Privacy Preserving Structure-from-Motion Supplementary Material

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

Supplementary Material

In the supplementary material we present

- Additional experiments and discussion regarding the initialization (Section 1),
- Results with COLMAP [2] on the Strecha [4] for comparison (Section 2)
- More qualitative reconstruction results and comparisons (Section 3)
- Qualitative results with the feature inversion method from [1] (Section 4)
- Detailed runtime comparison with COLMAP (Section 5)

See also the provided video where we show a qualitative comparison of applying the InvSfM method from Pittaluga [1], both on the reconstruction from COLMAP [2] and from our privacy preserving pipeline. The video shows results from the internet image collection dataset *the Alamo* from Wilson and Snavely [5].

Finally, in Figure 1 we provide additional illustration and explanation of the line to point projection constraint.

1 Additional Experiments for Initialization

In the main paper we propose an initialization scheme which first estimates the poses of four cameras in 2D w.r.t. the ground plane, followed by an upgrade step which estimates the out-of-plane translations for each camera. Each of the steps are performed by running minimal solvers in RANSAC for robust estimation. In our case all of the involved equations are linear and thus the corresponding solvers are very stable. To evaluate the numerical stability of the solvers we generated 1000 noise-free synthetic instances. Figure 2 shows histograms of the camera pose errors for the initialization solvers. We can see that both the 2D estimation and complete pipeline including the upgrade to full pinhole cameras are stable.

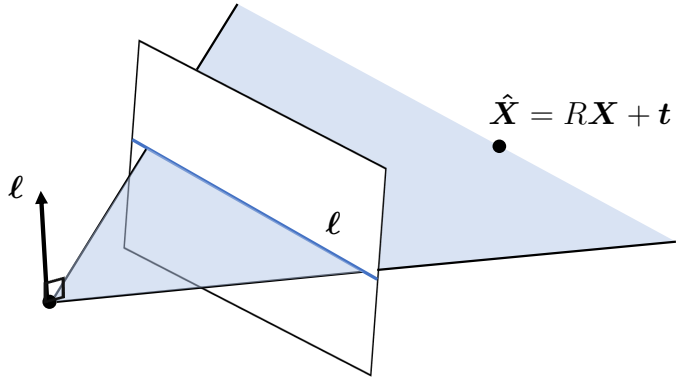

Fig. 1. *Line projection constraint.* The line $\boldsymbol{\ell}$ can be interpreted both as a 2D line in the image plane (in homogeneous representation) and as the normal to the plane going through the origin and tracing out the corresponding line in the image plane. If a 3D point projects onto this line, it must lie in this plane after transforming it into the camera coordinate system, i.e. $\boldsymbol{\ell}^T(R\boldsymbol{X} + \boldsymbol{t}) = 0$.

1.1 Degeneracies for Initialization

The 2D trifocal tensor estimation degenerates if two or more cameras coincide. In our setting this might happen even if the original cameras are distinct. Geometrically this occurs whenever the translation of the camera is along the gravity direction. In this case the corresponding gravity-aligned 2D cameras will have the same (2D-)camera center.

2 COLMAP Results on the Strecha Benchmark [4]

In the main paper we present reconstruction statistics and camera pose accuracies on the Strecha benchmark datasets [4]. As a reference we additionally provide the the same data for reconstructions with COLMAP in Table 1. Similarly to our method, we can see that COLMAP also obtains higher errors for the two castle datasets (*castle-P19*, *castle-P30*) .

3 Qualitative Results and Comparisons

In addition to the quantitative results in the paper, we show qualitative comparisons between standard SfM and our privacy preserving SfM on the remaining datasets provided by Speciale *et al.* [3] in Figure 3.

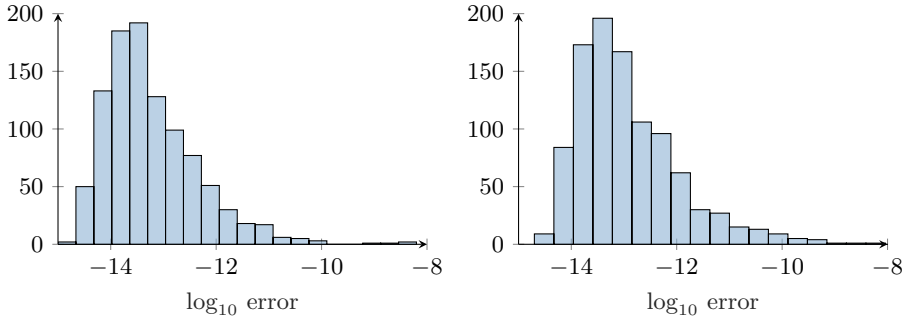

Fig. 2. Numerical stability of initialization. The histograms show the errors in the resulting camera poses, i.e. $\log_{10} \sum_{k=1}^{4} \|P_i - P_i^{GT}\|$. *Left:* Four-view 2D estimation (trifocal+resectioning) *Right:* Full initialization.

Table 1. Evaluation of camera pose accuracy on the Strecha benchmark [4] using COLMAP. This is for comparison with the line based results in Table 1 from the paper.

	#Images		#Points		Track Length	Rotation (deg)		Position (cm)	
Scene	Total	Reg.	3D	Obs.		Mean	Std.	Mean	Std.
castle-P19	19	19	11.3k	46.0k	4.06	0.15	0.07	4.96	5.35
castle-P30	30	30	22.1k	108.7k	4.91	0.06	0.03	2.88	1.68
entry-P10	10	10	8.1k	36.4k	4.51	0.04	0.01	0.63	0.29
fountain-P11	11	11	13.6k	64.5k	4.74	0.03	0.01	0.26	0.12
Herz-Jesu-P8	8	8	7.3k	29.6k	4.08	0.11	0.01	0.34	0.14
Herz-Jesu-P25	25	25	19.5k	112.6k	5.76	0.05	0.02	0.53	0.25

4 Feature Inversion Comparison

We provide additional examples of the feature inversion results with InvSfM [1] in Figure 4 and Figure 5. As the results in the main paper, COLMAP inversions are generated from all extracted image features, while for our method we project triangulated points into the camera.

5 Runtime Comparison

Especially for larger scenes the runtime of a reconstruction can grow significantly and can quickly become a limitation of Structure-from-Motion methods. We compare the runtimes of the different steps in our pipeline with COLMAP. The runtimes for feature extraction and matching are essentially the same for both methods and we therefore omit them. Generally, line-based constraints are weaker and more likely to be noisy, and take more time in all steps that include RANSAC (initialization, image registration, point triangulation), but also

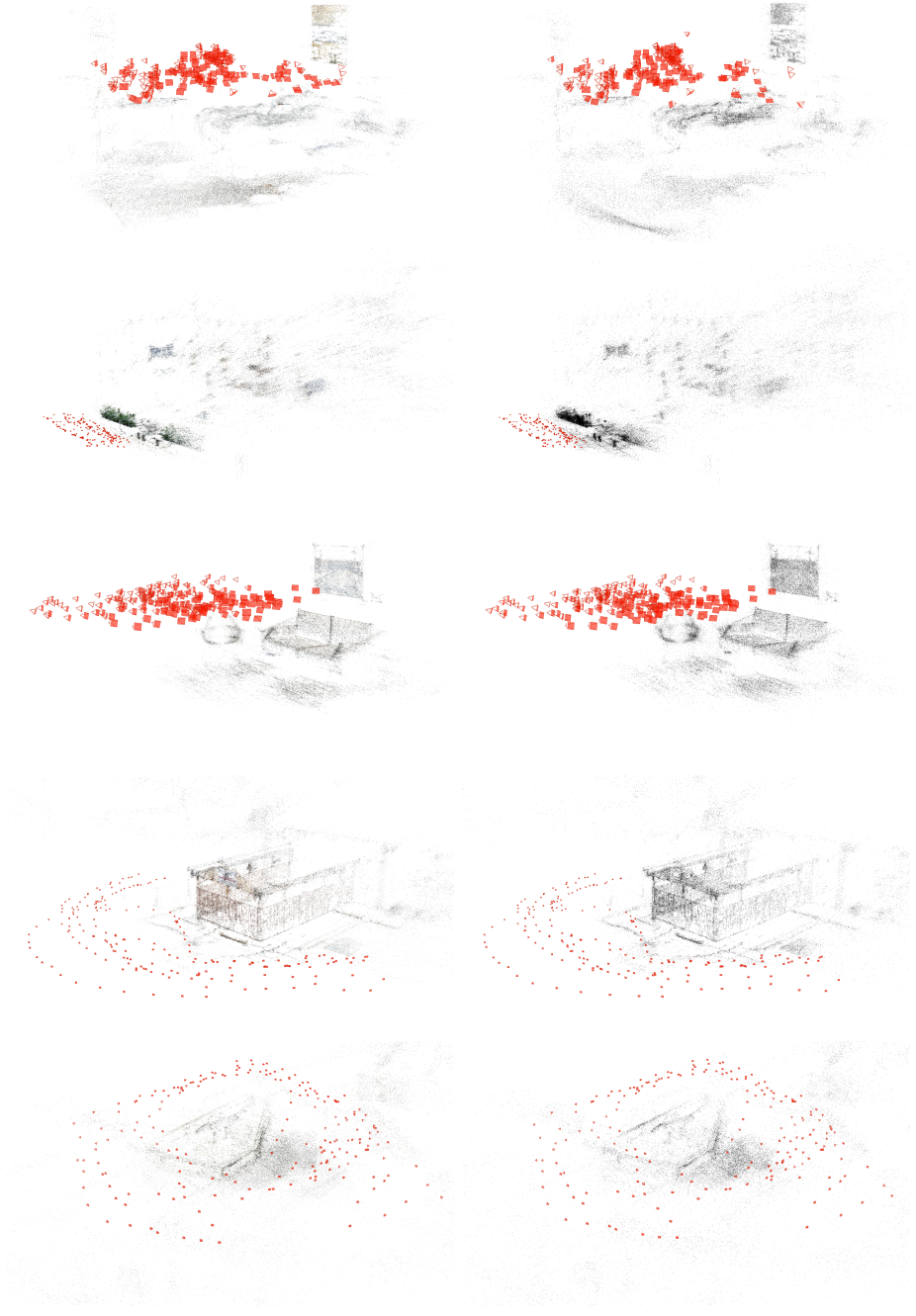

Fig. 3. Qualitative comparison between standard SfM implemented in COLMAP [2] (left) and our privacy preserving SfM (right)

in the first, local Bundle Adjustment after the image registration. However, in our line-based reconstructions there are generally less points triangulated. This makes the global Bundle Adjustment problem much simpler and leads to a significantly lower runtime compared to COLMAP. As this step usually dominates the runtime for large reconstructions, our total reconstruction times are lower than with COLMAP.

Of course, these runtimes are highly dependent on the particular parameters and thresholds used. For easier comparison we used the default COLMAP parameters for both methods. For the *Madrid Metropolis* dataset, these thresholds resulted in many images being incorrectly registered for our method which explains the longer runtimes.

Table 2. Runtime comparison between COLMAP and our privacy preserving implementation. We report the accumulated times spent in Initialization (Init.), Image registration (Reg.), Point triangulation (Tri.), Failed image registration attempts (Failed Reg.), and local and global Bundle Adjustment (Local/Global BA) as well as the total time spent for the reconstruction. The numbers of registered images and triangulated points differ from the ones reported in the paper due to different thresholds and random factors (*i.e.* RANSAC) during the reconstruction. Runtimes are reported in seconds and rounded as *COLMAP / Ours*.

Scene	#Reg. Images	#3D Points	Init.	Reg.	Tri.	Failed Reg.	Local BA	Global BA	**Total**
Alamo	883 / 821	181k / 80k	0 / 36	29 / 52	5 / 4	20 / 213	527 / 814	14.5k / 4.2k	**15.2k / 5.5k**
Gendarmenmarkt	1016 / 902	250k / 83k	0 / 4	20 / 98	10 / 7	7 / 165	353 / 721	9.9k / 1.4k	**10.5k / 2.5k**
Madrid Metropolis	475 / 1091	89k / 80k	0 / 11	13 / 517	16 / 64	24 / 187	188 / 1204	3.1k / 2.1k	**3.4k / 4.3k**
Tower of London	727 / 601	180k / 93k	0 / 20	17 / 27	5 / 4	10 / 135	342 / 604	5.5k / 1.5k	**6.0k / 2.3k**

Fig. 4. Comparison of InvSfm [1] results with and without privacy preserving SfM. *Top:*The Alamo dataset [5]. *Bottom:* Gendarmenmarkt dataset [5].

Fig. 5. Comparison of InvSfm [1] results with and without privacy preserving SfM. *Top:* Madrid Metropolis dataset [5]. *Bottom:* Tower of London dataset [5].

References

1. 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)
2. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Computer Vision and Pattern Recognition (CVPR) (2016)
3. 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)
4. 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)
5. Wilson, K., Snavely, N.: Robust global translations with 1DSfM. In: European Conference on Computer Vision (ECCV) (2014)