SeqLoc: Visual Localization with an Image Sequence

Liyuan Zhu
ETH Zurich
liyzhu@ethz.ch

Han Sun
ETH Zurich
hansun@ethz.ch

Jingyan Li
ETH Zurich
jingyli@ethz.ch

Abstract

Visual localization nowadays has gained rising popularity in navigation, Simultaneous Localization and Mapping(SLAM), AR/VR, and many robotics applications. Conventional methods only deal with single image localization, which consists of image retrieval, feature extraction, matching, and PnP pose estimation. In this paper, we propose a novel solution: localization with a short image sequence to leverage the redundant information in the sequence. Instead of establishing 2D-3D correspondences, we solve the pose estimation problem by introducing point cloud registration. The proposed method also provides a simple and efficient 3D-3D correspondence generation algorithm to solve the transformation between two SfM-based point clouds. On top of the point cloud registration based localization, we add a global bundle adjustment module to refine the pose estimate with additional constraints within the sequence. Experimental results show that our proposed method can reach centimeter-level accuracy and outperforms HLoc [24], one of the state-of-the-art visual localization methods, in terms of robustness. To benefit the community, our code is publicly available: github.com/Zhu-Liyuan/SeqLoc.

1. Introduction

Recently, visual localization has become a key technique for applications in augmented, mixed and virtual reality, robotics, autonomous driving etc. It aims to estimate the 6 Degree-of-Freedom (DoF) camera pose from which a given image was taken relative to a reference scene representation. In many of these application scenarios, images are captured sequentially. Many studies [16, 15, 4, 27, 28, 33] have been working on the localization of a single image, ignoring the information offered by the sequence. Other realizations of using sequences for visual localization base on the local camera tracking by visual and/or inertial odometry [31, 7], requiring relative motion constraints of the camera. In this work, we hope to achieve a robust visual localization by leveraging the redundant information in the image sequence without knowing relative motions of the sensor. The proposed pipeline *SeqLoc*, instead of establishing 2D-3D correspondences, solves the pose estimation problem by introducing point cloud registration. We also provide a simple and efficient 3D-3D correspondence generation algorithm to solve the transformation between two SfM-based point clouds. On top of the point cloud registration based localization, a global bundle adjustment module is added to refine the pose estimate with additional constraints within the sequence.

2. Related Work

Structure-based Visual Localization, as the traditional category of methods, first does 3D reconstruction by finding the pixel-wise image correspondences. This procedure is achieved by matching the local image feature descriptors. Then the 2D-3D correspondences between the query image and the reconstructed 3D map are established by descriptor matching as well. Once the 2D-3D correspondences are available, camera poses could be estimated via Perspective-n-Point(PnP) solvers. As one can imagine, although accurate, those methods suffer from high computation requirements due to the exhaustive feature matching. This also makes it hard for those methods to be scaled to large scenes as the model grows quite big with large areas to be covered. To deal with the model size, image retrieval [25, 1, 14] could be used to reduce the computation effort. Images that are captured at the neighboring location of the query image and share common local visual structures are retrieved as similar images. Only the 3D space defined by those most relevant images is searched instead of the whole area.

Learning-based Visual Localization approaches benefit from large-scale datasets, replacing some manually designed compoments in traditional localization pipelines with learnable data-driven structures. Some methods try to solve the visual localization problem in an end-to-end manner. Those methods take advantage of the development of deep neural networks(DNNs) and their success in vision tasks, assuming that low-level features extracted by DNNs encode abundant information that could be used as descriptors or directly for pose estimation. To be more spe-

cific, [16, 15, 4] directly estimate the absolute camera poses with a neural network end-to-end. However, these end-to-end methods fall short in accuracy. Instead of directly learning an entire pipeline, recent methods [27, 28, 33] shift the problem to learning feature descriptors or 2D-3D matches, combining with traditional structure-based or image retrieval pipelines.

Feature Extraction describes a pixel with designed descriptors, encoding context information from its neighborhood pixels or the whole input image. Such descriptors are expected to be informative and invariant under image transformation. Traditional methods [17] rely on hand-crafted descriptors to extract useful structures. Those descriptors are not task-specific and thus have limited ability to represent useful features for the challenge at hand. Recently, deep convolutional neural networks(CNNs) have had huge progress and proved their effectiveness in all kinds of vision tasks, e.g. image classification, semantic segmentation, and image retrieval. Those features are dense pixel-wise with the nature of CNN and learnable from task-specific datasets, thus could act as powerful representations for image matching and localization [6].

Feature Matching is to establish correspondences between two given feature sets, utilizing the spatial geometrical relations. Classical methods usually establish the preliminary feature correspondences by counting the similarity between local feature descriptors. Given a selected feature point in the reference image, methods like fixed threshold and nearest neighbor are employed to filter the most similar feature points in the query image. Afterward, false matches are removed with respect to certain geometric constraints, mainly implemented by a robust solver like RANSAC [10]. Recently, learning-based methods accomplish feature matching based on the extracted feature sets. Some methods are inspired by the classical RANSAC to estimate the transformation model like fundamental matrix [22] and epipolar geometry [3] from the feature points by CNN. Another direction is to train the model to identify true matches based on deep graph matching [9, 27]. SuperGlue [27], as an example, matches two sets of local features by jointly finding correspondences and rejecting non-matchable points by optimizing a differentiable optimal transport problem through a graph neural network.

Image Retrieval is to extract a global embedding of an image, such that given a query image, we can find images that are semantically matched or similar in visual content from a large image gallery according to the distance between embeddings [5]. Classical methods usually apply manual engineered features like SIFT [18] and aggregate the local features into a compact one through embedding methods like bag-of-visual-words (BoW) [20], VLAD [2] or Fisher vector [19]. Deep learning models have shown strong power in the field. NetVLAD [1], as an example, plugged VLAD as a layer into a CNN, offering end-to-end training on place recognition benchmarks. Some extensions like MobileNetVLAD [26] distilled the large network into a smaller one to support real-time localization tasks.

Point Cloud Registration (PCR) is to find the transformation estimation between two point clouds. The major challenges lie in the noise and outliers from the environment and sensors, the partial overlap between two point clouds, cross-registration from multiple sensors as well as the inconstant densities of points clouds [13]. Classical methods solve the problem as an optimization task [21] - minimizing a geometric projection error through two steps: finding correspondence and transformation estimation. Recent deep learning models can either learn the robust features from 3D point clouds [11] and set up correspondences between two feature sets by one-step estimation like RANSAC, or offer an end-to-end network [8] which optimizes the feature extraction together with the transformation estimation in the meantime.

3. Methodology

To improve the robustness and accuracy of visual localization, we resort to temporal information within an image sequence while a single image can only capture 2D features of the 3D scene. Our SeqLoc pipeline is displayed in Figure 1. We aim to obtain the 3D structure of the scene prior to pose estimation. Therefore, the small-scale local structure is firstly reconstructed within the image sequence. Then the goal of locating query images is switched to locating the local reconstructed point cloud in the built global 3D point cloud. PCR is introduced to estimate the transformation between the two point clouds. The transformation is applied to each camera to estimate the individual poses of all the images in the sequence. In the end, we refine the estimated poses using global bundle adjustment with additional sequential constraints.

3.1. SfM Reconstruction

Structure from Motion(SfM) is an important component in visual localization, which can recover the sparse 3D structure given a collection of images by minimizing reprojection errors. In **SeqLoc**, SfM is performed twice: in the reconstruction of the global 3D model and in the reconstruction of the image sequence. The SfM in **SeqLoc** consists of: i) feature extraction, ii) feature matching. and iii) bundle adjustment.

For i), SuperPoint[6], a self-supervised learning-based feature extractor outperforming the other state-of-the-art methods, extracts the **feature descriptors** of each image. For ii), we use SuperGlue[27], which is an attention-based graph neural network for feature matching with outstanding performances. To increase the efficiency of feature matching among all the images, the image retrieval technique

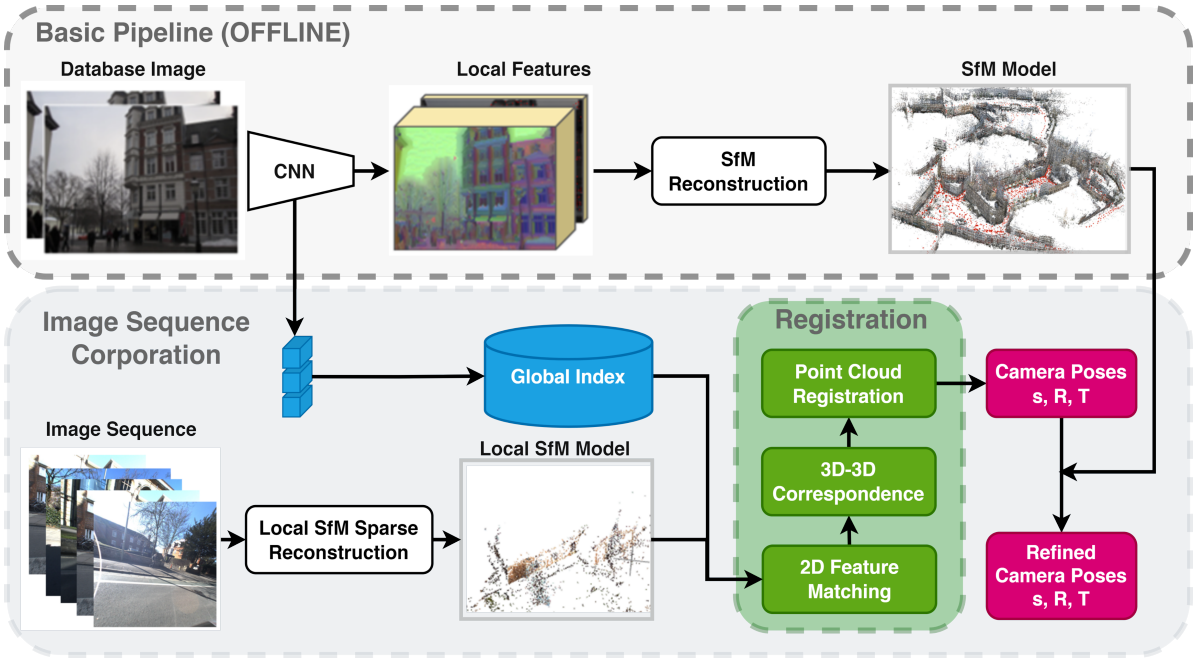

Figure 1. **SeqLoc**, adapted from HFNet[24]. The OFFLINE part to reconstruct the whole environment, is kept unchanged. We incorporate image sequences by local reconstruction, 3D-3D correspondences, Point Cloud Registration, and pose refinement.

NetVLAD[1] is utilized to only generate necessary image pairs to match, avoiding exhaustive matching. With the features and matches, structure-from-motion is realized by COLMAP[30], an incremental SfM reconstruction toolbox.

3.2. 3D-3D Correspondence Generation

After the sparse reconstruction from **SfM**, we obtain 3D point clouds of the environment and the image sequence, so-called **Global Reconstruction** and **Local Reconstruction**, respectively in the rest of the paper. Inspired by our motivation to leverage the redundant information in the sequence, the goal is transformed into locating the **Local Reconstruction** in the **Global Reconstruction**.

Similar to image localization, locating a point cloud also demands finding correspondences between two 3D structures. Common ways to link two point clouds is to generate point-wise 3D feature descriptors(e.g. FPFH[23]) and match the feature vectors to generate point pairs. Deep learning in recent years facilitates the development in PCR with learned descriptors and end-to-end registration networks[11, 12]. However, these methods are mostly developed for LiDAR data, not for SfM point clouds. Point clouds from SfM are not uniformly distributed. Point clouds from SfM have various point densities from place to place. Registration methods for LiDAR are likely to fail in this scenario.

With existing 2D features and correspondence graphs, we propose a method to generate 3D-3D correspondences without 3D descriptors. The relation of 2D points and 3D points in SfM is demonstrated in Figure 3. Our method is illustrated in Algorithm 1:

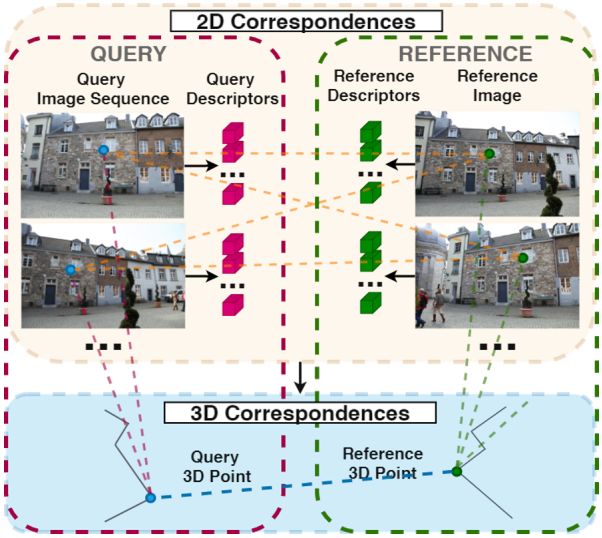

Figure 2. **Generation of 3D correspondences from 2D.** For each query 3D point from Local Reconstruction, we find its corresponding 2D points in the query image sequence. We then establish 2D-to-2D correspondences by matching query feature descriptors and the reference descriptors. Finally we can accomplish the 3D correspondences by link the reference descriptors to the 3D point in Global Reconstruction.

P_i^{ref}, P_i^q denote the i^{th} 3D point in **Global Reconstruction** and **Local Reconstruction**, respectively. $p_{i,n}^{ref}$

Algorithm 1 Generating 3D correspondences from 2D

$Corr \leftarrow Empty$
$R \leftarrow 0.6$
while P_i^q in P^q **do**
  $Ref \leftarrow Empty$
  while $p_{i,n}^q$ in P_i^q **do**
    Get match of p_i^q: p_j^{ref} ▷ (NetVLAD+SuperGlue)
    if p_j^{ref} is reconstructed **then**
      Get P_j^{ref} of p_j^{ref}
      $Ref \leftarrow P_j^{ref}$
    end if
    Compute histogram of Ref
    Get the most frequent P_t^{ref} and $freq_{max}$
    if $freq_{max} \geq R$ **then**
      $Corr \leftarrow (P_i^q, P_t^{ref})$
    end if

stands for the n^{th} corresponding 2D point of 3D point P_i^{ref}. $R \in [0, 1]$ is an empirical threshold to get reliable correspondence, higher value meaning stricter and fewer correspondences. The final output $Corr$ is a list of correspondences between **Global Reconstruction** and **Local Reconstruction**.

3.3. Camera Pose Estimation

With the 3D correspondences obtained from 3.2, we demonstrate our method to solve transformation between two point clouds and apply it to camera pose estimation. Estimating transformation between two point clouds from correspondences is known as **Point Cloud Registration**. Due to noisy measurements and various overlaps, the correspondences between the two are usually including outliers. Therefore, simple singular value decomposition (SVD) could not fulfill our demands for a good estimate. In addition, since SfM is not up to scale, the scale factor between two point clouds should be estimated as well.

TEASER++[32] is a fast and certifiable point cloud registration solver in the presence of large amounts of outlying correspondences. In TEASER++, PCR is reformulated into **a Truncated Least Squares (TLS)** with **Semidefinite Relaxation**, and a cascaded graph-theoretic framework. With TEASER++, we can get a reliable estimate of scale λ, translation $T_{_{pcr}}$ and R_{pcr}, where $_{pcr}$ is short for PCR. The transformation Reg is formulated as

$$Reg = [\lambda R_{pcr} | T_{pcr}] \tag{1}$$

The next step is to transform the camera poses from **Local Reconstruction** into **Global Reconstruction** using Reg. In COLMAP the[30], camera pose is formatted as a quaternion vector$Q_{vec}(Q_W, Q_X, Q_Y, Q_Z)$ and a translation vector $T_{vec}(T_X, T_Y, T_Z)$, as the projection from world to the camera coordinate system. We should first convert this representation into a world coordinate system:

$$Pose_{cam} = [R_q^T | -R_q^T T_{vec}] \tag{2}$$

Base on Eq.(2), all camera poses in **Local Reconstruction** are computed as $Pose_{cam_i}^{Local}, i \in [0, num_{cam}]$. Afterwards, $Pose_{cam_i}^{Local}$ are transformed into $Pose_{cam_i}^{Global}$:

$$Pose_{cam_i}^{Global} = Reg \times Pose_{cam_i}^{Local} \tag{3}$$

$Pose_{cam_i}^{Global}$ is then parsed into $Q_{vec,i}^{Global}$ and $T_{vec,i}^{Global}$:

$$Q_{vec,i}^{Global} = Rot2Qvec(Pose_{cam_i}^{Global}[:3, :3]^T) \tag{4}$$

$$T_{vec,i}^{Global} = -Pose_{cam_i}^{Global}[:3, :3]^T \times Pose_{cam_i}^{Global}[:3, 3] \tag{5}$$

3.4. Pose refinement

From **Camera Pose Estimation**, we could get a coarse pose estimation of query images. Such an estimate from PCR is an optimal solution between two point clouds. However, there is still room for improvement. We hereby propose a pose refinement module after PCR, based on bundle adjustment with extra constraints.

A camera projection model that projects a 3D point onto 2D image goes here:

$$z \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R} \mid \mathbf{T}] \begin{bmatrix} p_x \\ p_y \\ p_z \\ 1 \end{bmatrix} \tag{6}$$

Multiple 2D feature points and corresponding 3D points are observed. We reformulate Eq.(6) to Eq.(7) and stack the observation equations into Eq.(8) for multiple cameras.

$$\begin{aligned} \left(m_1^{\mathrm{T}} - u_i m_3^{\mathrm{T}}\right) \cdot P_i = 0 \\ \left(m_2^{\mathrm{T}} - v_i m_3^{\mathrm{T}}\right) \cdot P_i = 0 \end{aligned} \tag{7}$$

$$\begin{pmatrix} P_1^{\mathrm{T}} & 0^{\mathrm{T}} & -u_1 P_1^{\mathrm{T}} \\ 0^{\mathrm{T}} & P_1^{\mathrm{T}} & -v_1 P_1^{\mathrm{T}} \\ & & \vdots \\ P_n^{\mathrm{T}} & 0^{\mathrm{T}} & -u_n P_n^{\mathrm{T}} \\ 0^{\mathrm{T}} & P_n^{\mathrm{T}} & -v_n P_n^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \tag{8}$$

$$R_i, T_i = \mathrm{argmin}_{R_i, T_i} \sum_{i=1}^{n} \mid p^i - \pi\left(P_W^i, K_i, R_i, T_i\right) \mid^2 \tag{9}$$

In the bundle adjustment optimization, the sum of reprojection errors is minimized, where camera poses are adjusted and refined to find the optimal alignment of the image to

the reference 3D model. Based on estimated poses and precomputed correspondence graph, the related 3D points, 2D points, and camera intrinsics are retrieved from the database for refinement. We implement 2 modes of refinement: **Single Image Bundle Adjustment(SBA)** and **Global Bundle Adjustment(GBA)**. **SBA** only optimizes reprojection errors individually for every image one after another, where the refining process of different images are independent of each other. While **GBA** takes all images in the sequence into account and introduces additional constraints to the sequence center. The parameter setting for bundle adjustment is displayed in Table. 1.

4. Experiment

In this section, we illustrate our experimental setup, dataset, and evaluation metrics. We perform a sequence-based visual localization in a large-scale outdoor dataset.

4.1. Setup

In our implementation, SuperPoint, SuperGlue, and NetVLAD are all deep learning methods, demanding trained models for inference. As these models are provided by Hloc[24] toolbox[1], we do not train additional models. The provided SuperPoint model is trained on Aachen-Day-Night[29] with a Non Maximum Suppression (NMS) radius of 4 pixels and 4096 as a maximal number of detected keypoints. During feature matching, the SuperGlue pre-trained model runs Sinkhorn for 50 iterations. For structure from motion(SfM) with COLMAP[30], the original C++ implementation is capsulated into python[2]. The camera intrinsics are fixed during reconstruction and all the other parameters in SfM are set as default in COLMAP. The parameter configuration for TEASER++ is also set as default. The pose refinement module is developed based on COLMAP and pycolmap[3]. Our experiment is carried out on a workstation with 32GB of RAM, AMD Ryzen 5600x as CPU, and Nvidia Geforece RTX 3060 12G as GPU.

4.2. Dataset

Since image sequence based localization is a novel idea, there is no existing dataset designed for this task. We create a sub-sampled dataset derived from Aachen-Day-Night[29]. The images for reference and query in Aachen-Day-Night are captured by individual smartphone cameras, therefore we abandon the given query images and subsample new query images in a sequential manner from the reference image database. There are over 4000 images for building the **Global Reconstruction**. We compute the top-50 image retrieval scores for every image and randomly selected 3 to 10 images in its top-20 most similar images to form a pseudo image sequence. To keep the fairness of the experiment, the selected images are eliminated from the database and the reference point cloud is re-triangulated using the remaining images in the database to make sure the localization would not be benefited.

[1]The official implementation of Hloc[24] with SuperGlue[27]: github.com/cvg/Hierarchical-Localization/

[2]The official implementation of pycolmap: github.com/colmap/pycolmap

[3]The modified version of pycolmap with **GBA** : github.com/Zhu-Liyuan/pycolmap/

4.3. Evaluation Metrics

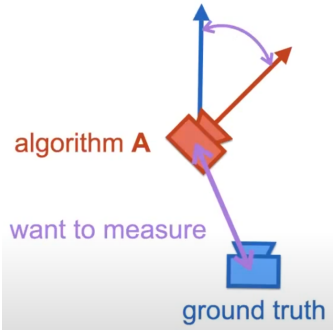

Figure 3. **Visualization of translation and rotation error.** Credit: *Eric Brachmann*.

To evaluate the accuracy of localization, the pose error is described as two parts: one for translation and one for rotation. The translation error is defined as the 2D norm of the estimated coordinate subtracting the ground truth. The rotation error is computed by the relative rotation between the estimate and ground truth. We use recall of error under $(0.25m, 5^\circ)$ as an assessment of localization robustness.

4.4. Results and Analysis

We set sequence localization with registration(w/o refinement) as our baseline and test **PCR** (Point Cloud Registration only), **PCR+SBA** (single image bundle adjustment), and **PCR+GBA** (global bundle adjustment) on the sequence data generated from 4.2and evaluate them with the metrics from 4.3.

In Figure. 4, the **Global Reconstruction**, **Local Reconstruction**, intermediate correspondences, and the final registration are displayed, from which good quality data association and registration could be observed.

From Table.2, **PCR+PBA** provides the best translational accuracy, improving PCR-only by 2 cm. In terms of rotation error, PCR with refinement outperforms PCR-only methods by 30%. Boxplot in Figure. 5 shows the error distribution. We also compare our method with state-of-the-art single image localization method Hloc. Both methods manage to locate most of the images. **SeqLoc** has a better recall, meaning it is more robust in the challenging outdoor environment.

Parameters	Fixed as Constant	To be Optimized
3D structure: P_i^{Global}	x	
Camera Orientation: $R_{cam,i}^{Global}$		x
Camera Intrinsics: $K_{cam,i}$	x	
*Center of Image Sequence $\sum_{i=1}^{n} T_{cam,i}^{Global}$	x	
*Orientation of Camera Center		x

Table 1. **Parameter setting for Pose Refinement**. * is only considered for **GBA** not for **SBA**.

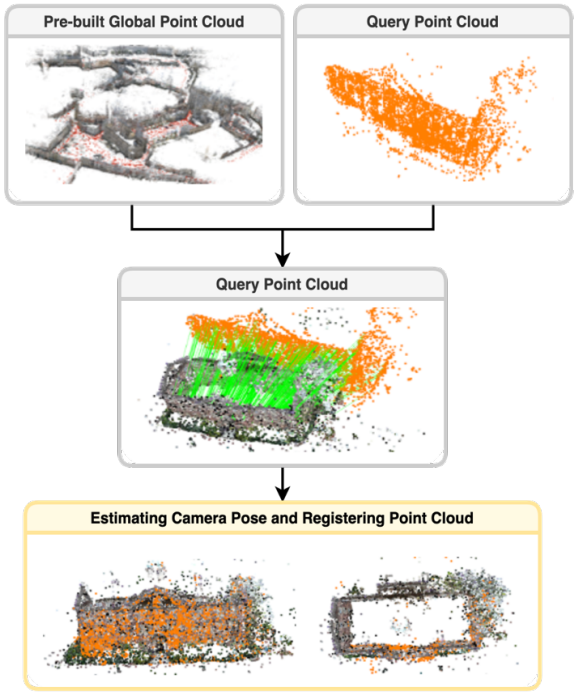

Figure 4. **Qualitative visualization of our workflow.**

	PCR	PCR+SBA	PCR+GBA
Mean Trans. err.(m)	0.1097	0.0916	**0.0824**
Mean Rot. err.(°)	0.1553	**0.1058**	0.1078

Table 2. **Evaluation of PCR, PCR+SBA, PCR+GBA on sequence data.**

Method	HLoc	PCR+GBA
$(0.25m, 5°)$ recall	98.1%	**99.07%**

Table 3. **Comparison between SeqLoc and Hloc.**

5. Conclusion and Outlook

In this work, we present **SeqLoc**, an image-sequence visual localization method, which associates the images in the sequence through a sparse reconstruction, then performs a point cloud to point cloud localization with pose refinement. We propose a simple algorithm to generate correspondences between SfM point clouds and make good use of the redundancy in the sequence by adding global constraints into global bundle adjustment. We show that **SeqLoc** can realize centimeter-level localization in the outdoor environment and provide better localization robustness compared to single image localization. Due to the lack of data for sequence localization, we demand more datasets and experiments to explore the pros and cons of sequence localization.

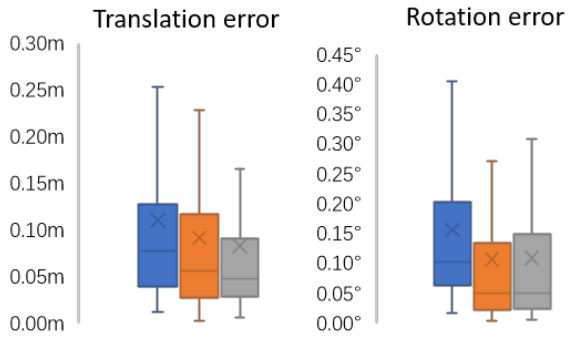

Figure 5. **Boxplot of the evaluation of three methods. Blue, orange and gray boxes denote PCR, PCR+SBA, PCR+GBA**, respectively.

To handle the challenge of changing environment, we have two potential solutions: i) Design the feature descriptor and matcher to be robust enough against variation; ii) Regularly update the 3D map to make sure the point cloud is up-to-date for localization. There is some recent progress in robust feature design creditable to deep learning. However, there is nearly no development for map update, which is profound and whose task is hard to define. With **SeqLoc**, we could generate a local point cloud during localization. The local point cloud is an advantage over single image localization, in which no 3d structure is reconstructed. The local point cloud can be used to update a local region in the 3D map. More exploration and experiments should be done in the future.

Acknowledgement

We really appreciate the support and advice from Dr. Iro Armeni and Dr. Daniel Barath for this project and Prof. Marc Pollefeys for opening 3D Vision at ETH Zurich, providing great research opportunities for us students.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 1, 2, 3

[2] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1578–1585, 2013. 2

[3] Eric Brachmann and Carsten Rother. Neural-guided ransac: Learning where to sample model hypotheses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4322–4331, 2019. 2

[4] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2018. 1, 2

[5] Wei Chen, Yang Liu, Weiping Wang, Erwin M Bakker, TK Georgiou, Paul Fieguth, Li Liu, and MSK Lew. Deep image retrieval: A survey. *ArXiv*, 2021. 2

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2

[7] Ryan C DuToit, Joel A Hesch, Esha D Nerurkar, and Stergios I Roumeliotis. Consistent map-based 3d localization on mobile devices. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 6253–6260. IEEE, 2017. 1

[8] Gil Elbaz, Tamar Avraham, and Anath Fischer. 3d point cloud registration for localization using a deep neural network auto-encoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2017. 2

[9] Matthias Fey, Jan E Lenssen, Christopher Morris, Jonathan Masci, and Nils M Kriege. Deep graph matching consensus. *arXiv preprint arXiv:2001.09621*, 2020. 2

[10] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[11] Zan Gojcic, Caifa Zhou, Jan D Wegner, and Andreas Wieser. The perfect match: 3d point cloud matching with smoothed densities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5545–5554, 2019. 2, 3

[12] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4267–4276, 2021. 3

[13] Xiaoshui Huang, Guofeng Mei, Jian Zhang, and Rana Abbas. A comprehensive survey on point cloud registration. *arXiv preprint arXiv:2103.02690*, 2021. 2

[14] Martin Humenberger, Yohann Cabon, Nicolas Guerin, Julien Morat, Jérôme Revaud, Philippe Rerole, Noé Pion, Cesar de Souza, Vincent Leroy, and Gabriela Csurka. Robust image retrieval-based visual localization using kapture. *arXiv preprint arXiv:2007.13867*, 2020. 1

[15] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5974–5983, 2017. 1, 2

[16] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015. 1, 2

[17] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999. 2

[18] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[19] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3384–3391. IEEE, 2010. 2

[20] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 2

[21] François Pomerleau, Francis Colas, and Roland Siegwart. A review of point cloud registration algorithms for mobile robotics. *Foundations and Trends in Robotics*, 4(1):1–104, 2015. 2

[22] René Ranftl and Vladlen Koltun. Deep fundamental matrix estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 284–299, 2018. 2

[23] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE international conference on robotics and automation*, pages 3212–3217. IEEE, 2009. 3

[24] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 1, 3, 5

[25] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 1

[26] Paul-Edouard Sarlin, Frederic Debraine, Marcin Dymczyk, Roland Siegwart, and Cesar Cadena. Leveraging deep visual descriptors for hierarchical efficient localization. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*,

volume 87 of *Proceedings of Machine Learning Research*, pages 456–465. PMLR, 29–31 Oct 2018. 2

[27] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1, 2, 5

[28] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, et al. Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3247–3257, 2021. 1, 2

[29] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *BMVC*, volume 1, page 4, 2012. 5

[30] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 3, 4, 5

[31] Erik Stenborg, Torsten Sattler, and Lars Hammarstrand. Using image sequences for long-term visual localization. In *2020 International Conference on 3D Vision (3DV)*, pages 938–948. IEEE, 2020. 1

[32] Heng Yang, Jingnan Shi, and Luca Carlone. Teaser: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 37(2):314–333, 2020. 4

[33] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. Sanet: Scene agnostic network for camera localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 42–51, 2019. 1, 2