A Two-stage Outlier Filtering Framework for City-Scale Localization using 3D SfM Point Clouds

Wentao Cheng, Kan Chen, Weisi Lin, *Fellow, IEEE,* Michael Goesele, Xinfeng Zhang, and Yabin Zhang

***Abstract*—3D Structure-based localization aims to estimate the 6-DOF camera pose of a query image by means of feature matches against a 3D Structure-from-Motion (SfM) point cloud. For city-scale SfM point clouds with tens of millions of points, it becomes more and more difficult to disambiguate matches. Therefore a 3D Structure-based localization method, which can efficiently handle matches with very large outlier ratios, is needed. We propose a two-stage outlier filtering framework for city-scale localization that leverages both visibility and geometry intrinsics of SfM point clouds. Firstly, we propose a visibility-based outlier filter, which is based on a bipartite visibility graph, to filter outliers on a coarse level. Secondly, we apply a geometry-based outlier filter to generate a set of fine-grained matches with a novel data-driven geometrical constraint for efficient inlier evaluation. The proposed two-stage outlier filtering framework only relies on intrinsic information of a SfM point cloud. It is thus widely applicable to be embedded into existing localization approaches. The experimental results on two real-world datasets demonstrate the effectiveness of the proposed two-stage outlier filtering framework for city-scale localization.**

***Index Terms*—City-scale localization, outlier filter, image-based localization, hybrid inlier evaluation.**

I. Introduction

ESTIMATING the camera pose of a query image *w.r.t.* a Structure-from-Motion (SfM) point cloud plays a key role in many computer vision tasks such as 3D reconstruction [1]–[3], image-based localization [4]–[6] and visual navigation for self-driving cars [7]. A typical 3D structure-based localization pipeline starts with establishing 2D-3D matches by finding correspondences between the feature descriptors (e.g., SIFT [8]) in a query image and the feature descriptors associated with 3D points in a SfM point cloud. The 6-DOF camera pose can be computed from 2D-3D matches by applying perspective-n-point pose solvers [9], [10] in RANSAC [11].

A conventional method to disambiguate matches is the widely used SIFT ratio test [8]. However, in a city-scale SfM point cloud that depicts urban scenes, the associated dense feature space consists of many nearly identical feature descriptors. It is therefore difficult for the SIFT ratio test to obtain sufficient high quality matches with a city-scale SfM point cloud. In order to better preserve correct matches, recent state-of-the-art works [5], [12] usually adopt a relaxed SIFT ratio test, yielding a large number of wrong matches. This makes RANSAC difficult to find a reliable solution with such large ratio of outliers, thereby results in a failure of localization.

In order to handle such cases with very large outlier ratios, many outlier filters are proposed to remove outliers based on visibility [4], [13] or geometry intrinsics [5], [12], [14] of SfM point clouds. However, due to accuracy and computation complexity limitations, they cannot well deal with city-scale localization problems. In this paper, we propose a two-stage outlier filtering framework that consists of an improved visibility-based outlier filter and a novel geometry-based outlier filter. The two-stage framework overcomes the limitations of both outlier filters with a coarse-to-fine design, and achieves both efficiency and accuracy in disambiguating matches with very large outlier ratios.

The visibility-based outlier filter, which consists of database image voting, re-ranking and match augmentation operations, is conducted on the image-level to remove outliers in a coarse level. A database image voting method is proposed based on the widely known knowledge that correct matches exhibit a strong co-visibility relationship [4], [13]. To further improve the filtering performance, we introduce a re-ranking scheme to eliminate falsely voted database images. Previous methods [5], [12], [14] assume that in the initialization step, the relaxed SIFT ratio test does not reject any correct matches. However, this assumption is untenable when dealing with the dense feature space of the city-scale SfM point cloud. To this end, we propose a match augmentation scheme to carefully recover rejected correct matches with the aid of selected database images. Although the proposed visibility-based outlier filter is efficient when dealing with extremely large outlier ratio scenarios, e.g., 99% outliers, the resultant matches may still contain a large number of outliers due to the limited accuracy of the database image voting procedure.

The second stage is a geometry-based outlier filter based on a novel data-driven geometrical constraint. Our key observation is that, in a city-scale SfM point cloud, there are many 3D points that can only be observed by nearby cameras due to strong view occlusions. We denote such 3D points as *locally visible points*. Based on this observation, we derive a geometrical constraint to restrict the position of camera that can observe the *locally visible points*. Previous geometry-based outlier filters either heavily rely on additional priors about the vertical direction and approximate height of a camera

W. Cheng, and W. Lin are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wcheng005@e.ntu.edu.sg; wslin@ntu.edu.sg).

K. Chen is with the Fraunhofer IDM research center, Nanyang Technological University, Singapore 639798 (e-mail: kchen1@e.ntu.edu.sg).

M. Goesele is with the Department of Computer Science, Technische Universität of Darmstadt, Germany 64283 (e-mail: michael.goesele@gris.informatik.tu-darmstadt.de).

X. Zhang is with the department of Computer Science, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhangx-inf07@gmail.com)

Y. Zhang is with the Median Lab, Tencent, Shenzhen (e-mail: zhan0398@ntu.edu.sg).

relative to a SfM point cloud [5], [12], or require a set of high quality matches for statistically pruning outliers [14]. The derived geometrical constraint in our method does not require any prior knowledge about the camera model. In addition, this geometrical constraint enables us to efficiently handle potential low quality matches which are generated by the visibility-based outlier filter in the first stage. The main contributions of this paper can be summarized as follows:

1) A two-stage outlier filtering framework is proposed that simultaneously leverages the merits of the visibility and the geometry intrinsics of a SfM point cloud for city-scale localization. The proposed framework removes outliers in a coarse-to-fine manner by sequentially applying the designed visibility and geometry based outlier filters.
2) We propose a visibility-based outlier filter, which utilizes the bipartite relationship between database images and 3D points in a SfM point cloud. Through database image re-ranking and match augmentation, the visibility-based outlier filter is able to preserve more correct matches without severely degrading the filtering quality.
3) We derive a novel data-driven geometrical constraint for *locally visible points*, which are widespread in city-scale SfM point clouds. Based on this constraint, we propose a geometry-based outlier filter in which matches with *locally visible points* and *non-locally visible points* are separately evaluated with a hybrid scheme. Comparing with the classic re-projection error measurement, the derived geometrical constraint exhibits a superior efficiency in handling matches with large outlier ratio.
4) The effectiveness and efficiency of the proposed two-stage framework and its individual modules are comprehensively analyzed. Based on the extensive experimental results, the matches generated by our method show a high reliability for successful 3D structure-based city-scale localization.

The rest of this paper is organized as follows: Section II reviews related work in city-scale localization. Section III gives an overview of the proposed two-stage outlier filtering framework. Section IV presents the proposed visibility-based outlier filter as the first stage. Section V presents the proposed geometry-based outlier filter as the second stage. Section VI shows comprehensive experimental results on two city-scale datasets. The conclusion is finally given in Section VII.

II. Related Work

A. 2D Image-based Localization

2D image-based localization methods find a query image's location based on the geotags of the most relevant images retrieved from image datasets [15]–[20]. Thanks to the scalability of image retrieval techniques such as the Bag-of-Words (BoW) model, 2D image-based localization is efficient in dealing with city-scale datasets. Recent BoW-based works have significantly improved the retrieval performance. Philbin *et al.* [21] refined the top retrieved image list with a spatial verification step. Feature descriptors with more ambiguities were down-weighted [18], [22], [23] or avoided [24], [25] in the index construction step. Sattler *et al.* [26] down-weighted matches that could be inliers in multiple places to tackle the geometrical burstness problem, i.e., similar geometrical configurations in both relevant and irrelevant database photos. Chen *et al.* [27] fused two types of street-level image representations and incorporated the GPS priors to improve city-scale localization. Liu *et al.* [28] represented local features using binary codes to achieve a fast geometrical verification. More compact representations such as VLAD [29], [30] or NetVLAD [31] have also been explored for 2D image-based localization.

B. 3D Structure-based Localization

The 6-DOF camera pose of an image can be computed based on a 3D point cloud reconstructed via SfM techniques. 2D-3D matches should first be established between the feature descriptors in an image and the feature descriptors associated with the 3D points. The 6-DOF camera pose can then be computed using the found 2D-3D matches. Concerning city-scale SfM point clouds, recent works can be divided into two major categories:

1) Towards efficient feature matching: There can be more than tens of millions of feature descriptors in a city-scale SfM point cloud, which makes feature matching time-consuming. In order to accelerate the feature matching process, Li *et al.* [32] employed a prioritized 3D-to-2D matching strategy which prioritized 3D points with higher degrees in the bipartite visibility graph. Sattler *et al.* [33] quantized the feature descriptors in a SfM point cloud into a compact visual vocabulary dictionary for efficient localization. Choudhary *et al.* [34] proposed to apply feature matching on a small subset of 3D points, which are potentially visible in the query image. Several bi-directional feature matching frameworks [4], [35] were proposed to improve the localization performance without sacrificing the run-time efficiency. The major shortcoming of the above approaches is that they rely on the SIFT ratio test to disambiguate matches, which may lose the discriminative power especially in the much larger city-scale SfM point clouds. Feature matching can also be accelerated by using a compact binary feature description with a supervised indexing method [36], or simplifying the city-scale SfM point cloud into a very compact model [37], [38]. However, information loss is inevitable in these methods.

2) Towards disambiguating matches: Traditional outlier filters based on feature appearance, e.g., the SIFT ratio test, have difficulties when handling matching ambiguity in city-scale SfM point cloud. Recent works [4], [5], [13], [14] attempted to relax the SIFT ratio test in order to preserve more correct matches. To deal with the resultant matches with very large outlier ratio, Li *et al.* [4] employed a RANSAC sampler by encoding the co-visibility relationship among correct matches. Sattler *et al.* [13] implicitly conducted feature matching by quantizing the feature descriptors in the SfM point cloud into a fine vocabulary. An image voting strategy based on hyperpoints was adopted to filter ambiguous matches.

Other approaches [5], [12], [14] filtered wrong matches using additional or intrinsic geometrical cues. Svärm *et al.* [12]

assumed that the camera's vertical direction and approximate height relative to the SfM point cloud were known in advance. They proposed an outlier filter by formulating a 2D registration problem under this assumption. Similarly, Zeisl *et al.* [5] used the same camera model assumption and derived a linear camera pose voting algorithm. To improve the efficiency of camera pose voting, they pre-filtered obvious wrong matches using local feature constraints such as feature scale and feature orientation. Camposeco *et al.* [14] introduced a novel pose solver using intrinsic angle constraints of SfM point clouds. The camera position can be quickly estimated using two matches with this pose solver. However, all camera position hypotheses should be computed i.e., 10^8 camera position hypotheses for 10^4 matches, to remove outliers. In addition, the final outlier filter requires a set of high quality matches to statistically remove outliers.

C. Hybrid Localization

Hybrid localization [39], [40] combines 2D image-based and 3D structure-based approaches to obtain a query image's 6-DOF camera pose. Irschara *et al.* [39] augmented the 3D database by synthesizing multiple views on the 3D scene and implicitly conducted 2D-3D feature matching through selected views. Recently, Sattler *et al.* [40] utilized advanced image retrieval techniques [17], [31] to find relevant database images for a query image. Instead of reconstructing a large-scale SfM point cloud, they proposed to reconstruct a local compact SfM point cloud using the retrieved database images. Even though the experimental results showed that the advanced image-retrieval techniques helped to obtain better matches than directly matching with a large-scale SfM point cloud, the local SfM point cloud reconstruction step significantly decreased the computational efficiency.

D. Learning-based approaches

Recent advancement in deep learning techniques has made it possible to process general 3D point clouds for reconstruction [41], semantic reasoning [42], and learning discriminative 3D descriptors [43]. In the context of 3D structure-based localization, Kendall *et al.* trained a convolutional neural network (CNN) that can regress a 6-DOF camera pose from an image [44]. Walch *et al.* improved the performance of camera pose regression by integrating Long-Short Term Memory (LSTM) units with CNN [45]. To better model the scene with CNN, Kendall *et al.* proposed several novel loss functions based on re-projection error and scene geometry [46]. However, these learning-based approaches are still less accurate than traditional 3D structure-based approaches. Instead of directly regressing camera pose, Brachmann *et al.* proposed an implicit CNN-based 2D-3D matching approach by regressing the 3D scene coordinate for an input image patch [47]. Though achieving high accuracy, this approach encountered a training failure for large-scale outdoor scenes.

III. Overview of The Proposed Method

Fig. 1 illustrates the complete localization pipeline with the proposed two-stage outlier filtering framework. The pipeline

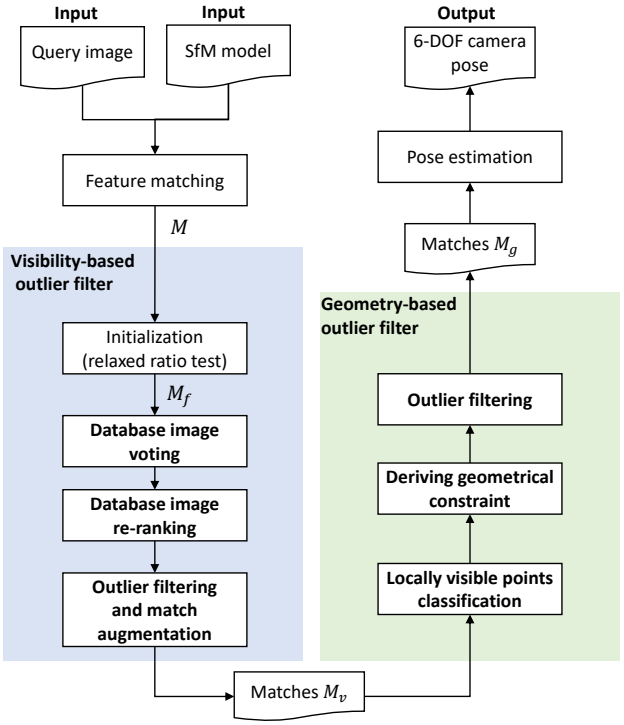


Fig. 1: The localization pipeline with the proposed two-stage outlier filtering framework (in bold font).

starts with a 1-to-N feature matching procedure [5] between a query image and a pre-computed SfM point cloud to obtain a set of 2D-3D matches $\mathcal{M}$. In the beginning of the visibility-based outlier filter, we use a relaxed SIFT ratio test as an initialization step to leverage its power of rejecting unreliable matches. By casting votes to database images using the initialized matches $\mathcal{M}_f$ and the bipartite visibility graph, the probability that a database image contains correct matches can be measured by its corresponding weighted votes. After database image re-ranking, wrong matches can be filtered using the top rank database images. Moreover, correct matches can also be augmented using the top rank database images. With the matches $\mathcal{M}_v$ obtained by the visibility-based outlier filter in the first stage, a subsequent geometry-based outlier filter is applied as the second stage. *Locally visible points* are classified and a novel geometrical constraint is derived based on *locally visible points*. The outliers can be further removed by integrating the derived geometrical constraint into a RANSAC-based pose estimation method. The final 6-DOF camera pose can be computed using the matches $\mathcal{M}_g$ generated from the geometry-based outlier filter.

IV. Visibility-based Outlier Filter

The pipeline of the proposed visibility-based outlier filter is illustrated in Fig. 2. In the following, we will describe the proposed visibility-based outlier filter in detail.

A. Initialization

In a SfM point cloud, each 3D point is associated with a set of 2D feature descriptors such as SIFT feature descriptors [8].

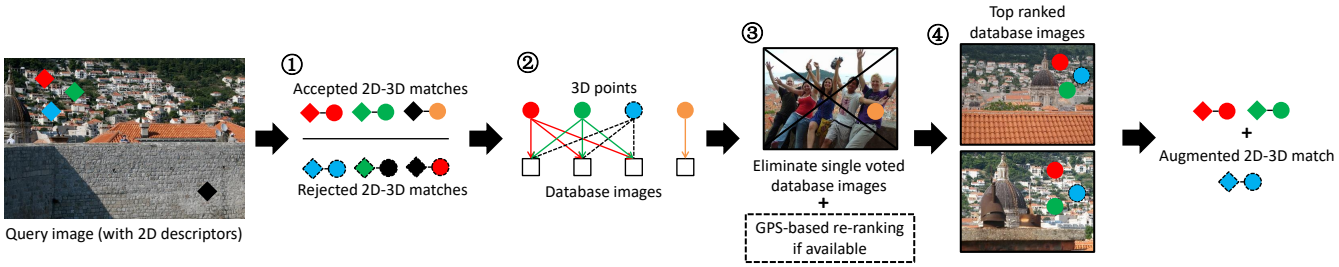


Fig. 2: The pipeline of the proposed visibility-based outlier filter. 1: initialization with a relaxed ratio test (Section IV-A). 2: database image voting with the bipartite visibility graph (Section IV-B). 3: re-ranking by eliminating single voted database images (Section IV-C). In addition, the ranking can be optionally refined if GPS data is available. 4: outlier filtering and match augmentation (Section IV-D).

2D-3D matches can be established by searching nearest neighbors in a SfM point cloud for each query feature descriptor. Let $\mathcal{M}$ be a set of initial 2D-3D matches established between a query image and a SfM point cloud. The visibility-based outlier filter starts with rejecting matches based on feature appearance. The SIFT ratio test is a widely utilized method to reject unreliable matches [5], [8], [12], [14], [35]: let p^{1st} and p^{2nd} be the first and second nearest neighbors in a SfM point cloud for a query feature q. A match is considered to be reliable if it satisfies the ratio test: $\left\|q-p^{1st}\right\|_2 / \left\|q-p^{2nd}\right\|_2 < \tau$. The threshold τ is usually set as 0.8 when matching between two images. However, due to the high density of feature space in a city-scale SfM point cloud, a correct match will often fail the SIFT ratio test and be rejected. In order to preserve more correct matches while rejecting wrong matches as much as possible, a relaxed SIFT ratio test should be applied. The relaxation can be done by either increasing the threshold of the SIFT ratio test or using an adaptive threshold [5]. Let $\mathcal{M}_f$ be the matches that are accepted by the relaxed SIFT ratio test. In practice, $\mathcal{M}_f$ contains much fewer ambiguous matches than the original matches $\mathcal{M}$. We therefore use $\mathcal{M}_f$ instead of $\mathcal{M}$ for the following database image voting procedure.

B. Database Image Voting

After obtaining the matches $\mathcal{M}_f$ with a relaxed SIFT ratio test, we aim to remove outliers by utilizing the visibility intrinsics of a SfM point cloud. In a SfM point cloud, the relationship between 3D points and database images can be modeled as a bipartite visibility graph $\mathcal{G} = (\mathcal{P}, \mathcal{D}, \mathcal{E})$. Each node $p \in \mathcal{P}$ represents a 3D point in the SfM point cloud, and each node $d \in \mathcal{D}$ represents a database image which is used to reconstruct the SfM point cloud. An edge $(p, d) \in \mathcal{E}$ exists if the 3D point p is visible in the database image d.

Leveraging the bipartite graph $\mathcal{G}$, each 2D-3D match $(q, p) \in \mathcal{M}_f$ can cast a vote to the database images that observe p. Thus, the votes for a database image d can be computed as follows:

$$\mathcal{V}(d) = \{(q,p) \mid (p,d) \in \mathcal{E}, (q,p) \in \mathcal{M}_f\}. \tag{1}$$

Ideally, a correct 2D-3D match (q, p) means that the query feature q should depict the same location as the 3D point p. Due to the continuity of geometry space, correct matches should be frequently co-visible. The co-visibility of correct matches makes the corresponding database images receiving high votes. Meanwhile, the weak co-visibility among wrong matches makes them randomly casted to irrelevant database images. In a city-scale dataset which contains a large number of database images, the votes that each database image can receive with wrong matches should be much smaller than the votes from correct matches. However, a database image that observes more 3D points is inherently more likely to receive votes from wrong matches. In order to avoid bias towards database images with more visible 3D points, the original vote $|\mathcal{V}(d)|$ should be weighted by the number of 3D points that are seen by the database image d. Let $\mathcal{F}(d) = \{p \mid (p,d) \in \mathcal{E}\}$ be the 3D points observed by the database image d. The weighted votes $\mathcal{W}(d)$ of the database image d can be calculated as follows:

$$\mathcal{W}(d) = \frac{|\mathcal{V}(d)|}{|\mathcal{F}(d)|}. \tag{2}$$

In real-world scenes, there are various kinds of repetitive patterns, e.g., doors or windows, in a local region. It is possible that a query feature may establish multiple locally ambiguous 2D-3D matches in repetitive patterns. The locally ambiguous matches will falsely increase the weighted votes of the corresponding database images especially when the votes contain few correct matches. Unfortunately, the relaxed SIFT ratio test used in the initialization cannot entirely remove such locally ambiguous matches. In order to reduce the influence of the locally ambiguous matches in the database image voting procedure, we use an approach similar to Sattler *et al.* [13] to enforce that a query feature casts one unique vote to the same database image. Considering a query feature q that establishes local ambiguous matches to the database image d as $\{(q,p) \mid (p,d) \in \mathcal{E}\}$, we randomly choose one match from the locally ambiguous matches for casting vote to make sure that $\forall (q',p') \in \mathcal{V}(d) \setminus (q,p) : q \neq q'$.

C. Database Image Re-ranking

A database image with more weighted votes indicates that the corresponding matches are more likely to be correct. Thus

the outlier filtering problem can be formulated as solving an image retrieval problem. Given a query image, the database images are ranked according to the corresponding weighted votes. Among the top rank database images, special attention should be paid on those, which receive only one single vote from the established matches. In a city-scale dataset, it is common that some database images can only see a small number of 3D points due to low image resolution or viewpoint uniqueness. For such database images, a single vote could produce a large weighted vote value and a top rank. To recap, our core idea is based on the fact that correct matches are frequently co-visible and thereby vote to the same database image. Since the single vote does not exhibit any co-visibility feature, we first eliminate all database images with $|\mathcal{V}(d)| \leq 1$ from the database image list. The top K database images $\mathcal{D}^K$ are selected for outlier filtering. In addition, for the datasets with additional prior information such as GPS data, we can use the available data to further refine the ranking of database images. The Euclidean distance between the query image and each database image can be estimated using the associated GPS tags. We only select the top K database images whose Euclidean distances to the query image are below a threshold. In this paper, we set this threshold as 300 meters as suggested by Zeisl *et al.* [5]. To avoid misunderstanding, all distances mentioned below are Euclidean distances.

D. Outlier Filter and Match Augmentation

In previous approaches [4], [5], [14], an inappropriate assumption is that the matches rejected by the relaxed SIFT ratio test are all wrong matches. Here, we point out that even though wrong matches take up the majority of the rejected matches by the relaxed SIFT ratio test, a portion of correct matches are mistakenly rejected. It is meaningful and beneficial to recover correct matches back to further improve the quality of the matches. After retrieving the top rank database images $\mathcal{D}^K$, a match in $\mathcal{M}_f$, which casts a vote to one of the database image in $\mathcal{D}^K$, can be safely selected into $\mathcal{M}_v$ as follows:

$$\mathcal{M}_v = \left\{(q,p) \mid (q,p) \in \mathcal{M}_f, (p,d) \in \mathcal{E} \wedge d \in \mathcal{D}^K\right\}. \quad (3)$$

Moreover, for a match in $\mathcal{M} \setminus \mathcal{M}_f$ which also casts a vote to one of the database image in $\mathcal{D}^K$, it can be recovered as long as the associated query feature has not been found in $\mathcal{M}_v$ yet. Therefore, for each match in $(q',p') \in \mathcal{M} \setminus \mathcal{M}_f$, we iteratively select it into $\mathcal{M}_v$ if $\forall (q,p) \in \mathcal{M}_v : q \neq q'$. Note that the recovered matches from $\mathcal{M} \setminus \mathcal{M}_f$ are not involved in the previous database image voting procedure.

V. Geometry-based outlier filter

Having obtained the matches $\mathcal{M}_v$ using the visibility-based outlier filter in the first stage, we now propose to further filter wrong matches using geometrical considerations. Our key observation is that visual occlusion is a common phenomenon in a city-scale SfM point cloud. Therefore, there are a large number of *locally visible points*, which are only observed by database images whose camera positions lie nearby. Fig. 3 illustrates a typical example of a *locally visible point*. The restriction of cameras observing *locally visible points* enables us to derive a novel geometrical constraint that is simply based on the camera position. Different from traditional re-projection error measurement, the proposed geometrical constraint can serve as a more robust inlier evaluation measurement in RANSAC-based pose estimation, especially under large outlier ratio scenario. In this section, we will describe the proposed geometrical constraint and its application in detail.

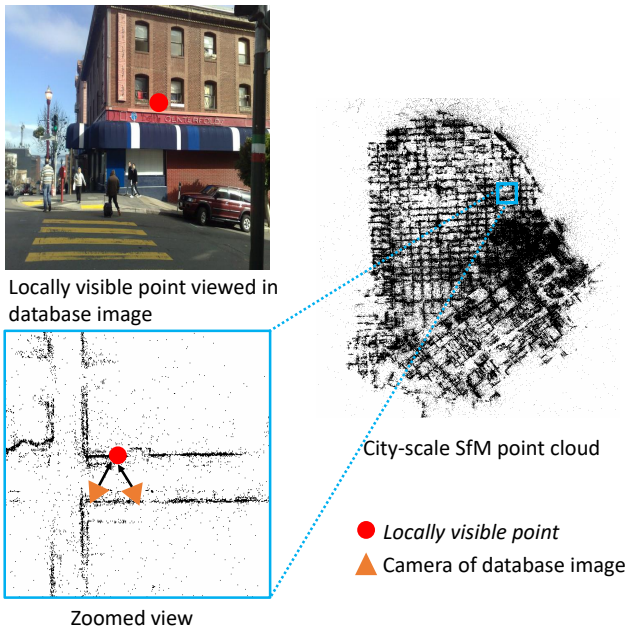


Fig. 3: An illustration of a *locally visible point* in the *San Francisco* dataset [4]. A *locally visible point* (red) is observed by nearby cameras (orange) of the database images.

A. A Data-driven Geometrical Constraint

In order to efficiently classify the *locally visible points*, we leverage the bipartite visibility graph $\mathcal{G}$. Let $\mathcal{I}(p)$ be the set of database images which observe the 3D point p as follows:

$$\mathcal{I}(p) = \{d \mid (p,d) \in \mathcal{E}\}. \quad (4)$$

Suppose $\mathcal{I}(p)$ is of size n, a 3D point p can be regarded as a *locally visible point* if the distance between p and the camera position of each database image in $\mathcal{I}(p)$ is below a defined distance threshold T_{local} as follows:

$$\forall c_i : \|c_i - p\|_2 \leq T_{local}, \quad (5)$$

where c_i represents the camera position of the i^{th} database image in $\mathcal{I}(p)$.

For each *locally visible point* in a SfM point cloud, we derive a geometrical constraint to restrict the position of a hypothetical camera, which can observe the *locally visible point*. We define a sphere of radius r around the *locally visible point* to represent the region that a hypothetical camera may appear. The radius should be smaller than the distance defined in Eq. 5 to ensure the locality. In addition, an adaptive radius can be defined based on the average camera-to-point distance from a *locally visible point* p to the camera position of each

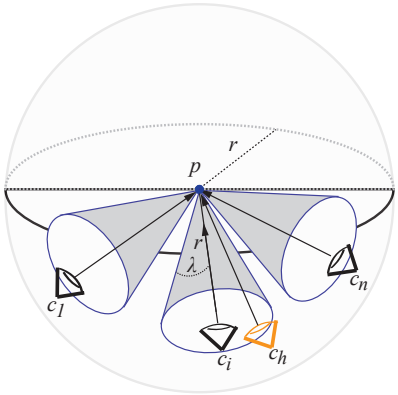


Fig. 4: The derived geometrical constraint for a *locally visible point* p. For each camera position of the database image which observes p, we define a cone with height r and angle λ. A hypothetical camera c_h which observes p should lie inside at least one of the defined cones.

database image in $\mathcal{I}(p)$. The average camera-to-point distance is calculated using the equation:

$$\text{dist}(p) = \frac{\sum_{i=1}^{n}(\|c_i - p\|_2)}{n} \tag{6}$$

where c_i represents the camera position of the i^{th} database image in $\mathcal{I}(p)$. For cases when the average camera-to-point distance is much smaller than the local distance threshold T_{local} in Eq. 5, we define an adaptive radius as $r = \alpha\text{dist}(p)$. Therefore, the radius of the sphere is $r = \min(\alpha\text{dist}(p), T_{local})$. In this paper, we empirically set $\alpha = 4$.

In addition, we apply the angle constraint [5], [13], [20] based on the view direction since the SIFT feature descriptor is variant to a significant viewpoint change. For each database image that can observe the *locally visible point* p, we compute the viewing direction as a normalized vector pointing from the camera position c_i to p. For a camera that observes p, the angle between the current viewing direction and the viewing direction from one of the database images should be smaller than an angle threshold λ. Therefore the final derived geometrical constraint $Constraint(c_h, p)$ can be defined as follows:

$$Constraint(c_h, p) = \begin{cases} \|c_h - p\|_2 < \min(\alpha\text{dist}(p), T_{local}) \\ \exists c_i : \measuredangle(\overrightarrow{c_h p}, \overrightarrow{c_i p}) < \lambda, \end{cases} \tag{7}$$

where c_i represents the camera position of the i^{th} database image in $\mathcal{I}(p)$. The derived geometrical constraint is illustrated in Fig. 4.

B. The Outlier Filter

In order to apply the derived geometrical constraint to filter outliers, a hypothetical camera position needs to be established. Assuming that the camera's internal calibration matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$ of a query image is known in advance, we utilize a P3P pose solver [10] to establish a hypothetical camera pose $\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}] \in \mathbb{R}^{3\times4}$, where $\mathbf{R}$ represents the rotation matrix and $\mathbf{t}$ is the translation vector. The hypothetical camera position can be computed as $\mathbf{c} = -\mathbf{R}\mathbf{t}$. Given the matches $\mathcal{M}_v$ generated after the visibility-based outlier filter, our goal is to find the camera position that is most likely to observe the 3D points associated with correct matches in $\mathcal{M}_v$. To this end, we adopt a standard RANSAC scheme [11] to verify multiple camera position hypotheses. In each RANSAC iteration, the matches corresponding to *locally visible points* are regarded as inliers if they satisfy the geometrical constraint in Eq. 7. The matches corresponding to *non-locally visible points* are evaluated using the traditional re-projection error measurement as follows:

$$\|q - \mathbf{P}p\|_2 \leq \gamma. \tag{8}$$

The match corresponding to a *non-locally visible point* can be regarded as an inlier if the re-projection error is below the pixel threshold γ. Using the above hybrid inlier evaluation scheme, the camera model $\mathbf{P}^*$ with the largest number of inliers is returned. The inliers of $\mathbf{P}^*$ therefore are selected as $\mathcal{M}_g$ for the final pose estimation. The geometry-based outlier filter is summarized in Algorithm. 1.

By incorporating the derived geometrical constraint for *locally visible points*, our geometry-based outlier filter is efficient in handling matches with large outlier ratio for two reasons. Firstly, traditional inlier evaluation method based on re-projection error requires an accurate 6-DOF camera pose. While in the proposed geometry-based outlier filter, the inlier evaluation for matches corresponding to *locally visible points* is relaxed since it only requires an approximate 3-DOF camera position. Secondly, for a query image that depicts a local scene, a P3P sample with three inliers is able to produce a theoretically correct camera position which lies nearby *locally visible points* corresponding to inliers. Inspired from Camposeco *et al.* [14], we observe that a P3P sample with only two inliers is able to produce an approximate camera position

Algorithm 1 The Geometry-based Outlier Filter

Require: $\mathcal{M}_v$, matches selected by the visibility-based outlier filter; $\mathcal{M}_\mathcal{L} \subseteq \mathcal{M}_v$, matches corresponding to the set of *locally visible points* $\mathcal{L}$.
Require: The camera internal matrix $\mathbf{K}$; re-projection error threshold γ; maximum RANSAC iterations $Iter$.

```
Inlier_max ← 0
for j = 0; j < Iter do
    Randomly sample three matches from M_v
    Compute the rotation matrix R and the translation vector t using P3P solver
    Obtain the projection matrix P = K[R|t] and the camera center c = −Rt
    Inliers I_1 = Re-projection(M_v \ M_L, P, γ)
    Inliers I_2 = Constraint(c, p), p ∈ L
    if |I| ≥ Inlier_max then
        P* ← P, Inlier_max ← |I|
    end if
    j ← j + 1
end for
return The inliers of P* as M_g
```

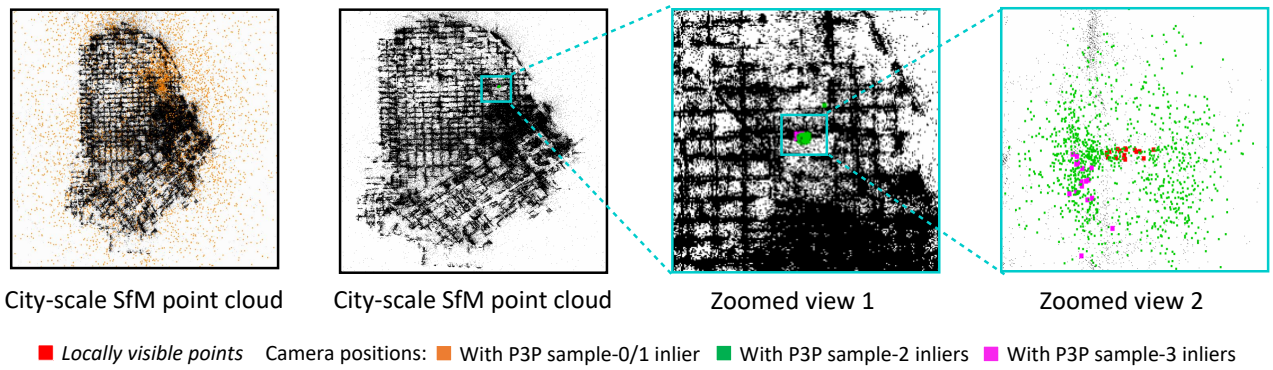


Fig. 5: The distribution of camera positions in the geometry-based outlier filter for a query image that depicts a local scene. The camera positions with P3P samples (0 or 1 inlier) distributed throughout the whole SfM point cloud. It looks like many of these are clearly wrong, e.g., in the ocean. The distribution shows that a P3P sample with 2 inliers, which is much easier to be obtained than a P3P sample with 3 inliers under large outlier ratio scenarios, can provide us an approximate camera position to apply the proposed geometrical constraint. The data was generated by randomly sampling 10^5 trials using the image in Fig. 3.

which lies nearby the theoretically correct camera positions. Fig. 5 shows an example of the camera position distribution with a query image that depicts a local scene. This relaxation on the number of inliers in a P3P sample significantly increases the probability of finding an approximate camera position to apply the derived geometrical constraint.

Suppose the inlier ratio of established 2D-3D matches is ε, the probability of obtaining a P3P sample with two inliers (P3P-2i) can be computed as ε^2. The traditional re-projection error measurement requires an accurate 6-DOF camera projection matrix, which is computed by a P3P sample with three inliers (P3P-3i). The probability of obtaining a P3P-3i sample can be computed as ε^3. Considering a large outlier ratio case, e.g., $\varepsilon < 0.1$, the probability of obtaining a P3P-2i sample is much larger than obtaining a P3P-3i sample by a factor of $1/\varepsilon$. Therefore, the proposed geometrical constraint for locally visible points is more robust under large outlier ratio scenario comparing with traditional re-projection error measurement.

VI. Experiments

We evaluate the proposed two-stage outlier filtering framework on two popular real-world datasets: the *San Francisco* dataset [4], [27] and the *Dubrovnik* dataset [32]. Table I summarizes the statistics of the datasets used in our experiments. The *San Francisco* dataset consists of 1.06 million street-view database images for image retrieval tasks. For 3D structure-based localization, we use the publicly available SF-0 SfM point cloud [4], which is built from 610k database images in the *San Francisco* dataset. The query images have a different spatial distribution compared to the database images, making feature matching in the *San Francisco* dataset difficult. In addition, each database image is associated with a precise GPS coordinate. The query images also are associated with GPS coordinates, in which some are not very precise. As far as we know, the *San Francisco* dataset is the most challenging dataset for 3D structure-based localization so far. Therefore, we mainly focus on evaluating our approach on this dataset.

TABLE I: The statistics of the datasets used in our experiments.

Dataset	Database images	3D points	Query images
San Francisco (SF-0)	610k	30.34M	803
Dubrovnik	6k	1.89M	800

The *Dubrovnik* dataset has been widely studied by [4], [32], [33], [35] and almost all query images can be localized. Similar to recent works [5], [14], we mainly focus on evaluating the pose accuracy on the *Dubrovnik* dataset. For a comprehensive comparison, we include the state-of-the-art approaches from three categories as follows:

- 3D structure-based approaches: Active search [35], Co-occurrence [4], KVD [12], CPV [5], Hyperpoints [13] and Toroidal [14].
- Hybrid localization approaches which combine 2D image-based and 3D structure-based approaches: DenseVLAD + SfM [40].
- Learning-based localization approach: PoseNet with novel geometrical loss functions (GLF), abbreviated as PoseNet (GLF) [46].

A. Evaluation on the San Francisco Dataset

1) Implementation Details: In the feature matching step, we use the FLANN library [48] for approximate nearest neighbor searching between a query image and the SF-0 SfM point cloud. For fair comparison, we follow the 1-to-N matching strategy used in existing works [5], [12]. For each query feature, at most 3 matches will be established. In the initialization, a match is verified with a variable search threshold, which is defined as 0.7 times the squared distance to the nearest neighbor in the query image itself. In the proposed visibility-based outlier filter, we empirically select the top 200 database images and perform the match augmentation scheme with the selected database images. In the proposed geometry-based outlier filter, we empirically set the distance threshold

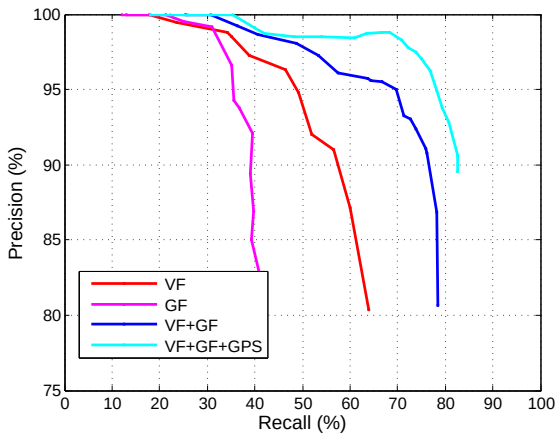


Fig. 6: The experimental results of our method on the *San Francisco* dataset.

TABLE II: The average match statistics of successfully localized query images in different stages of VF+GF in the *San Francisco* dataset.

Matches	$\mathcal{M}_f$	$\mathcal{M}_v$	$\mathcal{M}_g$	Final Inliers
Stage	Input	+VF	+VF+GF	+VF+GF+P3P
#Matches	4528	287	92	40
#Matches_correctIDs	76	102	84	38
%Matches_correctIDs	1.8%	30.4%	87.8%	90.2%

$T_{local} = 50$ meters and the angle threshold $\lambda = 60°$. Note that we used the same parameter setting in both the *San Francisco* dataset and the *Dubrovnik* dataset. We run a maximum of 1000 RANSAC iterations in the geometry-based outlier filter.

2) Evaluation Criteria: In the *San Francisco* dataset, all database images and query images are annotated with the ground truth building IDs. A query image is considered to be successfully localized if the final inliers are registered to the ground truth building IDs. We use the improved version of ground truth annotations reported by Arandjelović *et al.* [23]. Note that there are 66 query images whose ground building IDs are missing in the SF-0 point cloud. We use the same evaluation criteria as customary used in previous work [4], [5], [13], [27] that the performance is evaluated as the recall rate under 95% precision.

3) Overall Evaluation: Our method includes two major modules: the proposed visibility-based filter, abbreviated as VF and the proposed geometry-based filter, abbreviated as GF. In order to separately evaluate the impact of each module, we conduct several experiments on the *San Francisco* dataset with the following settings:

- VF: only use the visibility-based outlier filter.
- GF: only use the geometry-based outlier filter.
- VF+GF: use the visibility-based outlier filter and the subsequent geometry-based outlier filter.
- VF+GF+GPS: use the visibility-based filter and the subsequent geometry-based filter. Incorporate the GPS data in the visibility-based filter as described in Section IV-C.

After obtaining the set of matches using the above experimental settings, we use P3P-RANSAC [10] to compute the final 6-DOF camera pose. Fig. 6 reports the experimental results using the above settings. For each setting, multiple recall@precision results are generated by varying the inlier threshold to determine whether a query image is successfully localized. We notice that GF achieves the worst performance among all settings. The reason is that the original matches are very noisy, i.e., below 1% inlier ratio. RANSAC used in GF requires too many iterations to find a reliable solution with such extremely noisy matches. The significant gain of VF+GF over VF indicates that the matches generated from VF may still contain a large number of outliers, which GF can remove efficiently. With VF+GF, we achieve a 69.6% recall at 95% precision. By incorporating the provided GPS data, the relevance between then selected top rank database images and the query image has been significantly improved. VF+GF+GPS can provide us a 78.1% recall at 95% precision.

TABLE III: The comparison of our method with the state-of-the-art works on the *San Francisco* dataset. All the listed recall rates are measured at a 95% precision rate. The Vertical and Height assumptions mean that the camera's vertical direction with respect to the underlying SfM point cloud and the camera's approximate height are known in advance.

Method	Geometrical Assumptions	Recall Rate [%]	
		w/o GPS	w/ GPS
KVD [12]	Vertical+Height	68.0	-
CPV+P3P [5]	Vertical+Height	67.5	74.2
CPV [5]	Vertical+Height	68.7	73.7
Co-occurrence [4]	-	54.2	-
Hyperpoints [13]	-	61.9	-
Our method	-	**69.6**	**78.1**

In Table II, we report the average match statistics of successfully localized query images using our full prior-free pipeline VF+GF. Since it is difficult to determine the number of inliers in the original matches with extremely large outlier ratios, we use the number of matches which are registered to the correct building IDs as an approximate upper bound of inliers. The ratio of the matches with correct building IDs among the whole matches can be used to evaluate the quality of the matches. The matches $\mathcal{M}_f$ after the initialization step have a very large outlier ratio, which make the pose estimation difficult. After applying VF, the quality of matches is significantly improved from a 1.8% ratio to 30.4% ratio. With the matching augmentation procedure in VF, the number of matches with correct building IDs increases from 76 to 102. However, for some query images the matches still contain a large number of wrong matches, which make VF obtain a lower recall rate compared with VF+GF. Due to the relaxation in both hypothesis and verification phase of RANSAC, GF is able to efficiently handle the matches with large outlier ratio, which are generated by VF. After VF+GF, the matches with correct building IDs are well preserved and the ratio significantly increases from 30.4% to 87.8%.

4) Comparison with state-of-the-art: Table III reports the comparison between our method and the state-of-the-art approaches. The performance is evaluated by the recall at 95% precision, which was also used by related works [4], [5], [13], [27]. Our method outperforms state-of-the-art 3D Structure-based methods in scenarios without and with GPS. Without additional assumptions about the camera's vertical direction

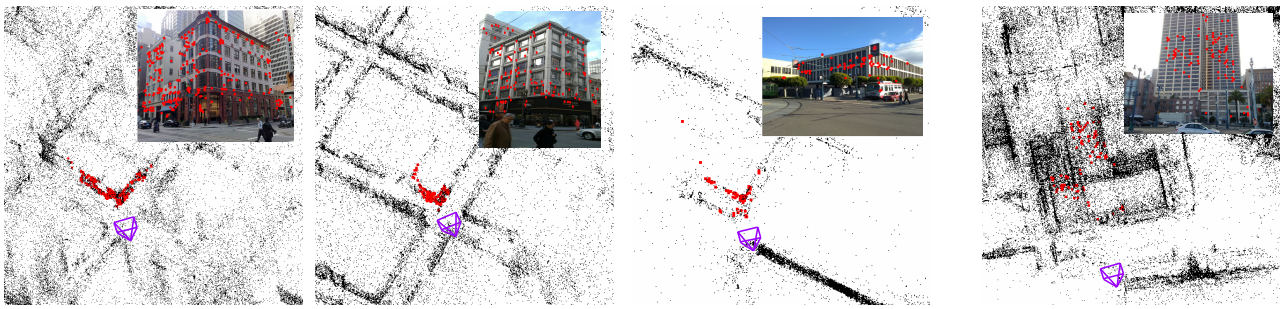
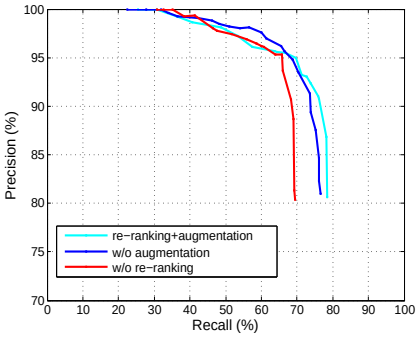


Fig. 7: The exemplary query images and the corresponding estimated 6-DOF camera poses in the SF-0 SfM point cloud for the *San Francisco* dataset.



(a) The performance comparison to evaluate the re-ranking scheme in Section IV-C and the matching augmentation scheme in Section IV-D.

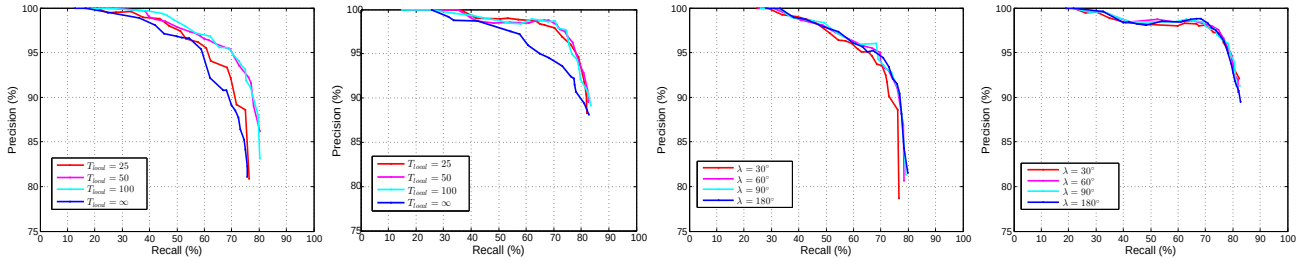
(b) First row: the query images whose ground truth building ID annotations exist in the SF-0 SfM point cloud, the bounding box shows the corresponding pier marks of different building IDs. Second row: the falsely localized query images caused by the matching augmentation scheme. Their ground truth building ID annotations are missing in the SF-0 SfM point cloud. They are falsely registered to the same building IDs as images in the first row due to nearly identical appearance.

Fig. 8: The ablation study of the proposed visibility-based outlier filter (VF) on the *San Francisco* dataset.

and approximate height relative to the SfM point cloud, our method (GF+VF) achieves a 69.6% recall at 95% precision. By incorporating the GPS data, the recall at 95% precision increases to 78.1%. The localization performance achieved by our method (VF+GF) proves that the visibility intrinsics and geometry intrinsics in a city-scale SfM point cloud are not mutually exclusive and can be combined to remove outliers. Comparing with the 2D image-based approaches [18], [23], [26], our method is able to provide a 6-DOF camera pose for a query image, as illustrated by Fig. 7. The Burstness [26] approach, which is 2D image-based, achieves a 72.4% recall at 95% precision. Note that in the Burstness approach [26], they leverage the original 1.06 million database images while the SF-0 SfM point cloud used in our method only contains the information of 610k database images. In addition, the GPS data of database images are leveraged for clustering locations.

5) Ablation Study of VF: To evaluate the impact of each individual component of the visibility-based outlier filter (VF), we conduct an ablation study on the *San Francisco* dataset with different VF schemes. Fig. 8a presents the experimental results of the re-ranking scheme in Section IV-C and the match augmentation scheme in Section IV-D. We can notice that the re-ranking scheme improves the performance significantly. This improvement indicates that the top rank database images after re-ranking are more relevant to the query image, and are more likely to contain correct matches. The improvement of the recall rate proves that the match augmentation method is able to recover correct matches back that were previously removed. However, there is a drop of precision rate in the high precision regime ($> 95\%$). We found that the majority of the additional falsely localized query images caused by the match augmentation scheme are due to missing building ID annotations as shown in Fig. 8b. The falsely localized query images with missing building IDs are registered to other locations with nearly identical appearances. In such cases, the 2D-3D matches recovered by the match augmentation step still have high reliability to ensure the consistency between the 2D query features and the features associated with the matched 3D points.

6) Ablation Study of GF: To evaluate the impact of each individual component of the geometry-based outlier filter (GF), we conduct an ablation study on the *San Francisco* dataset by varying the distance threshold T_{local} of the derived geometrical constraint in Eq. 7. Fig. 9a shows the experimental results using different T_{local} settings. All points are classified as *locally visible points* with $T_{local} = \infty$. We can notice that this setting significantly decreases the localization performance. The main reason is that for *non-locally visible points*



(a) T_{local}, without GPS (b) T_{local}, with GPS (c) λ, without GPS (d) λ, with GPS

Fig. 9: The ablation study of the proposed geometry-based outlier filter (GF) on the *San Francisco* dataset. Different distance thresholds T_{local} in meter in both with GPS and without GPS scenarios.

TABLE IV: The statistics of *locally visible points* with different distance thresholds T_{local} in the *San Francisco* and *Dubrovnik* dataset.

The San Francisco Dataset				
T_{local}	25	50	100	∞
#Locally visible points	16.68M	24.60M	28.21M	30.34M
%Locally visible points	55%	81%	93%	100%
The Dubrovnik Dataset				
T_{local}	25	50	100	∞
#Locally visible points	0.27M	0.87M	1.24M	1.89M
%Locally visible points	14%	46%	66%	100%

which can be seen by distant cameras, applying the derived geometrical constraint will result in that many wrong matches can easily satisfy the hybrid inlier evaluation measurement. The resultant matches with $T_{local} = \infty$ usually have a large outlier ratio, which make P3P-RANSAC difficult to obtain a reliable solution.

Looking at Fig. 9, it is necessary to define an appropriate distance threshold T_{local} to ensure that the derived geometrical constraint is accurate for evaluating inliers with respect to *locally visible points*. To achieve this goal, we evaluate three distance thresholds. The statistics of *locally visible points* in the *San Francisco* dataset is shown in Table IV. We can notice that by setting $T_{local} = 50m$, 81% of 3D points are classified as *locally visible points*, which is compliant with the characteristics of the *San Francisco* dataset since most of the database images depict street-view scenes. As can be seen in Fig. 9a, by setting $T_{local} = 50m$ or $T_{local} = 100m$, our method achieves a significantly gain in both recall and precision comparing with $T_{local} = \infty$. This proves the benefit of the hybrid inlier evaluation measurement in GF. By setting $T_{local} = 25m$, the localization performance is worse than $T_{local} = 50m$ or $T_{local} = 100m$. The reason is that under such setting, several points that should be *locally visible points* are classified as *non-locally visible points* instead, thereby need the classic re-projection error measurement. Comparing with the inlier evaluation measurement using the derived geometrical constraint, the efficiency of classic re-projection error measurement relies more heavily on the quality of matches.

We also evaluate different T_{local} settings when incorporating the GPS data in VF as shown in Fig. 9b. The matches generated by VF usually have a larger inlier ratio than without GPS scenario. Therefore, the difference of performance among

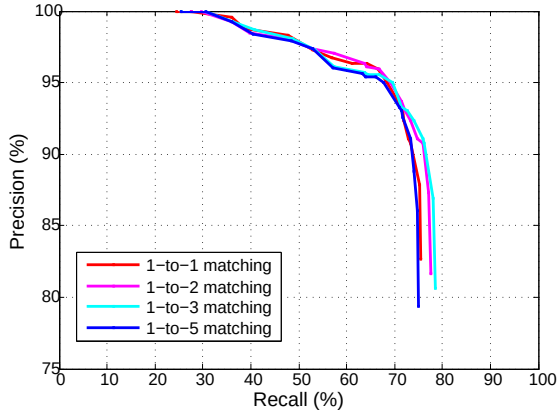


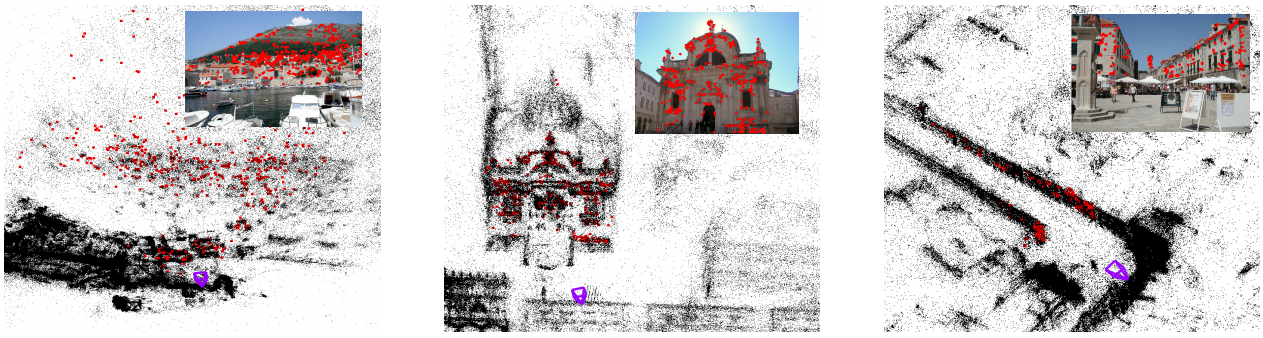
Fig. 10: The localization performances using different 1-to-N matching schemes.

$T_{local} = 25, 50, 100$ is smaller than the cases without GPS. Applying the derived geometrical constraint to all points with $T_{local} = \infty$ still achieves the worst localization performance. We also evaluate the impact of the angle constraint in GF by varying the angle threshold λ as shown in Fig. 9c and Fig. 9d. There is a noticeable performance drop when $\lambda = 30°$, which indicates that this threshold is too strict and may reject correct matches. From the ablation study, we can notice that the derived geometrical constraint based on the distances between the camera positions and the *locally visible points* plays a major role in the geometry-based outlier filter.

7) Scalability and Efficiency: To evaluate the scalability of our method, we conduct a experiment by varying the number of nearest neighbors in the 1-to-N matching scheme as shown in Fig. 10. As can be seen, finding only one nearest neighbor per query feature achieves the lowest recall due to insufficient correct matches. Among all the cases, $N = 2$ or $N = 3$ achieve the best performance, since these seem to provide a good balance between preserving correct matches and rejecting wrong matches. In general, our method shows its effectiveness in dealing with the very large outlier ratio scenario with multiple 1-to-N matching schemes. With 1-to-3 matching scheme, the computational time for the two-stage outlier filter (VF+GF) is close to 0.1 second.

TABLE V: The comparison between our method and the state-of-the-art works on the *Dubrovnik* dataset.

Method	Query Image Statistics			Localization Error e [meter]			Geometrical Assumptions	Time [sec]
	#Images	$e < 18.3$	$e > 400$	1^{st} Quarter	Median	3^{rd} Quarter		
P3P-RANSAC	628	596	11	1.30	5.46	8.28	-	11.8
Active search [35]	796	704	9	0.4	1.40	5.30	-	0.25
KVD [12]	798	771	3	-	0.56	-	Vertical and Height	5.06
CPV [5]	798	725	2	0.75	1.69	4.82	Vertical and Height	3.78
CPV+P3P [5]	796	744	7	0.19	0.56	2.09	Vertical and Height	-
CPV+P3P+BA [5]	794	749	13	0.18	0.47	1.73	Vertical and Height	-
Toroidal [14]	800	739	8	0.22	1.07	2.99	-	9.7
DenseVLAD + SfM [40]	-	-	-	0.30	1.00	5.10	-	∼200
PoseNet (GLF) [46]	-	-	-	-	7.9	-	-	0.005
Our method (Scheme 1)	794	745	4	0.29	0.69	2.15	-	2.6
Our method (Scheme 2)	797	749	3	0.28	0.70	2.10	-	1.4



Localization error = 2.46m Localization error = 0.24m Localization error = 0.92m

Fig. 11: The exemplary query images with corresponding estimated 6-DOF camera poses and localization errors of the *Dubrovnik* dataset.

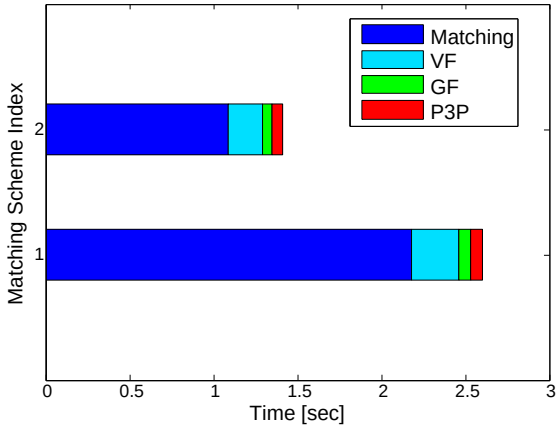


Fig. 12: The computational time of our method with Scheme 1 and Scheme 2 on the *Dubronik* dataset (also reported in Table V).

B. Evaluation on the Dubrovnik Dataset

In order to fairly compare with existing works, we adopt two feature matching schemes on the *Dubrovnik* dataset as follows:

- Scheme 1: the 1-to-3 matching scheme which is used in CPV [5]. Each query feature can find at most three nearest neighbors in the SfM point cloud. An adaptive distance threshold which is defined by 0.7 times the squared distance to the nearest neighbor in the underlying query image is set to reject ambiguous matches.
- Scheme 2: each query feature can only find at most one nearest neighbor in the SfM point cloud, a squared distance ratio is set as 0.9 in the SIFT ratio test to reject ambiguous matches. This matching is the same with KVD [12] and Torodial [14].

We use the same evaluation criteria as [5], [12], [14], [35] to evaluate the localization result: a query image is successfully localized if the best camera pose returned by RANSAC has more than 11 inliers. The re-projection error threshold is set as 6 pixels. The pose accuracy can be measured with the ground truth 6-DOF camera poses provided by Li *et al.* [32]. In the *Dubrovnik* dataset, we select the top 20 database images in the first stage to apply the visibility-based outlier filter. Table V shows the results of our method and other related works on the *Dubrovnik* dataset. Under Scheme 2, we achieve a slightly better performance in both successfully localized images and pose accuracy compared with Scheme 1. This indicates that the matches established with Scheme 2 contain sufficient correct matches in the *Dubrovnik* dataset for an accurate pose estimation. Comparing with other methods that do not need any additional geometrical priors [14], [35], [40], [46], we achieve the state-of-the-art performance on the median and 3^{rd} quarter pose accuracy, and a comparable performance on the 1^{st} quarter pose accuracy comparing with the Toroidal approach [14]. Comparing with the methods [5], [12] that rely on the assumption of the camera's vertical direction and approximate height, we are able to achieve competitive results.

Fig. 11 shows the exemplary estimated 6-DOF camera poses in the *Dubrovnik* dataset using our method. In addition, our method has the third lowest computational time among all existing methods. The Active search [35] method is efficient by establishing at most 100 2D-3D matches for each query image, which in the meantime reduces the pose accuracy. Fig. 12 gives the details of our method's computational time. As can be seen, the feature matching step occupies the majority of the computational time. The proposed visiblity-based outlier filter (VF) and the geometry-based outlier filter (GF) can be efficiently executed in less than half a second.

VII. Conclusion

In this paper, we have proposed a two-stage outlier filtering framework that consists of an improved visibility-based outlier filter and a subsequent novel geometry-based outlier filter. In the first stage, we have demonstrated that through database image re-ranking and match augmentation, the performance of the visibility-based outlier filter can be significantly boosted. In the second stage, we have derived a novel data-driven geometrical constraint that is useful in generating a set of fine-grained matches. With a comprehensive evaluation on two real-world city-scale SfM datasets, we have demonstrated the effectiveness and efficiency of the proposed two-stage outlier filtering framework in very large outlier ratio scenarios.

VIII. Acknowledgement

This research is partially supported by the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative. This research is also partially supported by Singapore Ministry of Education Tier-2 Fund MOE2016-T2-2-057(S).

References

[1] N. Snavely, S. M. Seitz, and R. Szeliski, "Photo tourism: exploring photo collections in 3d," in *ACM transactions on graphics (TOG)*, vol. 25, no. 3. ACM, 2006, pp. 835–846.

[2] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz, "Multi-view stereo for community photo collections," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[3] J. Heinly, J. L. Schonberger, E. Dunn, and J.-M. Frahm, "Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset)," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3287–3295.

[4] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3d point clouds," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 15–29.

[5] B. Zeisl, T. Sattler, and M. Pollefeys, "Camera pose voting for large-scale image-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2704–2712.

[6] 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, 2017.

[7] T. Sattler, W. Maddern, A. Torii, J. Sivic, T. Pajdla, M. Pollefeys, and M. Okutomi, "Benchmarking 6dof urban visual localization in changing conditions," vol. 2, 2018.

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

[9] M. Bujnak, Z. Kukelova, and T. Pajdla, "A general solution to the p4p problem for camera with unknown focal length," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[10] L. Kneip, D. Scaramuzza, and R. Siegwart, "A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2969–2976.

[11] M. A. Fischler, "Random sampling concensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[12] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson, "City-scale localization for cameras with known vertical direction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 7, pp. 1455–1461, 2017.

[13] T. Sattler, M. Havlena, F. Radenovic, K. Schindler, and M. Pollefeys, "Hyperpoints and fine vocabularies for large-scale location recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2102–2110.

[14] F. Camposeco, T. Sattler, A. Cohen, A. Geiger, and M. Pollefeys, "Toroidal constraints for two-point localization under high outlier ratios," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4545–4553.

[15] J. Li, X. Qian, Y. Y. Tang, L. Yang, and T. Mei, "Gps estimation for places of interest from social users' uploaded photos," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 2058–2071, 2013.

[16] R. Ji, L.-Y. Duan, J. Chen, T. Huang, and W. Gao, "Mining compact bag-of-patterns for low bit rate mobile visual search," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 3099–3113, 2014.

[17] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 1808–1817.

[18] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 883–890.

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

[20] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *European conference on computer vision*. Springer, 2008, pp. 304–317.

[21] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[22] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–7.

[23] R. Arandjelović and A. Zisserman, "Dislocation: Scalable descriptor distinctiveness for location recognition," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 188–204.

[24] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," in *European Conference on Computer Vision*. Springer, 2010, pp. 748–761.

[25] P. Turcot and D. G. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 2109–2116.

[26] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys, "Large-scale location recognition and the geometric burstiness problem," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1582–1590.

[27] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvä, K. Roimela, X. Chen, J. Bach, M. Pollefeys *et al.*, "City-scale landmark identification on mobile devices," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 737–744.

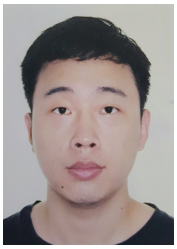
[28] Z. Liu, H. Li, W. Zhou, R. Zhao, and Q. Tian, "Contextual hashing for large-scale image search," *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1606–1614, 2014.

[29] 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, 2012.

[30] R. Arandjelovic and A. Zisserman, "All about vlad," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 1578–1585.

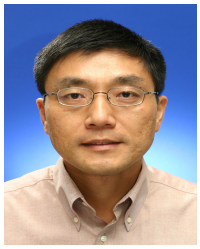
[31] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5297–5307.
[32] Y. Li, N. Snavely, and D. P. Huttenlocher, "Location recognition using prioritized feature matching," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 791–804.
[33] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2d-to-3d matching," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 667–674.
[34] S. Choudhary and P. Narayanan, "Visibility probability structure from sfm datasets and applications," in *European conference on computer vision*. Springer, 2012, pp. 130–143.
[35] T. Sattler, B. Leibe, and L. Kobbelt, "Improving image-based localization by active correspondence search." Springer, 2012, pp. 752–765.
[36] Y. Feng, L. Fan, and Y. Wu, "Fast localization in large-scale environments using supervised indexing of binary features," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 343–358, 2016.
[37] S. Cao and N. Snavely, "Minimal scene descriptions from structure from motion models," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 461–468.
[38] W. Cheng, W. Lin, X. Zhang, M. Goesele, and M.-T. Sun, "A data-driven point cloud simplification framework for city-scale image-based localization," *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 262–275, 2017.
[39] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof, "From structure-from-motion point clouds to fast location recognition," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2599–2606.
[40] 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 *CVPR 2017-IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
[41] Y. Yang, C. Feng, Y. Shen, and D. Tian, "Foldingnet: Point cloud auto-encoder via deep grid deformation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 3, 2018.
[42] H. Su, V. Jampani, D. Sun, S. Maji, E. Kalogerakis, M.-H. Yang, and J. Kautz, "Splatnet: Sparse lattice networks for point cloud processing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2530–2539.
[43] H. Deng, T. Birdal, and S. Ilic, "Ppfnet: Global context aware local features for robust 3d point matching," *Computer Vision and Pattern Recognition (CVPR). IEEE*, vol. 1, 2018.
[44] 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*, 2015, pp. 2938–2946.
[45] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using lstms for structured feature correlation," in *Int. Conf. Comput. Vis.(ICCV)*, 2017, pp. 627–637.
[46] A. Kendall, R. Cipolla *et al.*, "Geometric loss functions for camera pose regression with deep learning," in *Proc. CVPR*, vol. 3, 2017, p. 8.
[47] E. Brachmann and C. Rother, "Learning less is more-6d camera localization via 3d surface regression," in *Proc. CVPR*, vol. 8, 2018.
[48] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2227–2240, 2014.
[49] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image retrieval for image-based localization revisited." in *BMVC*, vol. 6, 2012, p. 7.



Wentao Cheng received the B.E. degree in Computer Science and Engineering from Harbin Institute of Technology in 2012. He is currently pursuing the Ph.D. degree from the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include image-based localization and 3D SfM point cloud simplification.

Kan Chen is a Research Fellow in Fraunhofer Singapore. Prior to that, he held various Game Development/Education roles, include Game Programmer in Koei Tecmo and Assistant Professor in Digipen Institute of Technology. He received a B.Comp (Honors) in Computer Science from National University of Singapore, an M.Eng and a Ph.D. in Computer Engineering from Nanyang Technological University. His research interests include Computer Graphics, Computer Vision and Human-Computer Interaction.



Weisi Lin (M92–SM98–F16) received his Ph.D. from Kings College, London University, U.K. He served as the Lab Head of Visual Processing, Institute for Infocomm Research, Singapore. Currently, he is an Associate Professor in the School of Computer Engineering. His areas of expertise include image processing, perceptual signal modeling, video compression, and multimedia communication, in which he has published 170 journal papers, 230+ conference papers, filed 7 patents, and authored 2 books. He is an AE for IEEE Trans. on Image Processing, and IEEE Trans. Circuits and Systems for Video Tech. He has been a Technical Program Chair for IEEE ICME 2013, PCM 2012, and QoMEX 2014. He chaired the IEEE MMTC Special Interest Group on QoE (2012-2014). He has been an invited/panelist/keynote/tutorial speaker in 20+ international conferences, as well as a Distinguished Lecturer of IEEE Circuits and Systems Society 2016-2017, and Asia-Pacific Signal and Information Processing Association (APSIPA), 2012-2013. He is a Fellow of IEEE and IET, and an Honorary Fellow of Singapore Institute of Engineering Technologists.

Michael Goesele received the Diploma degree in computer science from Ulm University and the Ph.D. degree from the MPI Informatik and Saarland University. After a postdoctoral stay at the University of Washington as Feodor Lynen Fellow funded by the Alexander von Humboldt Foundation, he joined the Department of Computer Science of Technische Universitt Darmstadt. He recently became a research scientist at Facebook

Xinfeng Zhang (M16) received the B.S. degree incomputer science from the Hebei University of Technology, Tianjin, China, in 2007, and the Ph.D. degreein computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2014. From 2014 to 2017, he was a Research Fellow with the Rapid-Rich Object SearchLab, Nanyang Technological University, Singapore. From 2017 to 2018, he was a Post-Doctoral Fellow with the School of Electrical Engineering System, Universityof Southern California, LosAngeles, CA, USA. He currently is a Research Fellow with the department of Computer Science, City University of Hong Kong. He has authored over 100 technical papers in important conferences and journals. His research interests include image and video processing, image and video compression.

Yabin Zhang received the B.E. degree in Electronic Information Engineering in the Honors School, Harbin Institute of Technology and the Ph.D. degree from the School of Computer Science and Engineering, Nanyang Technological University, Singapore in 2013 and 2018, respectively. He is currently a senior researcher in Media Lab, Tencent, Shenzhen. His research interests include video coding, image/video processing, image quality assessment and computer vision.