re-render the scene from the estimated pose, and computes the similarity function using densely extracted RootSIFT features between the query and rendered image. However, they only consider the 3D geometry visible for a single scanned location when generating the synthetic views, which consequently increasing the number of invalid pixels (i.e. without depth value) in rendered views. On the other hand, these approaches take almost all pixels into account when computing the similarity function, which is not robust to scenes with significant clutter or movable objects. Previous work use semantic consistency in pose estimation [33]-[35] or pose verification [13], which is usually measured between a 3D point in the map (or a pixel in the database image) and its 2D projection in the query image. InLoc+N+S [13] follows the convention of InLoc [12], and furthermore integrates modalities in terms of surface normals and semantic clues into the verification stage. However, surface normals have to be further predicted using additional approaches for query images whose depth information is not available.

To sum up, there is still room for improvement using view synthesis for pose verification in indoor localization problems. We select the final pose by taking advantage of the explicitly structured sitemap, which

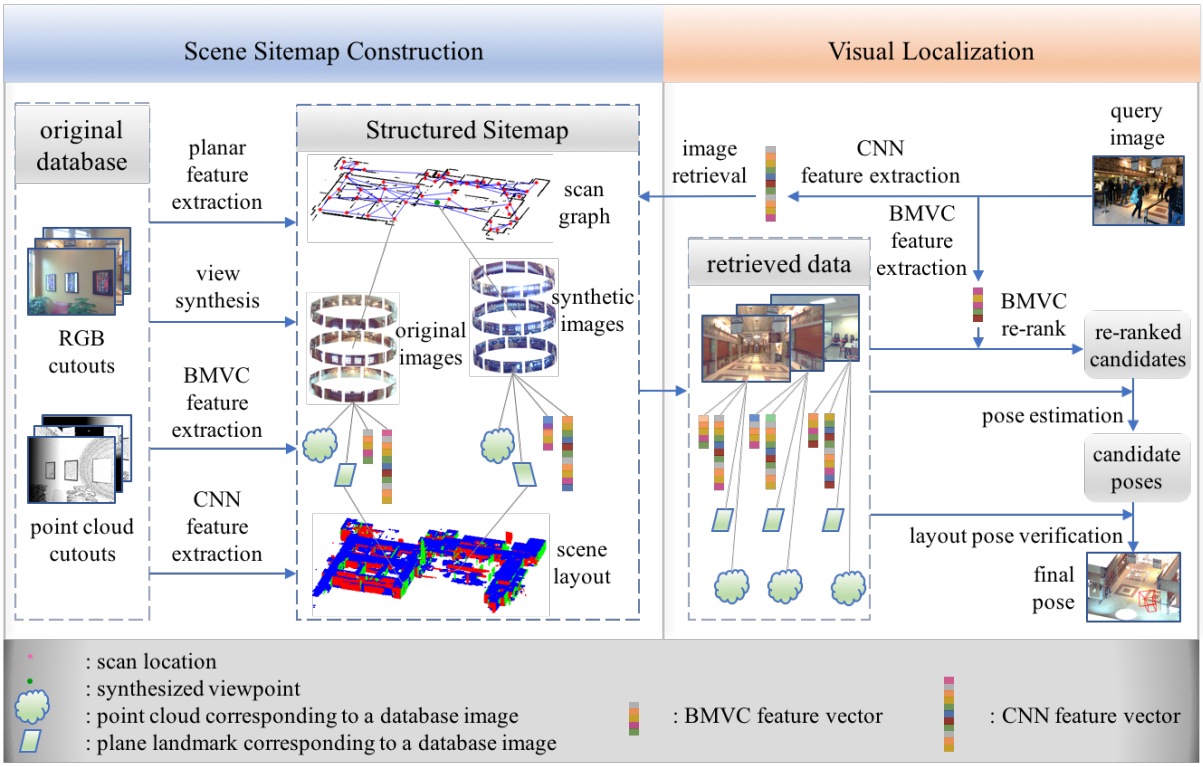

Fig.1. Overview of the proposed system. It includes two stages, off-line structured sitemap construction and on-line visual localization. In the off-line stage, we utilize view synthesis to enrich the original database and establish a structured sitemap. In the on-line stage, we propose a global image descriptor for re-ranking and a novel method to determine the final pose.

makes our pose verification method more effective under strong changes in viewpoint or image appearance.

3 Overview of ReLoc System

We propose an indoor visual localization system ReLoc based on hierarchical sitemap, which includes two stages, i.e. off-line structured sitemap construction and on-line visual localization, as shown in Figure 1.

The original database contains a set of perspective images created from RGB-D panoramic scans, each of which corresponds to a scan location in the scene. All panoramic scans have been registered to the scene, and the number of perspective images from each panorama are determined by the sampling stride in yaw and pitch directions.

The main task of the off-line stage is to construct sitemap based on those original database. It begins with appearance feature extraction. For a database image, extract both CNN-based feature descriptors and our BMVC descriptors. Besides, identify the planar surfaces (i.e. walls, ceilings and floors, etc.) from scanned point cloud, and then associate those planes with their corresponding images that the plane appears in. The synthetic views are generated from multiple synthesized viewpoints to enrich the database.

We construct a scan graph to organize the original database and extracted features as shown in Figure 1. Furthermore, to describe the overlap between different panoramic scans, we add links on two scan locations with at least two common planar surfaces. To distill the geometric features of the scene, the scene layout is constructed by the set of reliable planar surfaces. Both the scan graph and scene layout consist of our scene sitemap.

In the on-line stage, we use our descriptor BMVC to re-rank the shortlist of image candidates that have been

retrieved via CNN-based global descriptor, and then the bottom-ranked candidates are filtered out. Consequently, multiple camera pose hypotheses are estimated from top-ranked image candidates. Finally, in the pose-verification stage, a novel similarity function is employed to determine the final pose by taking advantage of the extracted scene layout conveniently.

4 Scene Sitemap Construction

Rather than directly using the database of RGB images and scanned point cloud, a hierarchical sitemap in the context of planar surfaces, synthetic views, BMVC features and CNN-based features is constructed additionally in the off-line stage.

4.1 Structured Sitemap

Our sitemap of a scene plays a key role in localization. Usually, a scene model consists of large point clouds and scanned images. The main goal of sitemap is to associate all the points data, images and extracted features to afford convenience to image-retrieval and localization by attaching a very light structure on the original data.

Our structured sitemap is constructed in off-line stage (refer to Figure 1). We introduce how to construct the scan graph at first. For building database, many scan locations were set up to capture RGB-D data. Therefore, we organize the RGB-D images of database according to the scan locations of the floors. Usually, for each individual scan location, there corresponds to a panoramic scan and multiple perspective database images. With the images and point clouds of this scene, the scan graph is initialized by adding those images to its corresponding scan location. For each image, its point cloud is linked to it naturally, and CNN features (we use NetVLAD here) and our BMVC features (explained in section 4.2) are extracted and also linked to this image.

The construction of scan graph is further strengthened with geometric description. We detect planar surfaces (explained in section 4.3) from point clouds, and then planar surfaces are registered on corresponding images on our scan graph as a geometric feature. Synthetic views (explained in section 4.4) are employed to compensate the shortage due to sparse of scan locations. Once a synthetic location is set up, the rendered views with images and point clouds are added with multiple extracted features into scan graph like a original data capture form a scan location. Obviously, it is convenient to locally add or delete synthesized viewpoints or scan locations, without changing the structure of the rest part of the sitemap.

We also construct scene layout to represent the geometry frame of a scene based on planar surfaces registered on images. Given two perspective images originated from different scan locations, if they see at least one same planar surfaces according to the plane association, an edge is connected between the two scan locations. This connection describes the overlap between different panoramic scans, which are thus useful for rendering the synthetic views.

Our scene sitemap is constructed by the scan graph and scene layout, which can benefit the pose verification discussed later. Since the data is very complex, the sitemap is actually hierarchical to present the association between different elements in the extended database. It should be mentioned that our sitemap mainly consists of links, and BMVC is also compact with low-dimension feature vector, thus the increasing of data of our sitemap is very limited.

4.2 Image Feature Extraction

For indoor environment, movable objects therein such as sofas, chairs and people act as noise in the task

of visual localization. Inspired by previous work outlined above in the field of retrieval, we leverage the statistic information of pixel pairs which might be invariant to noise like movable objects. Therefore, we propose BMVC, a global image descriptor, by integrating the probability distribution of sampled pixel pairs of an image as well as color histogram.

Below is how to extract BMVC descriptor from an image, which is illustrated in Figure 2.

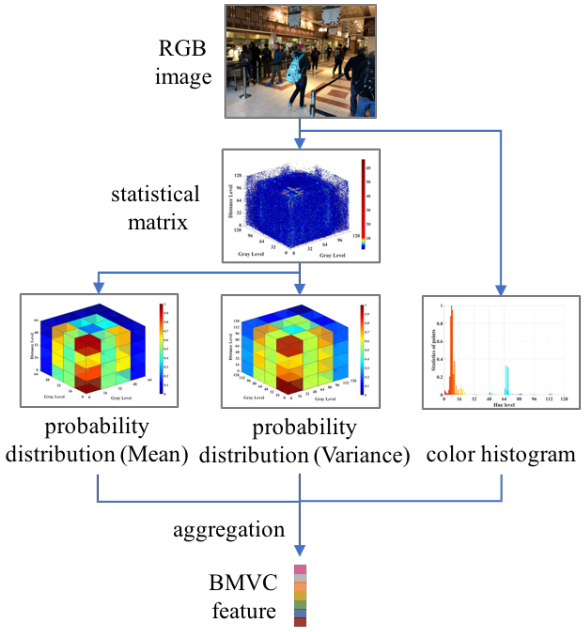

Fig.2. Flowchart of BMVC feature extraction.

1) Randomly sample a large number of point pairs from a candidate image. Record the Euclidean distance between each individual pair and the corresponding grayscale values and store them using a triplet, all of which are quantified to an integer ranging from 1 to 128.

2) Generate the 3D statistical matrix that is denoted by $V(grayLevel1, grayLevel2, distanceLevel)$. Each element in the matrix corresponds to the number of point pairs with corresponding grayscale and distance values.

3) Subdivide the matrix from $128 \times 128 \times 128$ to $4 \times 4 \times 4$, and calculate the mean and variance for each block. Concatenate the mean and variance values of these blocks to form a 128-dimension statistic vector which is denoted by

$$MV = [m_1, m_2, ..., m_{64}, v_1, v_2, ..., v_{64}].$$

4) Extract the color histogram for the sampled points in the HSV space. The 128-dimension color vector is denoted by

$$H = [h_1, h_2, ..., h_{128}].$$

5) Concatenate the statistic and color vector to form a 256-dimension feature vector, termed BMVC feature. It leverages the statistical information to describe the content of an image and is defined by

$$F = [m_1, m_2, ..., m_{64}, v_1, v_2, ..., v_{64}, h_1, h_2, ..., h_{128}].$$

Other than BMVC, we also extract CNN-based feature for each of database images. We use NetVLAD representation, which is extracted by a pre-trained Pitts30K [22] VGG-16 [18] model (other model could also be used), as a set of multi-scale features which enables for on-line image retrieval and pose estimation stages.

4.3 Plane Feature Extraction

Indoor environment is, normally, composed of planar surfaces (walls, ceilings and floors, etc.) which are aligned to one of three orthogonal directions. And these three orthogonal directions are referred to as the Manhattan Frame (MF) [41] of the scene. Based on the observation, we extract the planar surfaces in the scene to enrich our sitemap. Those planes improve the geometric representation efficiency of the scene by taking advantage of the plane constraints in an environment.

The scene structure can be easily explored by creating a common coordinate system for all spaces (e.g. rooms, hallways, etc.). Specifically, we define a Cartesian reference coordinate system on spaces. We choose

the z axis as the gravitational axis, with the normal direction of the floor taken as the positive direction. x axis and y axis are perpendicular to each other and could be defined by the floor-plan of the scene. A plane associated with multiple (at least two) images is regarded to be a planar feature.

For each RGB-D database image, we use Agglomerative Hierarchical Clustering (AHC) [42] to detect planar patch. A planar surface is extracted and parameterized in the Hessian form if its fitted normal is aligned to one of the global MF coordinate axes and its area is larger than a certain threshold. We search plane correspondences between adjacent images according to the following criteria: having overlap, the difference of normal angle is below a threshold as well as the difference of distance to origin is small (10° and 10 centimeters in our experiment). Thereby, an extracted plane have three possibilities, to be a new planar feature, to be associated with an existed planar feature, or to be removed if it is not seen by other images. In this manner, we generate the association between planar features, which are then registered on corresponding images on our sitemap.

4.4 View Synthesizing

The constructed sitemap enables us to enrich the original database by generating synthetic views, especially for places with too sparse scan locations. Below we explain the procedure from two aspects, sitemap-based synthetic viewpoint and synthetic view rendering.

First, synthetic viewpoint should be selected far enough from the scan locations to avoid unnecessary overlap of data. We identify the floors from the constructed scene layout according to the following criteria: the normal direction of a plane is positively aligned to the y axis of MF and the coordinate value in the z axis is within a certain range. The synthetic viewpoints are then sampled on the floor plane with a regular grid. If a grid point is too close to existed scan locations, we remove this point to avoid repetition. For each synthetic viewpoint, we sample 12 possible horizontal rotations and at 3 different pitches with 30 degrees at a time to complete a 360-degree sweep to decide a set of camera poses.

Then, we render synthetic images by leveraging multiple corresponding scans rather than individual scans to handle occlusions with the help of constructed scan-graph. Given a synthesized viewpoint and a camera pose, we find the nearest scan location in the same floor and all associated panoramic scans according to the sitemap. Then project the 3D points visible in these panoramic scans in the given pose. Further taking normal directions into account could handle occlusions to certain extent on these views. The rendered views also output RGB-D data.

This procedure results in thousands of images which depends on both the number of synthesized viewpoints and sampling stride of the camera rotation. After generating these views, we discard these less-informative images by visual inspection, and then add those images with rendered point clouds to our sitemap.

Indoor areas mainly include rooms, lobbies, stairways and hallways. We observed that the localization failures are more likely to happen in hallways. This is attributed to the "bottleneck" areas in some hallways with narrow width and sparse scan locations on them.

5 Sitemap-Boosted Localization

In the on-line localization stage, for a given query image we estimate the camera pose by taking advantage of the constructed sitemap.

5.1 Image Retrieval and Pose Estimation

First, retrieve a shortlist of candidate images visually similar to the query. Given a query image, extract NetVLAD (other CNN-based features could also be used) and BMVC descriptor from it. N nearest images among the database are retrieved via NetVLAD using normalized L2 distance (100 in our experiment). Corresponding elements for each retrieved image are also returned through the sitemap, i.e., the point cloud, image features and planar features.

Then, use BMVC descriptor to re-rank the shortlist of retrieved image candidates. Normalized Cosine distance are used to measure the dissimilarity of BMVC descriptors of the query and retrieved database image. Adding up the two distance values mentioned above and re-rank the image candidates according to the distance. Filter out K candidates which have low similarity in terms of both the image appearance and statistic information (20 in our experiment).

Finally, estimate camera pose. Similar to InLoc, intermediate convolutional layers (the conv3 and conv5 layer) are utilized for feature matching, followed by homography fitting which can result in inliers. According to the inlier count we obtain top k image candidates (10 in our experiment), which can form k image pairs together with the query photo. For each pair, standard P3P-RANSAC [39] is utilized to compute the camera pose with a given focal length of the query photo. These k candidate poses are then fed to the next pose verification stage.

5.2 Scene Layout-Enhanced Pose Verification

The pose verification module aims to select the final pose from k candidate poses. Our method builds on the Dense Pose Verification (DensePV) approach of InLoc.

We first review their DensePV algorithm. InLoc compares the query and each of its synthesized views using RootSIFT descriptor in a pixel-wise manner, and then scores the similarity to select the final pose among candidates. Different from DensePV using all pixels which have valid depth, we leverage additional geometry clues of scene layout to pick certain pixels. Figure 3 shows the procedure. For a given query image, as we explained before, we obtain k estimated poses, each of which corresponds to a retrieved database image. For each retrieved image, it is convenient to acquire the associated point cloud and planar features through the constructed sitemap. We project the corresponding colored point cloud in the pose i ($i = 1, 2, ...k$) to generate a synthetic view of the query image and we further use the extracted layout to enable for picking some certain pixels. Technically, we project the corresponding planar features in the pose i to form a rendered view of scene layout, through which we assign attention to those pixels on layout regions. For an image pair i demonstrated in Figure 3, the top image is rendered from candidate pose i, the bottom image is query photo, and the color-code indicates scene layout. Red, green, and blue color are assigned to x, y and z axis of MF.

We compare the query and synthesized image by taking care of both the geometry and appearance similarity. Firstly we compute the geometry similarity. Extract RootSIFT descriptors from both images for the pixels selected by layout planes. Local geometry similarity score is computed as the median of the inverse Euclidean distance between RootSIFT descriptors corresponding to the same pixel position. Additionally, we take a global similarity metric to investigate the similarity of the image appearance between the query and retrieved database image. The appearance similarity score is defined as the inverse Euclidean distance between NetVLAD descriptors of the two images. We get a final similarity score by adding up the two kinds of score for each pose candidate. Finally, the final camera

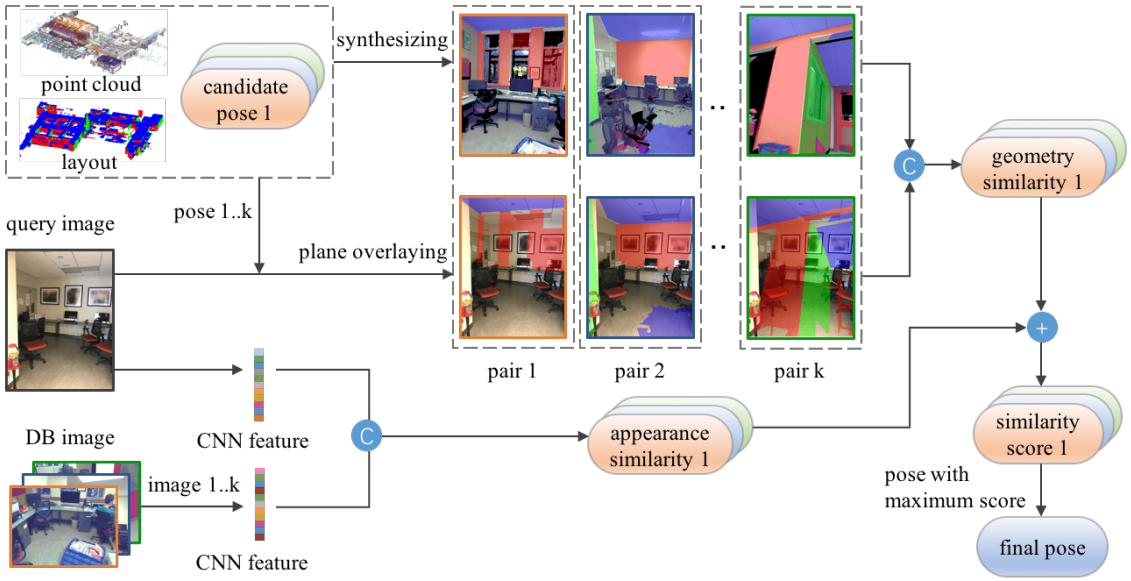

Fig.3. Flowchart of the proposed LayoutPV method. The final pose is selected from k given candidate poses through comparing both the geometric similarity and image appearance. In figure, Ⓒ means computing the inverse distance between descriptors, ⊕ means addition operation.

pose with the highest similarity score is verified among k candidate poses.

6 Experiment Results

In this section, we use the InLoc benchmark to evaluate the proposed approach ReLoc in three aspects: 1) the effectiveness of proposed BMVC descriptor in image retrieval; 2) the enhancements of view synthesis and geometric similarities for pose verification; 3) the final localization comparisons with the state-of-the-art methods.

The reason why we use Inloc benchmark in our experiments is that the benchmark is one of the large-scale indoor localization dataset and widely used in various methods, which is composed of 10k RGB-D images that are generated from panoramic scans of university buildings, and a query set of RGB images taken by mobile phones about a year after the database images acquired, thus the appearance variation between the dataset and query image makes the task of visual localization significantly challenging.

We implement our method on a PC with Intel Core i9-9820 and 64 GB RAM. In the sitemap construction module, we use 256-dimension BMVC descriptor and 4096-dimension NetVLAD representation pre-trained by Pitts30K.

6.1 Recall Evaluation on Image Retrieval

We study the performance in the context of image retrieval on large-scale InLoc dataset. We use evaluation metric Recall@N, which is the probability of queries that are correctly localized in the given N nearest neighbor database images returned by the module of image retrieval, to evaluate the performance of our image retrieval method.

Table 1. Comparison of the state-of-the-art approach w/o and w/ BMVC re-ranking for place recognition

Method	**N**		
	40	60	80
NetVLAD	83.2	88.0	91.7
NetVLAD+BMVC	84.6	90.3	92.2

Note: **N** used in the table represents "Number of top database candidate images".

The query is considered to be correctly localized if at least one relevant database image is reported in the top

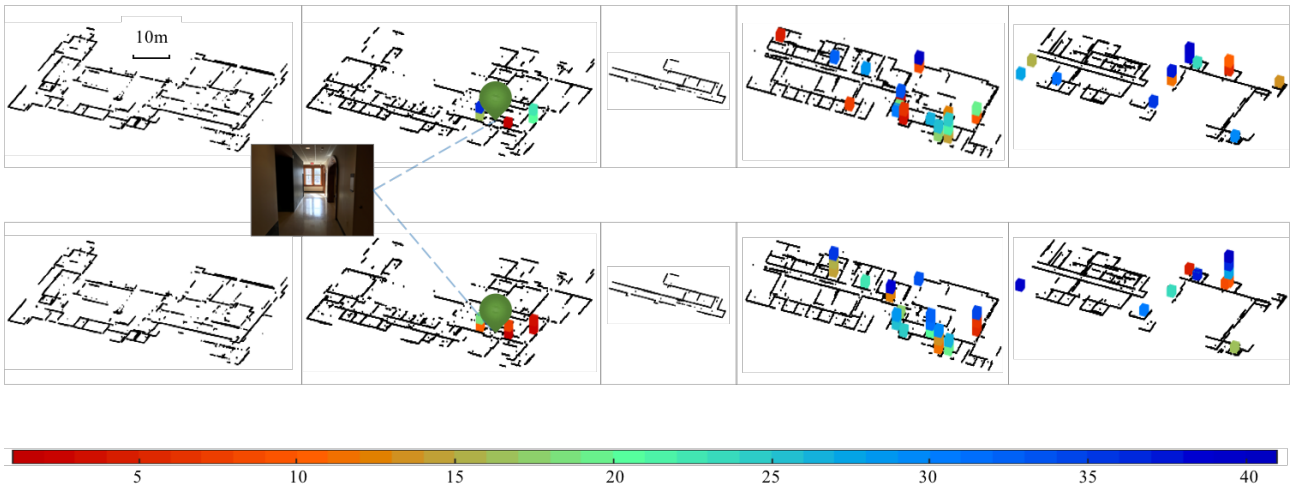

Fig.4. Spatial Distribution of returned database images using different methods. The top row shows top 40 candidate images retrieved using NetVLAD. The bottom row illustrates top 40 candidate images returned from NetVLAD followed by BMVC re-ranking. The colorbar corresponds to ranking. The darker the red color the higher the ranking, and the darker the blue color the lower the ranking. Green balloon location is the reference position where the query photo was taken. It can be seen that re-ranking via BMVC improves the ranking of nearest neighbours of the query photo.

N ranked database images. The relevance is determined by whether the query and database image see the same planar feature.

For each query image in the InLoc dataset, a ranked list of top 100 candidate images are returned via NetVLAD descriptor, and then all candidate images are re-ranked using our BMVC descriptor. We consider NetVLAD as a baseline, and compare it with NetVLAD followed by BMVC (NetVLAD+BMVC). Totally we evaluate 329 InLoc query images and report the Recall@N for the two methods in Table 1.

It is noteworthy that NetVLAD+BMVC outperforms NetVLAD at Recall@N with various thresholds. Furthermore, we present the spatial distribution of top 40 returned database images via NetVLAD versus NetVLAD followed by BMVC re-ranking using a challenging query image from the InLoc dataset as shown in Figure 4. We use vertical planes within the scene layout to depict the structure of the five floors. The top row shows top 40 candidate images retrieved using NetVLAD. The bottom row illustrates top 40 candidate images returned from NetVLAD followed by BMVC re-ranking. The colorbar corresponds to the order of ranking. The darker the red color the higher the ranking, and the darker the blue color the lower the ranking. Green balloon indicates the reference position where the query photo was taken. It is obvious that BMVC re-ranking improves the ranking of truly nearest neighbours of the query image.

The evaluation results on Recall@N demonstrate that statistic information based BMVC is capable to describe images, and has potential to make images more discriminative.

6.2 Comparisons of Pose Verification

In this part, we present a group of statistical results to show the advantages of proposed pose verification LayoutPV against the baseline method DensePV on the InLoc and extended dataset, to show the effectiveness of our view synthesis strategy and the using of planar geometries similarity when doing pose verification as shown in Figure 3.

Table 2 shows the results of first experiment. We compare the statistical correction rates of two methods under the same localization accuracy. In this exper-

iment we produce an extended dataset by adding 10 synthetic viewpoints, leading to add 2450 synthesized images in total finally after removing less-informative images by visual inspection. The extended dataset contains 12422 images in which 9972 images are originated from the InLoc dataset and others are rendered from synthesized viewpoints. To make a fair comparison, we use exactly the same set of candidate poses acquired by NetVLAD to test the two pose verification methods.

Table 2. Comparison with DensePV on the InLoC and extended dataset

	DensePV	LayoutPV
original database	69.9	**72.5**
extended database	74.6	**77.3**

Note: the rate (%) of correctly localized queries is within 1 meter distance threshold and 10° error threshold.

As shown in Table 2, compared with the original database, the localization performance of DensePV and LayoutPV on the extended dataset has increased in 4.7% and 4.8%, respectively. This can be attributed to using of view synthesis that enables to generate synthetic views more similar in appearance with the query image. Furthermore, LayoutPV has shown 2.6% and 2.7% performance gain compared to DensePV on the original dataset and the extended dataset respectively. This makes us believe that the improvement mainly comes from the leverage of scene layout, which measuring the geometric similarity between the query and rendered image other than only evaluating pixel level similarity. Technically, in our method LayoutPV, scene geometry acts as an attention mechanism which focuses on pixels where stable (e.g. tables, couches, and wardrobes) or fixed (e.g., walls, floors, and ceilings) objects are therein, while the DensePV takes all pixels with valid depth values into consideration, which is thus less robust compared with our method.

6.3 Comparisons on Localization Accuracy

In this part we evaluate our final indoor localization performance by measuring the differences in position and orientation between our result and ground truth. We report the percentage of query images whose poses differ by no more than X meters and Y degrees from the reference pose for different pairs of thresholds (X, Y). Here InLoc [12] and InLoc+N+S [13] serve as our main baseline. InLoc+N+S is a variant of InLoc which integrate normals and semantics to regress the camera pose. Other than InLoc and its variants, we also compare with other state-of-the-art localization approaches, i.e., Direct 2D-3D matching [6] and DisLoc [11]. Direct 2D-3D is a 3D structure-based image localization approach, using RootSIFT features associated with the scene point cloud. Disloc is a classical image retrieval-based localization method which represents images using bag-of-words.

Table 3 reports comparison results on the rate of correctly localized queries within different distance threshold and a 10° orientation error threshold. It is obvious that the proposed method ReLoc (with synthesized viewpoint and LayoutPV) constantly improves the localization accuracy by about 5.0% and 3.5% when compared to the state-of-the-art InLoc and InLoc+N+S, respectively. We furthermore show multiple qualitative examples of localization on the InLoc dataset in Figure 5. Column 1 in the figure shows the query photos, and columns 2 to 5 are the synthesized views rendered from corresponding camera poses acquired by different methods. The numbers under the rendered images indicate the position and orientation error with respect to the ground truth poses. ReLoc (column 5) achieves a higher accuracy of camera pose estimation when compared to ReLoc without generating synthesized viewpoints (column 4).

Table 3. Comparison with the state-of-the-art localization approaches on the InLoc dataset

Threshold	Method				
	Direct2D-3D [6]	DisLoc [11]	InLoc [12]	InLoc+N+S [13]	ReLoc
0.25m	11.9	13.0	39.8	41.0	**45.9**
0.50m	15.8	17.7	59.0	60.5	**64.0**
1.00m	22.5	22.3	69.0	72.3	**77.3**

Note: the rate (%) of correctly localized queries is within a 10° orientation error threshold.

7 Conclusion

We have proposed an indoor visual localization system ReLoc supported by a hierarchical sitemap, which is constructed in off-line stage and consists of a scan graph and scene layout. The scan graph organizes a hierarchical structure of image features and geometry features of pre-scanned and synthetic views, and also builds up the connections between scan locations with overlap of views. Furthermore, a scene layout is constructed based on the planar elements to represent the geometric frame of a scene. Therefore, the sitemap is efficient to associate all features of the scene.

In on-line stage, BMVC descriptor improved the proximity to the true location by re-ranking the shortlist of candidates that returned from image retrieval. The proposed pose verification method is also effective to measure the similarity only in the planar surface regions of the query photo and synthetic views from the estimated pose to verify the candidates and select the final pose, which alleviates the problems due to drastic variation in scene appearance over time. With hierarchical sitemap and improved pose verification approach, our ReLoc system enhanced the efficiency and accuracy of indoor visual localization.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M. Seitz, Richard Szeliski. Building Rome in a day. Commun. *ACM*, 2011, 54(10): 105-112.

[2] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, Christian Theobalt. BundleFusion. Real-time Globally Consistent 3D Reconstruction using On-the-fly Surface Re-integration. *ACM Transactions on Graphics*, 2017, 36(3):76a.

[3] Raul Mur-Artal, Juan D. Tardós. ORB-SLAM2. An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. Robotics*, 2017, 33(5): 1255-1262.

[4] Yunpeng Li, Noah Snavely, Dan Huttenlocher, Pascal Fua. Worldwide Pose Estimation Using 3D Point Clouds. In *12th European Conference on Computer Vision*, October 2012, pp.15-29.

[5] Bernhard Zeisl, Torsten Sattler, Marc Pollefeys. Camera Pose Voting for Large-Scale Image-Based Localization. In *IEEE International Conference on Computer Vision*, December 2015, pp.2704-2712.

[6] Torsten Sattler, Michal Havlena, Filip Radenovic, Konrad Schindler, Marc Pollefeys. Hyperpoints and Fine Vocabularies for Large-Scale Location Recognition. In *IEEE International Conference on Computer Vision*, December 2015, pp.2102-2110.

[7] Torsten Sattler, Bastian Leibe, Leif Kobbelt. Efficient and Effective Prioritized Matching for Large-Scale Image-Based Localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017, 39(9): 1744-1756.

[8] Relja Arandjelovic and Andrew Zisserman. All about VLAD. In Proc. CVPR, 2013.

[9] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, Tomás Pajdla. 24/7 place recognition by view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015, pp.1808-1817.

[10] Torsten Sattler, Michal Havlena, Konrad Schindler, Marc Pollefeys. Large-Scale Location Recognition and the Geometric Burstiness Problem. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp.1582-1590.

[11] Relja Arandjelovic, Andrew Zisserman. DisLocation: Scalable Descriptor Distinctiveness for Location Recognition. In *12th Asian Conference on Computer Vision*, November 2014, pp.188-204.

[12] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomás Pajdla, Akihiko Torii. InLoc: Indoor Visual Localization With Dense Matching and View Synthesis. In *IEEE Conference on*

Fig.5. Three Qualitative examples of visual localization on InLoc dataset. The query images of 3 examples are in left column, the right 4 columns show the localization results of InLoc, InLoc+N+S, and proposed method without and with view synthesis, respectively. It is obvious that the proposed method outperforms the others, and in which the view synthesis matters.

Computer Vision and Pattern Recognition, June 2018, pp.7199-7209.

[13] Hajime Taira, Ignacio Rocco, Jirí Sedlár, Masatoshi Okutomi, Josef Sivic, Tomás Pajdla, Torsten Sattler, Akihiko Torii. Is This the Right Place? Geometric-Semantic Pose Verification for Indoor Visual Localization. In *IEEE International Conference on Computer Vision*, October 2019, pp.4372-4382.

[14] Alex Kendall, Matthew Grimes, Roberto Cipolla. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *IEEE International Conference on Computer Vision*, December 2015, pp.2938-2946.

[15] Vassileios Balntas, Shuda Li, Victor Prisacariu. RelocNet: Continuous Metric Learning Relocalisation Using Neural Nets. In *15th European Conference on Computer Vision*, September 2018, pp.782-799.

[16] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, Torsten Sattler. D2-Net: A Trainable CNN for Joint Description and Detection of Local Features. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2019, pp.8092-8101.

[17] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, Laura Leal-Taixé. Understanding the Limitations of CNN-Based Absolute Camera Pose Regression. CVPR 2019: 3302-3312.

[18] Karen Simonyan, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR 2015

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp.770-778.

[20] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, Marcin Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. *CVPR*, 2019: 12716-12725.

[21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *CVPR*, 2018: 4510-4520.

[22] Relja Arandjelovic, Petr Gronát, Akihiko Torii, Tomás Pajdla, Josef Sivic. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, 40(6): 1437-1451.

[23] Wei Zhang, Jana Kosecka. Image Based Localization in Urban Environments. In *3rd International Symposium on 3D Data Processing, Visualization and Transmission*, June 2006, pp.33-40.

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

[25] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image Retrieval for Image-Based Localization Revisited. In Proc. BMVC, 2012.

[26] H. Badino, D. Huber, and T. Kanade. Visual topometric localization. In Proc. IV, 2011.

[27] Tommaso Cavallari, Stuart Golodetz, Nicholas A. Lord, Julien P. C. Valentin, Luigi di Stefano, Philip H. S. Torr. On-the-Fly Adaptation of Regression Forests for Online Camera Relocalisation. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp.218-227.

[28] Lili Meng, Jianhui Chen, Frederick Tung, James J. Little, Julien Valentin, Clarence W. de Silva. Backtracking regression forests for accurate camera relocalization. In *IEEE/RJS International Conference on Intelligent Robots and Systems*, September 2017, pp.6886-6893.

[29] Ronald Clark, Sen Wang, Andrew Markham, Niki Trigoni, Hongkai Wen. VidLoc: A Deep Spatio-Temporal Model for 6-DoF Video-Clip Relocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp.2652-2660.

[30] Newcombe R A, Izadi S, Hilliges O, et al. KinectFusion: Real-time dense surface mapping and tracking[C]//Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on. 127-136.

[31] Yuichi Taguchi, Yong-Dian Jian, Srikumar Ramalingam, Chen Feng. Point-plane SLAM for hand-held 3D sensors. In *IEEE International Conference on Robotics and Automation*, May 2013, pp.5182-5189.

[32] Pyojin Kim, Brian Coltin, H. Jin Kim. Linear RGB-D SLAM for Planar Environments. In *15th European Conference on Computer Vision*, September 2018, pp.350-366.

[33] Noha Radwan, Abhinav Valada, Wolfram Burgard. VLocNet++: Deep Multitask Learning for Semantic Visual Localization and Odometry. *IEEE Robotics Autom. Lett.*, 2018, 3(4): 4407-4414.

[34] Johannes L. Schönberger, Marc Pollefeys, Andreas Geiger, Torsten Sattler. Semantic Visual Localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018, pp.6896-6906.

[35] Tianxin Shi, Hainan Cui, Zhuo Song, Shuhan Shen. Dense Semantic 3D Map Based Long-Term Visual Localization with Hybrid Features. CoRR abs/2005.10766,2020.

[36] Mathieu Aubry, Bryan C. Russell, and Josef Sivic. Painting-to-3D Model Alignment via Discriminative Visual Elements. ACM Trans. Graph., 33(2):14:1-14:14,2014.

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

[38] Jisan Mahmud, Rajat Vikram Singh, Peri Akiva, Spondon Kundu, Kuan-Chuan Peng, Jan-Michael Frahm. ViewSynth: Learning Local Features from Depth using View Synthesis. BMVC 2020.

[39] Martin A. Fischler, Robert C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Commun. *ACM*, 1981, 24(6): 381-395.

[40] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-From-Motion Revisited. In Proc. CVPR, 2016.

[41] Bernard Ghanem, Ali K. Thabet, Juan Carlos Niebles, Fabian Caba Heilbron. Robust Manhattan Frame estimation from a single RGB-D image. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015, pp.3772-3780.

[42] Feng, C., Taguchi, Y., and Kamat, V. R. (2014). Fast Plane Extraction in Organized Point Clouds Using Agglomerative Hierarchical Clustering. Proceedings of the IEEE International Conference on Robotics and Automation, Hong Kong, China, 6218-6225.

[43] Venkat N. Gudivada. Content-Based image Retrieval Systems (Panel). In *ACM Conference on Computer Science*, 1995, pp.274.

[44] John A. Black Jr., Gamal Fahmy, Sethuraman Panchanathan. A Method for Evaluating the Performance of Content-Based Image Retrieval Systems. In *5th IEEE Southwest Symposium on Image Analysis and Interpretation*, April 2002, pp.96-100.

[45] Chuen-Horng Lin, Rong-Tai Chen, Yung-Kuan Chan. A smart content-based image retrieval system based on color and texture feature. *Image Vis. Comput.*, 2009, 27(6): 658-665.

[46] David G. Lowe. Distinctive image features from scale invariant keypoints. IJCV, 60(2):91-110, 2004.