Privacy Preserving Localization and Mapping from Uncalibrated Cameras

Marcel Geppert[1] Viktor Larsson[1] Pablo Speciale[2] Johannes L. Schönberger[2] Marc Pollefeys[1,2]
[1] Department of Computer Science, ETH Zurich [2] Microsoft

Abstract

Recent works on localization and mapping from privacy preserving line features have made significant progress towards addressing the privacy concerns arising from cloud-based solutions in mixed reality and robotics. The requirement for calibrated cameras is a fundamental limitation for these approaches, which prevents their application in many crowd-sourced mapping scenarios. In this paper, we propose a solution to the uncalibrated privacy preserving localization and mapping problem. Our approach simultaneously recovers the intrinsic and extrinsic calibration of a camera from line-features only. This enables uncalibrated devices to both localize themselves within an existing map as well as contribute to the map, while preserving the privacy of the image contents. Furthermore, we also derive a solution to bootstrapping maps from scratch using only uncalibrated devices. Our approach provides comparable performance to the calibrated scenario and the privacy compromising alternatives based on traditional point features.

1. Introduction

The recent trend towards cloud-based localization and mapping systems for mixed reality and robotics (*e.g*., Microsoft Azure Spatial Anchors [25], Facebook LiveMaps [1], or Google VPS [48]) is largely driven by the need for scalable solutions to enable multi-device experiences and crowd-sourced mapping. However, as these systems primarily rely on acquiring imagery of the environment, this development has raised significant privacy concerns by the public [27, 44, 49, 62]. Existing works on privacy preserving localization and mapping are based on the concept of lifting traditional point-based features to lines to conceal the appearance of images [10, 13, 55, 57, 58].

The fundamental limitation of these existing works is that they assume calibrated cameras, which prevents their applicability in a wide range of scenarios. In practice, the intrinsic calibration of cameras used in mixed reality devices and robots continuously changes over time due to environmental impact such as temperature change or dropping the device. Re-calibration of devices by the manufacturer is typically very costly or technically very challenging using automatic techniques. Sometimes, it can even be desirable to hide the exact calibration parameters of a camera to protect the identity of the device from fingerprinting attacks. Due to these issues, the resulting changes to the calibration parameters can be quite dramatic and negatively influence the accuracy of camera localization and mapping, or even prevent the successful use of the data entirely. Furthermore, in crowd-sourced mapping scenarios from mobile cameras [2, 53, 54, 56], prior focal length estimates are typically quite inaccurate (*e.g*., due to auto-focus) or completely incorrect due to manual editing of the imagery by the user. Recent interest in crowd-sourced mapping for autonomous driving from dashboard cameras [63] is another application that suffers from challenges to calibrate the cameras. The windshield acts as an additional lens and drastically impacts the effective calibration parameters of the acquired imagery.

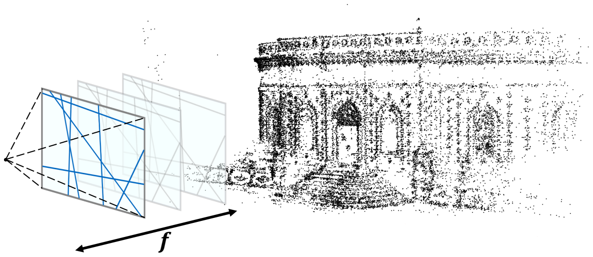

Figure 1: Our method jointly estimates absolute pose and focal length of uncalibrated cameras from privacy preserving line features. We also present a solution for bootstrapping maps from scratch using four views.

In this paper, we propose a principled solution to privacy preserving localization from uncalibrated cameras. Our approach self-calibrates the focal length and jointly estimates the absolute pose of a camera from only privacy preserving line features. Successfully localized cameras can contribute to existing maps in an incremental fashion without any further assumptions. In addition, we present a solution to initializing maps from scratch from only uncalibrated cameras. Our boostrapping approach requires a subset of consistently aligned lines similar to the work presented by Geppert *et al*. [13]. This assumption does not impact the amount of privacy preservation. We demonstrate the efficacy of our approach on a wide range of experiments on localization and the scenario of end-to-end crowd-sourced mapping. Our method achieves comparable performance to the calibrated case and to traditional point-based approaches.

The content in the paper is organized as follows: we first review related methods in the context of our work in Sec. 2 before describing our proposed solution to camera localization with unknown focal length from line features in Sec. 3.1. Sec. 3.2 then describes our approach to initializing maps from uncalibrated cameras, while Sec. 4 experimentally validates the efficacy of our proposed methods.

2. Related Work

In this section, we first review related work on privacy preserving methods in the context of localization and mapping. We then discuss existing approaches to uncalibrated localization and mapping with a particular focus on the absolute pose estimation problem.

2.1. Privacy Preserving Methods

With their work on privacy preserving image-based localization, Speciale *et al.* [57, 58] were the first to address privacy concerns in the context of localization and mapping. Their main idea is to lift 2D features in images and 3D points in the map to lines in order to conceal the appearance of the original images or the map, respectively. Localizing with 2D lines against a 3D point map unprojects the 2D lines to planes and solves a 3D point-to-plane alignment problem. A 2D points vs. 3D lines localization unprojects the 2D points to lines (rays) and can be seen as a generalized camera relative pose estimation. Note that it is not possible to combine the two approaches, *i.e.* lifting to lines in both 2D and 3D, since we cannot ensure that the random directions will be consistent between the image and the map. The following works by Geppert *et al.* [13] and Shibuya *et al.* [55] leverage the same concepts to tackle the full Structure-from-Motion (SfM) problem. Recently, Dusmanu *et al.* [10] propose an approach to preserve the privacy of local image features by lifting high-dimensional descriptors to affine subspaces. In our work, we rely on the same lifting-technique as previous approaches, but consider for localization and mapping for uncalibrated cameras.

2.2. Uncalibrated Localization and Mapping

Early methods on SfM for uncalibrated images [4, 12, 41, 46] served as the foundation for later works on mapping systems from unstructured and crowd-sourced imagery [52, 56] exhibiting extremely challenging scenarios for camera self-calibration. In the following years, research focused on scaling these approaches to thousands and millions of images [2, 19, 54] and further robustifying the self-calibration process [20, 21, 35, 39, 53, 60]. The arguably most reliable systems on self-calibrating SfM follow an incremental reconstruction paradigm [2, 19, 46, 53, 54, 56], while later global approaches also demonstrated impressive results [8, 60, 64]. Most recently, Geppert *et al.* [13] extend on traditional incremental systems [53] to build a privacy-preserving SfM system. The main limitation of their work lies in requiring calibrated images, which is a significant limitation in practice. We overcome this limitation by using a hybrid global-incremental approach, where we use global optimization during map initialization and then incrementally localize cameras into the map. Both the initialization as well as the localization stage are fully self-calibrating.

2.3. Absolute Pose Estimation

Recovering the camera pose w.r.t. a pre-built map is an important problem occurring in many vision applications, *e.g.*, SfM [53], visual localization [50], and SLAM [42]. If the camera's intrinsic calibration is known, this can be minimally estimated from three point-to-point correspondences [15, 43] (usually referred to as *P3P*). In the privacy preserving framework from Speciale *et al.* [58], the absolute pose problem is solved from 2D-line to 3D-point correspondences. While these offer weaker geometric constraints, requiring six correspondences instead of three, Speciale *et al.* show that it is possible to perform robust visual localization in this setting as well. The solver used in their work is based on the 3Q3 solver from Kukelova *et al.* [30].

If the intrinsic camera calibration is unknown, it can be estimated jointly with the pose using additional correspondences. If there is no constraints on the intrinsic parameters, the *Direct Linear Transform* (DLT) [18] can be used to linearly estimate the 3×4 pinhole camera matrix from at least 5.5 point correspondences.[1] If the camera has square-pixels (*i.e.*, zero skew and unit aspect ratio), this can be used to replace one point-correspondence, allowing estimation from 4.5 points. However, this introduces non-linear constraints on the camera matrix and the DLT approach no longer applies. Minimal solvers for this case were first introduced by Triggs [61] and recently improved in Larsson *et al.* [33].

For most consumer cameras, we further assume a centered principal point. Even if this is not exactly satisfied, small offsets in the principal point can be well compensated with small pose corrections in the x/y-translation. In this case, only the focal length remains to be estimated, and by centering the image coordinates, we can w.l.o.g. assume the calibration matrix to be diagonal, *i.e.*, $K = \text{diag}\,(f, f, 1)$. The pose estimation problem with unknown focal length was first solved by Bujnak *et al.* [5] from 4 point correspondences (P4Pf). Note that the problem is minimal with 3.5 points, and Josephson and Byröd [23] used this additional

[1] Here 0.5 means only using one of the two equations available from one of the 2D-3D correspondences.

0.5 point correspondence to estimate one radial distortion parameter (P4Pfr). This solver was later improved in the followup works by Bujnak *et al*. [6] and Larsson *et al*. [32]. In [29] Kukelova *et al*. presented a solver that relies on 5 point correspondences (P5Pf), but offsets the larger sample size by greatly improving the runtime. In addition to estimating focal length, this solver can also estimate up to three distortion parameters. This solver was later extended in Larsson *et al*. [34] to more general distortion models.

There has also been a series of works on improving upon the original P4Pf solver from Bujnak *et al*. [5]. Zheng *et al*. [68] used a similar approach but managed to greatly improve the runtime and stability. Wu [66] presented the first solver that was able to leverage the minimal set of 3.5 points instead 4. Kukelova *et al*. [30] showed that P4Pf can be formulated as a 3Q3 problem and proposed a very efficient solver for this. Larsson *et al*. [32] proposed an alternative P3.5Pf solver which avoids the degeneracies in [66]. If the focal length is approximately known (*e.g*., from EXIF data), Sattler *et al*. [51] showed that sampling based-approaches can also work very well. The work most related to ours is by Kuang and Åström [28], which considers absolute pose with unknown focal length from combinations of point-to-point, line-to-line, and so-called quiver correspondences. While they do not consider the case of 2D-line-to-3D-point correspondences necessary for the privacy preserving setting, it is possible to adapt their approach to this case. Still, their rotation parameterization degenerates for any 180° rotation and requires additional steps to handle these degeneracies.

The absolute pose problem has also been considered for the case of 2D-line-to-3D-line correspondences, sometimes referred to as *Perspective-n-Lines (PnL)* in the literature. The first solutions were given by Dhome et al. [9] and Chen [7]. Since then there have been various works which improve the solvers in different aspects. For example, [3] proposed a linear estimator that can use multiple correspondences. This was later improved in [40] which instead solve for the global minimizer of a non-linear cost. Xu et al. [67] investigated the special cases for P3L and provided a complete solution for these. The case of known vertical direction was studied in [36] for perspective cameras, and in [22] for generalized cameras.

3. Method

A localization and mapping pipeline generally consists of multiple building blocks: feature extraction and matching, image pose estimation, point triangulation, and bundle adjustment. To create a new map from scratch, an additional map initialization step is used. In our setup, both the feature extraction and matching part and bundle adjustment are trivial extensions of the standard case. We extract standard SIFT [38] features and simply lift the keypoints to lines. Hereby we usually create randomly oriented lines, but add consistently aligned lines for initialization as explained in Sec. 3.2. For bundle adjustment we replace the standard 2D point-to-point distance with a 2D point-to-line distance. For point triangulation we rely directly on the calibrated method presented in [13], using the already obtained estimates for the focal length and ignoring any additional image distortion. This solution lifts the 2D lines to 3D planes and finds the observed 3D points by intersecting corresponding planes of at least three images. Consequently, we focus our explanations on the remaining building blocks. We detail our new minimal solver for image pose estimation from line features with unknown focal length in Sec. 3.1. We then explain explain the map initialization technique for this setup in Sec. 3.2.

3.1. Privacy Preserving Localization with Unknown Focal Length

In the privacy preserving framework [13, 58], each 2D keypoint is replaced with a randomly oriented line passing through it. During pose estimation, each 2D line to 3D point correspondence constrains the camera pose as

$$\boldsymbol{\ell}^T(R\boldsymbol{X}+\boldsymbol{t}) = 0 \ , \tag{1}$$

where $\boldsymbol{\ell}$ is the homogeneous representation of the 2D line. This, however, assumes that the lines $\boldsymbol{\ell}$ are given in the normalized image plane requiring known calibration. If the lines are defined in image space, the corresponding equation is instead (disregarding any distortion)

$$\boldsymbol{\ell}^T K(R\boldsymbol{X}+\boldsymbol{t}) = 0 \ , \tag{2}$$

where K is the intrinsic calibration matrix. The extra non-linearly introduced by the K matrix prevents us from using the same 3Q3-based approach as in [13, 58].

Minimal Solver for Absolute Pose. In this section, we present a minimal solver for uncalibrated absolute pose estimation from line-to-point correspondences. We assume square pixels and centered principal point, *i.e*., only the focal length is unknown, $K = \text{diag}(f, f, 1)$. This is a common assumption in the literature (see *e.g*. [5, 29, 32, 34, 68]) that holds for most cameras.

Each 2D-line-to-3D-point correspondence only yields a single constraint and thus we need 7 correspondences to get a minimal problem. Rewriting equation (2) in terms of the camera matrix $P = K[R\ \boldsymbol{t}] \in \mathbb{R}^{3\times 4}$, we obtain

$$\boldsymbol{\ell}^T P \begin{bmatrix}\boldsymbol{X}\\ 1\end{bmatrix} = 0 \implies \left(\begin{bmatrix}\boldsymbol{X}^T & 1\end{bmatrix} \otimes \boldsymbol{\ell}^T\right)\text{vec}(P) = 0 \tag{3}$$

where $\otimes$ denotes the Kronecker product. Stacking the constraints from 7 correspondences, we get

$$A\boldsymbol{p} = 0, \quad \text{where} \quad A = \begin{bmatrix}\begin{bmatrix}\boldsymbol{X}_1^T & 1\end{bmatrix} \otimes \boldsymbol{\ell}_1^T \\ \vdots \\ \begin{bmatrix}\boldsymbol{X}_7^T & 1\end{bmatrix} \otimes \boldsymbol{\ell}_7^T\end{bmatrix} \in \mathbb{R}^{7\times 12} \ . \tag{4}$$

Note that if we had 11 point correspondences, we could solve linearly for the camera matrix P by finding the nullspace to A. This would be equivalent to DLT [18] for normal pose estimation and does not enforce any constraints on the intrinsic calibration. Instead, we recover a basis for the nullspace $N \in \mathbb{R}^{12\times 5}$. Any camera matrix which satisfies the 7 correspondences can then be written as some linear combination of these basis vectors, i.e. $\boldsymbol{p} = N\left(\boldsymbol{\alpha}^T, 1\right)^T$, where we have fixed the scale by setting the last coefficient to one. To recover the unknown coefficients $\boldsymbol{\alpha}$, we want to ensure that the first 3x3 block of P corresponds to $\operatorname{diag}(f, f, 1)\, R$, where R is a rotation matrix. This can be enforced by requiring the three rows to be pairwise orthogonal and that the first two rows have the same norm. These constraints were already used in [6]. However, as pointed out in [32], this introduces spurious complex solutions (where $p_{11}^2 + p_{12}^2 + p_{13}^2 = 0$), increasing the complexity of the solver. To avoid these spurious solutions, the authors proposed to use additional polynomial constraints [32]. We provide these constraints in the supplementary material for completeness. Inserting the nullspace parameterization into these equations yields new equations in $\boldsymbol{\alpha}$. Using the framework from [31], we create a Gröbner basis solver for these equations. Similar to the P3.5Pf solver [32], the problem has 10 solutions. At runtime, the solver requires linear elimination on a 25×35 matrix followed by solving a 10×10 eigenvalue problem.

3.2. Privacy Preserving Map Initialization

In practice, it is not always possible to extend an existing map, but a new map needs to be created from scratch (*e.g.*, when devices visit a location for the first time). Our proposed initialization scheme extends the approach in Geppert *et al.* [13] to also work in the uncalibrated setting. As in their work, we require correspondences in four views to estimate relative poses, but we do not require calibrated gravity measurements. While we still need to align features for the initialization there is no constraint on the direction as long as it is used consistently in all views. We detect lines in the images, robustly estimate vanishing points to find a dominant direction in the scene, and then create the lines to intersect both the corresponding keypoint and the vanishing point. In practice, the vertical direction is still often the easiest to detect robustly. We assume the images to be coarsely aligned with gravity (*i.e.*, within 45°) to distinguish between vertical or horizontal vanishing points in scenes with multiple dominant directions.

Our initialization begins by solving a projective 2D-SfM problem that is used to convert a subset of the line correspondences into pairwise 2D-2D point matches. These point matches are then used in an approach similar to non-incremental SfM methods [45, 60] to globally recover the poses for the four images used for initialization. We use a simple heuristic based on the number of feature matches to select the initialization image set. This is detailed in the supplementary material.

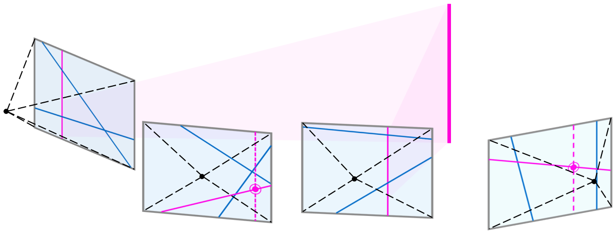

Figure 2: ***Initialization***. We solve a projective 2D-SfM problem with consistently aligned lines. Points in the projective reconstruction project onto aligned lines in the images. Intersecting random lines with the aligned projections allow us to recover the 2D key points for a subset of the original line correspondences.

Projective 2D-SfM on the Horizon. Similar to Geppert *et al.* [13], we require a mix of consistently aligned and randomly oriented lines. For simplicity we assume a vertical alignment based on a detected vertical vanishing point. Following [13], we can interpret the vertically aligned lines as 1D-measurements. Note that these vertically aligned lines are projections of vertical lines in 3D, and are generally not vertical in the image. We transform each image's features into a synthetic, horizontally aligned image, which now also has vertical lines in image space. From this we can extract the 1D measurements as the features' horizontal position in the image. The goal is now to solve a 2D SfM problem with the 1D measurements. Since the internal calibration of the cameras is unknown, we only aim to recover a projective reconstruction in this step. However, we will show that this is sufficient for our task.

Using RANSAC [11], we estimate the uncalibrated 2D trifocal tensor [47] for the first three images. Factorizing the tensor and triangulating points gives us a projective 2D reconstruction of the first three cameras. We resect the fourth image to this reconstruction using RANSAC, followed by projective bundle adjustment to refine the reconstruction.

To get more meaningful reprojection errors for the 1D cameras we propagate all projections back into the original image and compute the errors along a horizontal line, passing through the image center and orthogonal to the observed vertical direction.

Recovering Image Keypoints. While working in the 2D setup we can triangulate a (2D) point from two images (instead of 3 in the 3D case), but only from aligned features. With the relative 2D poses of 4 images known, we can select correspondences with 2 aligned and 2 randomly oriented lines. We triangulate the point in 2D from the aligned lines, and then project it back to the other 2 images. Propagating the projections back into the original image as before, we now have a random (measured) line and an aligned (projected) line for the correspondence in those two images. Consequently, the original 2D keypoint position can be recovered by simply intersecting both lines, which it is illustrated in Fig. 2. To obtain more stable keypoint positions, we enforce a minimal angle of 45° between aligned and random lines in images where both types are present.

Triangulating all possible points and projecting them in all images allows us to convert a subset of the line correspondences into pairwise 2D-2D point correspondences. Since we are only interested in the projections, we do not need to upgrade the 2D-projective reconstruction to metric (since the projections are independent of this).

Note that we are only able to recover 2D-points that could be triangulated at a later stage in the reconstruction pipeline. Thus, recovering the keypoint locations does not reveal any additional information compared to a keypointless initialization method.

Initialization of Image Rotations and Focal Lengths from Recovered Keypoints. We can now use the previously recovered keypoints to initialize the focal lengths and camera rotations. We estimate pairwise fundamental matrices using RANSAC between all four images used for initialization. While it is possible to recover focal lengths directly from each fundamental matrix [16, 17, 26], this problem is notoriously unstable in two views, partly due to the problem degenerating for intersecting principal axes [24]. Instead, we take an approach similar to Sweeney *et al.* [60].

For noise-free data, the essential matrix

$$E_{ij}(f_i, f_j) = \text{diag}(f_j, f_j, 1) F_{ij} \text{diag}(f_i, f_i, 1) \quad (5)$$

satisfies

$$\|E_{ij} E_{ij}^T\|^2 - \frac{1}{2}\|E_{ij}\|^4 = 0 \quad (6)$$

where f_i and f_j are the focal lengths. Similar to [60], we setup an optimization over the focal lengths, optimizing the pairwise consistency of the focal lengths over all pairs in the image set used for initialization. Specifically, we consider

$$\min_{f_1,\ldots,f_4} \sum_{i,j} \|\boldsymbol{r}_{ij}(f_i, f_j)\|^2 \quad (7)$$

$$\text{where } \boldsymbol{r}_{ij} = \left(\frac{\|E_{ij} E_{ij}^T\|^2}{\|E_{ij}\|^4} - \frac{1}{2}, \frac{\|E_{ij}\|^4}{\|E_{ij} E_{ij}^T\|^2} - 2 \right) \quad (8)$$

In [60], they directly minimized the residual from (6). However, we found that the scale-invariant version in (8) worked slightly better (see supplementary material). In our experiments we initialize the focal lengths as the image widths.

Once the focal lengths are recovered, we factorize the essential matrices to recover the relative orientations. We perform rotation averaging by randomly sampling minimum spanning trees (*c.f.* [14, 45]) to assign each rotation.

Estimation of Translations. Given the focal lengths and the camera rotations, we now aim to estimate the translations. This is performed in a similar style to the upgrade step in Geppert *et al.* [13], except that the translation is not constrained to be along the detected dominant direction.

Let $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_4$ be a line correspondence normalized by the estimated focal lengths. If this is an inlier correspondence, it should satisfy

$$\begin{bmatrix} \boldsymbol{\ell}_1^T R_1 & \boldsymbol{\ell}_1^T \boldsymbol{t}_1 \\ \vdots & \vdots \\ \boldsymbol{\ell}_4^T R_4 & \boldsymbol{\ell}_4^T \boldsymbol{t}_4 \end{bmatrix} \begin{pmatrix} \boldsymbol{X} \\ 1 \end{pmatrix} = 0 \quad (9)$$

for some 3D point $\boldsymbol{X}$. This implies that the 4×4 matrix above is rank-deficient and thus it's determinant vanishes. Setting $\boldsymbol{t}_1 = 0$ and expanding the determinant w.r.t. the other translation vectors, we obtain the following

$$\begin{bmatrix} D_{134}\boldsymbol{\ell}_2^T & -D_{124}\boldsymbol{\ell}_3^T & D_{123}\boldsymbol{\ell}_4^T \end{bmatrix} \begin{pmatrix} \boldsymbol{t}_2 \\ \boldsymbol{t}_3 \\ \boldsymbol{t}_4 \end{pmatrix} = 0 \quad (10)$$

$$\text{where } D_{ijk} = \det\left(R_i^T \boldsymbol{\ell}_i \ \ R_j^T \boldsymbol{\ell}_j \ \ R_k^T \boldsymbol{\ell}_k \right). \quad (11)$$

Each line correspondence gives us a linear constraint on the translation vectors. Collecting the constraints from at least 8 correspondences into a matrix, we can recover the translations via SVD. In case of more than 8 correspondences, we can solve it as a homogeneous least squares problem.

We wrap the above solver in a RANSAC loop; randomly selecting samples of 8 line correspondences and estimating translations from these. We validate the samples by triangulating the 3D points and measuring the line reprojection error in the images. Note that this step does not rely on the recovered 2D keypoints, but directly uses the original line correspondences.

Finally, we bundle adjust over all inliers in the 4 views to refine their camera parameters by minimizing the line reprojection error in the images.

4. Experiments

4.1. Evaluation of Absolute Pose Solver

In this section, we evaluate our proposed minimal solver for camera resectioning with unknown focal length from line correspondences (Sec. 3.1). For the experiments, we generate synthetic problems by uniformly sampling 2D points in a 2000x2000 image, a field-of-view in the interval [45,90] degrees, and computing the corresponding focal length in pixels. For each 2D point, we uniformly sample a depth value in the interval [0.1, 100] and back-project it. Next, we sample a random camera pose and apply the inverse transform to the 3D points. We then optionally add Gaussian zero-mean noise to each keypoint. Finally, we replace each keypoint with a random line passing through it.

	Runtime	Real solutions
P3.5Pf [32]	20.1 µs	4.4 / 10
P4Pf [30]	4.2 µs	3.9 / 8
P5Pf [29]	1.7 µs	2.2 / 4
L7Pf (Cayley) [28]	1513.1 µs	8.5 / 20
L7Pf (Nullspace)	23.7 µs	4.2 / 10

Table 1: ***Runtime analysis.*** The table shows the median runtime for the synthetic stability experiment as well as the average number of real solutions. Note that the Cayley parameterization returns duplicate solutions with flipped sign on the focal length which can easily be filtered. For comparison we also show the runtime of the state-of-the-art point-based solvers.

First, we evaluate the numerical stability of the solver on noise-free instances. We compare with a modified version of the solver from Kuang and Åstrom [28] that uses 7 line-to-point correspondences, denoted *L7Pf (Cayley)* in Fig. 3, which presents the residuals in the estimated focal length for 1000 instances. This shows that the nullspace-based parameterization presented in Sec. 3.1 provides a more stable estimate. Table 1 shows the average number of solutions and runtime for this experiment. We can see that the solver based on the Cayley parameterization (as in [28]) is significantly slower.

Next, we evaluate the noise-sensitivity of the proposed solvers. We generate instances with varying standard deviation for the keypoint-noise and report the errors in the estimated poses/focal length. For comparison, we also include the results with state-of-the-art point-based solvers: *P3.5Pf* (Larsson *et al.* [31]), *P4Pf* (Kukelova *et al.* [30]) and *P5Pf* (Kukelova *et al.* [29]), applied to the point correspondences. The results are shown in Fig. 4, and we can see that the line-based solvers have similar noise-sensitivity as the point-based pose solvers.

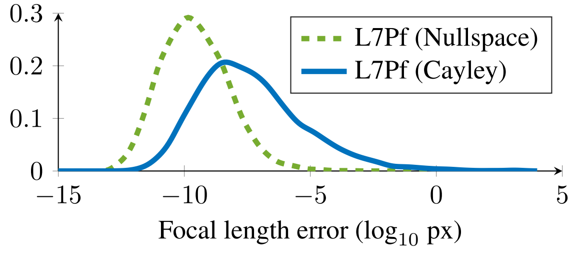

Figure 3: ***Numerical Stability***. The graph shows the distribution of the $\log_{10}$ focal length errors for noise free data.

4.2. Privacy Preserving Visual Localization

In this section, we evaluate privacy-preserving visual localization and self-calibration using the single-image evaluation protocol and setup from Speciale *et al.* [58]. They

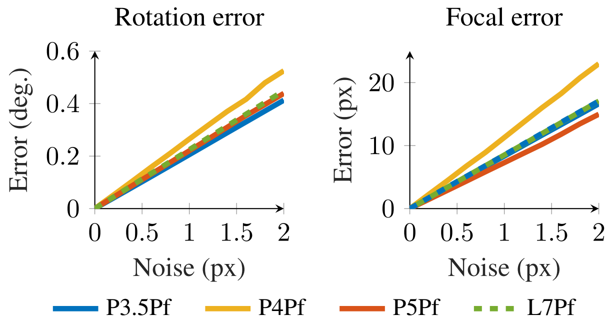

Figure 4: ***Noise Sensitivity.*** The graphs show the median errors in the rotation (*Left*) and the focal length (*Right*) for varying noise levels. The errors in the translation are qualitatively similar and can be found in the supplementary material.

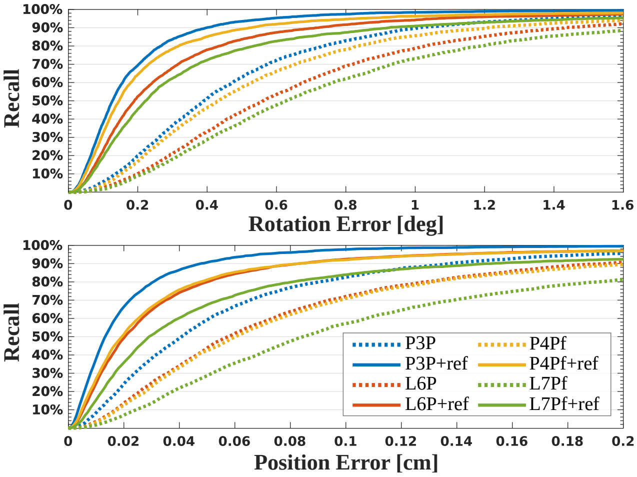

Figure 5: ***Cumulative Rotation and Position Errors***. Results for a mix of Mobile Phones and Microsoft HoloLens datasets from [58]. *+ref* indicates an additional, non-linear refinement.

evaluate on 15 datasets captured by a mix of mobile phones and Microsoft HoloLens. We replicate this experiment in the uncalibrated setting, estimating the focal length in addition to the camera pose. We compare both with methods that have access to the ground-truth intrinsic calibration (*P3P* and *L6P* [58]), as well as point-based solvers that estimate focal length (*P3.5Pf* [31], *P4Pf* [30] and *P5Pf* [29]). In our experiments, the point-based solvers for unknown focal length performed similarly and we only report the result for *P4Pf* [30] (see supplementary material). Fig. 5 shows the cumulative distributions of the rotation and translation errors. The figure shows that the uncalibrated line solver has reasonable performance compared to the corresponding calibrated solver *L6P* [58], which uses the ground-truth calibration. For an evaluation on Internet Photo Collection datasets on these four methods, please refer to Fig. 6 to see the performance as box plot graphs, validating again that the uncalibrated case behaves similarly to the correspond-

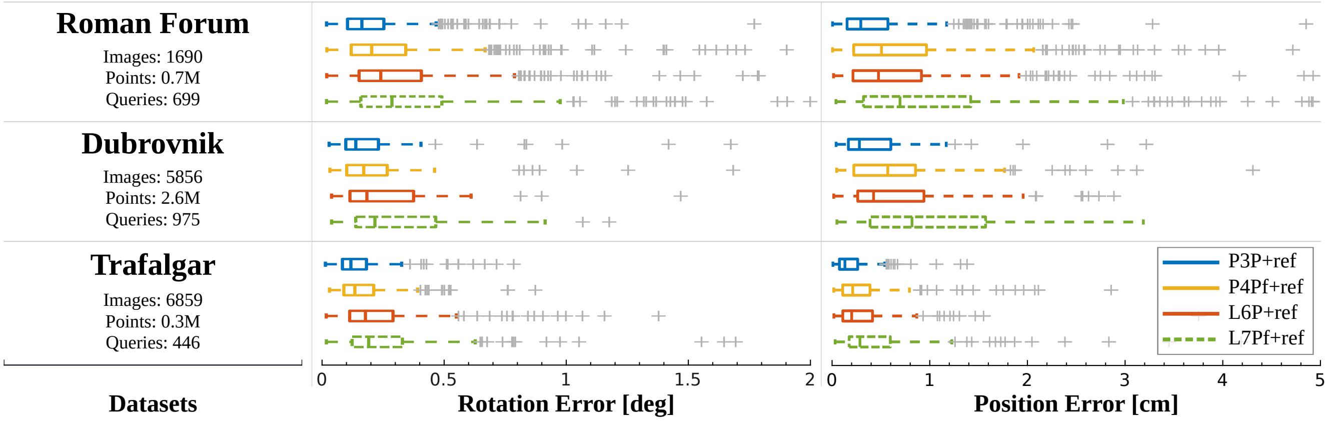

Figure 6: ***Internet Photo Collection.*** Localization results after non-linear refinement on three datasets from [37,65]. Comparison are made against a pre-built COLMAP reconstruction, where all the poses are refined together with bundle adjustment as opposed to single-image localization of the shown methods. The performance loss of the uncalibrated *L7Pf* solver – with more points in the RANSAC loop and parameters to be estimated – is similar to the loss from the traditional *P3P* method to its privacy-preserving counterpart *L6P*.

ing calibrated one. Since there is no ground-truth for these scenes, we compare against COLMAP [53] results.

4.3. Evaluation of Initialization

We evaluate our initialization method (Sec. 3.2) on a collection of Mobile Phone datasets from Speciale *et al.* [57]. Relative pose estimation from images is generally less stable than absolute pose estimation w.r.t. a point cloud. However, we do not require perfect relative image poses to initialize successfully. It is sufficient, if the initial scene structure is reconstructed well enough such that we can register more images. The additional constraints will stabilize the reconstruction using repeated bundle adjustments. To account for this initial uncertainty, we run the initialization 100 times for each scene and then continue to run the mapping pipeline up to 50 registered images. Fig. 7 shows the recall for different thresholds on the mean image position error of each test. We report tests where we cannot register 50 images as failure cases. The figure shows that the initialization scheme works very well for the scenes *Bedroom*, *Sofa* and *Stable*. For *Lobby*, only around 50% of the initialization trials are successful. This is caused by the relatively difficult scene with many repetitive structures and plants in the foreground. For the successful trials, the accuracy is still comparable to the other scenes. The camera poses in *Gatehouse* are generally slightly more noisy, especially along the principal axis direction, likely caused by the dominant planes in the scene. We provide results for the full reconstructions of the presented scenes in the supplementary material. In Sec. 4.4, we evaluate the full SfM pipeline on large crowd-sourced photo collections and show that we are able to initialize even in these challenging conditions with extremely heterogeneous image sets.

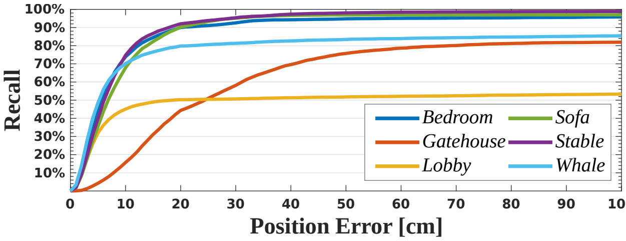

Figure 7: ***Initialization evaluation.*** The figure shows the cumulative position errors for the initialization evaluation.

4.4. Structure-from-Motion Evaluation

Finally, we evaluate the combination of all components by replicating two of the experiments presented in Geppert *et al.* [13] for calibrated privacy-preserving SfM. All experiments use a camera model with a single focal length and one radial distortion parameter (which is zero-initialized and refined in the bundle adjustment).

First, we consider the datasets from Strecha *et al.* [59]. While all images in these datasets were captured with a single camera, we do not enforce this during reconstruction, but allow each camera to have distinct intrinsic parameters. Table 2 shows the results. Compared to [13], we see slightly larger position errors and two additional scenes where not all images could be registered. The *castle-P30* scene has one image with correspondences mainly on a flat wall, which leads to a poorly constrained focal length, resulting in the large average errors.

To demonstrate the performance of our method with difficult input data, we use the internet datasets from Wilson and Snavely *et al.* [64]. As there is no reliable ground-truth available for these datasets, we do not report errors but reconstruction statistics, and compare these to the results with known camera calibrations reported by Geppert *et al.* [13].

	#Images		#Points		Track Length	Rotation (deg)			Position (cm)			Focal Length (%)		
Scene	Total	Reg.	3D	2D		Mean	Std.	Median	Mean	Std.	Median	Mean	Std.	Median
castle-P19	19	14	4.3k	24.1k	5.6	0.7	0.1	0.7	20.3	24.3	9.5	0.5	0.5	0.4
castle-P30	30	28	11.0k	74.8k	6.8	1.7	5.7	0.6	169.6	690.9	9.2	4.5	18.7	0.2
entry-P10	10	10	3.6k	22.4k	6.2	0.5	0.1	0.5	7.4	3.4	7.2	0.6	0.3	0.7
fountain-P11	11	10	6.8k	38.6k	5.7	0.4	0.0	0.4	0.8	0.3	0.7	0.3	0.2	0.3
Herz-Jesu-P8	8	8	3.4k	17.3k	5.2	0.5	0.0	0.5	1.4	0.7	1.5	0.2	0.1	0.2
Herz-Jesu-P25	25	25	11.1k	86.3k	7.8	0.4	0.1	0.4	2.8	2.0	2.6	0.4	0.2	0.4

Table 2: ***Camera Pose Accuracy***. Evaluation on the Strecha benchmark [59]. The large mean error for castle-P30 was primarily caused by a single outlier pose.

		#Registered Images		#Points		#Observations		Mean Track Length		Median Point Reproj. Error (px)		Median Line Reproj. Error (px)	
Scene	#Images	[13]	Ours	[13]	Ours	[13]	Ours	[13]	Ours	[13]	Ours	[13]	Ours
Alamo	2915	750	711	79k	84k	1730k	1722k	21.9	20.5	0.66	0.68	0.52	0.32
Gendarmenmarkt	1463	810	773	83k	77k	958k	853k	11.5	11.0	0.88	0.85	0.34	0.34
Madrid Metropolis	1344	377	365	43k	34k	447k	399k	10.4	11.7	1.13	0.70	0.40	0.29
Tower of London	1576	608	502	93k	87k	1122k	1008k	12.0	11.6	0.54	0.63	0.23	0.28

Table 3: ***Structure-from-Motion on Internet Photo Collections.*** The table shows reconstruction statistics for some of the 1D SfM datasets from Wilson and Snavely [64] compared to the calibrated privacy preserving reconstruction from [13]. *#Observations* is the number of 2D features that observe a 3D point. Note that the point reprojection error is only given as a reference and is not available to the algorithm.

While our method usually registers and triangulates slightly fewer images and points, the results are generally comparable. Note that the final results depend on the choice of parameters. We selected the thresholds to obtain clean reconstructions with few outlier images for all scenes. By relaxing some constraints and allowing more spurious image poses, the number of images and points could be increased significantly. We show the reconstructed point cloud based on the *Gendarmenmarkt* and *Tower of London* datasets [65] as qualitative results in Fig. 8 and provide the results for the remaining datasets in the supplementary material.

5. Conclusion

With this work, we make another major step towards enabling fully privacy preserving localization and mapping services. Removing the requirement for known calibration enables these services on the majority of devices without tedious and complicated calibration procedures. We present a new minimal solver for privacy preserving camera pose estimation with unknown focal length and demonstrate comparable performance to its classical counterparts. For a full solution to the SfM problem, we also present an initialization method using a hybrid global-incremental optimization approach. Experiments on multiple challenging localization and mapping scenarios underline the practical relevance of our work.

Acknowledgements: Viktor Larsson was supported by an ETH Zurich Postdoctoral Fellowship. We thank the anonymous reviewers for their valuable feedback.

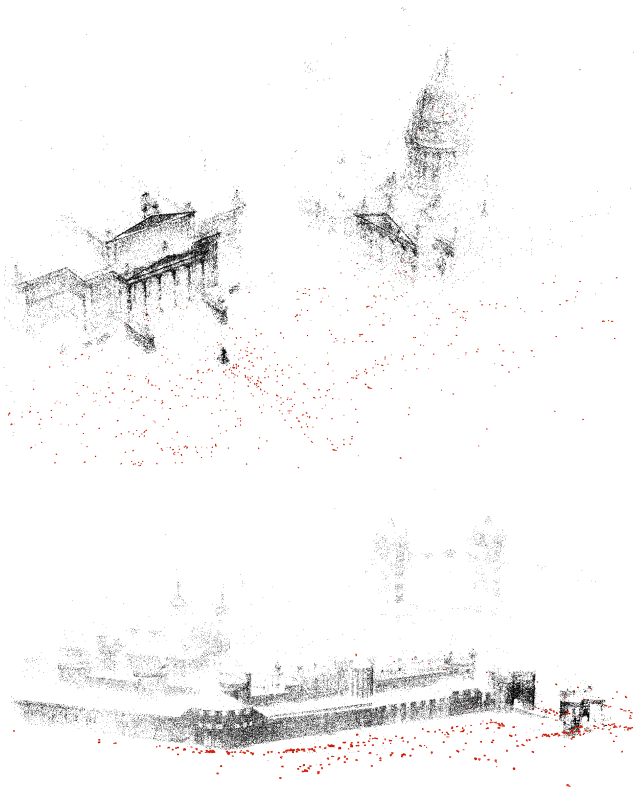

Figure 8: ***Qualitative result.*** The figure shows the reconstruction for the *Gendarmenmarkt* (*Top*) and *Tower of London* (*Bottom*) datasets [65] using our proposed method.

References

[1] Inside Facebook Reality Labs: Research updates and the future of social connection. https://tech.fb.com/inside-facebook-reality-labs-research-updates-and-the-future-of-social-connection/, 2019.

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven Seitz, and Richard Szeliski. Building rome in a day. In *International Conference on Computer Vision (ICCV)*, 2009.

[3] Adnan Ansar and Konstantinos Daniilidis. Linear pose estimation from points or lines. *Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 2003.

[4] Sylvain Bougnoux. From projective to euclidean space under any practical situation, a criticism of self-calibration. In *International Conference on Computer Vision (ICCV)*, 1998.

[5] Martin Bujnak, Zuzana Kukelova, and Tomas Pajdla. A general solution to the P4P problem for camera with unknown focal length. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[6] Martin Bujnak, Zuzana Kukelova, and Tomas Pajdla. New efficient solution to the absolute pose problem for camera with unknown focal length and radial distortion. In *Asian Conference on Computer Vision (ACCV)*, 2010.

[7] Homer H Chen. Pose determination from line-to-plane correspondences: existence condition and closed-form solutions. *IEEE Computer Architecture Letters*, 1991.

[8] David Crandall, Andrew Owens, Noah Snavely, and Daniel P. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[9] Michel Dhome, Marc Richetin, J-T Lapreste, and Gerard Rives. Determination of the attitude of 3d objects from a single perspective view. *Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 1989.

[10] Mihai Dusmanu, Johannes L. Schönberger, Sudipta N. Sinha, and Marc Pollefeys. Privacy-preserving image features via adversarial affine subspace embeddings. In *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[11] 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 (CACM)*, 1981.

[12] Andrew Fitzgibbon and Andrew Zisserman. Automatic camera recovery for closed or open image sequences. In *European Conference on Computer Vision (ECCV)*, 1998.

[13] Marcel Geppert, Viktor Larsson, Pablo Speciale, Johannes L. Schönberger, and Marc Pollefeys. Privacy preserving structure-from-motion. In *European Conference on Computer Vision (ECCV)*, 2020.

[14] Venu Madhav Govindu. Robustness in motion averaging. In *Asian Conference on Computer Vision (ACCV)*, 2006.

[15] Johann August Grunert. Das pothenotische problem in erweiterter gestalt nebst über seine anwendungen in geodäsie. *Grunerts Archiv fur Mathematik und Physik*, 1841.

[16] Richard Hartley. Estimation of relative camera positions for uncalibrated cameras. In *European Conference on Computer Vision (ECCV)*, 1992.

[17] Richard Hartley. Extraction of focal lengths from the fundamental matrix. *Unpublished manuscript*, 1993.

[18] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[19] Jared Heinly, Johannes L. Schönberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world* in six days *(as captured by the yahoo 100 million image dataset). In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[20] Je Hyeong Hong and Christopher Zach. pOSE: Pseudo object space error for initialization-free bundle adjustment. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[21] Je Hyeong Hong, Christopher Zach, Andrew Fitzgibbon, and Roberto Cipolla. Projective bundle adjustment from arbitrary initialization using the variable projection method. In *European Conference on Computer Vision (ECCV)*, 2016.

[22] Nora Horanyi and Zoltan Kato. Generalized pose estimation from line correspondences with known vertical direction. In *International Conference on 3D Vision (3DV)*, 2017.

[23] Klas Josephson and Martin Byröd. Pose estimation with radial distortion and unknown focal length. In *Computer Vision and Pattern Recognition (CVPR)*, 2009.

[24] Fredrik Kahl and Bill Triggs. Critical motions in euclidean structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, 1999.

[25] Neena Kamath. Announcing Azure Spatial Anchors for collaborative, cross-platform mixed reality apps. https://azure.microsoft.com/en-us/blog/announcing-azure-spatial-anchors-for-collaborative-cross-platform-mixed-reality-apps/, 2019.

[26] Kenichi Kanatani, Atsutada Nakatsuji, and Yasuyuki Sugaya. Stabilizing the focal length computation for 3-D reconstruction from two uncalibrated views. *International Journal of Computer Vision (IJCV)*, 2006.

[27] Alex Kipman. Azure Spatial Anchors approach to privacy and ethical design. https://www.linkedin.com/pulse/azure-spatial-anchors-approach-privacy-ethical-design-alex-kipman, 2019.

[28] Yubin Kuang and Kalle Astrom. Pose estimation with unknown focal length using points, directions and lines. In *International Conference on Computer Vision (ICCV)*, 2013.

[29] Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Real-time solution to the absolute pose problem with unknown radial distortion and focal length. In *International Conference on Computer Vision (ICCV)*, 2013.

[30] Zuzana Kukelova, Jan Heller, and Andrew Fitzgibbon. Efficient intersection of three quadrics and applications in computer vision. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[31] Viktor Larsson, Kalle Astrom, and Magnus Oskarsson. Efficient solvers for minimal problems by syzygy-based reduction. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[32] Viktor Larsson, Zuzana Kukelova, and Yinqiang Zheng. Making minimal solvers for absolute pose estimation compact and robust. In *International Conference on Computer Vision (ICCV)*, 2017.

[33] Viktor Larsson, Zuzana Kukelova, and Yinqiang Zheng. Camera pose estimation with unknown principal point. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[34] Viktor Larsson, Torsten Sattler, Zuzana Kukelova, and Marc Pollefeys. Revisiting radial distortion absolute pose. In *International Conference on Computer Vision (ICCV)*, 2019.

[35] Viktor Larsson, Nicolas Zobernig, Kasim Taskin, and Marc Pollefeys. Calibration-free structure-from-motion with calibrated radial trifocal tensors. In *European Conference on Computer Vision (ECCV)*, 2020.

[36] Louis Lecrosnier, Rémi Boutteau, Pascal Vasseur, Xavier Savatier, and Friedrich Fraundorfer. Camera pose estimation based on pnl with a known vertical direction. *IEEE Robotics and Automation Letters (RA-L)*, 2019.

[37] Yunpeng Li, Noah Snavely, and Daniel P Huttenlocher. Location recognition using prioritized feature matching. In *European Conference on Computer Vision (ECCV)*, 2010.

[38] David Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 2004.

[39] Ludovic Magerand and Alessio Del Bue. Revisiting projective structure for motion: A robust and efficient incremental solution. *Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 2018.

[40] Faraz M Mirzaei and Stergios I Roumeliotis. Globally optimal pose estimation from line correspondences. In *International Conference on Robotics and Automation (ICRA)*, 2011.

[41] Roger Mohr, Long Quan, and Françoise Veillon. Relative 3D reconstruction using multiple uncalibrated images. *International Journal of Robotics Research (IJRR)*, 1995.

[42] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 2015.

[43] Gaku Nakano. A simple direct solution to the perspective-three-point problem. In *British Machine Vision Conference (BMVC)*, 2019.

[44] Mary Lynne Nielsen. Augmented Reality and its Impact on the Internet, Security, and Privacy. https://beyondstandards.ieee.org/augmented-reality/augmented-reality-and-its-impact-on-the-internet-security-and-privacy/, 2015.

[45] Carl Olsson and Olof Enqvist. Stable structure from motion for unordered image collections. In *Scandinavian Conference on Image Analysis (SCIA)*, 2011.

[46] Marc Pollefeys. *Self-calibration and metric 3D reconstruction from uncalibrated image sequences*. PhD thesis, 1999.

[47] Long Quan and Takeo Kanade. Affine structure from line correspondences with uncalibrated affine cameras. *Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 1997.

[48] Tilman Reinhardt. Google Visual Positioning Service. https://ai.googleblog.com/2019/02/using-global-localization-to-improve.html, 2019.

[49] Franziska Roesner. Who Is Thinking About Security and Privacy for Augmented Reality? https://www.technologyreview.com/s/609143/who-is-thinking-about-security-and-privacy-for-augmented-reality/, 2017.

[50] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[51] Torsten Sattler, Chris Sweeney, and Marc Pollefeys. On sampling focal length values to solve the absolute pose problem. In *European Conference on Computer Vision (ECCV)*, 2014.

[52] Frederik Schaffalitzky and Andrew Zisserman. Multi-view matching for unordered image sets, or how do I organize my holiday snaps? In *European Conference on Computer Vision (ECCV)*, 2002.

[53] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[54] Johannes L. Schönberger, Filip Radenović, Ondrej Chum, and Jan-Michael Frahm. From Single Image Query to Detailed 3D Reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[55] Mikiya Shibuya, Shinya Sumikura, and Ken Sakurada. Privacy preserving visual SLAM. In *European Conference on Computer Vision (ECCV)*, 2020.

[56] Noah Snavely, Steven Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3D. *ACM Trans. Gr.*, 2006.

[57] Pablo Speciale, Johannes L. Schönberger, Sing Bing Kang, Sudipta Sinha, and Marc Pollefeys. Privacy Preserving Image-Based Localization. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[58] Pablo Speciale, Johannes L. Schönberger, Sudipta N. Sinha, and Marc Pollefeys. Privacy preserving image queries for camera localization. In *International Conference on Computer Vision (ICCV)*, 2019.

[59] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[60] Chris Sweeney, Torsten Sattler, Tobias Hollerer, Matthew Turk, and Marc Pollefeys. Optimizing the viewing graph for structure-from-motion. In *International Conference on Computer Vision (ICCV)*, 2015.

[61] Bill Triggs. Camera pose and calibration from 4 or 5 known 3d points. In *International Conference on Computer Vision (ICCV)*, 1999.

[62] Jan-Erik Vinje. Privacy Manifesto for AR Cloud Solutions. https://medium.com/openarcloud/privacy-manifesto-for-ar-cloud-solutions-9507543f50b6, 2018.

[63] Frederik Warburg, Søren Hauberg, Manuel López-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[64] Kyle Wilson and Noah Snavely. Robust global translations with 1DSFM. In *European Conference on Computer Vision (ECCV)*, 2014.

[65] Kyle Wilson and Noah Snavely. Robust global translations with 1DSfM. In *European Conference on Computer Vision (ECCV)*, 2014.

[66] Changchang Wu. P3.5P: Pose estimation with unknown focal length. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[67] Chi Xu, Lilian Zhang, Li Cheng, and Reinhard Koch. Pose estimation from line correspondences: A complete analysis and a series of solutions. *Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 2016.

[68] Yinqiang Zheng, Shigeki Sugimoto, Imari Sato, and Masatoshi Okutomi. A general and simple method for camera pose and focal length determination. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.