Scientific Research Publishing

Journal of Computer and Communications, 2024, 12, 47-66
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Visual Indoor Localization Method Based on Efficient Image Retrieval

Mengyan Lyu, Xinxin Guo, Kunpeng Zhang, Liye Zhang*

School of Computer Science and Technology, Shandong University of Technology, Zibo, China
Email: *zhangliye@sdut.edu

How to cite this paper: Lyu, M.Y., Guo, X.X., Zhang, K.P. and Zhang, L.Y. (2024) A Visual Indoor Localization Method Based on Efficient Image Retrieval. *Journal of Computer and Communications*, **12**, 47-66.
https://doi.org/10.4236/jcc.2024.122004

Received: January 17, 2024
Accepted: February 18, 2024
Published: February 21, 2024

Abstract

The task of indoor visual localization, utilizing camera visual information for user pose calculation, was a core component of Augmented Reality (AR) and Simultaneous Localization and Mapping (SLAM). Existing indoor localization technologies generally used scene-specific 3D representations or were trained on specific datasets, making it challenging to balance accuracy and cost when applied to new scenes. Addressing this issue, this paper proposed a universal indoor visual localization method based on efficient image retrieval. Initially, a Multi-Layer Perceptron (MLP) was employed to aggregate features from intermediate layers of a convolutional neural network, obtaining a global representation of the image. This approach ensured accurate and rapid retrieval of reference images. Subsequently, a new mechanism using Random Sample Consensus (RANSAC) was designed to resolve relative pose ambiguity caused by the essential matrix decomposition based on the five-point method. Finally, the absolute pose of the queried user image was computed, thereby achieving indoor user pose estimation. The proposed indoor localization method was characterized by its simplicity, flexibility, and excellent cross-scene generalization. Experimental results demonstrated a positioning error of 0.09 m and 2.14˚ on the 7Scenes dataset, and 0.15 m and 6.37˚ on the 12Scenes dataset. These results convincingly illustrated the outstanding performance of the proposed indoor localization method.

Keywords

Visual Indoor Positioning, Feature Point Matching, Image Retrieval, Position Calculation, Five-Point Method

1. Introduction

With the advancement of Location-Based Service (LBS) technology, the employment of Global Positioning System (GPS) for positioning can no longer ade-

quately meet the populace's requisites for indoor location information. Consequently, an escalating array of indoor positioning methodologies has been successively introduced, encompassing Wi-Fi positioning [1], ultrasound positioning [2], Ultra-Wideband (UWB) positioning [3], and geomagnetic positioning [4], among others. Nevertheless, within intricate indoor environments, these solutions may necessitate substantial manual configuration and supplementary infrastructure, potentially resulting in exorbitant costs and inadequate interference resilience. Conversely, vision-based indoor positioning approaches solely exploit image information to perceive the user's camera surroundings and compute their position and orientation. This approach exhibits advantages such as effortless deployment and economical costs, and has garnered extensive and profound research attention.

The existing main research methods for indoor visual localization are classified into structure-based and regression-based localization. After establishing the correlation between the features in the query image and the 3D structural features in the scene model, structure-based localization refers to the application of Perspective-n-Point (PnP) to solve the camera pose by reducing outliers in RANSAC. Match-based localization and scene coordinate regression-based localization are two further categories of structure-based localization techniques. The majority of matching-based localization techniques are transformed into feature descriptor matching jobs, which can then be further separated into direct matching and hierarchical matching based on how far apart the descriptors are represented. The 2D query image feature set and the 3D scene feature points are directly matched using the direct matching approach [5]. By searching the collection of picture features to compare with the scene database image features, hierarchical matching algorithms inadvertently create 2D-3D correspondences. In contrast, scene coordinate regression methods, which have received much attention in recent years, use compact random numbers to directly regress dense scene coordinate maps to directly predict the absolute 3D coordinates of image pixels, which are explicitly improved by using scene structures. The scene structure is generally represented using a 3D point cloud model, which is constructed based on Structure-from-Motion (SFM) or SLAM. However, the scene model is difficult to construct and the geometric alignment between the query image and the 3D model is difficult to solve.

In regression-based methods, end-to-end direct regression for localization predicts the reference image pose by means of a Convolutional Neural Network (CNN), continuously optimizes the network weights, and outputs the position and orientation information of the image directly to the regressor. Different file types, including single photos, image sequences, and movies, can be used as network inputs. End-to-end direct regression localization needs to be trained for a specific dataset in a multi-scene manner, and needs to be retrained when generalized to a new scene, which is less adaptive and prone to overfitting problems. Relative positional regression methods are based on image retrieval, predicting the relative pose between the query image and the most similar image in the data-

base, and finally obtaining the absolute pose of the query image.

In the past few years, although the field has reached a fairly mature level, it is still difficult to balance the computational cost, localization accuracy, and robustness. In order to better solve this problem, this paper designs an indoor localization method based on efficient image retrieval and relative position estimation. The proposed localization system consists of two phases, offline and online, the offline phase efficiently extracts and stores all the global features of the offline database images; in the online phase, for the target query, after efficiently returning the query results, the implicit matching between image pairs is utilized to return to the essential matrix to compute the position information. In this paper, we try the current more advanced feature extraction matching model and propose a robust position calculation method based on relative position regression. The scheme is able to achieve high-precision localization on multiple indoor datasets without targeting specific datasets, and the adaptability is much better than other localization systems. The indoor localization system based on efficient image retrieval uses only RGB images and position information, does not rely on 3D models, and uses a server-hosted image database for computational operations. The main contributions are as follows:

1) Proposing an indoor localization approach based on efficient image retrieval. Rapidly matching query images with database images to obtain a set of similar image pairs for localization calculations.

2) Vision-based indoor localization algorithms do not require a 3D model and recover the user's camera position from only a few sets of 2D-2D matches, reducing database processing in the offline phase.

3) Using a learnable feature detection and matching method to decompose the essence matrix, we propose a new RANSAC mechanism to solve the relative positional ambiguity problem and recover the user's absolute poses.

2. Related Work

2.1. Image Retrieval

Image Retrieval Techniques Image retrieval tasks aim at querying the content similar to the input image from an image database, as the basis of visual indoor localization, which can effectively improve the efficiency of visual matching. Early research on image retrieval is Text-Based Image Retrieval (TBIR), which mainly includes Page-Rank methods, probabilistic methods, classification or clustering methods, lexical annotation methods, etc. TBIR retrieval is fast and accurate, but it requires a lot of manpower and time, which is not able to satisfy the ever-changing retrieval needs. Content-Based Image Retrieval (CBIR) task extracts image features by mathematically describing the visual content of an image. CBIR early relied on local feature aggregation methods, the most representative of which are visual word representations of images and their extensions, such as Fisher vectors [6] and Vector of Locally Aggregated Descriptors (VLADs) [7]. After 2012, the dominant role of SIFT [8] is gradually replaced by data-driven

Deep Neural Networks (DNNs). The representative NetVLAD [9] constructs a global image descriptor for instance-level image retrieval by applying a pooling mechanism on the activation of the last convolutional feature map in a convolutional neural network. Another widely used method, such as MAC [10], focuses on the region of interest on the feature map and selects just the most active neurons using optimal pooling on each distinct feature map. The retrieval effect of convolutional neural network in deep learning algorithm is the most outstanding, it uses the combination of multiple convolutional layers and pooling layer to get the visual features of the image, and combines with the feedback and classification techniques to achieve better retrieval results. In Literature [11], SFM information is used to fine-tune a pretrained classification network guided by database images and a pooling layer based on generalized means with learnable parameters is proposed to effectively improve the retrieval performance.

2.2. Structure-Based Approach

Structural feature-based visual localization uses sparse feature matching to obtain 2D-3D correspondences and robust optimization to recover the camera pose. Matching-based localization establishes the connection between the query object and the scene image using feature descriptors, and each 3D point in 3D scene models typically receives one or more local descriptors. Direct matching methods that require searching for query features at each 3D point are very inefficient and show fragile robustness to repeated local features. Coarse-to-fine hierarchical localization is based on image retrieval, which achieves accurate localization of large-scale datasets by searching for the smallest subset of scene models and computing the correspondence between the target query and the smallest subset of scenes. The hierarchical localization process requires accurate extraction of local features of the query for similar scene image matching. Melekhov *et al.* [12] proposed a DGC-Net localization method, based on the framework of CNN, which exploits the advantages of the optical flow approach from coarse to fine, and achieves dense and subpixel-accurate localization computation in complex environments by extending the optical flow to the case of large transformations, with a strong supervised training in terms of ground-truth labels per pixel. The inherent hierarchical nature of network features is exploited in ASLFeat [13], which proposes a new multiscale detection mechanism to improve the ability of local shape modeling, to obtain stronger geometric invariance, and to locate the keypoints more accurately.

The scene coordinate regression approach directly predicts the correspondence between the query image and the 3D scene space, which works well on small datasets but does not scale well to larger, more complex scenes. Literature [14] designed a lightweight visual localization network that uses knowledge distillation to efficiently extract deep local features for accurate localization, however, this approach requires a large number of images and dense point cloud information

from Light Detection and Ranging (LiDAR) sensors.

2.3. Regression-Based Approach

The methodology of direct regression-based visual localization involves learning the complete localization pipeline for 2D-3D matching. The PoseNet [15] approach is the first to directly regress camera pose prediction from a single image using a Convolutional Neural Network (CNN). It employs the Structure-from-Motion (SFM) technique to automatically generate training labels, thereby alleviating the burden of manual annotation. ANNet [16] uses discriminator networks and adversarial learning to implicitly learn the joint distribution of images and their corresponding camera poses to further refine the image-based position estimation and further improve the localization accuracy. In Literature [17], camera pose autoencoder is introduced to improve camera position estimation by using multi-layer perceptron. FeatLoc [18] uses sparse feature descriptors directly to train network models through data augmentation, mitigating the effects of light changes or environmental gradients.

To boost scalability, relative pose regression is trained on typically numerous unseen scenes. After determining the relative pose of the reference image, absolute bit-position information in the world coordinate system is obtained by spatial coordinate translation. The relative pose regression method utilizes a multi-stage strategy that generalizes well to new scenes. NN-Net [19] pioneered the use of Siamese CNN to predict the relative pose between two input images. Literature [20] proposes a localization method that decouples the scene by regressing the essential matrix without adjusting the parameters. Literature [21] proposes a graphical neural network with image representation nodes and peer-to-peer representation of edge images for relative positional regression.

In this paper, we periodically combine image retrieval and position calculation to design a visual localization scheme based on efficient image retrieval and generalized camera position solving. For a user query image, a pre-trained retrieval model is first utilized to efficiently return relevant database images, and then the absolute position is solved based on the feature correspondences. The scheme does not use 3D structural model information about the scene and can be easily applied to new indoor scenes.

3. System Models and Methods

This section presents the overall process framework of indoor localization based on efficient image retrieval.

3.1. System Models

In this paper, a generalized indoor localization model based on a single RGB image is proposed to greatly reduce the image retrieval time with guaranteed retrieval accuracy while obtaining more accurate localization results. The proposed indoor visual localization method consists of two phases: the offline phase of in-

door image data acquisition processing and database construction, and the online phase of the user's image after retrieval of the bit position calculation, as shown in Figure 1.

- During the offline phase, all images from indoor scenes are processed using a pre-trained image retrieval model to extract global descriptors, which are then used to construct an offline feature database.
- During the online phase, the same global feature extraction process is applied to query images. The similarity between the query image's global feature vector and each feature vector in the offline database is computed. The top five reference images are iteratively selected based on their decreasing similarity scores. For each query-nearest neighbor image pair, the essential matrix E is computed using the five-point algorithm. By solving for the essential matrix E and removing the ambiguity in relative pose, the relative pose between the two images is obtained. Finally, using the retrieved database images with known absolute poses and the relative poses, the absolute pose of the query image, *i.e.* the indoor user's pose, is estimated.

3.2. Offline Data Preparation

The offline phase is performed by a camera or other mobile platform for RGB image capture as well as positional recording. Specifically, each RGB image in the database created in the offline phase has its corresponding real pose label: 3D spatial coordinates representing the absolute position (x, y, z) indicating the position and quaternions (w, p, q, r) indicating the absolute orientation. In this paper, quaternions are used to represent the user's camera orientation, this is because quaternions use only four-dimensional vectors, which perfectly solves

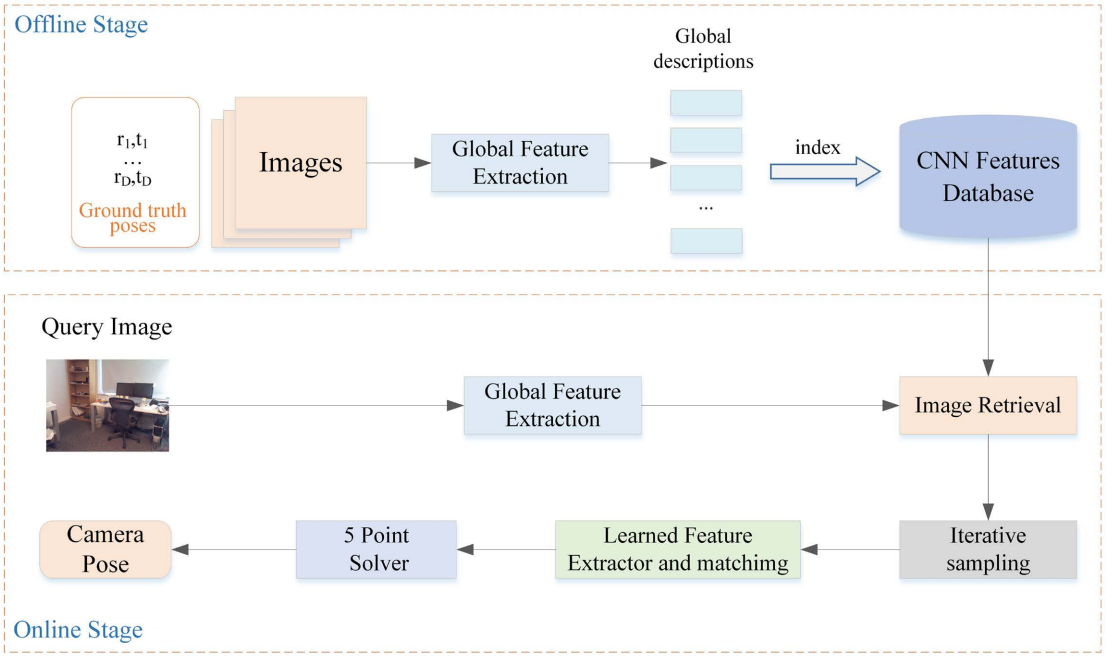

Figure 1. Visual indoor localization pipeline.

the singularity problem and requires less storage space compared to the commonly used 3 × 3 rotation matrices to represent the object orientation. The dataset S all contains n different indoor scenes: $S=\{S_1,S_2,\cdots,S_n\}$. For each scene S_p a global representation of that scene is created: the image name, the positional information, and the extracted global descriptors, as in Table 1.

3.3. Image Retrieval Model

Regarding the image retrieval module, a novel image feature aggregation method based on MLP is adopted [22]. Through a succession of feature mixers, each individual feature map derived from the chopped Resnet backbone is combined with spatial relationships using this method's pre-trained neural network. A compact representation space is then used to obtain the projected output, resulting in global image descriptors used for image retrieval. The specific structure of this network model is shown in Figure 2.

An RGB image is input and the middle layer feature map $F\in R^{c\times h\times w}$ is first extracted using the pre-trained ResNet model on ImageNet, which differs from the existing technology NetVLAD by treating the tensor F as c 2D features X_i of size $h\times w$. Then, the feature mapping is built to give each 2D feature a 1D representation: $F\in R^{c\times n}$, which is input to a feature blender consisting of L-cascaded MLPs of the same structure, defined as follows:

$$X^i+W_2\left(\sigma\left(W_1X^i\right)\right)\to X^i. \tag{1}$$

where W_1 and W_2 are the two fully connected layer weights in the feature blender, and σ refers to the ReLU activation operation. After one Feature-Mixer output $Z\in R^{c\times n}$, it continues to be delivered to the second Feature-Mixer block, and finally, two fully connected layers are added for depth projection and line-by-line projection, weighted pooling operations to control the size of the obtained

Table 1. Offline database.

Scene labels	S_1	...	S_1
RGB images	$I_{11},I_{12},\cdots,I_{1k1}$	...	$I_{n1},I_{n2},\cdots,I_{nkn}$
Position information	$P_{11},P_{12},\cdots,P_{1k1}$	...	$P_{n1},P_{n2},\cdots,P_{nkn}$
Global descriptors	$D_{11},D_{12},\cdots,D_{1k1}$	...	$D_{n1},D_{n2},\cdots,D_{nkn}$

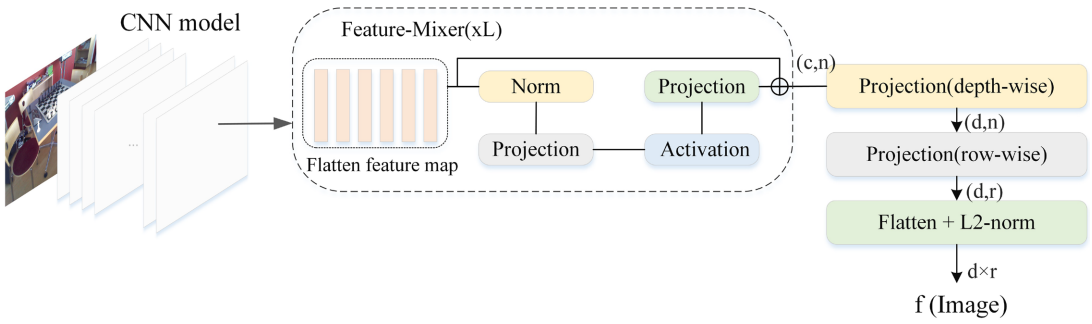

Figure 2. Image retrieval model.

global descriptors. In brief, we obtain 512-dimensional global descriptors for the feature maps obtained from pre-trained ResNet backbone clipping using several MLP Feature-Mixer blocks in a large visual recognition dataset GSV-Cities [23], retrained using multiple similarity loss.

Iterative Selection: The commonly observed phenomenon of the top k retrieved images exhibiting highly similar poses can result in suboptimal performance when estimating the camera position for subsequent triangulation. Therefore, we adopt an iterative approach where we iteratively select and sample five retrieved images based on their decreasing similarity scores.

The dataset images undergo local feature extraction and are globally aggregated into several fixed-length global descriptors. The same method is applied to extract features from the query image, resulting in a feature vector of the same length. The similarity between the query image vector V_q and the image feature vector V_i is calculated, and the results are sorted in descending order based on the similarity. A certain number of database images are then output, where $i \in [1, n]$ and n represents the number of images in the database. The specific state is defined as follows:

$$V_i = \left[V_i^0, V_i^1, \cdots, V_i^{1024} \right]. \tag{2}$$

$$V_q = \left[V_q^0, V_q^1, \cdots, V_q^{1024} \right]. \tag{3}$$

$$score_i = V_i * V_q^{\mathrm{T}}. \tag{4}$$

3.4. Feature Detection and Matching

After image retrieval, relative pose estimation is performed after obtaining a set of iteratively selected query image geotagged image pairs, which consists of four main steps: extraction of keypoints and descriptors, feature-point matching, and false-match rejection; recovery of relative poses between image pairs from the essentiality matrix; and recovery of the absolute camera poses of the query images. Firstly, SuperPoint [24] + LightGlue feature point detection and matching is applied to each query-database image pair. LightGlue is a simple and effective improvement to SuperGlue [25], which is adaptive and can be flexibly adjusted according to the difficulty of the image pairs, and it is also more efficient and accurate in terms of memory and computation. Note that the user query image and the image in the database may be captured by different cameras, and the performance of the localization could be impacted by the variations in the intrinsic properties of the cameras. Therefore, to get more precise localization findings, the camera needs to be precalibrated.

3.5. Position Estimation

By employing image retrieval, the database image with the highest similarity score to the query image is returned. Based on the epipolar constraint, the rotational and translational relationship between the three-dimensional camera coordinate systems of two images can be computed through feature point matching between

the two images. The epipolar constraint reflects the pose relationship between the query camera and the database camera, as shown in Figure 3. In this context, R represents the relative rotation matrix between the two cameras, and t represents the relative translation vector. $O_Q X_Q Y_Q Z_Q$ denotes the camera coordinate system of the query image, while $O_D X_D Y_D Z_D$ represents the camera coordinate system of the database image.

3.5.1. Relative Pose Estimation

For a calibrated camera, the geometric relationship between two images, I_q and I_p, can be estimated using feature point matching based on the epipolar constraint. E can be used to describe this relationship and the expression for E is as follows:

$$E = [t]_x R. \tag{5}$$

The matrices t and R represent the relative translation and relative rotation between the two images, $[\ ]_x$ is an antisym metric matrix, and its calculation formula is as follows:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{bmatrix}. \tag{6}$$

A pair of matched point pairs x_1 and x_2 under the normalized plane, based on the pair of polar geometric constraints is formulated as follows:

$$[u_1u_2, u_1v_2, u_1, v_1u_2, v_1v_2, v_1, u_2, v_2, 1]E = 0. \tag{7}$$

Similarly representing other point pairs, the relative poses R, t between the target-database images can be found using Singular Value Decomposition (SVD), and in general, E can be solved for four poses: (R, t), $(R, -t)$, (R', t), $(R', -t)$. This paper proposes a novel RANSAC method to address pose ambiguity, instead of the traditional feature matching-based approaches that find the correct relative pose among four candidates. Specifically, considering the positions of points triangulated from multiple directions as $t_1, t_2, \cdots, t_n$ the sign of any angle t_i can

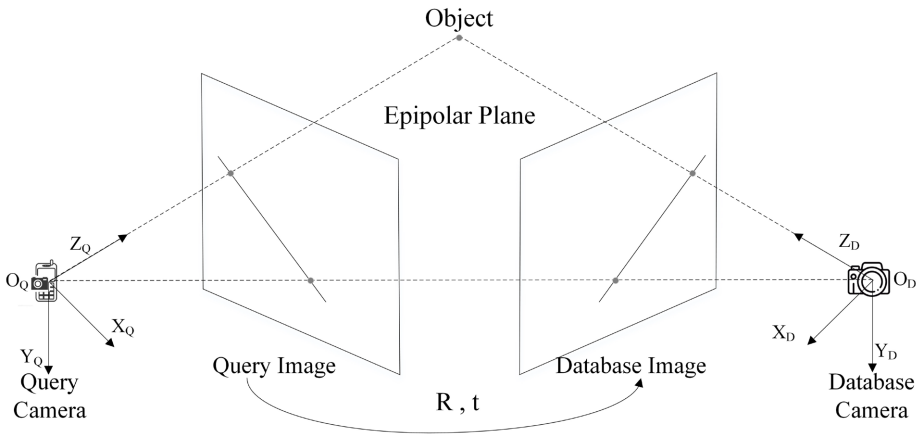

Figure 3. Epipolar constraint relationship.

be inverted without changing, hence only the rotation needs to be determined. As a result, the absolute pose of the target image can be determined by $n \geq 2$ image pairs.

The transformation matrix ground truth T_{12} between the two images is defined as follows:

$$T_{12} = \begin{bmatrix} R_1^{\mathrm{T}} R_2 & R_1^{\mathrm{T}} \left(t_1 - t_2 \right) \\ 0 & 1 \end{bmatrix}. \tag{8}$$

where the absolute poses of image I_1 are and the absolute poses of image I_2 are, it should be noted that the relative transformation $R_1^{\mathrm{T}} R_2$, $R_1^{\mathrm{T}} \left(t_1 - t_2 \right)$ of I_1 from to I_2 is the transformation in the I_2 camera coordinate system.

3.5.2. Absolute Pose Estimation of Query Images

According to triangulation, there are four possible relative rotations between the query image I_q and its two nearest neighbor database images I_i and I_j: R_i, R_i', R_j, R_j', corresponding to the four absolute poses of I_q: $R_i R_{Ii}$, $R_i' R_{Ii}$, $R_j R_{Ij}$, $R_j' R_{Ij}$. R_{Ii}, R_{Ij} are the ground truth of database images I_i, I_j. In theory, among the four absolute poses, two of them are identical, while the rest differ significantly. This means that a hypothesis for an absolute pose is determined based on two nearest neighboring images. In actuality, the relative rotation from each pair that corresponds to the two absolute postures with the smallest angular difference is taken into account as the real one.

Using the two picture pairings (I_q, I_i), (I_q, I_j), calculate the query image's absolute rotation and the bitmap of the query image is calculated from the intersection of the two rays by triangulation. The rays l_1, l_2 are denoted as:

$$l_1 = c_{I_i} + \lambda_i R_{I_i}^{\mathrm{T}} R_i t_i . \tag{9}$$

$$l_2 = c_{I_j} + \lambda_j R_{I_j}^{\mathrm{T}} R_j t_j . \tag{10}$$

where $\lambda_i, \lambda_j \in R$ define the positions of points along the rays. Only when the centers of the three cameras are noncollinear are the results of triangulation specified. $c_I = -R_I^{\mathrm{T}} t_I$ is the global coordinate of the camera center. In our experiments, we use the five queried nearest neighbor database images to calculate the final pose.

In other words, given a pair of images (I_q, I_i), (I_q, I_j), a pose hypothesis (R_{I_q}, t_{I_q}) is obtained. For any pair of query database images (I_q, I_m), four potential solutions need to be found to determine the rotation matrix R_m that best approximates $R_m R_{I_m}$ is closest to R_{I_q}, and the relative translation from I_q to I_m is defined as:

$$t_{pre} = R_{I_m} \left(c_{I_q} - c_{I_m} \right). \tag{11}$$

$$\alpha = \cos^{-1} \left(\frac{t_m^{\mathrm{T}} t_{pre}}{\left\| t_m \right\|_2 \left\| t_{pre} \right\|_2} \right). \tag{12}$$

Equation (12) represents the definition of threshold α, whereby it is consi-

dered an inlier when the angle between the reference image and the predicted translation direction is less than α, as depicted in Figure 4. By counting all the inliers corresponding to the pose hypotheses in all image pairs, the hypothesis with the highest number of inliers is selected as the output.

3.5.3. Evaluation Metrics

In vision-based indoor localization tasks, evaluating the performance of the proposed user camera pose estimation method involves comparing the poses computed by the estimation method with the ground truth poses, and measuring the proximity of the estimation results to the ground truth. Specifically, the pose accuracy is measured by the deviation between the estimated pose and the ground truth pose, *i.e.* the absolute pose error of the query image.

Absolute attitude error is measured by a combination of absolute position error and orientation error, where the position error is expressed as the Euclidean distance in m between the estimated position of the query image and the recorded true value, as denoted below:

$$t_{abs\_err} = \left\| t_{abs\_gt} - t_{abs\_pre} \right\|_2 \tag{13}$$

The absolute directional error is expressed in degrees and represents the minimum angle of rotation required to align the directional true value and the calculated direction, as expressed below:

$$rot_{abs\_err} = 2\arccos\left| q_{abs\_gt}\, q_{abs\_pre} \right| \frac{180^\circ}{\pi} \tag{14}$$

where the quaternion q_{abs_gt} is the truth value of the recorded query image orientation and the quaternion q_{abs_pre} is the computed orientation of the query image: rot_{abs_pre} is the error between the predicted absolute orientation and the truth value, and arccosis the inverse cosine computed in the inverse trigonometric function.

4. Experiments

This section presents an evaluation of an indoor visual localization method based

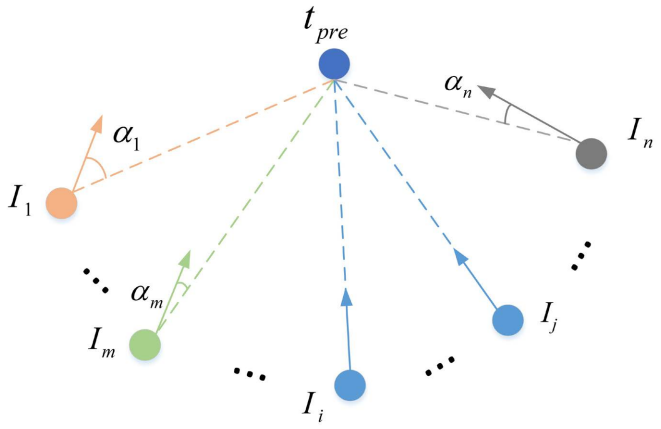

Figure 4. Estimation of query image translation.

on efficient image retrieval and matrix factorization using the essential matrix. The effectiveness and versatility of the RANSAC-based indoor localization approach are demonstrated through measurements of absolute position error in meters and absolute azimuthal error in degrees on two publicly available indoor datasets.

4.1. Datasets

7Scenes [26] was recorded by a handheld Kinect RGB-D camera and contains 7 scenes with a total of 43,000 images. All scenes were shot in an office building, and each scene usually consists of a room with a spatial extent of less than 4 meters, which contains many blurred and untextured features that are very challenging.

12Scenes [27] is a dataset of four large scenes (12 rooms) captured using the Structure.io depth sensor and iPad color camera, pushing the boundaries of RGB-D and RGB camera repositioning, and recording a significantly larger environment than the 7Scenes dataset. A total of 22,628 images were recorded after removing the 233 with anomalous bit-pose labels.

The images of the two datasets are shown in **Figure 5** and **Figure 6**.

Figure 5. Images in the 7Scenes dataset.

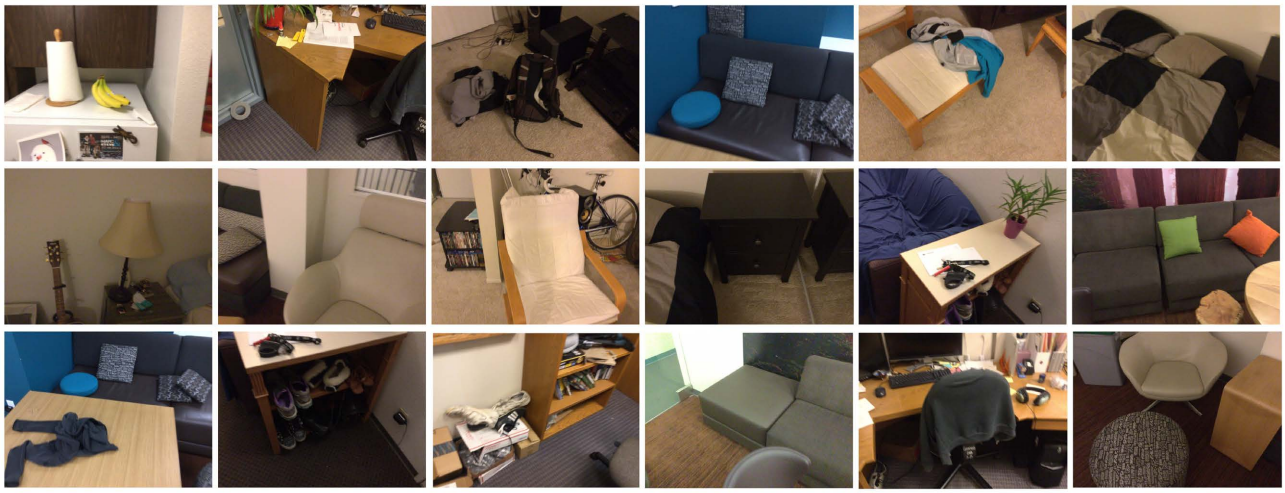

Figure 6. Images in the 12Scenes dataset.

4.2. Image Retrieval Performance

Unlike the classical image global aggregation methods, in this paper, a feature map of size 256 × 20 × 20 obtained by clipping the ResNet intermediate layer (the second ResNet residual block) pre-trained in ImageNet is re-trained on a large visual recognition dataset using the global spatial feature relations using four MLP feature mixing blocks with multi-similarity loss.

The network, with an initial learning rate of 0.05, momentum of 0.9, and weight decay of 0.001, was optimized using stochastic Stochastic Gradient Descent (SGD) and was trained for a total of 80 periods. The image retrieval model extracts the feature mapping from the middle layer, reducing the number of parameters by at least half (the last layer contains the majority of the pre-training backbone's parameters).

The image retrieval model trained in this paper can be better adapted to large scene datasets with significant variations for use as a large-scale visual scene recognition task, as tested using the Pitts250k-test database [28] and the MSLS database [29] with a wide range of illumination viewpoint variations. **Table 2** reports the performance of the proposed MLP-based image retrieval method for recall@k, and it can be seen that the proposed method is used in Pitts250k-test recall@1 up to 93.2% is significantly improved compared to both the widely used Generalized Mean (GeM) [11] method. The paper [30] converts the training into a classification problem, avoiding the expensive mining required by the commonly used comparison learning and achieving better results.

It is worth noting that in the proposed visual localization based image retrieval process, since the feature extraction and matching cycles can seriously affect the results of pose estimation, the image pairs should share as many feature points as possible, which makes it essential for the query image and the retrieved image to share some common regions, *i.e.* the purpose of retrieval is to spend a shorter time to find, with guaranteed retrieval results, the global level that is most similar, rather than local information. At the same time, for the query, it is not the most similar 5 database images that are returned, but only the 5 database images with enough visual overlap need to be satisfied. The MLP-based image retrieval model largely shortens the offline database construction time as well as the online retrieval time to satisfy the query requirements of the localization process. Therefore, the number of matching features between images is calculated

Table 2. Comparison of several methodologies in well-known benchmarks trained with ResNet-50 on the same dataset. The proposed MLP based image retrieval method gets the best performance.

Methods	Pitts250k-test			MSLS		
	R@1	R@5	R@10	R@1	R@5	R@10
GeM [11]	82.9	92.1	94.3	76.5	85.7	88.2
CosePlace [30]	91.5	96.9	97.9	84.5	90.1	91.8
Ours	93.2	97.9	98.6	84.1	91.8	94.3

to evaluate the results of image retrieval. The number of satisfactory matches with a confidence level greater than 0.2 is returned by SuperPoint feature extraction and LightGlue matching, as shown in Table 3.

Figure 7 shows the visualization of Superpoint feature ex traction with LightGlue matching on 7Scenes and 12Scenes. Then the appropriate threshold value is chosen by grid search selection: distinguish between inner and outer points, remove the outer points, update the set of inner points and then compute the essence matrix of the image pairs using the 5-point method in RANSAC for the obtained well-matched point pairs, and then remove the positional ambiguity and obtain the relative position to recover the absolute position according to the proposed RANSAC mechanism.

The computation of the MLP feature mixer based image retrieval method is mainly a matrix multiplication of the fully connected layers, which accelerates the computation and reduces the memory usage as compared to the complex self-attention, and it takes only 7 milliseconds to generate a global description of an image.

During the online period, after the query-database image pairs are obtained by image retrieval, a suitable threshold value is selected by grid search selection: inner and outer points are distinguished, outer points are removed, and the set of inner points is updated to the obtained good matching point pairs to compute the essence matrix of the image pairs by using the 5-point method in RANSAC, and then the relative pose is obtained to recover the absolute pose by removing the pose ambiguity according to the proposed RANSAC mechanism.

Table 3. Average number of good matches in two indoor datasets.

NN-search	1	2	3	4	5	Average
7Scenes	508.6	450.7	424.2	387.1	332.6	420.6
12Scenes	488.3	412.6	387.1	340.5	297.3	371.4

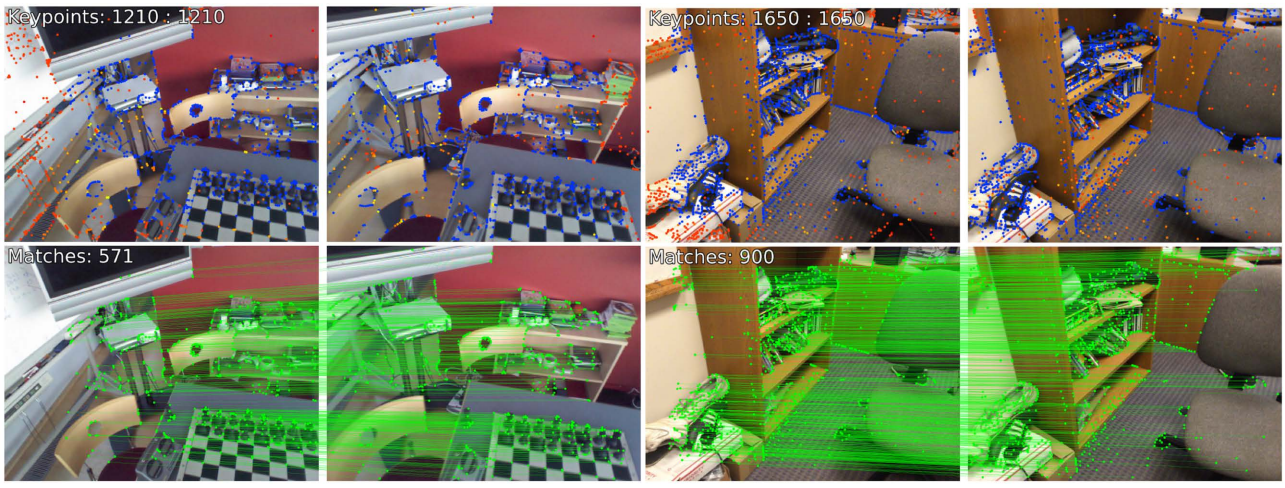

Figure 7. Feature extraction and matching effect.

4.3. Localization Accuracy

We compared the proposed model with many recent camera repositioning methods. These methods are classified into two major categories: 1) Absolute Pose Regression (APR) localization methods, and 2) Relative Pose Regression (RPR) localization methods. The proposed methods belong to the 2nd category. **Table 4** showcases the localization performance of the proposed methodology on the 7Scenes test dataset.

Experimental results show that our method reduces both positional and angular errors. The localization method proposed in this paper method basically minimizes the position and orientation errors in each indoor scene of the 7Scenes dataset, with an average position error of 0.09 m and an average orientation error of 2.14˚ on this dataset. However, due to the large difference in the environment of each indoor scene, the difference in localization error also shows a large difference, with the best performance in Heads, with a position error of only 3 cm and an orientation error of 2.16˚, while for the Stairs scene the localization performs poorly, with an error as high as 0.21 m and 3.47˚, which we believe is most likely due to the excessive repetitive structures in the staircase images, and the distinguishability of the extracted feature points poorly. We will investigate this in future work.

12Scenes is another indoor scene dataset, and the recorded indoor environments are significantly larger than 7Scenes. Since there are fewer studies related to the 12Scenes dataset, this paper spends a lot of time reproducing the classical relative bit-pose regression networks, NN-Net [19] as well as NC-EssNet [20], in strict accordance with the criteria of the paper, as shown in **Table 5**.

As shown in **Figure 7**, the indoor images recorded by the 12Scenes dataset have poor lighting conditions and more blurred images, which pose a challenge to the localization task. The experimental results of the 12Scenes dataset in **Table 5** are obviously worse than the localization performance of 7Scenes, but the results show that the proposed indoor localization method still obtains a significant improvement compared to other methods, with an average position error of

Table 4. Median position and rotation errors for different relocation methods on the 7Scenes dataset.

Scene	NN-Net [19]	NC-Esset [20]	GRNet [21]	FeatLoc [18]	Ours
Chess	0.13 m, 6.5˚	0.12 m, 5.6˚	0.08 m, 12.4˚	0.07 m, 3.66˚	0.05 m, 1.51˚
Fire	0.26,12.7˚	0.26,12.7˚	0.21 m, 7.5˚	0.17 m, 5.95˚	0.07 m, 1.96˚
Heads	0.14 m, 12.3˚	0.14 m, 10.7˚	0.13 m, 8.7˚	0.10 m, 7.57˚	0.03 m, 2.16˚
Office	0.21 m, 7.4˚	0.20 m, 6.7˚	0.15 m, 4.1˚	0.16 m, 5.20˚	0.07 m, 1.64˚
Pumpkin	0.24 m, 6.4˚	0.22 m, 5.7˚	0.15 m, 3.5˚	0.11 m, 3.86˚	0.10 m, 2.06˚
RedKitchen	0.24 m, 8.0˚	0.22 m, 6.3˚	0.19 m, 3.7˚	0.20 m, 6.43˚	0.08 m, 2.16˚
Stairs	0.27 m, 11.8˚	0.31 m, 7.9˚	0.22 m, 6.5˚	0.16 m, 8.57˚	0.21 m, 3.47˚
Average	0.21 m, 9.3˚	0.21 m, 7.5˚	0.16 m, 5.2˚	0.14 m, 5.89˚	0.09 m, 2.14˚

Table 5. Median position and rotation errors for different relocation methods on the 12Scenes dataset.

Scene	Volume	NN-Net [19]	NC-EssNet [20]	FeatLoc [18]	Ours
apt1_kitchen	33 m^3	0.22 m, 6.76˚	0.15 m, 11.53˚	0.32 m, 5.19˚	0.14 m, 7.58˚
apt1_living	30 m^3	0.25 m, 5.45˚	0.19 m, 6.57˚	0.26 m, 0.14˚	0.18 m, 5.89˚
apt2_bed	14 m^3	0.46 m, 6.13˚	0.21 m, 6.70˚	0.37 m, 5.39˚	0.08 m, 4.32˚
apt2_kitchen	21 m^3	0.83 m, 36.03˚	0.18 m, 9.39˚	0.73 m, 6.37˚	0.09 m, 8.61˚
apt2_living	42 m^3	0.23 m, 5.24˚	0.17 m, 8.18˚	0.40 m, 5.71˚	0.13 m, 5.73˚
apt2_luke	53 m^3	0.54 m, 6.26˚	0.23 m, 8.03˚	0.33 m, 4.85˚	0.17 m, 7.83˚
office1_gates362	29 m^3	0.27 m, 5.27˚	0.16 m, 5.47˚	0.52 m, 5.22˚	0.14 m, 4.73˚
office1_gates381	44 m^3	0.44 m, 7.27˚	0.28 m, 12.00˚	0.42 m, 6.23˚	0.24 m, 7.93˚
office1_lounge	38 m^3	0.53 m, 5.72˚	0.31 m, 7.01˚	0.39 m, 4.50˚	0.26 m, 6.26˚
office1_manolis	50 m^3	0.27 m, 5.66˚	0.19 m, 6.81˚	0.30 m, 4.67˚	0.15 m, 6.82˚
office2_5a	38 m^3	0.29 m, 5.50˚	0.20 m, 5.09˚	0.31 m, 4.32˚	0.11 m, 5.03˚
office2_5b	79 m^3	0.29 m, 5.07˚	0.21 m, 5.78˚	0.23 m, 4.14˚	0.15 m, 5.74˚
Average	39 m^3	0.35 m, 8.28˚	0.21 m, 8.04˚	0.38 m, 5.04˚	0.15 m, 6.37˚

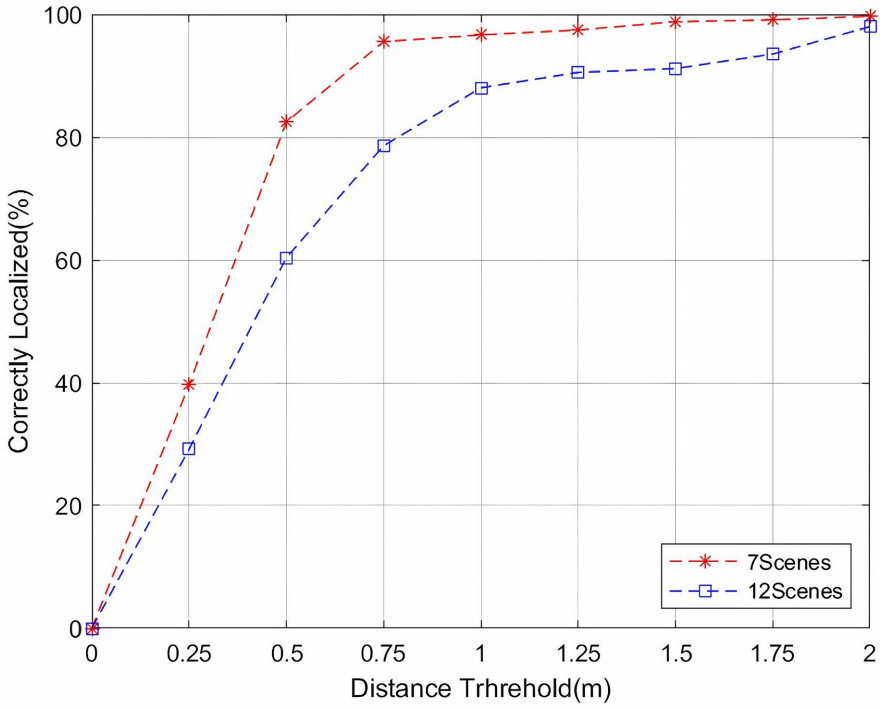

Figure 8. Position error cumulative distribution function.

0.15 m and an average orientation error of 6.37˚. Compared with the other visual localization methods mentioned above, the best results are achieved in terms of position error, but the performance is slightly worse in solving the camera orientation.

Figure 8 and Figure 9 show the performance of the proposed indoor localization

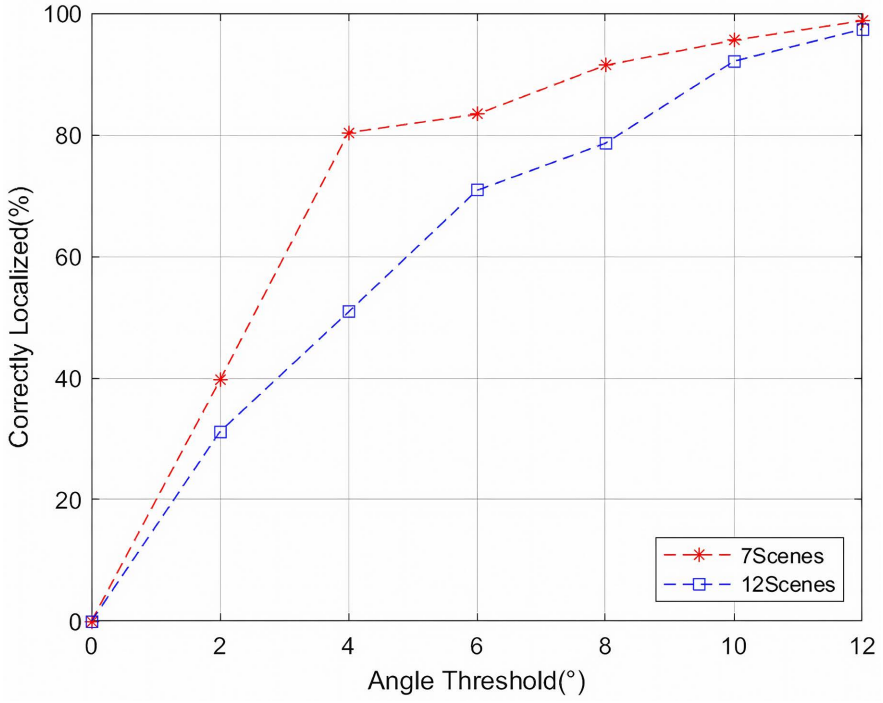

Figure 9. Directional error cumulative distribution function.

method more visually. Among them, 82.47% of the query images in the 7Scenes dataset have a localization error of less than 0.5 meters and 80% of the query images have an orientation error of less than 4 degrees, while more than 90% of the query images in the 12Scenes dataset have a localization accuracy of less than meters and a maximum orientation error of less than 12 degrees. Compared with the 7Scenes localization results, the 12Secens database has poorer localization results due to the fact that the indoor environment recorded in 12Scenes is significantly larger than that in the 7Scenes dataset, and also the images in the dataset have poorer lighting conditions and more blurred images, which poses a challenge to the localization task.

5. Conclusions

In this paper, an indoor visual localization method based on efficient image retrieval and relative position calculation is proposed. A novel image retrieval method based on CNN cropping and MLP aggregation is used to generate compact global descriptions by learning global spatial relations iteratively for the feature mapping of the pre-trained network, while the computational process of the retrieval method based on MLP aggregation is highly efficient due to the fact that, unlike the self-attention mechanism where the complexity scales into a quadratic scale, the fully-connected layer is mainly a matrix multiplication operation. The offline phase takes only 7 ms to generate a global description of an image.

The online localization phase performs efficient image retrieval by pre-training the retrieval model to obtain matching images with pose labels to construct query-database image pairs. A set of CNN features with original images and

poses can represent the whole region. Then, a feature-point correspondence strategy is applied to solve the relative pose ambiguity problem by a novel RANSAC mechanism to estimate the exact location and orientation of the query image. Experimental results conducted on two publicly available indoor localization datasets show that our monocular vision-based indoor pose estimation method produces highly accurate localization results. The proposed indoor method is suitable for scenes lacking depth information and has excellent cross-scene generalization capabilities without the need for complex preprocessing in the offline phase and without relying on the 3D scene structural model. In this paper, we argue that image data with poor lighting conditions and blurred images can have a large negative impact on the localization results, which serves as a reminder for our future work.

Acknowledgements

This paper was supported by the National Natural Science Foundation of China under the Youth Foundation Program (62001272).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Lim, C.H., Wan, Y., Ng, B.P. and See, C.M.S. (2007) A Real-Time Indoor Wifi Localization System Utilizing Smart Antennas. *IEEE Transactions on Consumer Electronics*, **53**, 618-622. https://doi.org/10.1109/TCE.2007.381737

[2] Hazas, M. and Hopper, A. (2006) Broadband Ultrasonic Location Systems for Improved Indoor Positioning. *IEEE Transactions on Mobile Computing*, **5**, 536-547. https://doi.org/10.1109/TMC.2006.57

[3] De Angelis, A., Dwivedi, S. and Händel, P. (2013) Characterization of a Flexible Uwb Sensor for Indoor Localization. *IEEE Transactions on Instrumentation and Measurement*, **62**, 905-913. https://doi.org/10.1109/TIM.2013.2243501

[4] Subbu, K.P., Gozick, B. and Dantu, R. (2013) LocateMe: Magnetic-Fields-Based Indoor Localization Using Smartphones. *ACM Transactions on Intelligent Systems and Technology*, **4**, 1-27. https://doi.org/10.1145/2508037.2508054

[5] Liu, L., Li, H. and Dai, Y. (2017) Efficient Global 2D-3D Matching for Camera Localization in a Large-Scale 3D Map. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 2391-2400. https://doi.org/10.1109/ICCV.2017.260

[6] Perronnin, F. and Dance, C. (2007) Fisher Kernels on Visual Vocabularies for Image Categorization. 2007 *IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, 17-22 June 2007, 1-8. https://doi.org/10.1109/CVPR.2007.383266

[7] Jégou, H., Douze, M., Schmid, C. and Pérez, P. (2010) Aggregating Local Descriptors into a Compact Image Representation. 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, 13-18 June 2010, 3304-3311. https://doi.org/10.1109/CVPR.2010.5540039

[8] Lowe, D.G. (2004) Distinctive Image Features from Scale-Invariant Keypoints. *In-*

ternational Journal of Computer Vision, **60**, 91-110.
https://doi.org/10.1023/B:VISI.0000029664.99615.94

[9] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T. and Sivic, J. (2017) NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 1437-1451.
https://doi.org/10.1109/TPAMI.2017.2711011

[10] Babenko, A. and Lempitsky, V. (2015) Aggregating Local Deep Features for Image Retrieval. *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 1269-1277.

[11] Radenović, F., Tolias, G. and Chum, O. (2018) Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **41**, 1655-1668. https://doi.org/10.1109/TPAMI.2018.2846566

[12] Melekhov, I., Tiulpin, A., Sattler, T., *et al.* (2019) DGC-NET: Dense Geometric Correspondence Network. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, 7-11 January 2019, 1034-1042.
https://doi.org/10.1109/WACV.2019.00115

[13] Luo, Z., Zhou, L., Bai, X., *et al.* (2020) ASLFeat: Learning Local Features of Accurate Shape and Localization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 6588-6597.
https://doi.org/10.1109/CVPR42600.2020.00662

[14] Shi, C., Li, J., Gong, J., Yang, B. and Zhang, G. (2022) An Improved Lightweight Deep Neural Network with Knowledge Distillation for Local Feature Extraction and Visual Localization Using Images and LiDAR Point Clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, **184**, 177-188.
https://doi.org/10.1016/j.isprsjprs.2021.12.011

[15] Kendall, A., Grimes, M. and Cipolla, R. (2015) PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. *Proceedings of the IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 7-13 December 2015, 2938-2946.
https://doi.org/10.1109/ICCV.2015.336

[16] Bui, M., Baur, C., Navab, N., Ilic, S. and Albarqouni, S. (2019) Adversarial Networks for Camera Pose Regression and Refinement. 2019 *IEEE/CVF International Conference on Computer Vision Workshop* (*ICCVW*), Seoul, 27-28 October 2019, 3778-3787. https://doi.org/10.1109/ICCVW.2019.00470

[17] Shavit, Y. and Keller, Y. (2022) Camera Pose Auto-Encoders for Improving Pose Regression. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M. and Hassner, T., Eds., *Computer Vision—ECCV* 2022, Springer, Cham, 140-157.
https://doi.org/10.1007/978-3-031-20080-9_9

[18] Bach, T.B., Dinh, T.T. and Lee, J.H. (2022) FeatLoc: Absolute Pose Regressor for Indoor 2D Sparse Features with Simplistic View Synthesizing. *ISPRS Journal of Photogrammetry and Remote Sensing*, **189**, 50-62.
https://doi.org/10.1016/j.isprsjprs.2022.04.021

[19] Laskar, Z., Melekhov, I., Kalia, S. and Kannala, J. (2017) Camera Relocalization by Computing Pairwise Relative Poses Using Convolutional Neural Network. *Proceedings of the IEEE International Conference on Computer Vision Workshop*s, Venice, 22-29 October 2017, 920-929. https://doi.org/10.1109/ICCVW.2017.113

[20] Zhou, Q., Sattler, T., Pollefeys, M. and Leal-Taixe, L. (2020) To Learn or Not to Learn: Visual Localization from Essential Matrices. 2020 *IEEE International Conference on Robotics and Automation*, Paris, 31 May-31 August 2020, 3319-3326.
https://doi.org/10.1109/ICRA40945.2020.9196607

[21] Turkoglu, M.O., Brachmann, E., Schindler, K., Brostow, G.J. and Monszpart, A. (2021) Visual Camera Re-Localization Using Graph Neural Networks and Relative Pose Supervision. 2021 *International Conference on* 3*D Vision* (3*DV*), London, 1-3 December 2021, 145-155. https://doi.org/10.1109/3DV53792.2021.00025

[22] Ali-Bey, A., Chaib-Draa, B. and Giguere, P. (2023) Mixvpr: Feature Mixing for Visual Place Recognition. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, 2-7 January 2023, 2997-3006. https://doi.org/10.1109/WACV56688.2023.00301

[23] Ali-bey, A., Chaib-draa, B. and Giguère, P. (2022) GSV-Cities: Toward Appropriate Supervised Visual Place Recognition. *Neurocomputing*, **513**, 194-203. https://doi.org/10.1016/j.neucom.2022.09.127

[24] DeTone, D., Malisiewicz, T. and Rabinovich, A. (2018) Superpoint: Self-Supervised Interest Point Detection and Description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, 18-22 June 2018, 224-236. https://doi.org/10.1109/CVPRW.2018.00060

[25] Sarlin, P.E., DeTone, D., Malisiewicz, T. and Rabinovich, A. (2020) Superglue: Learning Feature Matching with Graph Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 4937-4946. https://doi.org/10.1109/CVPR42600.2020.00499

[26] Shotton, J., Glocker, B., Zach, C., *et al.* (2013) Scene Coordinate Regression Forests for Camera Relocalization in RGB-D images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, 23-28 June 2013, 2930-2937. https://doi.org/10.1109/CVPR.2013.377

[27] Valentin, J., Dai, A., Nießner, M., *et al.* (2016) Learning to Navigate the Energy Landscape. 2016 *Fourth International Conference on* 3*D Vision* (3*DV*), Stanford, 25-28 October 2016, 323-332. https://doi.org/10.1109/3DV.2016.41

[28] Torii, A., Sivic, J., Pajdla, T. and Okutomi, M. (2013) Visual Place Recognition with Repetitive Structures. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, 23-28 June 2013, 883-890. https://doi.org/10.1109/CVPR.2013.119

[29] Warburg, F., Hauberg, S., Lopez-Antequera, M., *et al.* (2020) Mapillary Street-Level Sequences: A Dataset for Lifelong Place Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 2623-2632. https://doi.org/10.1109/CVPR42600.2020.00270

[30] Berton, G., Masone, C. and Caputo, B. (2022) Rethinking Visual Geo-Localization for Large-Scale Applications. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 4868-4878. https://doi.org/10.1109/CVPR52688.2022.00483