Privacy Preserving Structure-from-Motion

Marcel Geppert[1], Viktor Larsson[1], Pablo Speciale[2], Johannes L. Schönberger[2], and Marc Pollefeys[1,2]

[1] Department of Computer Science, ETH Zurich, Switzerland
[2] Microsoft, Switzerland

Abstract. Over the last years, visual localization and mapping solutions have been adopted by an increasing number of mixed reality and robotics systems. The recent trend towards cloud-based localization and mapping systems has raised significant privacy concerns. These are mainly grounded by the fact that these services require users to upload visual data to their servers, which can reveal potentially confidential information, even if only derived image features are uploaded. Recent research addresses some of these concerns for the task of image-based localization by concealing the geometry of the query images and database maps. The core idea of the approach is to lift 2D/3D feature points to random lines, while still providing sufficient constraints for camera pose estimation. In this paper, we further build upon this idea and propose solutions to the different core algorithms of an incremental Structure-from-Motion pipeline based on random line features. With this work, we make another fundamental step towards enabling privacy preserving cloud-based mapping solutions. Various experiments on challenging real-world datasets demonstrate the practicality of our approach achieving comparable results to standard Structure-from-Motion systems.

1 Introduction

Driven by the quickly growing mixed reality and robotics markets, there has been significant commercial interest in image-based localization and mapping solutions. Over the last years, several companies have launched their cloud services, including Microsoft Azure Spatial Anchors [7], Google's Visual Positioning System [50] underlying the Google Maps AR navigation [6], 6D.AI [41], and Scape Technologies [35]. For these services to function, they require users to upload image information to their servers, which can reveal potentially private user information to the service provider or an adversary. As Dosovitskiy *et al.* [14] and Pittaluga *et al.* [26] strikingly demonstrated, this is the case even when only uploading local features extracted from the image to the cloud.

In this emerging field, privacy concerns have been initially widely ignored by both consumers and the industry, while recent motions in the community [20,24, 31,48] spurred fundamental research to find technical solutions to address these concerns [42,43]. In their pioneering work, Speciale *et al.* propose an approach to enable privacy preserving image-based localization services. The core idea of their method is to obfuscate the geometry of the query images and maps in

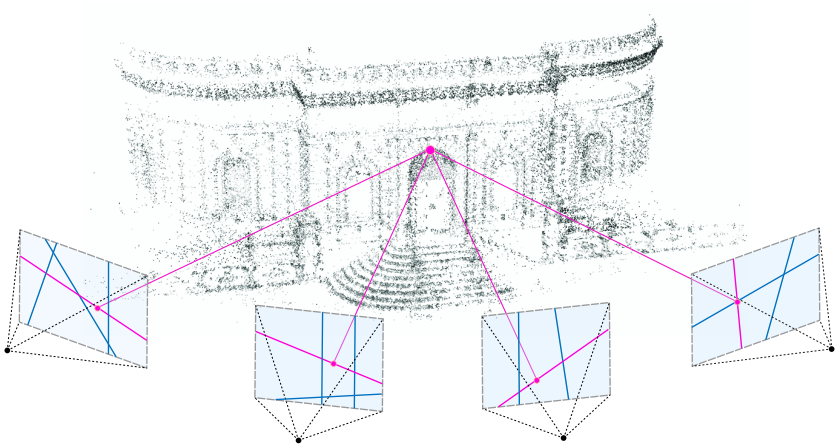

Fig. 1: Privacy Preserving SfM. Our proposed method takes privacy preserving feature lines as input and seeds the reconstruction from four views with at least eight corresponding line features. The intersection of at least three line features produces a point triangulation. Camera resectioning requiring at least six 2D line to 3D point correspondences is based on Speciale *et al.* [43].

a way that hides private user information but still provides sufficient geometric constraints to enable camera localization. Specifically, they lift 2D image features to random 2D lines to preserve the privacy in query images sent by the client for server-based localization. In addition, they also show how the same concept can be applied in the 3D domain to obfuscate the geometry of maps [42], where they lift the 3D points of Structure-from-Motion maps to random 3D lines to enable the sharing of private maps with a service provider or another client for the purpose of localization.

The fundamental limitation with their approach is that the map reconstruction stage must be performed on the client side, thereby prohibiting server-based mapping solutions. In the context of mixed reality and robotics, client devices typically have low compute capabilities and thus offloading the map reconstruction stage to the cloud is often required. Furthermore, crowd-based mapping solutions become increasingly relevant, especially as an increasing number of heterogeneous clients navigate through the same space and mapping large-scale spaces becomes infeasible for a single agent alone. In these scenarios, a server-based mapping solution is required to merge the visual data from multiple clients into a single consistent map.

In this paper, we address this limitation and propose an approach to perform privacy preserving Structure-from-Motion (SfM). Our approach is the first to enable cloud-based mapping solutions which do not sacrifice the privacy of users by hiding the privacy concerning contents of the input images. Equivalent to Speciale *et al.*, only consistently observed, triangulated and therefore static scene structure is revealed during our reconstruction process, while privacy concerning transient structure only consistently visible in less than three views (*e.g.*, moving people) are concealed and cannot be reconstructed. The proposed method is

based on the fundamental ideas from Speciale *et al.* in that we derive the necessary geometric constraints to perform end-to-end SfM from input images with obfuscated random 2D feature lines (see Figure 1). This representation hides the appearance information in the extracted 2D features from feature inversion techniques such as [26]. Only in combination with the 3D structure some of the scene appearance can be recovered. In detail, we make the following contributions: (1) We present an end-to-end privacy preserving SfM pipeline based on line features. Our pipeline builds upon COLMAP [36] and the entire source code of our system will be released as open source. (2) For each of the main processing stages of an incremental SfM system (initialization, camera resectioning, triangulation, bundle adjustment), we propose its equivalent counterpart in the privacy preserving setting. (3) We derive a practical minimal solver to initialize our incremental SfM pipeline from four views. The underlying geometric constraints are based on the theory of trifocal tensors and, by exploiting gravity information, we are able to decompose the problem into feasible subproblems. (4) We demonstrate robust and efficient performance of our system on challenging datasets and achieve comparable results with the traditional point feature based baseline.

2 Related Work

In this section, we review related works on privacy preserving methods with a focus on privacy preserving localization approaches. In addition, we also discuss adversarial methods to reconstruct images from its features. Background on SfM in general will be discussed more broadly in the following section.

Privacy Preserving Databases. Querying data in databases without leaking side information has been studied in [12]. Furthermore, various works focused on the specific problems of location privacy [3,5,17,45], differential privacy [15], k-anonymity [34], or learning data-driven models [1,18,55]. However, these approaches are not applicable to geometric computer vision problems, in particular to SfM and image-based localization.

Privacy-Aware Vision Recognition and Hardware Sensors. Privacy-aware techniques were also explored for image retrieval [38], video surveillance [46], biometric verification [47], and face recognition [16]. Learning methods on encrypted data [1,18,55] and adversarial training for vision/action recognition [25, 54] have also received much attention. Recent works on privacy in vision include anonymization for activity recognition [30,32]. Furthermore, there are works on privacy preserving optics [27, 28] and lens-less coded aperture camera systems that preserve privacy by making the images/video incomprehensible while still allowing action recognition [51]. Zhao *et al.* [58] demonstrated how to localize objects indoors using active cameras and Shariati *et al.* [37] performed ego-motion estimation using low-resolution cameras. However, these privacy preserving techniques cannot be used for SfM from regular images, since they focus either on recognition tasks or rely on special hardware to achieve privacy.

Recovering Images from Features. The main privacy concern we aim to avoid in our pipeline is the reconstruction of images using only its features. Weinzaepfel *et al.* [52] were the first ones to try to invert SIFT features, followed by Vondrick *et al.* [49] interpreting HOG features, bag-of-words features [19], and finally CNN features [23, 56, 57]. Currently, there are methods that can recover remarkably accurate images from extracted SIFT features [13, 14]. More recently, and to raise awareness in the community about the privacy implications, Pittaluga *et al.* [26] demonstrated that 3D point clouds of scenes reconstructed using SfM techniques retain enough information to reconstruct detailed images of the scene, even after the source images are discarded. This enables an adversary to recover confidential content, emphasizing the privacy risks of transmitting and permanently storing such data.

Privacy Preserving Localization. In response to such an attack, there have been a series of papers presented by Speciale *et al.* [42, 43] investigating new camera pose estimation algorithms that can safeguard the users privacy. While there were some existing approaches for recognizing objects in images and videos in a privacy-aware manner [9,10,30,32], those methods cannot be used for camera pose estimation or other geometric computer vision tasks used in mixed reality and robotics. The fundamental limitation of their approach is that the mapping stage (*i.e.*, Structure-from-Motion) must be performed on the client side. This prevents deploying their solution to a wide range of practical applications, where a server-based mapping solution is required. To address this limitation, we build upon their idea of obfuscating 2D feature points by lifting them to 2D random lines [43]. Instead of just enabling privacy aware image-based localization, our goal is to extend this task to create privacy preserving SfM models from the obfuscated images. Concurrent to our work, Shibuya *et al.* [40] extended the line-based map protection approach and developed a privacy preserving Visual SLAM system that uses mixed point- and lineclouds as maps to conceal the map geometry to the user.

3 Method

In this section, we describe our proposed solution to privacy preserving SfM. Our approach is similar to other incremental methods for SfM that operate by alternating between registering new images to the reconstruction (so-called *resectioning*) and triangulating new 3D points. The main steps of these pipelines are **initialization** (Section 3.1), **triangulation** (Section 3.2), **camera resectioning** (Section 3.3) and **bundle adjustment** (Section 3.4). In the following sections, we detail how we adapt each of these steps to the privacy preserving setting. The main difficulty is in dealing with the weaker geometric constraints induced by lifting 2D feature points $\boldsymbol{x}$ to random 2D feature lines $\boldsymbol{\ell}$ analogous to Speciale *et al.* [43], which renders the initialization stage especially challenging.

3.1 Initialization

From only corresponding 2D line features it is not possible to perform initialization from two views. To see this, note that two backprojected 2D lines become 3D planes and will always intersect in a 3D line, regardless of how the cameras are posed, thus the line-line correspondences provide no constraint on the two-view relative pose. Even with three views, the 3D planes will always intersect in a point and therefore not provide any constraints. In fact, the first constraints appear in four views. The relative poses of four images are described by the quadrifocal tensor (see *e.g.* [39]). While it is in theory possible to estimate the quadrifocal tensor from line correspondences, there is currently no tractable method for doing so due to the high complexity of the quadrifocal tensor's internal constraints.

Instead, we present an alternative approach for performing robust initialization from line-based correspondences. The method is based on the assumption of knowing the gravity direction of the images used for initialization, which is reasonable in practice as virtually any device nowadays comes with an intertial measurement unit. Furthermore, we leverage the fact that we have control over the process in which we create the random lines. The core idea is to align a random subset of the line features with the gravity direction. These lines are now consistently oriented w.r.t. the world frame, *i.e.*, the planes of backprojected lines should now intersect in a (gravity-oriented) 3D line if the camera poses are correct. This yields additional constraints on the relative poses which we can use to simplify the complexity of the estimation problem. Furthermore, we show that the gravity-aligned feature lines allow us to decompose the initialization problem such that we first solve a two-dimensional SfM problem, followed by upgrading the cameras into three dimensions.

Reduction to Two Dimensions. We assume the cameras have been rotated such that the y-axis coincides with the known gravity direction. Once the cameras' coordinate systems are gravity-aligned, each camera only has four degrees of freedom left: rotation θ around y-axis and translation components $\begin{pmatrix} t_x & t_y & t_z \end{pmatrix}^T$.

Consider the constraint posed by the vertical line $\boldsymbol{\ell} = (-1, 0, x)$ passing though the 2D point (x, y)

$$\begin{pmatrix} -1, 0, x \end{pmatrix} \left(\begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix} \right) = 0 \ . \tag{1}$$

Note that the gravity aligned lines do not place any constraints on either Y nor t_y. This is since they only translate either the 3D point $\begin{pmatrix} X, Y, Z \end{pmatrix}$ or the camera along the gravity direction. As such, we can rewrite Equation (1) as

$$\begin{pmatrix} -1, x \end{pmatrix} \left(\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{pmatrix} X \\ Z \end{pmatrix} + \begin{pmatrix} t_x \\ t_z \end{pmatrix} \right) = 0 \ . \tag{2}$$

This equation is exactly the projection equations for a 2D-to-1D camera, i.e.

$$\lambda \begin{pmatrix} x \\ 1 \end{pmatrix} = R_{2\times 2} \begin{pmatrix} X \\ Z \end{pmatrix} + \boldsymbol{t}_{2\times 1} \ . \tag{3}$$

Using this insight, we can decompose the problem into first solving a 2D relative pose problem using the gravity-aligned correspondences. The solution to this problem yields the full camera orientation θ as well as the two translation components t_x, t_z orthogonal to gravity. The only remaining unknown is the gravity-aligned translation t_y which is unobservable from the gravity-aligned lines. To recover these, we use additional line correspondences, which are randomly oriented in the images.

Relative Pose of Three Gravity Oriented Views from Vertical Lines. In the 2D setting, it is not possible to estimate the relative pose from only two views. The relative pose of three views was first solved in [29] by means of the 2D trifocal tensor. Since then there have been multiple papers using the trifocal tensor to perform 2D planar motion estimation, see e.g. [4, 11, 33].

The 2D trifocal tensor is a $2 \times 2 \times 2$ tensor, which constrains the 1D image measurements $\boldsymbol{x}_i \in \mathbb{P}^1,\ i = 1, 2, 3$ as

$$\left[\boldsymbol{x}_1^T T_1 \boldsymbol{x}_2,\ \boldsymbol{x}_1^T T_2 \boldsymbol{x}_2\right] \boldsymbol{x}_3 = 0 \ , \tag{4}$$

where T_1 and T_2 are 2×2 slices of the tensor. Equation (4) yields a linear constraint on the trifocal tensor, which means that it can be linearly estimated from 7 points (since it is homogeneous). In the case of calibrated 2D cameras, i.e. the first 2×2-block is a rotation matrix, there exist internal constraints on the tensor. These constraints were first identified by Åström and Oskarsson [8] and are linear constraints in the tensor elements, given in Equation (6) and (7). This allows the tensor to be estimated from only five point correspondences.

Given more than five point correspondences, it is possible to solve for the trifocal tensor by solving a homogeneous linear least-squares problem with homogeneous linear constraints, i.e.

$$\min_T \sum_i \left(\left[\boldsymbol{x}_{1i}^T T_1 \boldsymbol{x}_{2i},\ \boldsymbol{x}_{1i}^T T_2 \boldsymbol{x}_{2i}\right] \boldsymbol{x}_{3i}\right)^2 \tag{5}$$

$$\text{s.t. } T_{111} - T_{122} - T_{212} - T_{221} = 0 \tag{6}$$

$$T_{112} + T_{121} + T_{211} - T_{222} = 0 \ , \tag{7}$$

which admits a closed form solution using SVD. Once the 2D trifocal tensor has been estimated, we can factorize it to recover the 2D cameras (see [29] for details). Note this factorization only gives us the pose of the original cameras up to an unknown translation along the gravity direction.

Resectioning a Fourth View in 2D. Once the poses of the first three cameras are determined, they can be used to triangulate the 2D points (recovering the X and Z coordinate of the 3D points). These can then be used to estimate the 2D pose of a fourth camera by solving the optimization problem

$$\min_{\theta, t_x, t_z} \sum_{i=1}^{N} \left((1, -x_i) \left(\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{pmatrix} X_i \\ Z_i \end{pmatrix} + \begin{pmatrix} t_x \\ t_z \end{pmatrix} \right) \right)^2 \ . \tag{8}$$

Substituting $a = \cos\theta$ and $b = \sin\theta$ we get the equivalent problem

$$\min_{a,b,t_x,t_z} \left\| A \begin{bmatrix} a & b & t_x & t_z \end{bmatrix}^T \right\|^2 \text{ s.t. } a^2 + b^2 = 1 \ , \tag{9}$$

which is a homogeneous least-squares problem with a norm constraint. The optimal solution can be obtained using singular value decomposition after eliminating the translation. Since the cost is homogeneous, we get another solution corresponding to flipping the sign of the singular vector. To decide between these two solutions, we check cheirality. Note that this solution also works for this minimal case (with three correspondences). In this case, A is a 3×4 matrix.

Solving for Out-of-Plane Translation. From the 2D estimation in the previous step, we know the relative poses of the cameras except for their translation in the y-direction. Fixing the coordinate system such that the first camera is at the origin, we now aim to recover the three remaining translation parameters, i.e. we want to find t_{y2}, t_{y3} and t_{y4} where $\boldsymbol{t}_i = \left(t_{xi}, t_{yi}, t_{zi}\right)$ is the full translation vector for the ith camera. Let $\boldsymbol{X}$ be an unknown 3D point and $\boldsymbol{\ell}_i$ be the corresponding line feature in the ith view. These should then satisfy

$$\boldsymbol{\ell}_i^T \left(R_i \boldsymbol{X} + \boldsymbol{t}_i\right) = 0, \quad i = 1, 2, 3, 4 \tag{10}$$

Collecting these constraints in a matrix we can rewrite this as

$$\begin{bmatrix} \boldsymbol{\ell}_1^T & 0 \\ \boldsymbol{\ell}_2^T R_2 & \boldsymbol{\ell}_2^T \boldsymbol{t}_2 \\ \boldsymbol{\ell}_3^T R_3 & \boldsymbol{\ell}_3^T \boldsymbol{t}_3 \\ \boldsymbol{\ell}_4^T R_4 & \boldsymbol{\ell}_4^T \boldsymbol{t}_4 \end{bmatrix} \begin{pmatrix} \boldsymbol{X} \\ 1 \end{pmatrix} = 0. \tag{11}$$

If (11) is satisfied, the 4×4 matrix is rank deficient and thus the determinant must vanish. Since all of the unknowns appear in a single column, the determinant gives us a linear constraint on the unknown translation components. Thus from three sets of line correspondences it is possible to linearly estimate the out-of-plane translations.

Note that if any line $\boldsymbol{\ell}_i$ is gravity-aligned, then $\boldsymbol{\ell}_i^T \boldsymbol{t}_i$ becomes independent of the y-translation for that camera. However, as long as at most two of the lines are gravity-aligned, we still get constraints on the relative translations for the other cameras. Geometrically, two aligned lines restrict the 3D point to a vertical line in 3D. Any line in the other views then gives the 3D point by back-projecting the non-aligned lines. The translations of the two views are then constrained by the fact that they should have the same intersection on the vertical 3D line. In the case where three or four of the lines are gravity aligned, then no constraint on the translations can be derived from (11). We discuss the degenerate configurations for the proposed initialization procedure and provide additional evaluations of the initialization with synthetic data in the supplementary material.

Aligned Feature Selection. Our initialization method relies on consistent tracks of gravity-aligned features in four images. At the same time, potentially

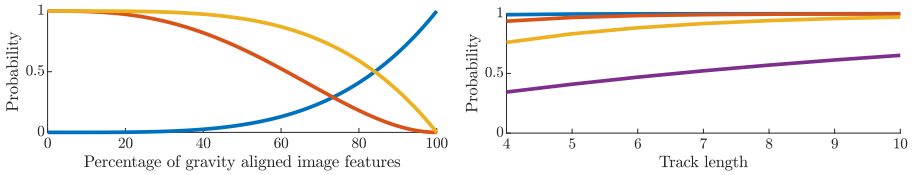

Fig. 2: Our method requires different combinations of gravity-aligned and random features at different stages of the pipeline. *Left:* Depending on the percentage of gravity-aligned features, we show the probabilities to obtain these combinations in a feature track in 4 images: All gravity-aligned (blue, required for 2D pose estimation), at least 2 randomly aligned (red, required for 3D pose upgrade), or at least 1 randomly aligned (yellow, required for point triangulation). *Right:* Probability to have at least 1 randomly aligned feature depending on track length for 30% (blue), 50% (red), 70% (yellow), and 90% (purple) aligned features.

collecting images from many different users over time, we cannot communicate between devices to agree on certain subsets of features to be gravity aligned. Instead, we randomly select a certain percentage of features in each image to be gravity aligned. Note that this strategy does not reduce the level of privacy preservation as compared to using random directions for all lines. Assuming a track of feature matches in four images, Figure 2 (Left) shows the probabilities that all features in the track are gravity-aligned (suitable for 2D pose estimation), at least two of them are randomly aligned (suitable for 3D pose upgrade), or at least one is randomly aligned (suitable for point triangulation). Our goal is to obtain enough feature tracks suitable for initialization while keeping most tracks suitable for triangulation in the following steps of the pipeline. Based on these results, we empirically select a random subset of 50% of the features to be gravity aligned for all of our experiments. Lower values could lead to insufficient initialization tracks in some challenging datasets. As shown in Figure 2 (*Left*), the impact on tracks for triangulation is negligible, and decreases further for longer tracks as shown in Figure 2 (*Right*).

3.2 Triangulation

Each 2D-to-3D correspondence places a single constraint on the 3D point as shown for the initialization images in Equation (10). Geometrically, this constraint can be interpreted as requiring the 3D point $\boldsymbol{X}$ to lie on the plane of the backprojected 2D line. We illustrate this geometric interpretation in in Figure 3 and the supplementary material. Since we have three degrees-of-freedom in the 3D point, we can perform triangulation given at least three correspondences. The constraints are linear in $\boldsymbol{X}$ and can be solved easily.

Note that with 2D lines, the triangulation is exactly minimal with three correspondences. This means that the 3D point will always have zero reprojection error for these three lines, thus it is not possible to determine if any of the matches were outliers or not. As such, in our SfM pipeline, we filter all 3D points which have three or fewer inliers.

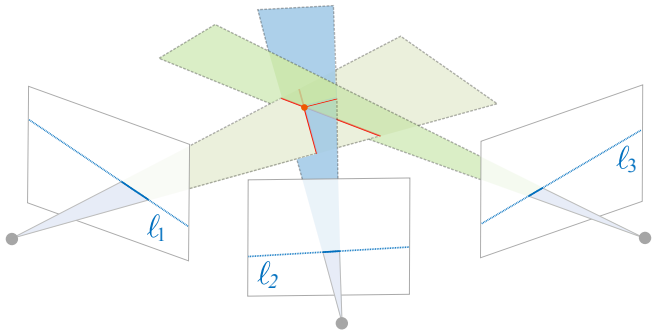

Fig. 3: Triangulation. Each 2D line backprojects into a plane. Intersecting the back-projected planes from multiple corresponding lines allows us to triangulate 3D points.

3.3 Camera Resectioning

Camera resectioning from 2D line features w.r.t. a 3D point cloud was previously introduced by Speciale *et al.* [43]. Geometrically this is performed by aligning the the planes of the backprojected 2D lines with the corresponding 3D points. This gives a single constraint from each 2D line to 3D point correspondence compared to two in the traditional point to point case. Thus estimating the image pose from line features requires at least six correspondences (*P6L*) instead of three with point features (*P3P*). With some reformulation, we can solve the *P6L* problem efficiently using the *E3Q3* solver from Kukelova *et al.* [21].

3.4 Bundle Adjustment

An essential component of any incremental SfM pipeline is the joint non-linear refinement of the structure and the camera poses, *i.e.*, bundle adjustment, to reduce accumulated errors from the triangulation and resectioning steps. The bundle adjustment stage becomes especially important in our pipeline, where we rely on weaker geometric constraints. For each 2D line to 3D point correspondence, we minimize the orthogonal distance from the projected point to the 2D line, *i.e.*, for the jth point seen in the ith image, we have

$$r_{ij}^2 = \frac{\left(\boldsymbol{n}^T \pi(R_i \boldsymbol{X}_j + t_i) + \alpha\right)^2}{\boldsymbol{n}^T \boldsymbol{n}} \quad \text{where} \quad \boldsymbol{\ell}_{ij} = (\boldsymbol{n}, \alpha)^T \ , \tag{12}$$

where $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ is the standard pinhole projection. In our bundle adjustment, we then minimize the reprojection errors over all current observations

$$\min_{\{R_i, t_i\}, \{X_j\}} \sum_{i,j} r_{ij}^2 \ . \tag{13}$$

3.5 Implementation Details

We implemented our proposed SfM method by extending the open-source framework COLMAP [36]. In summary, we replaced all the core processing steps from

relying on point features to using random line features as presented in the previous sections. Furthermore, we decreased some of the reprojection thresholds to increase the robustness of the system. This is mainly to address the fact that the projected point to line distance in the image is generally an underestimate of the traditional point to point distance. In addition, we are careful to reject spurious correspondences projecting outside of the image, as there is a significantly higher chance of a wrong feature match causing a 3D point to accidentally project onto a random feature line. In particular, this happens frequently with repetitive scene structure. Furthermore, our implementation currently assumes calibrated cameras. Similarly to COLMAP, the bundle adjustment is implemented using the open-source library Ceres [2].

For the initialization stage, we restrict the search for selecting the four initial views to the subset of images with gravity information. Note that only the small subset of images used for initialization must have gravity information while there are no such requirements from the other stages of our pipeline. First, we find suitable images sets by assembling all purely gravity aligned or purely random 4 view tracks from the pairwise feature matches. Note that this process can be very costly when searching through the whole matching graph. We therefore randomly select 10 images and use them as starting point for the search, meaning that all considered image sets will contain at least one of these images. We then select the 10 image sets with the most gravity aligned tracks and perform the proposed four-view initialization strategy for each one of them, using robust LO-RANSAC [22] loops for both the 2D pose estimation and the upgrade to 3D poses, respectively. We remove unstable geometric configurations by using a threshold on the mean minimum triangulation angle of the inlier tracks. From the remaining configurations, we select the one with the highest inlier ratio in the 3D pose upgrade as seed for the reconstruction.

In the original COLMAP pipeline, two-view relative pose estimation is used for geometric verification during the pairwise matching step. In our setting this geometric verification is not possible since we are using 2D line features, leading to higher outlier ratios during the remaining steps of the pipeline. In practice we did not observe any negative impact by this modification. The entire source code of the privacy-aware version of COLMAP will be released as open source.

4 Experiments

We evaluate the performance of our proposed SfM pipeline on benchmark datasets with ground-truth [42,44] as well as challenging large-scale internet datasets [53]. The results demonstrate that our privacy preserving system achieves comparable results to the state-of-the-art traditional SfM pipelines.

4.1 Evaluation of Camera Pose Accuracy

To quantitatively evaluate the effect of the weaker geometric constraints on the accuracy of the reconstruction, we use the well-known benchmark dataset

Table 1: Evaluation of camera pose accuracy on the Strecha benchmark [44].

	#Images		#Points		Track Length	Rotation (deg)		Position (cm)	
Scene	Total	Reg.	3D	2D		Mean	Std.	Mean	Std.
castle-P19	19	15	4.3k	24.2k	5.7	0.29	0.20	10.80	17.47
castle-P30	30	30	11.5k	78.7k	6.8	0.08	0.03	4.00	2.73
entry-P10	10	10	4.0k	24.5k	6.1	0.05	0.01	0.71	0.26
fountain-P11	11	11	7.9k	46.2k	5.8	0.03	0.01	0.30	0.14
Herz-Jesu-P8	8	8	3.4k	17.6k	5.2	0.21	0.03	0.53	0.30
Herz-Jesu-P25	25	25	11.1k	86.8k	7.8	0.04	0.02	0.58	0.23

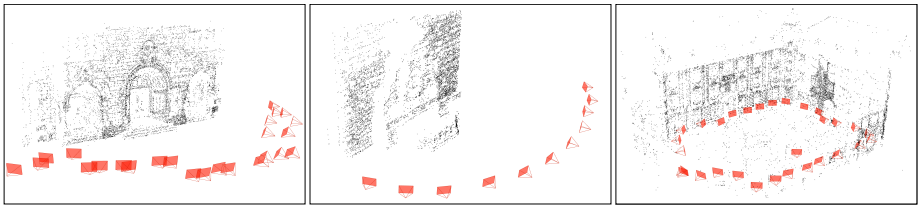

Fig. 4: Qualitative results. *Herz-Jesu-P25*, *Fountain-P11* and *Castle-P30* datasets from Strecha *et al.* [44].

from Strecha *et al.* [44], which comes with accurate ground-truth camera poses. Since the dataset does not contain real measured gravity direction, we generate perfect synthetic gravity using the ground-truth camera poses. Table 1 shows the reconstruction statistics and camera pose errors. Figure 4 shows qualitative results. Note that not all images could be registered for the *castle-P19* dataset due to insufficient overlap between a subset of the images, such that no four views were available to triangulate sufficient points to resection the missing camera views. Generally, our method was able to accurately register images with a mean rotation error below 1° and a mean position error below 1cm, except for the two *castle* datasets. For these two datasets, standard COLMAP also performs significantly worse with 5cm and 3cm position error for *castle-P19* and *castle-P30*, respectively (see supplementary material for full results with COLMAP).

4.2 Evaluation of Initialization Scheme

Our method requires known gravity for the images used for initialization. In practice, gravity directions can be obtained from inertial measurement units or through vanishing point detection. As such, gravity direction measurements are noisy. In this section, we demonstrate that our method is robust to the noise generally present in real-world data. Towards this end, we evaluate the performance of our pipeline on the dataset from Speciale *et al.* [42], which consists of six datasets with images captured by a Google Pixel smartphone with gravity directions extracted from EXIF data. In our experimental setup, we randomly select 15 quadruplets of images from each dataset, then run our proposed minimal solver inside LO-RANSAC, triangulate the inlier features, and then perform

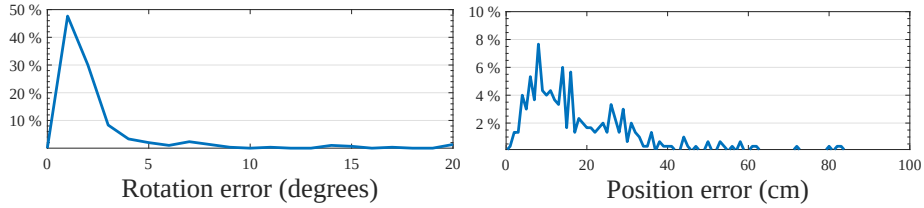

Fig. 5: Evaluation of initialization accuracy with real gravity.

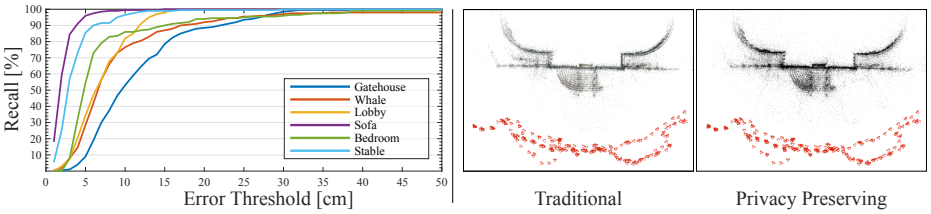

Fig. 6: Comparison against traditional SfM. We compare the accuracy of our results against standard COLMAP on the benchmark dataset by Speciale *et al.* [42].

a non-linear refinement using bundle adjustment. For each quadruplet, we compare the accuracy of our initialization of the four camera views with a pseudo ground truth generated by running COLMAP [36] with the original point correspondences on the full image sequences. Figure 5 shows histograms for the rotation and translation errors. Our approach consistently produces sufficiently accurate initializations using real-world data. Note that after initialization, gravity measurements are not required for the remaining images.

4.3 Comparison with Traditional Structure-from-Motion

Our entire pipeline is based on the incremental SfM pipeline COLMAP [36], which we use as a baseline in this experiment. Due to the weaker constraints used by our system, we do not expect to outperform COLMAP in terms of reconstruction quality, which can be considered as an upper bound for our method. As such, we use six datasets provided by Speciale *et al.* [42] with 200 images per scene and real-world gravity from a smartphone camera. The scenes each span a size in the order of 10s of meters and we consider standard COLMAP output as ground-truth. Figure 6 summarizes the results and our method consistently achieves a recall of 90% at an error threshold of 25cm. Furthermore, we are able to register all the cameras at a maximum error threshold of 50cm.

4.4 Structure-from-Motion on Internet Images

Finally, we evaluate our SfM pipeline on unstructured, large-scale image datasets crowd-sourced from the internet. This experiment is especially relevant, as one of the target applications of our system is privacy preserving crowd-sourced mapping in the cloud. We consider the *Madrid Metropolis*, *Alamo*, *Tower of*

Table 2: Comparison of reconstruction statistics for large-scale internet datasets with columns formatted as *Traditional / Privacy Preserving*.

Scene	#Imgs	#Reg Imgs	#Pts	#Obs	Mean Track Len	Median Pt Reproj Err	Median Line Reproj Err
Tower of London	1576	577/608	146k/93k	1235k/1122k	8.4/12.0	0.38/0.54	−/0.23
Gendarmenmarkt	1463	825/810	180k/83k	1185k/ 958k	6.6/11.5	0.48/0.88	−/0.34
Madrid Metropolis	1344	279/377	59k/43k	352k/ 447k	5.9/10.4	0.45/1.13	−/0.40
Alamo	2915	703/750	137k/79k	1763k/1730k	12.8/21.9	0.50/0.66	−/0.52

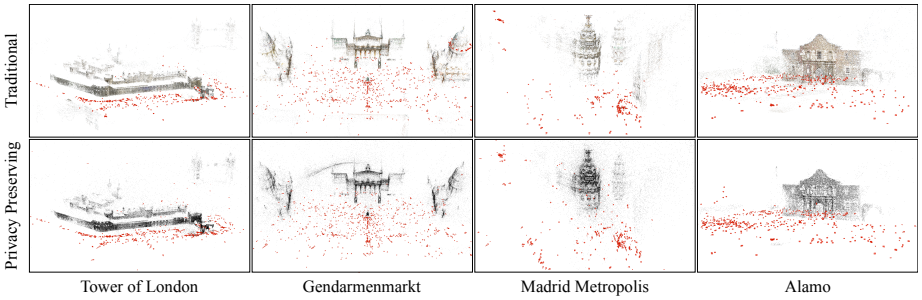

Fig. 7: Qualitative comparison. Internet datasets from Wilson & Snavely [53].

London and *Gendarmenmarkt* datasets from Wilson & Snavely [53]. Similarly to Section 4.1 we generate synthetic gravity measurements for the initialization. Coarse camera calibrations are obtained from EXIF tags and optimized as part of the COLMAP reconstruction. Therefore, accurate calibrations are only available for images that could be registered by COLMAP. Still, we run our method with all input images and coarse calibrations where necessary. Since there is no reliable ground truth reconstructions for these datasets, we only report general reconstruction statistics (see Table 2) and show qualitative results (see Figure 7). Additional registered images with our method compared to COLMAP are caused by the weaker constraints and likely to be noise. The results show that our privacy aware system achieves competitive results as compared to standard COLMAP, which underlines the practical relevance of our proposed method.

4.5 Qualitative comparison of Feature Inversion results

We perform a qualitative analysis of the feature inversion results with InvSfM [25] from the COLMAP model and ours. For COLMAP, we use all extracted SIFT features and their keypoint positions as this is the information that needs to be shared for traditional SfM. For our method, the keypoint positions are not available and the image is rendered by projecting 3D points into a virtual camera with the respective pose. Figure 8 shows a comparison of the inversion results. While the COLMAP inversion contains lots of details and reveals the persons in the scene, only the building can be reconstructed from the available information in our method. We provide more results in the supplementary material.

Fig. 8: Comparison of the feature inversion with InvSfM [26] on the *Alamo* dataset [53]. *Left:* Original image *Middle:* from COLMAP reconstruction. *Right:* from privacy preserving reconstruction.

5 Conclusion

In this paper, we presented the first privacy preserving SfM pipeline. Our method builds upon recent work to conceal image information using random feature lines. We derive a novel solution to estimate the camera geometry of four views from only line features and integrate it into the incremental SfM paradigm alongside the privacy preserving variants of triangulation, camera resectioning, and bundle adjustment. With this work, we make a fundamental step towards enabling privacy aware cloud-based mapping solutions without the risk of users revealing potentially confidential information to the mapping service provider or an attacker. Numerous experiments demonstrate that our system achieves comparable results to standard SfM systems despite effectively using only half of the geometric constraints, which underlines the high practical relevance of our work. However, this work alone can not solve the problem of protecting people's privacy in all situations. We assume that privacy concerning content is dynamic, *i.e.*, we do not encounter sequences of four or more images where sensitive content is seen consistently. While this assumptions usually holds for internet image collections, it quickly breaks when many images of the same scene in a short time frame are available. This could either happen when many users capture a situation at the same time, or image sequences are captured with high frame rate. Especially this second case is highly relevant for our work, as this is the case for a device constantly localizing in a new or changing environment that requires constant updates to the map. Also, our method does not handle the case when privacy concerning content is static. Still, this is a common case, *e.g.*, for users mapping their private apartments.

Acknowledgements. Viktor Larsson was supported by an ETH Zurich Postdoctoral Fellowship.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Conference on Computer and Communications Security (ACM CCS) (2016) 3

2. Agarwal, S., Mierle, K., Others: Ceres solver. http://ceres-solver.org 10
3. Andrés, M.E., Bordenabe, N.E., Chatzikokolakis, K., Palamidessi, C.: Geo-indistinguishability: Differential privacy for location-based systems. In: Conference on Computer & communications security (SIGSAC) (2013) 3
4. Aranda, M., López-Nicolás, G., Sagüés, C.: Omnidirectional visual homing using the 1D trifocal tensor. In: International Conference on Robotics and Automation (ICRA) (2010) 6
5. Ardagna, C.A., Cremonini, M., Damiani, E., Di Vimercati, S.D.C., Samarati, P.: Location privacy protection through obfuscation-based techniques. In: IFIP Annual Conference on Data and Applications Security and Privacy (2007) 3
6. ARNav: Google AR Navigation. https://ai.googleblog.com/2019/02/using-global-localization-to-improve.html (2019) 1
7. ASA: Azure Spatial Anchors. https://azure.microsoft.com/en-us/services/spatial-anchors/ (2019) 1
8. Åström, K., Oskarsson, M.: Solutions and ambiguities of the structure and motion problem for 1d retinal vision. Journal of Mathematical Imaging and Vision (2000) 6
9. Avidan, S., Butman, M.: Blind vision. In: European Conference on Computer Vision (ECCV) (2006) 4
10. Avidan, S., Butman, M.: Efficient methods for privacy preserving face detection. In: Advances in Neural Information Processing Systems (2007) 4
11. Dellaert, F., Stroupe, A.W.: Linear 2D localization and mapping for single and multiple robot scenarios. In: International Conference on Robotics and Automation (ICRA) (2002) 6
12. Dinur, I., Nissim, K.: Revealing information while preserving privacy. In: Symposium on Principles of Database Systems (2003) 3
13. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: Advances in Neural Information Processing Systems (2016) 4
14. Dosovitskiy, A., Brox, T.: Inverting visual representations with convolutional networks. In: Computer Vision and Pattern Recognition (CVPR) (2016) 1, 4
15. Dwork, C.: Differential privacy: A survey of results. In: International Conference on Theory and Applications of Models of Computation (2008) 3
16. Erkin, Z., Franz, M., Guajardo, J., Katzenbeisser, S., Lagendijk, I., Toft, T.: Privacy-preserving face recognition. In: International Symposium on Privacy Enhancing Technologies Symposium (2009) 3
17. Gedik, B., Liu, L.: Protecting location privacy with personalized K-anonymity: Architecture and algorithms. IEEE Transactions on Mobile Computing (2008) 3
18. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., Wernsing, J.: Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In: International Conference on Machine Learning (2016) 3
19. Kato, H., Harada, T.: Image reconstruction from bag-of-visual-words. In: Computer Vision and Pattern Recognition (CVPR) (2014) 4
20. Kipman, A.: Azure Spatial Anchors approach to privacy and ethical design. https://www.linkedin.com/pulse/azure-spatial-anchors-approach-privacy-ethical-design-alex-kipman/ (2019) 1
21. Kukelova, Z., Heller, J., Fitzgibbon, A.W.: Efficient intersection of three quadrics and applications in computer vision. In: Computer Vision and Pattern Recognition (CVPR) (2016) 9
22. Lebeda, K., Matas, J., Chum, O.: Fixing the Locally Optimized RANSAC. In: British Machine Vision Conference (BMVC) (2012) 10

23. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Computer Vision and Pattern Recognition (CVPR) (2015) 4
24. Nielsen, M.L.: 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) 1
25. Pittaluga, F., Koppal, S., Chakrabarti, A.: Learning privacy preserving encodings through adversarial training. In: Winter Conference on Applications of Computer Vision (WACV) (2019) 3, 13
26. Pittaluga, F., Koppal, S., Kang, S.B., Sinha, S.N.: Revealing scenes by inverting structure from motion reconstructions. In: Computer Vision and Pattern Recognition (CVPR) (2019) 1, 3, 4, 14
27. Pittaluga, F., Koppal, S.J.: Privacy preserving optics for miniature vision sensors. In: Computer Vision and Pattern Recognition (CVPR) (2015) 3
28. Pittaluga, F., Koppal, S.J.: Pre-capture privacy for small vision sensors. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2017) 3
29. Quan, L., Kanade, T.: Affine structure from line correspondences with uncalibrated affine cameras. Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (1997) 6
30. Ren, Z., Lee, Y.J., Ryoo, M.S.: Learning to anonymize faces for privacy preserving action detection. European Conference on Computer Vision (ECCV) (2018) 3, 4
31. Roesner, F.: 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) 1
32. Ryoo, M.S., Rothrock, B., Fleming, C., Yang, H.J.: Privacy-preserving human activity recognition from extreme low resolution. In: Proceedings of the Thirty-First Conference on Artificial Intelligence (AAAI) (2017) 3, 4
33. Sagues, C., Murillo, A., Guerrero, J.J., Goedemé, T., Tuytelaars, T., Gool, L.V.: Localization with omnidirectional images using the radial trifocal tensor. In: International Conference on Robotics and Automation (ICRA) (2006) 6
34. Samarati, P.: Protecting respondents identities in microdata release. IEEE transactions on Knowledge and Data Engineering (2001) 3
35. Scape: Scape Technologies. https://scape.io/ (2019) 1
36. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Computer Vision and Pattern Recognition (CVPR) (2016) 3, 9, 12
37. Shariati, A., Holz, C., Sinha, S.: Towards privacy-preserving ego-motion estimation using an extremely low-resolution camera. IEEE Robotics and Automation Letters (RAL) (2020) 3
38. Shashank, J., Kowshik, P., Srinathan, K., Jawahar, C.: Private content based image retrieval. In: Computer Vision and Pattern Recognition (CVPR) (2008) 3
39. Shashua, A., Wolf, L.: On the structure and properties of the quadrifocal tensor. In: European Conference on Computer Vision (ECCV) (2000) 5
40. Shibuya, M., Sumikura, S., Sakurada, K.: Privacy preserving visual SLAM. In: European Conference on Computer Vision (ECCV) (2020) 4
41. Six.D AI: http://6d.ai/ (2018) 1
42. Speciale, P., Schönberger, J.L., Kang, S.B., Sinha, S.N., Pollefeys, M.: Privacy preserving image-based localization. In: Computer Vision and Pattern Recognition (CVPR) (2019) 1, 2, 4, 10, 11, 12
43. Speciale, P., Schönberger, J.L., Sinha, S.N., Pollefeys, M.: Privacy preserving image queries for camera localization. In: International Conference on Computer Vision (ICCV) (2019) 1, 2, 4, 9

44. Strecha, C., von Hansen, W., Gool, L.V., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: Computer Vision and Pattern Recognition (CVPR) (2008) 10, 11
45. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (2002) 3
46. Upmanyu, M., Namboodiri, A.M., Srinathan, K., Jawahar, C.: Efficient privacy preserving video surveillance. In: International Conference on Computer Vision (ICCV) (2009) 3
47. Upmanyu, M., Namboodiri, A.M., Srinathan, K., Jawahar, C.: Blind authentication: A secure crypto-biometric verification protocol. IEEE Transactions on Information Forensics and Security (2010) 3
48. Vinje, J.E.: Privacy Manifesto for AR Cloud Solutions. https://medium.com/openarcloud/privacy-manifesto-for-ar-cloud-solutions-9507543f50b6 (2018) 1
49. Vondrick, C., Khosla, A., Malisiewicz, T., Torralba, A.: HOGgles: Visualizing object detection features. In: Computer Vision and Pattern Recognition (CVPR) (2013) 4
50. VPS: Google Visual Positioning System. https://www.engadget.com/2018/05/08/g/ (2018) 1
51. Wang, Z., Vineet, V., Pittaluga, F., Sinha, S., Cossairt, O., Kang, S.B.: Privacy-preserving action recognition using coded aperture videos. Computer Vision and Pattern Recognition (CVPR) Workshops (2019) 3
52. Weinzaepfel, P., Jégou, H., Pérez, P.: Reconstructing an image from its local descriptors. In: Computer Vision and Pattern Recognition (CVPR) (2011) 4
53. Wilson, K., Snavely, N.: Robust global translations with 1DSfM. In: European Conference on Computer Vision (ECCV) (2014) 10, 13, 14
54. Wu, Z., Wang, Z., Wang, Z., Jin, H.: Towards privacy-preserving visual recognition via adversarial training: A pilot study. In: European Conference on Computer Vision (ECCV) (2018) 3
55. Yonetani, R., Boddeti, V.N., Kitani, K.M., Sato, Y.: Privacy-preserving visual learning using doubly permuted homomorphic encryption. In: International Conference on Computer Vision (ICCV) (2017) 3
56. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. In: ICML Workshop on Deep Learning (2015) 4
57. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision (ECCV) (2014) 4
58. Zhao, J., Frumkin, N., Konrad, J., Ishwar, P.: Privacy-preserving indoor localization via active scene illumination. In: Computer Vision and Pattern Recognition (CVPR) Workshops (2018) 3