2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Retrieval and Localization with Observation Constraints

Yuhao Zhou†, Huanhuan Fan†, Shuang Gao†, Yuchen Yang, Xudong Zhang*, Jijunnan Li, Yandong Guo

***Abstract*— Accurate visual re-localization is very critical to many artificial intelligence applications, such as augmented reality, virtual reality, robotics and autonomous driving. To accomplish this task, we propose an integrated visual re-localization method called RLOCS by combining image retrieval, semantic consistency and geometry verification to achieve accurate estimations. The localization pipeline is designed as a coarse-to-fine paradigm. In the retrieval part, we cascade the architecture of ResNet101-GeM-ArcFace and employ DBSCAN followed by spatial verification to obtain a better initial coarse pose. We design a module called observation constraints, which combines geometry information and semantic consistency for filtering outliers. Comprehensive experiments are conducted on open datasets, including retrieval on R-Oxford5k and R-Paris6k, semantic segmentation on Cityscapes, localization on Aachen Day-Night and InLoc. By creatively modifying separate modules in the total pipeline, our method achieves many performance improvements on the challenging localization benchmarks.**

I. INTRODUCTION

Visual Localization serves as the fundamental capability of numerous vision applications, including augmented reality, intelligent robotics and autonomous driving navigation [1, 2]. This approach's core task is to estimate the 6-degrees of freedom (DoF), i.e., the position and orientation of a query RGB image in a known 3-Dimensional (3D) coordinate environment.

The presentation of the environment can be a map reconstructed by Structure From Motion (SFM) [3]–[5], a database of images [6, 7], or even regression Convolutional Neural Network (CNN) [8]. In detail, the SFM based map is typically used to describe the position of landmarks [9, 10], i.e., 3D points and structures in the environments, which are pre-collected and extracted from the database images. During localization, correspondences between 2D keypoints and 3D landmarks are established to recover the query image's 6-DoF pose using Perspective-n-Point (PnP) [11] within a RANSAC loop [12, 13]. To avoid costly timing on searching and matching in irrelevant mapping areas, image retrieval is used to select the most relevant database images [7, 14]. Local feature matching is then established between the query image and the area defined by retrieved database images.

Since the correspondences between query and database images need to be established in visual localization tasks, environment changes, such as weather, illumination, or seasonal changes, present critical challenges for local feature descriptors.

† Authors contributed equally to this work.
∗ indicates corresponding author. Contact: zhangxudong@oppo.com
All authors are with OPPO Research Institute, Shanghai, China.

The traditional local features and descriptors, e.g., SIFT [15], BRIEF [16], or ORB [17], which have been carefully designed for uniform intensity changes and slight variations of viewpoints, were shown to be highly sensitive to massive changes in lighting and seasonal conditions. An attempt at overcoming conditions-changing is training convolutional neural networks (CNNs) to produce more robust feature descriptors [18]–[21], instead of using handcraft features. Although CNNs are shown to have great improvements compared to SIFT and other handcrafted features, they were not designed to handle all the types of variations described above. As feature detectors and descriptors are less repeatable and reliable, localization pipelines then struggle to find enough query-to-database correspondings to recover successful pose estimation. Therefore, developing more robust localization pipelines that work well across a wider range of environmental conditions is desirable.

In this paper, we present a localization pipeline, Retrieval and Localization with Observation ConstraintS (RLOCS), which utilizes image retrieval to acquire coarse initial localization poses and combines geometric and semantic information to refine the localization results.

The core idea of RLOCS is to employ a natural coarse-to-fine strategy for recovering 6-DoF poses of query images in the related pre-built SFM model. In detail, RLOCS leverages both global descriptors for image-retrievals and local features for semantic-matching to establish a localization pipeline. We show that RLOCS, using CNN-based image retrieval method and hybrid local descriptors, enables robustness and reliable results under many challenging conditions. Our global descriptors outperform most previous results in the retrieval task, and the learning-based local features improve the accuracy of pose estimation.

Meanwhile, inspired by previous work on pose verification via observation [7, 22], we propose a 6-DoF pose optimization method based on observation constraints of the query image. The pose optimization starts with a standard PnP within a RANSAC loop and obtains an initial pose estimation. The 3D points with locations and descriptors are then collected using the initial pose and observation constraints. Consequently, matches between 2D points from the query image and 3D points from the SFM model are established to refine the initial pose. More details of pose optimization methods are discussed in the following sections.

In summary, our contributions include the following key enhancements on the retrieval-based visual localization pipeline:

1. A better retrieval CNN is proposed, followed by a

clustering and a spatial verification method known as re-rank. We regress the coarse localization initial pose by a 2D-2D matching module.

2. A coarse-to-fine two-stage localization pipeline using observation constraints as the back-end fine-tuning optimization method is utilized, which contains geometry attributes and semantic information.

3. An efficient and accurate semantic segmentation CNN structure is adopted and optimized to achieve better semantic precision, which finally benefits the localization accuracy.

II. Related Work

In this section, we will discuss recent approaches related to different components of our works: large-scale image retrieval tasks, semantic segmentation and visual localization.

Large-scale Image Retrieval. As a classical problem in computer vision, large-scale image retrieval has been widely analyzed in the past decades. For statistically quantifying the performance of different methods, several standard datasets have been published and widely used, e.g., R-Oxford5k, R-Paris6k [23], and Google Landmarks dataset v1, v2 [24, 25]. The diversity of environments and challenges of illumination changes are presented in these large datasets. In the retrieval tasks, approaches can be divided into two categories, local features and global features. As to the local features, some methods aggregate local features into global descriptors, such as VLAD [26] and FV [27], while others form a feature model for searching and query, like BOW [28] and other related methods [29, 30]. For the global features, with the rapid development of deep learning, CNN-based methods outperform most of the hand-crafted methods with higher recall and accuracy. Using these CNN-based global descriptors, [31] makes use of deep local features as a drop-in replacement for hand-crafted features in conventional aggregation such as VLAD. [24] relies on CNN to produce attentive local features and regresses them into global indexes. [32] unifies local and global descriptors into a single CNN with generalized mean pooling and attention selection modules.

Semantic Segmentation. Plenty of researchers have analyzed many strategies and structures on semantic segmentation CNNs in recent years. [33] adopts dilation convolutional layers to achieve larger receptive area, while [34] utilizes large kernel convolutional layers. [35] relies on the spatial attention module to enlarge the structure information, while [36] adaptively integrates local and global features by spatial-wise and channel-wise attention modules. [37] exploits shortcuts between multiple layers to avoid degradation problems, and [38] proves the better performance with the help of shortcut connections between layers in a feed-forward fashion. Unlike previous works, we integrated several modules to lift the semantic segmentation precision and analyzed the positive inference on visual localization brought by the semantic information.

Visual Localization. Among all the image-based localization approaches, structure-based ones are one of the most extensively discussed methods, which extract 2D key-points of a query image and find matching relationships between 3D points constructed by SFM [39]. [40] expands from original 2D-to-3D matches to 3D-to-2D matches to realize a better localization performance. [22] adds semantic information to help filter some outliers of key-points matching. [14] merges local features extractor and global one into a single CNN to realize higher efficiency and robustness. [41] relies on renderings of a 3D model to produce better local features matching and seeks a better final estimated pose iteratively.

III. Visual Re-Localization Pipeline

Our Re-localization method, named RLOCS, consists of four parts: mapping with observation constraints, image retrieval, initial pose estimation and iterative pose optimization. Fig.1 illustrates the pipeline based on a standard retrieval-based framework in [42].

Observation constraints mapping. We run COLMAP [4], a superior SFM algorithm to reconstruct a 3D model and camera poses of the database images captured at the target scene. After triangulation, 3D map points will be produced based on epipolar geometry theory. More semantic and geometric information, regarded as observation measurements, can be calculated and tagged on every 3D point for checking and filtering in the visual localization process.

Image retrieval. The landmark image search is performed by matching the query with the database images using the global descriptor calculated by CNN described detailedly in III-A. Through this retrieval scheme, a fixed number of similar database images are collected. DBSCAN [43] and a re-rank [32] procedure are adopted to fine-tune the retrieval results.

Initial pose estimation. For every candidate, the semantic consistency check is applied to filter out inaccurate key-points matches. The filtered 2D-3D matches will determine the 6-DoF camera pose by solving a PnP geometric consistency check inside a RANSAC loop. This algorithm is called Feature Match(FM)-PnP module in our method.

Pose iterative optimization. Given the coarse 6-DoF pose, more geometric and semantic observation constraints are applied to select the 3D points. After selection, K-Nearest Neighbor (KNN) matching enables us to get the 2D-3D matches. Such process is conducted iteratively to make the final estimated pose closer to the ground truth.

In the total localization pipeline, our contributions can be concluded into three main parts, i.e., image retrieval, semantic segmentation, and observation constraints.

A. Image Retrieval

In RLOCS, a coarse retrieval task is first performed by matching the query with the database images using global descriptors in both SFM mapping and localization pipeline. We leverage recent improvements in global feature designs, such as a Generalized Mean pooling (GeM) [44] layer and ArcFace [45] loss to generate effectively aggregated global descriptors. The strategy of global retrieval is illustrated in Fig. 2.

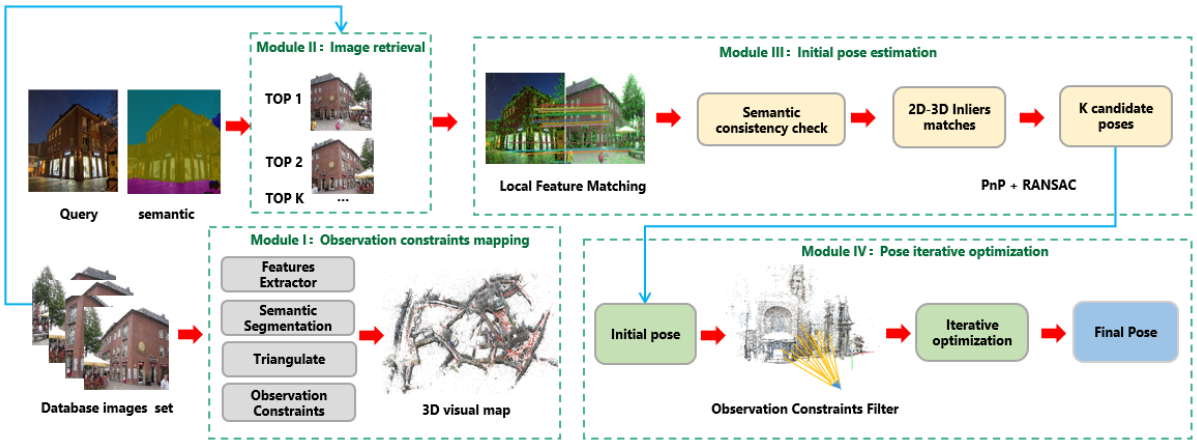

Fig. 1. **Visual Re-localization Pipeline**. The input data contains both the database images set for mapping and a query image for Re-localization. Module I is the construction of observation constraints map, including SuperPoint and R2D2 local feature extraction and matching, semantic segmentation, triangulation of matched image pairs, BA optimization and calculation of observation constraints. The output of Module I is a semantic 3D map with observation constraints. Module II is the image retrieval part, returning K images from database images set, which are most similar to the query image. Module III is to obtain a rough pose. The process includes the 2D-2D matching between query and candidate images, semantic consistency verification, pose clustering and solving PnP using 2D-3D inlier matches. Module IV is the iterative optimization.

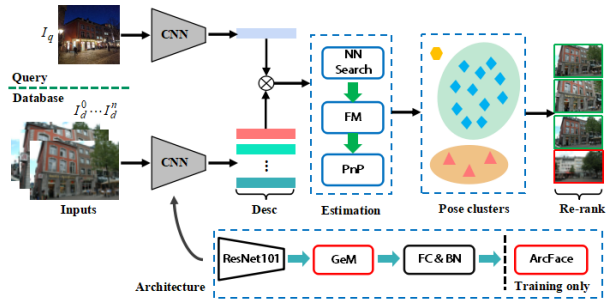

Fig. 2. **Illustration of retrieval strategy in RLOCS.** Firstly, global features are extracted from the cascaded architecture of ResNet101-GeM-ArcFace. These features are then fed into NN search in database images. For every retrieved candidate from the database, an FM-PnP pipeline is performed for solving coarse 6-DoF poses of the query image. And DBSCAN clustering method, followed by the re-rank procedure, is applied for seeking more accurate retrievals results.

Given a query image, a basic backbone CNN is first adopted to obtain the feature map, representing deep activations. GeM pooling is applied to weigh each feature map's contributions and aggregate the activations into a fixed-length global descriptor. In the work of [44], GeM is shown superior performance than other pooling methods, such as regional max-pooling (R-Mac) [46] and sum-pooled convolutional features (SPoC) [47]. The defination of GeM can be described as

$$f_c^g = \left(\frac{1}{|X_c|}\sum_{x\in X_c} x^p\right)^{\frac{1}{p}}, \tag{1}$$

where x is the feature at each location of X_c, which is extracted from the backbone and p denotes the generalized mean power parameter. Note that the p of GeM pooling is set to 3.0 and fixed during our training process [44]. Dimension reduction is then adopted behind the GeM pooling, adding a fully connected layer cascaded with one-dimensional Batch Normalization, which is crucial to alleviate the risk of overfitting and reduce the dimensional noise.

For better performance on global feature learning, we utilize the ArcFace [45] margin-based loss as the training components, which has achieved impressive results by including smaller intra-class variance in face recognition. The ArcFace is defined as

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}, \tag{2}$$

where $x_i \in \Re^d$ denotes the deep features of the i-th sample in d dimensions, belonging to the y_i-th class. W_{y_i} denotes the weights term of y_i-th class. $W_j \in \Re^d$ and $b_j \in \Re^d$ are the j-th column of the weights and the bias term, respectively. In this work, we follow [45] and train our retrieval model with image-level annotations on the Google Landmarks dataset v2 [25]. For evaluation components, we adopt the learned fixed-length global descriptor following by an L2-normalization and Principal Component Analysis (PCA) process for all the query and database images.

Secondly, a KNN search is performed by matching query images with the candidates using our global descriptors. However, each landmark category in the database may contain diverse samples, such as variation of viewpoints and illumination. These query images are tough to identify only using context-level global features. Therefore we employ a back-end discriminative clustering method to exploit the 6-DoF poses from the database. In detail, an FM-PnP strategy is firstly performed for initial 6-DoF poses between query images and the retrieved top-k images from the database. These poses are then clustered based on the inliers and distances using DBSCAN [43], following by a PnP spatial verification algorithm [32].

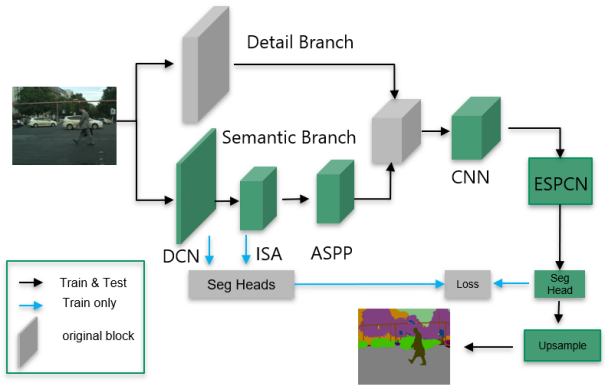

Fig. 3. **Illustration of proposed semantic segmentation architecture.** Feature maps are separated into semantic and detail branches. DCN, ISA and ASPP are applied to acquire a larger receptive area, gaining more global and shape information in the semantic branch. And ESPCN is used to recover stable and uniform semantic masks.

B. Semantic Segmentation

Considering that semantic segmentation is relatively stable under illuminational changes, we take advantage of semantic segmentation algorithms as one of the observation constraints during the mapping and localization process, to improve the pose estimations' accuracy.

For pixel-level semantic segmentation tasks, efficiency and accuracy are both significant. In the work of [48], a state-of-the-art network, BiSeNet-V2, meets both the high-speed and accuracy demands in our localization pipeline.

Two branches, i.e., detail branch and semantic branch, are inherited from [48]. Detail branch is meant to produce low-level features with shallow CNN layers. To enhance the receptive field and capture rich contextual information in the semantic branch, we adopt the Interlaced Sparse Self-Attention (ISA) [49] spatial attention mechanism in the semantic branch to enhance the receptive field. And the additional Atrous Spatial Pyramid Pooling (ASPP) [50] module together with ISA helps to solve the multi-scale problems for objects. We use Deformable Convolutional Networks (DCN) [51] module to modify the semantic branch backbone to make the branch paying more attention to the shapes of different objects. Afterward, the aggregation layer manages to merge both detail and semantic branch into a feature map.

With the help of Efficient Sub-Pixel Convolutional Neural Network (ESPCN), the feature map can be upsampled by a factor of 4. Ending with an additional 3x3 convolution layer, which is cascaded with a bilinear upsampling layer of factor 2, ensures us to get a stable and uniform semantic mask.

C. Observation Constraints and Pose Iterative Optimization

The original 3D point clouds reconstructed by COLMAP [4] only contain the primary attributes, such as position coordinates and color. In this paper, we add more attributes named observation constraints based on the idea of semantic consistency [22], including semantic and geometric constraints. The additional attributes of every 3D point will guarantee the filtering process before 2D-3D matching in the localization part. As to local features, we inherit both Super-Point [19] and R2D2 [20] to extract 2D points and descriptors.

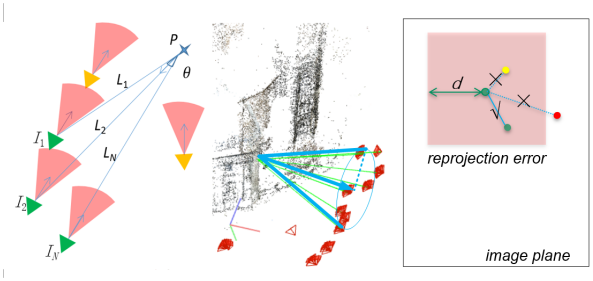

Fig. 4. **Illustration of Observation Constraints**. The left figure is the schematic diagram and N is the number of images associated with the 3D point P. L_i is the distance between point P and the optical center of the camera I_i. The middle figure shows a 3D point visual field which is a cone area in the map. The right one shows how the reprojection error and the semantic label function as observation constraints. d is the error threshold and points' color represents the semantic labels.

The additional information on each point consists of maximum visible distance, mean visible direction, maximum visible angle, semantic label and reprojection error. During the SFM process, we backtrace each 3D point to find the 2D feature points and the corresponding images that participate in the triangulation. Supposing the 3D point P has N track elements, the corresponding images are denoted as $I_1, I_2, ..., I_N$. As the schematic diagram illustrated in Fig.4, the visible field of a 3D point is the cone area. Whether a query image can see the 3D point depends on whether its pose is in the visible field.

The maximum visible distance L, the mean visible direction $\vec{n}$ and the maximum visible angle θ can be formulated as follows:

$$L = \max_i \|X - C_i\|_2 \quad i \in [1...N], \tag{3}$$

$$\vec{n} = \frac{1}{N}\sum_{i=1}^{N} \frac{\overrightarrow{C_i} - \overrightarrow{X}}{\sqrt{\|C_i - X\|^2}}, \tag{4}$$

$$\theta = 2\max_i(arccos(\vec{n}\cdot\frac{\overrightarrow{C_i} - \overrightarrow{X}}{\sqrt{\|C_i - X\|^2}})) \quad i \in [1...N], \tag{5}$$

where X denotes position coordinates of point P, and C_i denotes the camera optical center position of image I_i. $\overrightarrow{X}$ and $\overrightarrow{C_i}$ are the vector representations of X and C_i respectively. The visible field of point P is a cone area that is determined by L, $\vec{n}$ and θ. The L is the cone bus, which means the farthest distance where the point P can be seen. The $\vec{n}$ represents the average value of the direction from point P to the corresponding camera optical center C_i as the normal of P. The θ equals two times the maximum angle between the normal $\vec{n}$ and all of the $\overrightarrow{PI_i}$. We adopt

a voting strategy and determine the 3D semantic label with the largest occurrence frequency of correlated 2D points.

Each initial pose is obtained by 2D-3D inliers and PnP process described as Module III in Fig.1. Then the pose iterative optimization is performed by using observation constraints. The specific steps are listed as follows:

1) Select 3D points with initial pose and the visible field determined by L, $\vec{n}$, θ.
2) Produce global 2D-3D matches by KNN and semantic consistency check.
3) Filter outliers of 2D-3D matches within a certain threshold of re-projection error.
4) Compute the 6-DoF camera pose by solving a PnP algorithm inside a RANSAC loop.
5) Update the pose of the query if the convergence condition is fulfilled.

The convergence condition is determined by the uncertainty quantification of the iterative pose. Based on the Monte Carlo Sampling [52], we first randomly sample $k\%$ (e.g., $k = 30, 50, 70$) sub-matches from all the global 2D-3D matches. The sub-poses are then calculated from sub-matches using the PnP algorithm. The standard deviation between all the sub-poses and current pose is defined as the sampling uncertainty. It is supposed to get smaller during the iteration loops. Otherwise, the optimization stops and the final prediction is produced.

IV. Experimental Evaluation

The performance of whole visual re-localization is discussed in the following part. Experiments are executed on one NVIDIA Tesla V100 with the CUDA 10.0 and Intel(R) Xeon(R) Gold 6142 CPU @ 2.60GHz. On average, RLOCS consumes 198ms per frame.

A. Ablation Study

Image Retrieval. For large-scale image retrieval tasks, we conduct our evaluation of the proposed ResNet101-GeM-ArcFace pipeline on *R-Oxford5k* and *R-Paris6k* datasets. The Google Landmarks dataset v2 [25] is the training set. Table I shows that our method outperforms some of the state-of-art retrieval methods statistically.

TABLE I
RETRIEVAL RESULT(mAP) ON *R-Oxford5k* AND *R-Paris6k* WITH BOTH MEDIUM AND HARD EVALUATION PROTOCOLS.

Methods	*R-Oxford5k*		*R-Paris6k*	
	Medium	Hard	Medium	Hard
NetVLAD [31]	63.5	-	73.5	-
ResNet101-RMAC [53]	60.9	32.4	78.9	59.4
ResNet101-GeM-AP [54]	67.5	42.8	80.1	60.5
DELG [32]	69.7	45.1	81.6	63.4
Ours	**72.5**	**55.9**	**85.8**	**71.8**

While DELG utilizes ResNet50 to extract low-level feature maps, ours relies on larger ResNet101 to acquire more features to get better retrieval accuracy. As GeM pooling layers are meant to maintain more information than original max-pooling layers during the CNN inference, RLOCS using GeM as a pooling layer outperforms the ResNet101-RMAC whose pooling layers are derived from the max-pooling layers. Evidence [45] also indicates that margin-based loss will efficiently escalate the discriminative power between different classes. Thus RLOCS outperforms both [53] and [54] on these test datasets.

Semantic Segmentation. Cityscapes dataset [55] is a widely used semantic segmentation benchmark that contains urban street scenes. The final results on its test sets show that the combination of ASPP, DCN, ESPCN, and ISA modules can a achieve better mean Intersection of Union (mIoU) with reasonable speed. Statistics show that progressively adding these modules results in gradually increasing accuracy. Simultaneously, due to the additional calculation introduced by these modules in the feature extraction parts, the inference time increases as shown in table II. Compared with the state-of-the-art real-time semantic segmentation model on Cityscapes test dataset, we achieve better accuracy and faster speed, demonstrated in Table III.

TABLE II
ABLATION STUDY OF PROPOSED MODULES ON *Cityscapes* DATASET

ASPP	DCN	ESPCN	ISA	val mIoU(%)	test mIoU(%)	Time(ms)
✓	-	-	-	74.0	71.2	**10.4**
✓	✓	-	-	75.0	72.4	11.0
✓	✓	✓	-	77.2	75.9	12.8
✓	✓	✓	✓	**78.5**	**76.5**	14.1

TABLE III
SEMANTIC SEGMENTATION RESULTS ON *Cityscapes* TESTSET

Methods	val mIoU(%)	test mIoU(%)	Time(ms)
BiSeNet V2-Large [48]	75.8	75.3	21.1
SwiftNetRN-18 [56]	-	75.5	25.0
U-HarDNet-70 [57]	75.4	75.9	18.8
Ours	**78.5**	**76.5**	**14.1**

B. Localization Performance

We evaluate our Re-localization method using the online benchmark, which calculates the percentage of query images within three different thresholds of rotation and translation error. Two types of datasets, Aachen Day-Night [58] and InLoc [7], including indoor and outdoor scenes, are used for validation.

The improvement in localization accuracy brought by our image retrieval method proves the effectiveness of our

scheme. We conduct the experiments by changing different retrieval schemes followed by the same Re-localization pipeline on Aachen Day-Night dataset. Compared with DELG, one of the state-of-the-art retrieval methods illustrated in Table I, after applying the same DBSCAN clustering scheme, our method still performs better. The results are shown in Table IV. Our method has an absolute advantage on the night dataset, while on day datasets, ours is also the best one under the largest threshold. As retrieval gives a coarse initial pose, a more accurate and robust method will relieve much pressure on backend localization procedures and has more significant benefits on the larger accuracy threshold.

TABLE IV
RE-LOCALIZATION ACCURACY ON *Aachen Day-Night v1.1* USING DIFFERENT RETRIEVAL SCHEMES

Methods	Accuracy(0.25m, 2°)/(0.5m, 5°)/(5m, 10°)	
	Day	Night
DELG	88.8 / **95.9** / 98.8	69.6 / 84.8 / 94.8
DELG + DBSCAN	**89.2** / 95.5 / 98.5	72.3 / 88.0 / **98.4**
Ours	88.8 / 95.4 / **99.0**	**74.3** / **90.1** / **98.4**

Considering semantic segmentation is a part of the observation constraints for the pose optimization, we validate their inference to the localization accuracy and other geometric attributes. We adopt the retrieval results in Table IV as the initial poses to be optimized by using observation constraints on the Aachen dataset. On the InLoc dataset, we validate the influence introduced by observation constraints with and without semantic segmentation, as shown in Table V. We found that almost all the accuracy is improved under the observation constraints optimization, and further improved under the one with semantic information.

TABLE V
RE-LOCALIZATION ACCURACY ON *InLoc* WITH OBSERVATION CONSTRAINTS

Methods	Accuracy(0.25m, 2°)/(0.5m, 5°)/(5m, 10°)	
	duc1	duc2
BaseLine	41.9 / 68.2 / 84.3	50.4 / 76.3 / 80.2
+OC(w/o semantic)	**47.0** / 68.7 / **84.8**	57.3 / 76.3 / **80.9**
+OC(w/ semantic)	**47.0** / **71.2** / **84.8**	**58.8** / **77.9** / **80.9**

Totally, we compare our proposed pipeline with some existing state-of-the-art approaches at the Long-Term Visual Localization benchmark 2020 [1]. We capture the latest results of various typical approaches from `visuallocalization.net/benchmark/` and show in Table VI. Statistically, our methods show better accuracy compared to some of the localization methods on Aachen Day-Night and InLoc datasets, especially on harder night subsets and indoor datasets, including many occlusions. Alabtion study on observation constraints has been conducted in VI to show the improvements brought by our proposed method.

TABLE VI
EVALUATION OF STATE-OF-THE-ART APPROACHES ON *Aachen Day-Night v1.0, v1.1* AND *InLoc*

Aachen Day-Night v1.0	*day* (0.25m, 2°)/(0.5m, 5°)/(5m, 10°)	*night* (0.25m, 2°)/(0.5m, 5°)/(5m, 10°)
Active Search v1.1 [40]	57.3 / 83.7 / 96.6	19.4 / 30.6 / 43.9
NetVLAD + D2-Net [21]	84.8 / 92.6 / 97.5	84.7 / 90.8 / 96.9
DenseVLAD + D2-Net [21]	83.1 / 90.9 / 95.5	74.5 / 85.7 / 90.8
KAPTURE-R2D2-APGeM [59]	88.7 / 95.8 / 98.8	81.6 / 88.8 / 96.9
SuperPoint + SuperGlue [60]	**89.6** / **95.4** /98.8	**86.7** / **93.9** / **100.0**
Ours(w/o OC)	85.7 / 93.7 / 98.9	81.6 / 91.8 / 100.0
Ours(w/ OC)	88.8 / **95.4** / **99.0**	85.7 / **93.9** / **100.0**
Aachen Day-Night v1.1	*day* (0.25m, 2°)/(0.5m, 5°)/(5m, 10°)	*night* (0.25m, 2°)/(0.5m, 5°)/(5m, 10°)
Isrf-5k-o2s [61]	87.1 / 94.7 / 98.3	74.3 / 86.9 / 97.4
LISRD+SuperPoint [62]	- / - / -	72.3 / 86.4 / 97.4
KAPTURE-R2D2-APGeM [59]	**90.0** / **96.2** / **99.5**	72.3 / 86.4 / 97.9
SuperPoint + SuperGlue [60]	89.8 / 68.7 / 80.8	**77.0** / **90.6** / **100.0**
Ours(w/o OC)	85.7 / 93.7 / 98.9	74.3 / 90.1 / 98.4
Ours(w/ OC)	88.8 / 95.4 / 99.0	74.3 / **90.6** / 98.4
InLoc	*duc1* (0.25m, 10°)/(0.5m, 10°)/(5m, 10°)	*duc2* (0.25m, 10°)/(0.5m, 10°)/(5m, 10°)
HF-Net [14]	39.9 / 55.6 / 67.2	37.4 / 57.3 / 70.2
Isrf-5k-o2s [61]	39.4 / 58.1 / 70.2	41.2 / 61.1 / 69.5
Sparse-NCNet [63]	47.0 / 67.2 / 79.8	43.5 / 64.9 / 80.2
D2-Net [21]	42.9 / 63.1 / 75.3	40.5 / 61.8 / 77.9
KAPTURE-R2D2-FUSION [59]	41.4 / 60.1 / 73.7	47.3 / 67.2 / 73.3
SuperPoint + SuperGlue [60]	**49.0** / 68.7 / 80.8	53.4 / 77.1 / **82.4**
Ours(w/o OC)	41.9 / 68.2 / 84.3	50.4 / 76.3 / 80.2
Ours(w/ OC)	47.0 / **71.2** / **84.8**	**58.8** / **77.9** / 80.9

V. CONCLUSIONS

In conclusion, an integrated visual re-localization method named RLOCS is proposed. A more accurate and robust retrieval CNN is designed, and coarse initial localization poses are produced by DBSCAN clustering, spatial verification and 2D-2D matching. Furthermore, an optimization scheme called observation constraints containing semantic segmentation and other geometry attributes is adopted to iteratively fine-tune the poses. Abundant experiments are conducted on Aachen Day-Night and InLoc datasets to prove our method's effectiveness, with comparisons to some state-of-the-art visual localization methods. The entire pipeline has great expansibility and potential, such as further improving the image retrieval CNN or semantic CNN and including more geometric hints as additional constraints, like depths or normals of every 3D point.

REFERENCES

[1] H. Lim, S. Sinha, M. Cohen, and M. Uyttendaele, "Real-time image-based 6-dof localization in large-scale environments," 2013.

[2] R. O. Castle, G. Klein, and D. W. Murray, "Video-rate localization in multiple maps for wearable augmented reality," in *12th IEEE International Symposium on Wearable Computers (ISWC 2008), September 28 - October 1, 2008, Pittsburgh, PA, USA*, 2008.

[3] W.-T. Wang, Y.-L. Wu, C.-Y. Tang, and M.-K. Hor, "Adaptive density-based spatial clustering of applications with noise (dbscan) according to data," in *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1, pp. 445–451, IEEE, 2015.

[4] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016.

[5] S. Ullman, "The interpretation of structure from motion," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 203, no. 1153, pp. 405–426, 1979.

[6] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, and T. Pajdla, "Are large-scale 3d models really necessary for accurate visual localization?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1637–1646, 2017.

[7] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7199–7209, 2018.

[8] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, pp. 2938–2946, 2015.

[9] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial intelligence review*, vol. 43, no. 1, pp. 55–81, 2015.

[10] C. Valgren and A. J. Lilienthal, "Sift, surf & seasons: Appearance-based long-term localization in outdoor environments," *Robotics and Autonomous Systems*, vol. 58, no. 2, pp. 149–156, 2010.

[11] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 8, pp. 930–943, 2003.

[12] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[13] O. Chum, J. Matas, and J. Kittler, "Locally optimized ransac," in *Joint Pattern Recognition Symposium*, pp. 236–243, Springer, 2003.

[14] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "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, 2019.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[16] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *European conference on computer vision*, pp. 778–792, Springer, 2010.

[17] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*, pp. 2564–2571, Ieee, 2011.

[18] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *European Conference on Computer Vision*, pp. 467–483, Springer, 2016.

[19] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018.

[20] J. Revaud, P. Weinzaepfel, C. De Souza, N. Pion, G. Csurka, Y. Cabon, and M. Humenberger, "R2d2: Repeatable and reliable detector and descriptor," *arXiv preprint arXiv:1906.06195*, 2019.

[21] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint description and detection of local features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8092–8101, 2019.

[22] C. Toft, E. Stenborg, L. Hammarstrand, L. Brynte, M. Pollefeys, T. Sattler, and F. Kahl, "Semantic match consistency for long-term visual localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 383–399, 2018.

[23] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting oxford and paris: Large-scale image retrieval benchmarking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5706–5715, 2018.

[24] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *Proceedings of the IEEE international conference on computer vision*, pp. 3456–3465, 2017.

[25] T. Weyand, A. Araujo, B. Cao, and J. Sim, "Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2575–2584, 2020.

[26] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3304–3311, IEEE, 2010.

[27] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 9, pp. 1704–1716, 2011.

[28] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *null*, p. 1470, IEEE, 2003.

[29] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8, IEEE, 2007.

[30] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8, IEEE, 2008.

[31] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5297–5307, 2016.

[32] B. Cao, A. Araujo, and J. Sim, "Unifying deep local and global features for image search," *arXiv*, pp. arXiv–2001, 2020.

[33] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017.

[34] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters — improve semantic segmentation by global convolutional network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[35] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *European Conference on Computer Vision*, 2018.

[36] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," 2018.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016.

[38] G. Huang, Z. Liu, V. D. M. Laurens, and K. Q. Weinberger, "Densely connected convolutional networks," 2016.

[39] L. Svarm, O. Enqvist, F. Kahl, and M. Oskarsson, "City-scale localization for cameras with known vertical direction," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1455–1461, 2016.

[40] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 9, pp. 1744–1756, 2016.

[41] Z. Zhang, T. Sattler, and D. Scaramuzza, "Reference pose generation for visual localization via learned features and view synthesis," 2020.

[42] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2599–2606, IEEE, 2009.

[43] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial–temporal data," *Data & knowledge engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[44] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[45] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

[46] Z. Lin, Z. Yang, F. Huang, and J. Chen, "Regional maximum activations of convolutions with attention for cross-domain beauty and personal care product retrieval," in *Proceedings of the 26th ACM international conference on Multimedia*, pp. 2073–2077, 2018.

[47] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proceedings of the IEEE international conference on computer vision*, pp. 1269–1277, 2015.

[48] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *arXiv preprint arXiv:2004.02147*, 2020.

[49] L. Huang, Y. Yuan, J. Guo, C. Zhang, X. Chen, and J. Wang, "Interlaced sparse self-attention for semantic segmentation," *arXiv preprint arXiv:1907.12273*, 2019.

[50] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[51] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 764–773, 2017.

[52] I. M. Sobol, *A primer for the Monte Carlo method*. CRC press, 1994.

[53] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "End-to-end learning of deep visual representations for image retrieval," *International Journal of Computer Vision*, vol. 124, no. 2, pp. 237–254, 2017.

[54] J. Revaud, J. Almazán, R. S. Rezende, and C. R. d. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5107–5116, 2019.

[55] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

[56] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 12607–12616, 2019.

[57] P. Chao, C.-Y. Kao, Y.-S. Ruan, C.-H. Huang, and Y.-L. Lin, "Hardnet: A low memory traffic network," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3552–3561, 2019.

[58] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, *et al.*, "Benchmarking 6dof outdoor visual localization in changing conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8601–8610, 2018.

[59] M. Humenberger, Y. Cabon, N. Guerin, J. Morat, J. Revaud, P. Rerole, N. Pion, C. de Souza, V. Leroy, and G. Csurka, "Robust image retrieval-based visual localization using kapture," *arXiv preprint arXiv:2007.13867*, 2020.

[60] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947, 2020.

[61] I. Melekhov, G. J. Brostow, J. Kannala, and D. Turmukhambetov, "Image stylization for robust features," *arXiv preprint arXiv:2008.06959*, 2020.

[62] R. Pautrat, V. Larsson, M. R. Oswald, and M. Pollefeys, "Online invariance selection for local feature descriptors," *arXiv preprint arXiv:2007.08988*, 2020.

[63] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic, "Neighbourhood consensus networks," *arXiv preprint arXiv:1810.10510*, 2018.