A review of Visual-Based Localization

Xing Xin[1]
College of Systems Engineering
National University of Defense Technology
Changsha, China
1186130830@qq.com

Jie Jiang*
College of Systems Engineering
National University of Defense Technology
Changsha, China
Correspondence:
jiejiang@nudt.edu.cn

Yin Zou
College of Systems Engineering
National University of Defense Technology
Changsha, China
960956278@qq.com

ABSTRACT query images:查询图像

The visual-based localization (VBL) obtains the corresponding pose estimation in the localization system by utilizing various useful information in the surrounding environment, such as images, point cloud models, geometric information, semantic information. In recent years, visual-based localization (VBL) has been widely concerned by scientists, mainly because the commonly used GPS localization system cannot be effectively used in various environments. When GPS localization fails in some scenes such as very messy environments and severe signal occlusion, we can consider using visual-based localization to obtain the pose of the query images. Visual-based localization (VBL) has been widely used in the field of visual tasks, such as augmented reality, unmanned vehicle navigation, robotics, closed-loop detection, SFM (Structure from Motion) models. After years of development, the methods of visual-based localization (VBL) have been enriched and developed, In order to better understand the latest developments in VBL, overall research status and possible future development trends, we need make a systematic detailed classification of VBL. Although the predecessors have summarized the methods of VBL, due to the many new breakthroughs in VBL in recent years, the original summary is not perfect enough. So this paper will make a new and more detailed review of VBL in recent years. This paper divides the visual-based localization methods into three categories: image-based localization, localization based on learning model and localization based on 3D structure. And we also detail the principle, development of methods and the advantages and disadvantages of each method and future development trends.

occlusion :阻塞

augmented: 增广的,扩张的

CCS CONCEPTS

• General and reference ~Reference work • General and reference ~Surveys and overviews

KEYWORDS

Visual-based localization, Image-based localization, Localization based on learning model, Localization based on 3D structure.

RICAI '19, September 20–22, 2019, Shanghai, China

ACM ISBN 978-1-4503-7298-5/19/09…$15.00
https://doi.org/10.1145/3366194.3366211

1 Introduction

1.1 Overview

Visual-based localization (VBL) obtains the information around the query images by images, 3D point cloud and other data, and then obtains the pose of the query images (the direction of the images and the location of the images space) [1]. In recent years, the research and use of various data (images, point clouds, etc.) are gradually improving, and the benchmark datasets for VBL such as the Aachen Day-Night dataset are continuously constructed and enriched. And, the commonly used localization system (GPS localization system) cannot achieve accurate localization in all scenes, it is easy to be invalid in some scenes such as very messy environment and severe signal occlusion. For example, in some complicated and dense urban environment, there are a large number of buildings, the satellite signal occlusion is very serious, and the effect of GPS is not good, so VBL has gradually become a hot research topic. Moreover, VBL can play a role in many visual tasks, such as robot localization and navigation, navigation of autonomous vehicles, augmented reality, closed-loop detection, SFM(Structure from Motion model), re-localization in SLAM(Simultaneous Localization and Mapping） [2 , 3, 4, 5]. VBL needs to obtain information in the scene environment for pose estimation. In the real environment of the scene, there are many kinds of information such as image information, point cloud information, semantic information, geometric information, which leads to various localization methods. For these methods, the absence of a complete system classification is not conducive to learning research. In order to better understand the latest developments in VBL, overall research status and possible future development trends, we need make a systematic detailed classification of VBL. Although the predecessors [1] have summarized the methods of VBL, and they divided the VBL into two categories, there are many new

conducive: 有利于

breakthroughs in VBL in recent years, the original summary is not perfect enough. Therefore, this paper summarizes research results in the field of visual localization in recent years, and conducts a more detailed classification and summary of visual localization methods. At the same time of summarizing the methods, we also introduced the advantages and disadvantages and development of various methods, and compared various methods to illustrate some possible development trends of VBL in the future.

Basis :依据

1.2 Classification and Basis

In this paper, the VBL methods are divided into the following categories: 1. image-based localization, 2.localization based on learning model, 3. localization based on 3D structure. The classification of this paper is mainly based on the core principles of various methods: First, a large class of methods is based on the idea of image retrieval, they turn the task of localization into an image retrieval task. This paper classifies it as an image-based method; some methods are based on the idea of machine learning, the pose estimation is obtained by learning the scene model obtained from the training. This paper classifies it as localization based on learning model; one type of method is based on the idea of three-dimensional structure, and the pose of the camera is obtained by constructing a three-dimensional model and establishing a matching of the two-dimensional feature points of the image with the three-dimensional feature points of the 3D model.

retrieval :检索；

建立图像的二维特征点与三维特征的匹配来获得相机的姿态3D模型的要点。

The image-based localization method mainly retrieves the most similar photo from the query image in the image database and uses it to estimate the pose. They are commonly used for position recognition [6] and ring closure detection [7]. Their advantage is that they can also be robust against changing conditions [6, 8, 9]. The main idea of the method is to obtain a series of images of the scene and construct an image database, then the images currently used for querying and matching and the images in the image database are retrieved and queried, then we get the images with the highest degree of correlation, thereby obtaining the pose of the query image. The image-based localization method actually turns a positioning task into an image retrieval task, and finally obtains the pose of the query image. The method is relatively simple, and the method is relatively easy to implement. However, the localization accuracy is not competitive with the accuracy of the other two methods. However, although this method is relatively inferior in fine precision, (two kinds of data evaluation are commonly used in evaluating the accuracy of VBL: the estimated position and direction error; the percentage of localization success within the threshold. The second data usually has three Thresholds: fine precision (0.25m, 2°); (0.5m, 5°); coarse precision (5m, 10°)). However, in some special scenes, such as some environments with large environmental changes, strong illumination changes, and strong visual point changes, the coarse precision localization of this method is better than the coarse precision localization of method based on 3D structure [9], In these scenes, the robustness of the method is better than that based on 3D structure.

The main idea of localization method based on learning model is to learn and train model, such as training a regression model or learning CNN structure, and then get posture estimation through learning and training model, which is a localization method that directly to learn the corresponding image pose through the input query image data. There are also many branches of the localization method based on learning model, this paper mainly introduces the method of using the regression forest thought and the method of using CNN network. The CNN-based method can also have better effects when the environment scene changes, has better scalability. The idea of regression forest is to get the pose estimation of query image directly from images through pre-trained regression forest model of scene. The method of regression forest needs to get the depth information of image, the method can adapt to the scene without texture or with weak texture. Compared with the method based on keypoints, this method can have a large amount of point information to verify the camera pose estimation. At the same time, it does not need a series of very accurate description and matching of feature points, so that the localization speed is faster.

As for the localization based on 3D structure, the main idea is that we first construct a 3D point cloud model using the SFM algorithm [10], and then use the feature descriptor method to extract the 2D features of the image. The commonly extracted features is local features such as SIFT [11], LIFT [12], then we establish the association between the 2D features and the 3D model and match the 2D feature points extracted from the image with the 3D points in the model. Then the resulting of 2D-3D matching is placed into a RANSAC (Random Sample Consensus) cycle, and the n-point pose solver is used to estimate the pose estimation in this recirculation. This method is still the most advanced visual localization method at present. However, In the case of large changes in the scene environment, a large number of repeated structures in the scene, weak textures of the scene, no texture structure, strong illumination changes, motion blur, strong viewpoint changes, etc., the accuracy of the method will be greatly reduced, and the method may even fail.

1.3 Challenges

At present, although the VBL method has been enriched, VBL still has some challenging problems waiting to be solved, for example, in the scenes with large changes in scene appearance, scenes with strong viewpoint changes, scenes with strong illumination changes, and textureless scenes, the accuracy of localization needs to be improved. The image-based method and the method based on learning model are relatively robust in these scenes. The method based on 3D structure has poor localization effect in such scenes. For methods based on learning model, the data required is too large, the cost of training is too high, and the need for offline training and practicality is also a problem that needs to be solved.

1.4 Related Work

VBL is a hot research topic, and many contributions have introduced the VBL approach. Some of the latest contributions are

that Nathan Piasco [1] introduced a summary of VBL's numerous methods and the data representations required by VBL such as features, descriptors, cross-data challenges and the benefits of using heterogeneous data in challenging scenes. This contribution mainly divides VBL into two types: direct method and indirect method. And they mainly focus on urban localization. Brejcha [13] introduced various VBL-related work, which is mainly based on the environment of various specific VBL methods to divide VBL into multiple categories. Zamir et al. [14] collected articles from relevant fields at the time and mapped a large panorama of VBL, confirming the great development of VBL and confirming the growing importance of this field in current research.

1.5 Structure Layout

The rest of the paper is divided into three parts: on the one hand, we summarize the three major methods of VBL in detail, introduce the image-based localization method in the second section, and introduce the localization based on learning model in the third section, introduce the localization based on 3D structure in the fourth section. And we also introduce the development, advantages and disadvantages, and applicable scene conditions of various methods. On the other hand, detailed qualitative and quantitative comparisons are made on various methods. We compare the various methods on multiple datasets; finally, we summarized the work and discussed the future trends of the various methods.

2 The Image-based Localization

This type of method estimates the pose by a photo retrieved from the image database that is most similar to the query image. The image-based localization essentially converts the positioning task into an image retrieval question and provides rough posture information about the query position. For the image retrieval problem, it can generally be divided into two steps: a description of the data of the image database and the query image, and a similarity relationship between the data of the two sets. At the same time, the three more critical and important constraints to be considered for image retrieval are the accuracy of the search, the efficiency of the search, and the amount of memory used.

Relja Arandjelovi´c et al. [15] proposed a new query expansion method for image retrieval query. By using the reverse index, it can realize immediate retrieval. Douze et al. [16] proposed the VLAD idea to aggregate local image descriptors into finite dimensional vectors. The image retrieval efficiency is improved by jointly optimizing the dimensionality reduction and indexing algorithm. Inspired by this idea, various VBL methods have been developed. For example, Relja Arandjelovi et al. combined the Vector of Locally Aggregated Descriptors with CNN to develop a generic NetVLAD layer, which can be better inserted into various CNN architectures. The DenseVLAD method proposed by Akihiko Torii et al. [17] uses a local gradient-based dense sampling descriptor to represent images on multiple scales, combining the effective synthesis of the views with a compact indexable image representation. Kim et al [18] introduced the PBVLAD method for local fusion of SIFT features detected in MSER. Herve et al. [19] optimized the vector representation of images by reducing the effects of false matching on image similarity and limiting the interference of descriptors during the aggregation phase.

The FAB-MAP method proposed by M. Cummins et al. [20] is based on the Bag-ofWords (BoW), which processes and models different visual words and their concurrency probability. It is essentially a probability recognition method based on position representation. In general, image-based localization methods are used for position recognition because of its accuracy. In order to improve its the positioning accuracy, Charbel Azz et al. [21] proposed GIST-based Search Space Reduction (GSSR) method, which mainly use the global descriptors, especially GIST descriptors, introduces a new similarity measure for matching key frames, then matches with a limited number of 3D points.

Image-based localization can also be combined with CNN networks for more advanced performance. Filip Radenovi´c et al. [22] used CNN to simulate image matching based on spatial verification and local feature retrieval to retrieve images from large numbers of unordered images in a fully automated style, then the ultimate performance of the search is improved. In addition, in the stage of data description of the image database and image, the larger the description data of the image, the more details will be, the matching effect will be better, but the cost will be higher, Therefore, reducing the dimension of the descriptor by means of feature aggregation is helpful. Josef Sivic et al. [23] proposed a vector quantization method to reduce the dimension of descriptors. The corresponding dictionary and visual words are established by aggregation, and the feature package (BoF) is associated with the dimension vector of the dictionary containing the words. In this way, the similarity between the database image and the query image description data can be efficiently calculated. In view of the difficulty in assigning the above methods, the soft assignment method proposed by Philbin et al. [24] uses k closest visual words and linearly combines them to associate features. J´egou et al. [25] created binary tags and associated each feature with a tag to more detail its position in the visual vocabulary. Compared with the previous method, this method has a better improvement in speed and query precision, and this method is still used.

In some environments, especially urban environments, there are the large number of repetitive structures, and it brings interference to recognition, description and matching, which makes the localization more difficult. Torii et al. [26] introduced meta-features for this problem. The meta-features contain multiple similar descriptors, so that the meta-features are comparable in the descriptor vector. This method enables the intensive extraction of local features in this case which has a large number of repetitive structure. This method achieves good localization results in urban environments.

For the similarity correlation phase between image descriptors, firstly, when the amount of data is not particularly large, brute force search can be considered, such as Sunderhauf et al. [27] brute force comparison of the retrieval process based on local and mixed characteristics. When the dimension of the feature is too

large, we can consider using approximate proximity search, and get a better search speed by sacrificing a part of the effect. In addition to using approximate proximity, we can also consider using machine learning. For example, Cao and Snavely [28] used the SVM method, firstly clustered the image database based on the similarity of the images. For similar images, the SVM was used for training for each cluster obtained. Experiments show that the method is effective and robust.

3 Localization Based on Learning Model

Localization based on learning model is to learn and train regression model, the CNN network structure, etc., and use these models to directly obtain the pose of the corresponding image through the data of the input query image. There are also many branches in the localization based on learning model. This paper mainly introduces the method of using CNN network and the method of using regression forest.

3.1 The Regression Forest

The idea of regression forest is to obtain the pose estimation of the query image directly from the images by pre-training the regression forest of the scene.

Shotton et al. [29] first used regression idea and RGB-D data to construct a regression forest, and then encode each pixel in the forest that is related to the environment and their global location. In the query process, depth information is needed, so some pixels obtained by the depth camera are added to the regression forest, and then multiple pose hypotheses for each pixel are obtained, and perform random consistency optimization. It is worth noting that one limitation of this approach is the need to train the regression forest for 3D scenes in advance. Guzman-Rivera et al. [30] made some improvements to the above method, the authors obtained multiple candidate results through the trainer for further screening. That is, using multiple regression forests, a large number of pose hypothesis data are obtained through the forest, and then cluster the pose hypothesis data obtained from these regressions and then obtain the average pose. In this way, the error is minimized.

Valentin et al. [31] further introduced the multiscale navigation map into the original localization method to obtain the initial candidate pose. Then they introduced a mixed Gaussian function, used this function to express the uncertainty of the prediction process, and then modeled this uncertainty of the predicted point, added the uncertainty information to the pose regression estimation to improve the results of the pose estimation. Meng, L et al. [32] considered not using images other than RGB images in the query process, and adopted the nearest neighbor search method based on the thin SIFI feature to compensate for the loss of accuracy in the pose refinement. E. Brachmann et al. [33] constructed a classification regression forest with a stacked structure by adapting to random forests. In this method, although the depth information of the image is still needed in the training phase, the depth information in the test time is no longer needed. Massiceti et al. [34] combined the regression forest with the neural network and tried to use the neural network in regression forest. Since it is enough to take advantage of the performance advantages of neural networks to enhance the dense regression process, it can maintain the advantage of predictive efficiency of random forests.

One limitation of using regression forests is the need to train a scene's regression forest in advance. In order to break this limitation, Glocker et al. [35] proposed a method of using regression ferns. According to the principle of initial randomization, the regression fern generates corresponding descriptors, and then generates a table for searching and maintains it. In this way, the features of the image are associated with the three-dimensional pose, which can fastly associate RGB-D images and features. The method does not need to generate the regression forest in advance, which improves the limitation of the pre-training regression forest of the scene to a certain extent, but the accuracy of the method is not as accurate as the above method. Cavallari et al. [36] improved it on the basis of regression forest and constructed an adaptation forest. This method does not require pre-training in the scene where pose estimation is to be performed. The main idea is to first perform pre-training in a common scene, then it get a regression forest of a common scene, and then retain the backbone of the regression forest and remove the leaves containing the specific information of the scene. Then if we need to estimate pose in a specific scene, we only need to import an example to get the regression forest under this scene. This method breaks the limitation of pre-training in a specific scene while ensuring the positioning accuracy, and realizes real real-time re-localization.

3.2 The CNN Network

The CNN-based method has received extensive attention in recent times and has been greatly developed. This type of method can also have good effects when the environment scene changes, and has better scalability. Since the CNN network has achieved good results in image classification [37, 38] and target detection [39], it has gradually been applied to the localization method. In order to obtain an accurate estimate of the position and attitude of the 6DoF camera, Kendall et al. [40, 41] proposed the PoseNet, which uses a pair of images to train, automatically regress the camera's position and pose, and builds the pose estimation problem and turns it into a regression problem. The method takes a single RGB image whose size is 224x224, takes an end-to-end approach, and return the camera's 6-degree-of-freedom pose in a scene using the images. PoseNet has better accommodation for the environment. However, due to the lack of sufficient training data, the pose estimated by PoseNet is less accurate than other structure-based methods. It is worth mentioning that the localization accuracy of the CNN-based method is generally not as good as the traditional three-dimensional structure-based method, However, the CNN-based method is more robust to scenes with larger changes, scenes with more repeating structures, weak textures, and motion blur than those based on 3D structure methods. And the CNN-based method still has a lot of space for improvement and is a good research direction.

In order to improve the accuracy of positioning, we can consider from the perspective of loss function. Alex Kendall et al. [42] considered improving the cost function of PoseNet with high cost and studied a loss function of some re-projection errors based on geometry and scene. They simulate and utilize the methods in the multiview geometry standard system, and directly train using the scene geometry, then they use geometric loss functions to improve performance. F. Walch et al. [43, 44] introduced the LSTM structural unit and constructed a CNN+ LSTM architecture, and used LSTM to reduce the dimension of the feature vector. Since the dimensions of images passing through the FC layer in the CNN are generally large, it is easy to cause over-fitting of the data, which makes the prediction less precise. The LSTM structural unit can be used on the output of the FC layer, then the feature vector can be structurally dimensionality reduction, and can get better association of FC layer and convolution layer features, and get the most useful feature association. Liu et al. [45] considered using depth information to improve the accuracy of pose estimation, they introduced depth map information into CNN, and combined with it, returned to the camera pose in the case of complete invisibility.

The number of training examples available in CNN is limited. So Jia et al. [46] used the dense point cloud model in the SFM model to replenish images by rendering artificial images, so that to increase the available training examples to enhance the process of image pose estimation. Contreras [47] used the CNN network to construct a map whoes size is fixed and improved it by constantly adding new tracks. Moreover, Contreras has also reduced the original size of CNN by a factor of three, and make it has good positioning effect both outdoors and indoors. Alex Kendall et al. [10] proposed a real-time relocation method, Bayesian PoseNet, which uses the Bayesian convolutional neural network to obtain the uncertainty measure of the model. This measure is used to estimate the credibility of the model data and the pose. The method is trained in an end-to-end manner and does not require an additional optimization process. Interestingly, Weyand et al. [48] used CNN to translate positioning problems into a classification task. Give the network an image, then the PlaNET layer in the network can return a map for subsequent queries and it extend the original network using the LSTM layer.

Valada [49] proposed the VLocNet network structure, and in the process of geometric consistency training, optimized the loss function of the search space to obtain accurate pose estimation by using relative motion information. On this basis, Noha Radwan et al. [50] proposed the VLocNet ++ network structure based on deep learning. The advantage of this network is multi-task learning, the pose estimation, mileage estimation and semantic segmentation can be obtained simultaneously by using the MTL framework. At the same time, we also use these multiple tasks to connect with each other, promote each other, and strengthen each other's results. A new adaptive weighted fusion layer network is added to the network. The network structure combines the world's geometric information and semantic information as well as motion-related information, and embeds this information into the pose regression network to enhance the regression results. This method achieves the most advanced results on the 7scenes datasets.

Paul-Edouard Sarlin et al. [51] proposed a coarse-to-fine hierarchical localization method, HFNet, based on global CNN. The hierarchical feature network (HF-Net) utilizes local and global features in the network. The method firstly performs coarse localization, obtains the corresponding candidate pose hypothesis through global image retrieval, and then performs fine localization: local feature matching is performed in these candidate pose hypotheses, and then accurate pose estimation results are obtained. This layered idea allows for precise localization in scenes with large changes in the appearance of the environment.

Inspired by the idea of generating against network (GAN), Mai Bui et al. [52] introduced a new network framework similar to GAN's framework, mainly with the camera pose regression network and pose discriminator network. The pose regression network is similar to the GAN generator for regress the pose. The pose discriminator network is similar to the GAN discriminator for distinguishing the correct pose from the wrong pose. The network is mainly for localization based on RGB image, give an RGB image, and the camera pose is obtained by using the previous regression network. The latter discriminator network is trained by using the extracted RGB image features, and then the discriminator network is used for pose identification.

Torsten Sattler et al. [53] introduced the absolute pose regression (APR) visual localization method, which is computationally efficient and usually requires a powerful GPU. Usually, it uses some basic networks such as VGG and ResNet to extract features. These features will be embedded in a high dimensional space to obtain a corresponding pose estimate. The APR method can use geometric reprojection errors, odometer constraints, and the like. One limitation of the APR approach is that it is difficult to generalize and it is difficult to get good results outside of the training data.

3.3 The Challenges

The localization based on the learning model requires a large amount of training data, their cost of retraining is high, and offline training is required. It is difficult to locate online and the practicality is not high. Further research on these aspects can be conducted in later studies.

Most of the existing state-of-the-art localization methods based on machine learning require offline training scenes. Pre-offline training will make the method less practical. In order to improve the problem, Cavallari et al. [36] improved it on the basis of regression forest and constructed an adaptation forest. This method does not require pre-training in the scene where pose estimation is to be performed. The main idea is to first perform pre-training in a common scene, then it get a regression forest of a common scene, and then retain the backbone of the regression forest and remove the leaves containing the specific information of the scene. Then if we need to estimate pose in a specific scene, we only need to import an example to get the regression forest under this scene.

Although Cavalari [36] proposed adaptive regression forest to achieve real-time adjustment of the regression forest to online use, but their forest use hand-crafted features designed for indoor environments, using this forest for outdoor online is time consuming and costly. Tommaso Cavallari et al. [54] conducted research on online localization problems and proposed adaptive scene coordinate regression (SCoRe) to achieve online localization. The network structure of SCoRe proposed by the author can use the network to predict the three-dimensional points of a scene first, and then obtain the information of the new scene through the predicted information. It can realize online localization in multiple scenes, and can achieve good localization results in both indoor and outdoor environments.

4 Localization Based on 3D Structure

4.1 The Development

The localization based on 3D structure is the most advanced VBL method so far. The main principle of this method based on 3D structure [10, 55, 56, 57] is to first construct 3D point cloud model using SFM algorithm. Then through a series of methods, such as the feature descriptor method, they extracted the two-dimensional features of the images. For example, local features such as SIFT [11] and LIFT [12] are commonly used. Then they established the association between the two-dimensional feature and the three-dimensional model by matching the two-dimensional feature points extracted from the image with the three-dimensional points in the model, then the obtained 2D-3D matching is put into a RANSAC cycle [58], they used the n-point pose solver [59] to estimate the corresponding pose in this RANSAC cycle. In this process, only a large number of correct matches are found in the matching phase, and the pose estimation in the latter stage can be successful. Therefore, this kind of method relies on the extraction and matching of feature points. The feature detector or descriptor can not achieve good results under some conditions, which may make the accuracy of the localization method worse or even cause the localization to fail. For example, some special scenes: scene environment changes Large, scenes with a large number of repetitive structures, weak textures of the scene, no texture structure, strong illumination changes, motion blur, strong viewpoint changes, etc., will make the extraction and matching of feature points more difficult, this method does not achieve the expected results. For the descriptor matching process, the commonly used methods include: priority method [60], efficient matching scheme [10] and geometric outlier filtering [61, 62]. Furthermore, B. Zeisl et al. [63] also considered the use of depth information to eliminate the effects of perspective distortion before extracting the descriptor process.

Irschara et al. [64] first established a point cloud structure based on the SFM method, using the two-dimensional features of the image to match the three-dimensional points, and directly obtaining the corresponding content in the 3D model through the feature index of the vocabulary tree, instead of linking the images database. Based on the above methods, Sattler et al. [65] proposed a vocabulary-based priority search (VPS) method based on the original features of the above methods and the BoF matching method. And, in subsequent work [61] they increased the VPS framework with features of based on structure points. At the same time, Sattler et al. [66] also introduced a visibility chart based on the original method, which rejects the wrong match to improve the positioning accuracy.

In order to improve the speed of 2d-3d matching, Heisterklaus et al. [67] introduced MPEG compression method when processing visual documents, which makes the matching system faster. Donoser et al. [68] trained the random ferns at the top of each point using descriptor redundancy associated with 3D points, and make the speed of 2d-3d matching faster. Feng et al. [69] used a fast point extractor and adopted a binary descriptor in the method, which greatly speeded up the calculation without affecting the accuracy of the pose estimation.

The scene contains many information that helps to strengthen localization, such as semantic information and geometric information. Recently, the idea of integrating these semantic information and geometric information into the visual localization process has become popular. For example, Cohen et al. [70,71] prove that the traditional SFM model is not coherent enough so that cannot reconstruct a complete architectural model. The author uses symmetry and semantic information to infer the model and obtain the possible geometric relationships between the models. Then they stitched the inconsistent SFM model and reconstructed the complete architectural model. Yu et al. [72] used object detectors to extract semantic information and used the information for the iterative process of the ICP algorithm, but one limitation is that these semantic feature informations are manually selected.

Another commonly used method is to incorporate semantic information and geometric information into the feature matching phase of the localization. In this stage, semantic and geometric information can be used to detect and match objects [73, 74], and semantic and geometric information can also be used to enhance feature descriptors [75, 7]. The semantic information in the method of enhancing the feature descriptor only serves as an auxiliary function, and is a weak signal. The main need is that the original descriptor should not be too weak. Johannes L. et al. [76] no longer uses geometric semantic information as a weak signal, but used a new generated model learning descriptor to obtain a descriptor based on the generated model. The descriptors are no longer based on the discriminant model. They added all the information needed to get the original scene into the Euclidean space, embed high-level coded 3D geometry and semantic information to make it a powerful descriptor. The advantage of use of this geometric and semantic combination of powerful descriptors is that this method can obtain accurate localization even in scenes with strong viewpoint changes and changing scenes. Toft C et al. [77] used the semantic information of images and scenes to evaluate each 2d-3d match, and then selected the matching pairs with higher evaluation scores to improve the localization performance. The main idea is to give a semantic label to the 3D point in the SFM model, and at the same time, the query image is semantically segmented, and the feature points of query image are also semantically tagged. In the 2d-3d matching

process, they calculated the score of the matching pair based on the similarity of the labels of the two features of the matching pair. The higher the score, the greater the probability that this match will be correct.

The active search (AS) method proposed by T. Sattler et al. [60] is based on the method of priority matching. This method is relatively efficient. Marcel Geppert et al. [78] proposed a variant of active search based on the above-mentioned active search method, and designed a priority function, so that the features and descriptors that are more likely to generate correct 2D-3D matches can be prioritized. After a predetermined number of sufficient matching correspondences are found, the matching process of 2d and 3d is ended, and the posture estimation is started using the RANSAC cycle.

For more complex scenes such as large urban environments, 2D-3D matching is usually not so unique due to the repetitive structure, which makes the matching phase easy to produce more mismatches. Therefore, Sattler et al. [66] used the co-visibility information between the three-dimensional points to remove a portion of the incorrect matching pairs after the matching pairs were obtained. Similarly, Li et al. [10] also used co-visibility information, combined with RANSAC's sampling strategy, to enable the removal of matches that are unlikely to be correct. Svarm et al. [62] added the step of outlier filtering, which was used to filter out those values that were significantly abnormal. The City-Scale Localization (CSL) [62] method proposed by Svarm for large urban environments has a good effect on stretchability. In addition, Li et al. [79] considered using BRISK descriptors and used graph matching methods to help distinguish similar key points to reduce mismatches. There are also some hybrid methods. For example, Mur-Artal [7] uses word bag recognition to get some possible candidates, and then integrates into the ORB-SLAM system. Then uses the PnP pose solver to estimate the pose, and uses RANSAC to cyclically screening the pose.

Eric Brachmann et al. [80] proposed a new completely differentiable localization pipeline. There is only one component that can be learned, in which soft inlier count is used for hypothesis scoring, and entropy control method is used to automatically adjust the score size. The method can automatically discover the geometric information of the three-dimensional scene during the localization process, and then use the information to optimize the scene coordinate re-projection error.

4.2 The Challenges

The biggest challenge of the localization method based on 3D structure is that it is difficult to adapt to scenes with large changes in the appearance of the environment, a large number of repeated structures, weak textures of the scene, no texture structure, strong illumination changes, motion blur, and strong viewpoint changes. In these scenes, the localization effect is greatly reduced.

For such problems, Johannes L. et al. [76] incorporated geometric semantic information into descriptors, and obtain a descriptor based on the generated model, which was added to the Euclidean space. All the information needed for the scene was embedded high-level encoded 3D geometry and semantic information, making it a powerful descriptor. The use of this geometric and semantic combination of powerful descriptors allows accurate localization even in scenes with strong viewpoint changes and changing scenes. Toft C et al. [77] used the semantic information of images and scenes to evaluate each 2d-3d match, In the 2d-3d matching process, the scores of the matches are scored according to the similarity between the two labels. The higher the score, the greater the probability that the matching is correct. The method can still obtain better results when the appearance of the scene changes greatly.

Uzair Nadeem et al. [81] proposed that a feature descriptor extracted from a query image and a feature descriptor in a three-dimensional point cloud obtained by other methods can be directly matched without using the SFM model to obtain a corresponding pose. The author mainly considers that with the development of 3D scanning technology, it is now possible to directly obtain the 3D point cloud model of the corresponding scene by directly using some 3D scanning software such as Microsoft Kinect, LIDAR or Faro 3D scanner or Matterport scanner. Moreover, the model of this method is better than the 3D model obtained by the SFM method. However, many existing methods are based on the SFM model, and it is not possible to directly use the three-dimensional point cloud obtained by the scanner. Therefore, the author proposes a new structure that does not utilize the SFM model to match the two-dimensional features and the three-dimensional features. Firstly, use a 2D-3D feature matching dataset to train a new matching classifier: Descriptor-Matcher. Then use this Descriptor-Matcher device to match the two-dimensional feature descriptor obtained from the query image with the three-dimensional feature descriptor in the three-dimensional point cloud model of the scene obtained from the scanner, and then use the obtained 2D-3D matching information and pose estimation algorithm to obtain the corresponding camera pose estimate. This approach breaks some SFM-based limitations, such as requiring images with sufficient texture and no excessive repetitive structures. Moreover, the method is more simple to localization, and since it can generate point cloud models of various environments by various scanners, the method can obtain good localization results indoors, outdoors, and in some complicated and challenging environments.

5 Method Comparison

This paper will analysis and compare these three categories of methods in detail from various aspects.

5.1 The Qualitative Comparison

Firstly, in order to deepen the understanding of these methods, this paper makes a qualitative comparison of these three methods, mainly from the accuracy of localization, the robustness of localization (mainly for scene changes), and the required resources (image data and computing resources, etc.), the cost (localization cost, relocation cost, etc.), time-consuming, online, scene size, application range, etc. We make a detailed comparison

of these methods. And, the results of the comparison are shown in Table 1 below.

As shown in the table1, we can obtain:

1. Accuracy of localization: In these three methods, the method based on 3D structure as a whole has the highest precision. The principle of image-based method is image retrieval. It is based on the pose of the most similar image retrieved. The retrieved image is not necessarily very close to the query image. So, this method in principle determine the localization accuracy is the worst. Localization based on Learning-based methods such as the CNN network have recently received widespread attention, but the overall accuracy is still not as good as the method based on 3D structure. (However, there are already CNN-based methods that achieve the best localization accuracy on some small indoor datasets, such as the VLocNet++ network [50])
2. The robustness: The comparison of location robustness in this paper is mainly for the robust adaptability of the scene environment. In the localization process, the adaptability to the change of the scene environment is also a very important factor. The method based on 3D structure is the worst in terms of environmental change robustness. In the scenes where the appearance of the environment changes greatly, the localization accuracy based on 3D structure method will be seriously degraded. Compared with the other two methods, the localization based on learning model has the strongest adaptability to the change of the scene environment, and the robustness is the best. The image-based method is relatively inferior.
3. The required resources: In the localization process, the required localization resources include image data required for localization and computing resources such as powerful GPU. The CNN network requires a large amount of training data when training the model, At the same time, because the training data is too large, CNN often needs a powerful GPU to support it. Therefore, the resource required by this method is the largest, and the method based on 3D structure requires a certain image data to construct a three-dimensional model, Therefore, the required resources of the method are second, and the image-based method only needs an image database for image retrieval, so the method requires minimal resources;
4. The cost: The main focus of this paper is the cost in the localization process and the cost of relocalization. The cost required for CNN network localization process is large, and most of the methods require offline training scene model. The cost of retraining during localization is very expensive, so the cost of the method is the largest, and the cost of the method based on 3D structure is relatively low. The image-based method is the least costly;
5. The time-consuming: The method based on 3D structure takes a long time to build a 3D point cloud model, so it takes the longest time. Although the CNN network needs a lot of data to train, it usually has powerful GPU support, so their time-consuming is generally the least, and the image-based method is second;
6. Online: Most of the methods based on learning model follow the setting of the offline training scene model, so it is difficult or even impossible to locate online. However, some people have proposed a CNN-based online localization method for this problem [36, 55], which has improved this problem to some extent. Relatively speaking, the method based on 3D structure has the best effect.
7. The scenes scale: The scale of the scene applicable to the CNN network is small. The localization of the CNN network on a small scale such as the 7 scenes dataset can be highly accurate, but it is difficult to generalize to scenes other than the training data, which makes the application scope of the method is limited. The method based on 3D structure can obtain good localization results in small-scale scenes and large scene scales. It has strong generalization ability and the widest application range.

	The accuracy	The robustness	The required resources	The cost	The time-consuming	Online	The scene size	The application range
The image-based localization	C	B	C	C	B	B	B	B
Localization based on learning model	B	A	A	A	A	C	C	C
Localization based on 3D structure	A	C	B	B	C	A	A	A

Table 1. This paper mainly compares these three methods from the eight aspects shown in the table, mainly using hierarchical ordering to distinguish different methods: A indicates that the method is the best in this respect; B indicates that the method is second; C indicates that this method is the worst in this respect. For example, the method based on 3D structure is A in terms of precision and is C in terms of robustness, indicating that this method is the best among the three methods in terms of accuracy, and is the worst among the three in terms of robustness.

5.2 The Quantitative Comparison

In Section 5.1, the qualitative comparison of the three methods is introduced in more detail. Next, in order to understand the differences between the various methods in more detail, this paper will focus on the localization accuracy and the accuracy of the scene change and make a detailed quantitative comparison.

5.2.*1* The accuracy of localization.

For all kinds of methods, the most important indicator is the accuracy of localization. Therefore, this paper first compares the accuracy of localization. In this paper, 13 methods are selected to compare the localization accuracy. The selected standard is the most classic methods and the latest methods for selecting various methods. The image-based method includes DenseVLAD[17]; the methods based on learning model includes: PoseNet[40], Bayesian

	The image-based method	The methods based on learning model							The methods based on 3D structure				
	DenseVLAD	PoseNet	Bayesian PoesNet	LSTM-pose	Geom.Loss-Net	VLocNet	VLocNet++	GAN-CNN	Active Search	2D-3D Directly matching	DSAC	InLoc	DSAC++
7 Scenes													
Chess	0.21m,12.5°	0.32m,8.12°	0.37m,7.24°	0.24m,5.8°	0.13m,4.5°	0.036m,1.71°	0.023m,1.44°	0.12m,4.8°	0.04m,1.96°		0.02m,1.2°	0.03m,1.05°	0.02m,0.5°
Fire	0.33m,13.8°	0.47m,14.4°	0.43m,13.7°	0.34m,11.9°	0.27m,11.3°	0.04m,5.34°	0.018m,1.39°	0.27m,11.6°	0.03m,1.5°	0.02m,1.56°	0.04m,1.5°	0.03m,1.07°	0.02m,0.9°
Head	0.15m,14.9°	0.29m,12°	0.31m,12°	0.21m,13.7°	0.17m,13°	0.046m,6.64°	0.016m,0.99°	0.16m,12.4°	0.02m,1.45°	0.01m,1.82°	0.03m,2.7°	0.02m,1.16°	0.01m,0.8°
Office	0.28m,11.2°	0.48m,7.68°	0.48m,8.04°	0.3m,8.1°	0.19m,5.6°	0.039m,1.95°	0.024m,1.14°	0.19m,6.8°	0.09m,3.6°		0.04m,1.6°	0.03m,1.05°	0.03m,0.7°
Pumpkin	0.31m,11.3°	0.47m,8.42°	0.61m,7.08°	0.33m,7°	0.26m,4.8°	0.037m,2.28°	0.024m,1.45°	0.21m,5.2°	0.08m,3.1°	0.01m,0.66°	0.05m,2°	0.05m,1.55°	0.04m,1.1°
Kitchen	0.3m,12.3°	0.59m,8.64°	0.58m,7.54°	0.37m,8.8°	0.23m,5.4°	0.039m,2.2°	0.025m,2.27°	0.25m,6°	0.07m,3.37°		0.05m,2°	0.04m,1.31°	0.04m,1.1°
Stairs	0.25m,15.8°	0.47m,13.8°	0.48m,13.1°	0.4m,13.7°	0.35m,12.4°	0.097m,6.48°	0.021m,1.08°	0.28m,8.4°	0.03m,2.2°		1.17m,33.1°	0.09m,2.47°	0.09m,2.6°
Cambridge													
Shop Facade	1.11m,7.61°	1.46m,8.08°	1.25m,7.54°	1.18m,7.4°	1.05m,4°	0.59m,3.53°			0.12m,0.4°	0.23m,1.6°	0.09m,0.4°		0.06m,0.3°
Old Hospital	4.01m,7.13°	2.31m,5.38°	2.57m,5.14°	1.51m,4.3°	2.17m,2.9°	1.07m,2.41°			0.44m,1°	0.65m,1.14°	0.33m,0.6°		0.2m,0.3°
K. College	2.8m,5.72°	1.92m,5.4°	1.74m,4.06°	0.99m,3.65°	0.99m,1.1°				0.42m,0.55°		0.3m,0.5°		0.18m,0.3°
St M. Church	2.31m,8°	2.65m,8.48°	2.11m,8.38°	1.52m,6.7°	1.49m,3.4°				0.19m,0.5°		0.55m,1.6°		0.13m,0.4°
Street	5.16m,23.5°				20.7m,25.7°				0.85m,0.8°				

Table 2. This paper mainly compares 13 localization methods, which are tested on the indoor small-scale dataset 7sences and the outdoor large-scale dataset Cambridge. For example, the test result of DenseVLAD on Chess is (0.21m, 12.5°), which means that the position error of the method on Chess is 0.21m and the direction error of the method on Chess is 12.5°.

PoesNet[48], LSTM-pose[43], Geom.Loss-Net[42], VLocNet[49], VLocNet++ [50]and CNN method GAN-CNN [52]constructed based on GAN-based ideas; The methods based on 3D structure are: Active Search[60], 2D-3D directly matching method[81], DSAC[79], InLoc[61], DSAC++ [80]method. The comparison results of the 13 methods are shown in Table 2.

It can be seen from the table that in terms of localization accuracy, the method based on 3D structure is better than the method based on learning model, and the method based on learning model is better than the image-based method.

5.2.*2* The accuracy of localization under scene changes.

Another important indicator for VBL is the environmental robustness, that is, the ability of the VBL to adapt to the environment when the scene environment changes. This paper mainly compares the localization accuracy of several methods under the change of scene environment. The criterion for selecting the VBL method is to select a method that has certain adaptability to scene changes. The localization results of these methods under scene changes are shown in Table 3 below.

This paper is tested on three challenging datasets, the Aachen Day-Night, RobotCar Seasons, and CMU Seasons datasets. And, The image-based method includes :DenseVLAD[17],NetVLAD[6]; the methods based on learning model includes: NV+SP[51], HF-Net [51] ; The methods based on 3D structure are: Active Search(AS)[60], City Scale Localization (CSL)[62], Semantic Match Consistency (SMC) [77]. It can be seen from the table 3 that the localization based on learning model has better environmental robustness when the appearance of the scene changes greatly.

6 Conclusion

This paper mainly divides VBL into three types of methods, and makes a detailed summary, evaluates the advantages and disadvantages of various methods and applicable scenes, and makes detailed comparisons between these methods in qualitative and quantitative aspects. The image-based localization is not competitive with the other two methods in localization accuracy, but the method is simple, fast and easy to implement, and as shown in Table 3, when the scene environment changes very largely, the localization effect on the coarse precision is competitive, which can be used for localization with low precision requirements. It can be used for detection and recognition. For this kind of method, future research can consider the accuracy of image retrieval. The overall accuracy of the localization based on learning model is not as good as that based on 3D structure, but

CNN network has a good development prospect. The CNN methods are also gradually surpassing the method based on 3D structure in accuracy. It is also superior to the method based on 3D

		Aachen		RobotCar				CMU	
		day	night	dusk	sun	night	night-rain	urban	suburban
		0.25/0.5/5.0 2/5/10	0.5/1.0/5.0 2/5/10	0.25/0.5/5.0 2/5/10	0.25/0.5/5.0 2/5/10	0.25/0.5/5.0 2/5/10	0.25/0.5/5.0 2/5/10	0.25/0.5/5.0 2/5/10	0.25/0.5/5.0 2/5/10
The image-based method	**DenseVLAD**	0.0 / 0.1 / 22.8	0.0 / 2.0 / 14.3	10.2 / 38.8 / 94.2	5.7 / 16.3 / 80.2	0.9 / 3.4 / 19.9	1.1 / 5.5 / 25.5	22.2 / 48.7 / 92.8	9.9 / 26.6 / 85.2
	NetVLAD	0.0 / 0.2 / 18.9	0.0 / 2.0 / 12.2	7.4 / 29.7 / 92.9	5.7 / 16.5 / 86.7	0.2 / 1.8 / 15.5	0.5 / 2.7 / 16.4	17.4 / 40.3 / 93.2	7.7 / 21.0 / 80.5
The methods based on 3D structure	**AS**	57.3 / 83.7 / 96.6	19.4 / 30.6 / 43.9	44.7 / 74.6 / 95.9	25.0 / 46.5 / 69.1	0.5 / 1.1 / 3.4	1.4 / 3.0 / 5.2	55.2 / 60.3 / 65.1	20.7 / 25.9 / 29.9
	CSL	52.3 / 80.0 / 94.3	24.5 / 33.7 / 49.0	56.6 / 82.7 / 95.9	28.0 / 47.0 / 70.4	0.2 / 0.9 / 5.3	0.9 / 4.3 / 9.1	36.7 / 42.0 / 53.1	8.6 / 11.7 / 21.1
	SMC			53.8 / 83.0 / 97.7	46.7 / 74.6 / 95.9	6.2 / 18.5 / 44.3	8.0 / 26.4 / 46.4	75.0 / 82.1 / 87.8	44.0 / 53.6 / 63.7
The methods based on learning model	**NV+SIFT**	82.8 / 88.1 / 93.1	30.6 / 43.9 / 58.2	55.6 / 83.5 / 95.3	46.3 / 67.4 / 90.9	4.1 / 9.1 / 24.4	2.3 / 10.2 / 20.5	63.9 / 71.9 / 92.8	28.7 / 39.0 / 82.1
	NV+SP	79.7 / 88.0 / 93.7	40.8 / 56.1 / 74.5	54.8 / 83.0 / 96.2	51.7 / 73.9 / 92.4	6.6 / 17.1 / 32.2	5.2 / 17.0 / 26.6	91.7 / 94.6 / 97.7	74.6 / 81.6 / 91.4
	HF-Net	75.7 / 84.3 / 90.9	40.8 / 55.1 / 72.4	53.9 / 81.5 / 94.2	48.5 / 69.1 / 85.7	2.7 / 6.6 / 15.8	4.7 / 16.8 / 21.8	90.4 / 93.1 / 96.1	71.8 / 78.2 / 87.1

Table 3. Evaluation of the localization on the Aachen Day-Night, RobotCar Seasons, and CMU Seasons datasets. We report the recall [%] at different distance and orientation thresholds.

structure in terms of robustness to the environment. In future research, we can consider how to achieve better online localization, how to improve the accuracy and reduce the cost; the localization based on the 3D structure has advantages in overall precision, but the robustness to the environment is poor. In the future, it can be considered how to make the localization method can obtain high precision when the scene changes greatly. At the same time, in the process of summarizing, it is found that combining the CNN method with the three-dimensional structure method can be a very good study directions, such as the DSAC++ [80]method, a hybrid method that combines CNN and three-dimensional structure can achieve better localization results. In the following research, we can consider combining the two methods to obtain higher precision while maintaining the advantages of CNN and three-dimensional structure.

ACKNOWLEDGMENTS

Thanks to the guidance of the instructor and the help of the schoolmate, and thanks to the National Natural Science Foundation of China under Project 61873274 for funding.

REFERENCES

[1] Piasco, N., Sidibé, D., Demonceaux, C. and Gouet-Brunet, V, 2018. A survey on visual-based localization: On the benefit of heterogeneous data. Pattern Recognition, 74, 90-109.

[2] Schonberger, J. L., and Frahm, J. M. 2016. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4104-4113).

[3] Lynen, S., Sattler, T., Bosse, M., Hesch, J. A., Pollefeys, M., and Siegwart, R. 2015. Get Out of My Lab: Large-scale, Real-Time Visual-Inertial Localization. In Robotics: Science and Systems (Vol. 1).

[4] Schreiber, M., Knöppel, C., and Franke, U. 2013. Laneloc: Lane marking based localization using highly accurate maps. In 2013 IEEE Intelligent Vehicles Symposium (IV) (pp. 449-454).

[5] Dubé, R., Dugas, D., Stumm, E., Nieto, J., Siegwart, R., and Cadena, C. 2016. Segmatch: Segment based loop-closure for 3d point clouds. In ICRA, arXiv preprint arXiv:1609.07720.

[6] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5297-5307).

[7] Mur-Artal, R., Montiel, J. M. M., and Tardos, J. D. 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE transactions on robotics, 31(5), 1147-1163.

[8] Chen, Z., Jacobson, A., Sünderhauf, N., and Milford, M. 2017. Deep learning features at scale for visual place recognition. In 2017 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3223-3230).

[9] Sattler, T., Maddern, W., Toft, C., Torii, A. and Kahl, F. 2018. Benchmarking 6dof outdoor visual localization in changing conditions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8601-8610).

[10] Li, Y., Snavely, N., Huttenlocher, D., and Fua, P. 2012. Worldwide pose estimation using 3d point clouds. In European conference on computer vision (pp. 15-29).

[11] Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision, 60(2), 91-110.

[12] Yi, K. M., Trulls, E., Lepetit, V., and Fua, P. 2016. Lift: Learned invariant feature transform. In European Conference on Computer Vision (pp. 467-483).

[13] Brejcha, J., and Čadík, M. 2017. State-of-the-art in visual geo-localization. Pattern Analysis and Applications, 20(3), 613-637.

[14] Hakeem, A., Zamir, L., Van Gool, Shah, M., and Szeliski, R. 2016. Large-scale visual geo-localization. Cham: Springer.

[15] Arandjelović, R., and Zisserman, A. 2012. Three things everyone should know to improve object retrieval. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 2911-2918).

[16] Jégou, H., Douze, M., Schmid, C., and Pérez, P. 2010. Aggregating local descriptors into a compact image representation. In CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition (pp. 3304-3311).

[17] Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., and Pajdla, T. 2015. 24/7 place recognition by view synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1808-1817).

[18] Kim, H. J., Dunn, and E., Frahm, J.-M., 2015. Predicting good features for image geo-localization using per-bundle VLAD. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Vol. 11-18-Dece. pp. 1170–1178.

[19] Jégou, H., and Zisserman, A. 2014. Triangulation embedding and democratic aggregation for image search. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3310-3317).

[20] Cummins, M., and Newman, P. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. IJRR, 27(6):647–665, 2008.

[21] Azzi, C., Asmar, D., Fakih, A., and Zelek, J., 2016. Filtering 3D Keypoints Using GIST For Accurate Image-Based Localization. In: British Machine Vision Conference (BMVC). No. 2. pp. 1–12.

[22] Radenoviˊc, F., Tolias, G., and Chum, O., 2016. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). Vol. 9905. pp. 3–20

[23] Sivic, J., and Zisserman, A. 2003. Video Google: A text retrieval approach to object matching in videos. In null (pp. 1470-1477).

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

[25] J´egou, H., Perronnin, F., Douze, M., Sanchez, J., and Schmid, C., 2012. Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 1–12

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

[27] Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., and Milford, M. 2015. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. Proceedings of Robotics: Science and Systems XII.

[28] Cao, S., and Snavely, N., 2015. Graph-Based Discriminative Learning for Location Recognition. International Journal of Computer Vision (IJCV) 112 (2), 239–254

[29] Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., and Fitzgibbon, A., 2013. Scene coordinate regression forests for camera relocalization in RGB-D images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2930–2937

[30] Guzman-Rivera, A., Pushmeet, K., Glocker, B., and Shotton, J., 2014. Multi-Output Learning for Camera Relocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–6.

[31] Valentin, J., Fitzgibbon, A., Shotton, J., and Torr, P. H. S., 2015. Exploiting Uncertainty in Regression Forests for Accurate Camera Relocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4400–4408.

[32] Meng, L., Chen, J., Tung, F., Little, J., and de Silva, C. W., 2016. Exploiting Random RGB and Sparse Features for Camera Pose Estimation. In: British Machine Vision Conference (BMVC). pp. 1–12.

[33] Brachmann, E., Michel, F., Krull, A., Ying Yang, M., and Gumhold, S. 2016. Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3364-3372).

[34] Massiceti, D., Krull, A., Brachmann, E., Rother, C., and Torr, P. H. 2017. Random forests versus Neural Networks—What's best for camera localization? In 2017 IEEE International Conference on Robotics and Automation (ICRA) (pp. 5118-5125).

[35] Glocker, B., Shotton, J., Criminisi, A., and Izadi, S. 2014. Real-time RGB-D camera relocalization via randomized ferns for keyframe encoding. IEEE transactions on visualization and computer graphics, 21(5), 571-583.

[36] Cavallari, T., Golodetz, S., Lord, N. A., and Valentin, J., 2017. On-the-Fly Adaptation of Regression Forests for Online Camera Relocalisation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[37] Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[38] He, K., Zhang, X., Ren, S., and Sun, J. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[39] Girshick, R. 2015. Fast R-CNN. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[40] Kendall, A., and Cipolla, R. 2016. Modelling uncertainty in deep learning for camera relocalization. In 2016 IEEE international conference on Robotics and Automation (ICRA) (pp. 4762-4769).

[41] Kendall, A., Grimes, M., and Cipolla, R. 2015. Convolutional networks for real-time 6-DOF camera relocalization. CoRR abs/1505.07427.

[42] Kendall, A., and Cipolla, R. 2017. Geometric loss functions for camera pose regression with deep learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5974-5983).

[43] Walch, F., Hazirbas, C., Leal-Taixé, L., Sattler, T., and Cremers, D. 2016. Image-based localization with spatial lstms. arXiv preprint arXiv:1611.07890, 2(6).

[44] Walch, F., 2016. Deep Learning for Image-Based Localization. Ph.D. thesis, Technical University of Munich.

[45] Liu, Z., Duan, L.-Y., Chen, J., and Huang, T., 2016. Depth-Based Local Feature Selection for Mobile Visual Search. In: Proceedings of the IEEE International Conference on Image Processing (ICIP).

[46] Jia, D., Su, Y., and Li, C., 2016. Deep Convolutional Neural Network for 6-DOF Image Localization. arXiv preprint (413113), 1790–1798.

[47] Contreras, L., and Mayol-Cuevas, W. 2017. Towards CNN Map Compression for camera relocalisation. arXiv preprint arXiv:1703.00845.

[48] Weyand, T., Kostrikov, I., and Philbin, J., 2016. PlaNet - Photo Geolocation with Convolutional Neural Networks. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV). Vol. 9905. pp. 37–55.

[49] Valada, A., Radwan, N., and Burgard, W. 2018, May. Deep auxiliary learning for visual localization and odometry. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 6939-6946).

[50] Radwan, N., Valada, A., and Burgard, W. 2018. Vlocnet++: Deep multitask learning for semantic visual localization and odometry. IEEE Robotics and Automation Letters, 3(4), 4407-4414.

[51] Sarlin, P. E., Cadena, C., Siegwart, R., and Dymczyk, M. 2019. From coarse to fine: Robust hierarchical localization at large scale. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 12716-12725).

[52] Bui, M., Baur, C., Navab, N., Ilic, S., and Albarqouni, S. 2019. Adversarial Joint Image and Pose Distribution Learning for Camera Pose Regression and Refinement. arXiv preprint arXiv:1903.06646.

[53] Sattler, T., Zhou, Q., Pollefeys, M., and Leal-Taixe, L. 2019. Understanding the Limitations of CNN-based Absolute Camera Pose Regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3302-3312).

[54] Cavallari, T., Bertinetto, L., Mukhoti, J., Torr, P., and Golodetz, S. 2019. Let's Take This Online: Adapting Scene Coordinate Regression Network Predictions for Online RGB-D Camera Relocalisation. arXiv preprint arXiv:1906.08744.

[55] Zeisl, B., Sattler, T., and Pollefeys, M. 2015. Camera Pose Voting for Large-Scale ImageBased Localization. In Proceedings of the IEEE International Conference on Computer Vision.

[56] Liu, L., Li, H., and Dai, Y. 2017. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2372-2381).

[57] Sattler, T., Torii, A., Sivic, J., Pollefeys, M., and Taira, H. 2017. Are large-scale 3d models really necessary for accurate visual localization? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1637-1646).

[58] Fischler, M. A., and Bolles, R. C. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24(6), 381-395.

[59] Kneip, L., Scaramuzza, D., and Siegwart, R. 2011. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In CVPR 2011 (pp. 2969-2976).

[60] Sattler, T., Leibe, B., and Kobbelt, L. 2016. Efficient & effective prioritized matching for large-scale image-based localization. IEEE transactions on pattern analysis and machine intelligence, 39(9), 1744-1756.

[61] Camposeco, F., Sattler, T., Cohen, A., and Pollefeys, M. 2017. Toroidal constraints for two-point localization under high outlier ratios. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4545-4553).

[62] Svärm, L., Enqvist, O., and Oskarsson, M. 2016. City-scale localization for cameras with known vertical direction. IEEE transactions on pattern analysis and machine intelligence, 39(7), 1455-1461.

[63] Zeisl, B., Koser, K., and Pollefeys, M. 2013. Automatic registration of RGB-D scans via salient directions. In Proceedings of the IEEE international conference on computer vision (pp. 2808-2815).

[64] Irschara, A., Zach, C., Frahm, J.-m., and Bischof, H., 2009. From Structure-from-Motion Point Clouds to Fast Location Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[65] Sattler, T., Leibe, B., and Kobbelt, L., 2011. Fast image-based localization using direct 2D-to-3D matching. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 667–674.

[66] Sattler, T., Havlena, M., Radenovi´c, F., Schindler, K., and Pollefeys, M., 2015. Hyperpoints and fine vocabularies for large-scale location recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). Vol. 11-18-Dece. pp. 2102–2106.

[67] Heisterklaus, I., Qian, N., and Miller, A., 2014. Image-based pose estimation using a compact 3D model. In: IEEE International Conference on Consumer Electronics Berlin (ICCE-Berlin). pp. 327–330.

[68] Donoser, M., and Schmalstieg, D., 2014. Discriminative feature-to-point matching in image-based localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 516–523.

[69] Feng, Y., Fan, L., and Wu, Y., 2016. Fast Localization in Large-Scale Environments Using Supervised Indexing of Binary Features. IEEE Transactions on Image Processing (ToIP) 25 (1), 343–358.

[70] Cohen, A., Sattler, T., and Pollefeys, M. 2015. Merging the unmatchable: Stitching visually disconnected sfm models. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2129-2137).

[71] Cohen, A., Schönberger, J. L., Speciale, P., Sattler, T., Frahm, J. M., and Pollefeys, M. 2016. Indoor-outdoor 3d reconstruction alignment. In European Conference on Computer Vision (pp. 285-300).

[72] Yu, F., Xiao, J., and Funkhouser, T. 2015. Semantic alignment of LiDAR data at city scale. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1722-1731).

[73] Atanasov, N., Zhu, M., Daniilidis, K., and Pappas, G. J. 2016. Localization from semantic observations via the matrix permanent. The International Journal of Robotics Research, 35(1-3), 73-99.

[74] Schonberger, J. L., Hardmeier, H., Sattler, T., and Pollefeys, M. 2017. Comparative evaluation of hand-crafted and learned local features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1482-1491).
[75] Hornung, A., Wurm, K. M., Bennewitz, M., Stachniss, C., and Burgard, W. 2013. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. Autonomous robots, 34(3), 189-206.
[76] Schönberger, J. L., Pollefeys, M., Geiger, A., and Sattler, T. 2018. Semantic visual localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6896-6906).
[77] Toft, C., Stenborg, E., Hammarstrand, L., Brynte, L., Pollefeys, M., Sattler, T., and Kahl, F. 2018. Semantic match consistency for long-term visual localization. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 383-399).
[78] Geppert, M., Liu, P., Cui, Z., Pollefeys, M., and Sattler, T. 2018. Efficient 2D-3D Matching for Multi-Camera Visual Localization. arXiv preprint arXiv:1809.06445.
[79] Li, S., and Calway, A. 2015. RGBD relocalisation using pairwise geometry and concise key point sets. In 2015 IEEE International Conference on Robotics and Automation (ICRA) (pp. 6374-6379).
[80] Brachmann, E., and Rother, C. 2018. Learning less is more-6d camera localization via 3d surface regression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4654-4662).
[81] Nadeem, U., Jalwana, M. A., Bennamoun, M., Togneri, R., and Sohel, F. 2019. Direct Image to Point Cloud Descriptors Matching for 6-DOF Camera Localization in Dense 3D Point Cloud. arXiv preprint arXiv:1906.06064.