Uncalibrated Structure from Motion on a Sphere

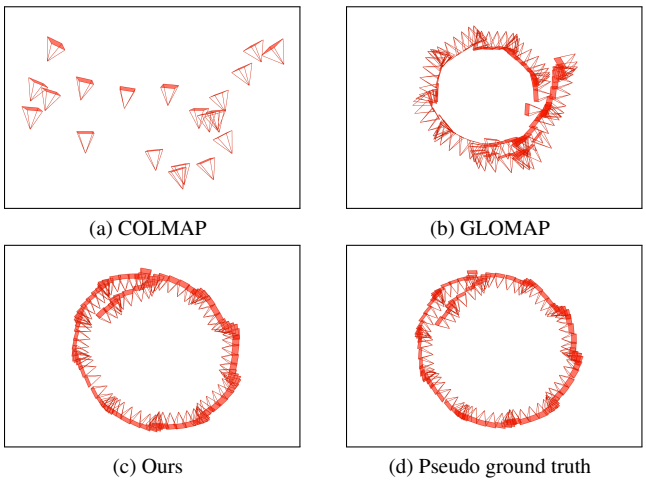
Jonathan Ventura
California Polytechnic State University
jventu09@calpoly.edu

Viktor Larsson
Lund University
viktor.larsson@math.lth.se

Fredrik Kahl
Chalmers University of Technology
fredrik.kahl@chalmers.se

Abstract

Spherical motion is a special case of camera motion where the camera moves on the imaginary surface of a sphere with the optical axis normal to the surface. Common sources of spherical motion are a person capturing a stereo panorama with a phone held in an outstretched hand, or a hemi-spherical camera rig used for multi-view scene capture. However, traditional structure-from-motion pipelines tend to fail on spherical camera motion sequences, especially when the camera is facing outward. Building upon prior work addressing the calibrated case, we explore uncalibrated reconstruction from spherical motion, assuming a fixed but unknown focal length parameter. We show that, although two-view spherical motion is always a critical case, self-calibration is possible from three or more views. Through analysis of the relationship between focal length and spherical relative pose, we devise a global structure-from-motion approach for uncalibrated reconstruction. We demonstrate the effectiveness of our approach on real-world captures in various settings, even when the camera motion deviates from perfect spherical motion. Code and data for our method are available at https://github.com/jonathanventura/spherical-sfm.



(a) COLMAP (b) GLOMAP (c) Ours (d) Pseudo ground truth

Figure 1. Comparison of our SFM results to COLMAP [29] and GLOMAP [26] on handheld camera sequence *arboretum4* from our Phone Sweep dataset. (a) COLMAP is unable to find a good camera pair for initialization and produces an incomplete and inaccurate reconstruction. (b) GLOMAP erroneously estimates the motion to be inward-facing and thus produces an inside-out reconstruction. (c) Our SFM method accurately estimates all of the camera poses, as can be seen by comparison to (d) the pseudo ground truth camera poses obtained by running COLMAP with additional images to ensure a successful reconstruction.

1. Introduction

Given a collection of input images of a scene or object, the problem of structure from motion (SFM) is to estimate both the camera parameters and the 3D structure of the scene [23, 25]. Ventura [37] introduced a special case of SFM where the cameras lie on the surface of an imaginary sphere, with the optical axis parallel to the surface normal. In an inward-facing configuration, each camera faces to the center of the sphere, as in the case of a camera gantry for object scanning. In an outward-facing configuration, each camera faces away from the center, as in the case of a person turning while holding a phone [2–4, 27, 28], or a hemi-spherical multi-camera rig for one-shot scene capture [5, 7].

General structure-from-motion systems such as COLMAP [29] and GLOMAP [26] tend to fail on image collections in an outward-facing spherical configuration, often producing either partial reconstructions or highly inaccurate reconstructions [2]. See Figure 1 for an example.

Building upon previous work on spherical SFM [36, 37], in this work we explore the uncalibrated case, assuming a single unknown focal length for all cameras. It is well known that any two-view camera configuration with intersecting optical axes is a critical motion for self-calibration [17, 19, 34]; thus two-view spherical motion is always a critical motion. However, we show that self-calibration is always possible from three or more distinct viewpoints with spherical motion. We show that essential matrices and fundamental matrices are in fact interchangeable under spher-

ical motion with a single unknown focal length, which allows us to re-use Ventura's calibrated minimal solver [37] for the uncalibrated case. Furthermore, we analyze the geometry of two-view uncalibrated spherical motion to determine the relationship between focal length and relative rotation for a given fundamental matrix. Based on these results, we develop a global SFM pipeline in which we first solve for all pairwise fundamental matrices, then find the focal length setting that maximizes agreement among the pairwise rotations through parameter search and refinement, and finally perform point triangulation and bundle adjustment to optimize the result. By removing the spherical constraint in the final step, we are able to accurately perform 3D reconstruction even when the camera motion deviates from the spherical assumption.

To evaluate our proposed approach and compare to baseline methods, we developed a new dataset called PhoneSweep, consisting of thirteen sequences captured by hand with two different smartphones and including pseudo-ground truth produced using traditional SFM methods. We also evaluated our approach using data from a hemispherical camera rig. Our evaluation demonstrates the clear advantage of our approach over traditional SFM systems in accurately processing sequences with near-spherical camera motion. Furthermore, we demonstrate the usefulness of our technique for downstream applications such as dense multi-view stereo and view synthesis.

In summary, our novel contributions are as follows:

- Theoretical analysis of uncalibrated spherical two-view and multi-view geometry;
- Global initialization procedure for uncalibrated spherical SFM;
- PhoneSweep dataset with pseudo-ground truth for evaluating spherical SFM methods;
- Comparative evaluation with traditional SFM techniques;
- Demonstration of view synthesis and dense depth estimation using our SFM results.

2. Related work

2.1. General structure-from-motion

General approaches to SFM can be categorized as taking either an incremental (cf. [29, 31]) or global (cf. [26, 41]) approach. COLMAP [29] is a widely used incremental SFM system which first estimates two-view geometry for an initial image pair and then incrementally incorporates other images into the reconstruction. In the uncalibrated case without any priors on the camera calibration, COLMAP initializes the focal length using a heuristic and then refines it as the reconstruction proceeds. GLOMAP [26] is a recent global SFM system which provides competitive accuracy compared to COLMAP but is usually much faster. GLOMAP randomly initializes all of the camera poses and 3D points and then iteratively optimizes them. In the uncalibrated case, GLOMAP applies view graph optimization over image triplets [35] to find an initial estimate for the focal length.

2.2. Spherical structure-from-motion

Ventura [37] introduced a three-point minimal solver for the essential matrix under the spherical motion assumption and showed how to uniquely decompose the essential matrix into a relative pose solution, assuming knowledge of whether the motion is inward- or outward-facing. They integrated this solver into a global structure-from-motion pipeline which uses rotation averaging [9, 12] to initialize the camera poses. In the paper we show that each spherical fundamental matrix is also a spherical essential matrix (belonging to a different focal length). This non-trivial observation allows us to use the three-point solver [37], which was previously only used for calibrated reconstruction, to also estimate the spherical fundamental matrix.

Baker et al. [2] refined the approach of Ventura [37] and demonstrated its use for stereo panorama creation [28] from handheld camera video. Other applications include simultaneous localization and mapping (SLAM) [3] and augmented reality [4]. Joo et al. [15, 16] extended the spherical motion concept to "spherical joint" motion, where the camera is at a fixed offset from the sphere surface, as in the case of a camera on a selfie stick.

While these methods assumed calibrated cameras, Sweeney et al. [36] proposed minimal solvers and a global method for uncalibrated spherical SFM. Their method however uses a rotation-only solver [8] to determine the calibration parameters, limiting its applicabilty to scenes with distant points. Zhang et al. [43] introduced a technique called DFR for uncalibrated stereo pair rectification under a latitudinal motion constraint; however their method only estimates rectifying homographies, not the camera motion itself, and is limited to latitudinal motion.

Larsson et al. [21] introduced a three-view solver for uncalibrated relative pose based on the 1D radial camera model and integrated it into an incremental SFM pipeline [29]. Hruby et al. [14] later introduced a four-view solver which improves the efficiency of the SFM initialization. Their techniques are applicable for spherical camera motion; however, in our evaluation our approach achieves both higher accuracy and faster run times.

3. Multi-view geometry of calibrated and uncalibrated spherical motion

3.1. Calibrated two-view geometry

Here we review the basics of calibrated two-view geometry with spherical motion [37]. A unique property of spherical motion is that, although the camera does observe parallax

since it is translating, the translation is entirely determined by the camera rotation. As a result, the absolute pose and relative pose under spherical motion are completely specified by three rotational degrees of freedom.

Suppose we have two outward-facing cameras on the unit sphere with extrinsics $\mathsf{P}^1 = [\mathsf{R}^1 \mid -\mathbf{z}]$ and $\mathsf{P}_2 = [\mathsf{R}^2 \mid -\mathbf{z}]$ where $\mathbf{z} = [0\ 0\ 1]^\top$. Writing the relative rotation as $\mathsf{R} = \mathsf{R}^2{\mathsf{R}^1}^T$, the relative pose P bringing points from camera 1 to 2 is

$$\mathsf{P} = [\ \mathsf{R} \mid \mathsf{R}_{:,3} - \mathbf{z}\ ] \tag{1}$$

where $\mathsf{R}_{:,3}$ is the third column of R.

The essential matrix E relates corresponding homogeneous points $\mathbf{u}$ and $\mathbf{v}$ in cameras 1 and 2, respectively, such that $\mathbf{v}^\top \mathsf{E}\mathbf{u} = 0$. The essential matrix for outward-facing cameras undergoing spherical motion is

$$\mathsf{E} \sim [\mathsf{R}_{:,3} - \mathbf{z}]_\times \mathsf{R} \tag{2}$$

where $[\mathbf{a}]_\times$ is the skew-symmetric matrix such that $[\mathbf{a}]_\times \mathbf{b} = \mathbf{a} \times \mathbf{b}\ \forall\ \mathbf{b}$. In the case of inward-facing cameras, the translation vector is negated. Since the essential matrix is only defined up to scale, this means the essential matrices for inward-facing and outward-facing cameras undergoing the same relative rotation are equivalent.

3.2. Spherical essential matrix parameterizations

Here we introduce a new form for the spherical essential matrix E, which directly expresses E using elements of R:

$$\mathsf{E} \sim \begin{bmatrix} R_{2,1} + R_{1,2} & R_{2,2} - R_{1,1} & R_{2,3} \\ R_{2,2} - R_{1,1} & -R_{2,1} - R_{1,2} & -R_{1,3} \\ R_{3,2} & -R_{3,1} & 0 \end{bmatrix}. \tag{3}$$

See the supplemental material (SM) for a full derivation.

Any rotation R can be decomposed into a rotation R_{xy} of angle θ_{xy} about a unit norm vector $\mathbf{r}_{xy}$ in the x-y plane and a rotation R_z about the z axis s.t. $\mathsf{R}(\mathbf{r}_{xy}, \theta_{xy}, \theta_z) = \mathsf{R}_{xy}(\mathbf{r}_{xy}, \theta_{xy})\mathsf{R}_z(\theta_z)$. The rotations can be written as follows:

$$\mathsf{R}_{xy} = \begin{bmatrix} (c_{xy}-1)r_y^2 + 1 & -r_x r_y (c_{xy} - 1) & r_y s_{xy} \\ r_x r_y (c_{xy} - 1) & (c_{xy}-1)r_x^2 + 1 & -r_x s_{xy} \\ -r_y s_{xy} & r_x s_{xy} & c_{xy} \end{bmatrix} \tag{4}$$

where $c_{xy} = \cos(\theta_{xy})$ and $s_{xy} = \sin(\theta_{xy})$, and

$$\mathsf{R}_z = \begin{bmatrix} c_z & -s_z & 0 \\ s_z & c_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{5}$$

where $c_z = \cos(\theta_z)$ and $s_z = \sin(\theta_z)$. In this paper we assume that $\theta_{xy} \neq 0$, since otherwise there is no camera translation and we are unable to estimate two-view geometry.

Plugging (4) and (5) into (3) we obtain a second form for the essential matrix, which will be useful when we next derive the relationship between focal length and rotation in two-view geometry:

$$\mathsf{E}(r_{xy}, \theta_{xy}, \theta_z) \sim \begin{bmatrix} (c_{xy}-1)S(r_x, r_y, \theta_z) & \begin{matrix} -s_{xy} r_x \\ -s_{xy} r_y \end{matrix} \\ \begin{matrix} s_{xy}(c_z r_x + r_y s_z) & s_{xy}(c_z r_y - r_x s_z) \end{matrix} & 0 \end{bmatrix} \tag{6}$$

where

$$S(r_x, r_y, \theta_z) = \begin{bmatrix} s_z r_x^2 - 2c_z r_x r_y - s_z r_y^2 & c_z r_x^2 + 2 s_z r_x r_y - c_z r_y^2 \\ c_z r_x^2 + 2 s_z r_x r_y - c_z r_y^2 & -s_z r_x^2 + 2 c_z r_x r_y + s_z r_y^2 \end{bmatrix}.$$

3.3. Uncalibrated two-view geometry

In the uncalibrated case we assume a single unknown focal length f and known principal point and skew, such that the intrinsics matrix is $\mathsf{K}(f) = \mathrm{diag}(f, f, 1)$. Corresponding homogeneous points $\mathbf{x}, \mathbf{y}$ in two uncalibrated views are related by the fundamental matrix F such that $\mathbf{y}^T \mathsf{F} \mathbf{x} = 0$.

The fundamental matrix is related to the essential one by

$$\mathsf{F}(f, r_{xy}, \theta_{xy}, \theta_z) \equiv \mathsf{K}(f)^{-T} \mathsf{E}(r_{xy}, \theta_{xy}, \theta_z) \mathsf{K}(f)^{-1} \tag{7}$$

$$\sim \begin{bmatrix} E_{11}/f & E_{12}/f & E_{13} \\ E_{21}/f & E_{22}/f & E_{23} \\ E_{31} & E_{32} & 0 \end{bmatrix}. \tag{8}$$

Since two-view spherical motion is always a critical motion [18, 33], from any fundamental matrix arising from spherical motion, we cannot uniquely determine the focal length which leads to a metric reconstruction. Accordingly, the following proposition describes how a 1D family of fundamental matrices are equivalent to a given essential matrix.

Proposition 1. *Assuming a single unknown focal length, any spherical essential matrix $E(r_{xy}, \theta_{xy}, \theta_z)$ is equivalent to a family of spherical fundamental matrices parameterized by the focal length f:*

$$E(r_{xy}, \theta_{xy}, \theta_z) \sim F(f, r_{xy}, \theta'_{xy}, \theta_z) \tag{9}$$

for any choice of f, where

$$\begin{aligned} \theta'_{xy}(f, \theta_{xy}) = atan2(2f sin(\theta_{xy}), \\ (1 + f^2) cos(\theta_{xy}) + (1 - f^2)). \end{aligned} \tag{10}$$

In the SM we prove Proposition 1 and the following converse proposition:

Proposition 2. *Any spherical fundamental matrix with constant focal length is equivalent to a spherical essential matrix.*

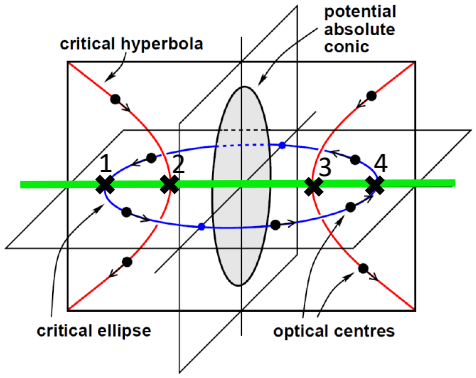


Figure 2. The critical motions for unknown and possibly varying focal lengths derived in [19, 34]. The motions are critical if the camera centers lie on either the critical hyperbola or the critical ellipse with the principal axes tangent to the hyperbola or ellipse respectively.

Since any spherical fundamental matrix is also a spherical essential matrix, it follows that any minimal solver for the spherical essential matrix [37] can be applied, without modification, to estimate the spherical fundamental matrix.

3.4. Three views and more

Now consider n cameras $\mathsf{P}^1, \mathsf{P}^2, \ldots, \mathsf{P}^n$ where $n > 2$. We will assume that there are at least three distinct optical axes. As before, we also assume a spherical camera motion and that the only self-calibration parameter is a single, unknown focal length, common for all cameras.

A well-established result (see [19, 34]) in the self-calibration literature is that the only critical motions for unknown but possibly varying focal lengths are the following: (i) camera motions with one optical axis, or (ii) motion on two conics (one ellipse and one hyperbola) whose supporting planes are orthogonal and where the optical axis is tangent to the conic at each position, see Figure 2. The potential (or false) absolute conic is on the plane orthogonal to these two supporting planes. If we can show that neither of these cases are feasible for spherical camera motion, then we have also shown that there are no critical motions for the constant focal length case that we are interested in. Note that assuming constant focal length is an additional constraint and therefore the critical motions will be a subset of those for varying focal lengths.

First, case (i) is by assumption excluded so it is not critical. Now, consider case (ii). If it were to be critical for a spherical motion, then the optical axes must intersect. In the following discussion, we will consider the locus of points that have one or more optical axes intersecting. We will show that there is no single point that has strictly more than two optical axes intersecting.

As the optical axes are tangent to the conics, optical axes intersections only occur at the supporting planes of the conics. First, consider the planar elliptical curve. For any point $\mathbf{p}$ outside the ellipse in the supporting plane, there are exactly two real tangent lines to the ellipse that intersect at $\mathbf{p}$. This follows from the pole-polar relationship in projective geometry [30]: The polar line of pole $\mathbf{p}$ intersects the conic curve in exactly two tangency points. For points inside the elliptic curve, there are no real polar lines and hence no optical axes intersections. Second, consider the supporting plane of the hyperbola. Analogously, for a point in this plane, there are at most two optical axes emanating from the hyperbola. The only possibility that there would be strictly more than two optical axes intersecting at a single point $\mathbf{p}$ would be on the intersection line of the two supporting planes. In Figure 2, the intersection line is marked in green.

Along this line, there are four points that meet the two conics which can be ordered: first an ellipse point (nr. 1), followed by two hyperbola points (nr. 2 and 3), and then the other ellipse point (nr. 4). In Figure 2, these four points are numbered and marked with crosses. The ordering is always the same – see the formula in [19]. In the first interval (before point nr. 1), there can only be two optical axes intersecting as the polar lines to the hyperbola are not real. Then, from point nr. 1 to nr. 2, there are no optical axes intersections, and then, from nr. 2 to nr. 3, for points on the finite interval inside the hyperbola, there are again exactly two intersecting optical axes. The remaining intervals have by symmetry no more than two intersection points. In conclusion, nowhere along this line are there more than two optical axes intersecting and therefore the potential absolute conic cannot be critical for a spherical camera motion with three or more cameras.

Proposition 3. *Multi-view spherical camera motion with at least three distinct optical axes never constitutes a critical motion.*

4. Uncalibrated global structure-from-motion

Buildings on the results presented above, we present a global SFM approach for uncalibrated reconstruction from a spherical motion video. The pipeline is summarized in the following steps:

1. Estimate the spherical fundamental matrix F^{ij} between each pair of images i, j (Section 4.1).
2. Initialize the focal length and camera rotations through one-dimensional search and minimization over the focal length and then refine the focal length and rotations using non-linear optimization (Section 4.2).
3. Triangulate points and apply spherical followed by general bundle adjustment (Section 4.3).

4.1. Pairwise fundamental matrix estimation

Between each pair of views i, j we match feature points and attempt to estimate the spherical fundamental matrix F^{ij} using LO-RANSAC [11, 22] with the spherical essential matrix solver [37]. We accept any image pair with greater than τ_{num} inliers using an inlier threshold of τ_{inlier} to form the set of image matches $\mathcal{M}$.

For numerical stability we normalize point observations by an initial intrinsics matrix K_{init} with focal length f_{init}. The essential matrix E^{ij} is then computed as $\mathsf{K}_{init}^T \mathsf{F}^{ij} \mathsf{K}_{init}$.

4.2. Camera initialization and refinement

Given the essential matrix E^{ij} we decompose the underlying rotation into $\mathbf{r}_{xy}^{ij}, \theta_{xy}^{ij}, \theta_z^{ij}$ (see SM). From (9) we parameterize the family of rotations $\mathsf{R}^{ij}(f)$ between views i and j by the focal length f:

$$\mathsf{R}^{ij}(f) \equiv \mathsf{R}(r_{xy}^{ij}, \theta'_{xy}(f, \theta_{xy}^{ij}), \theta_z^{ij}). \tag{11}$$

We define a cost function $C(\mathsf{R}^1, \ldots, \mathsf{R}^n, f)$ that evaluates the agreement between the absolute rotations and the relative rotations according to the choice of focal length f:

$$C(\mathsf{R}^1, \ldots, \mathsf{R}^n, f) = \sum_{i,j \in \mathcal{M}} d^{\mathsf{R}} \left(\mathsf{R}^{ij}(f), \mathsf{R}^j \mathsf{R}^{i^{-1}} \right) \tag{12}$$

where $d^{\mathsf{R}}(\mathsf{R}_1, \mathsf{R}_2) = \rho_{\mathsf{R}}(|| \log_{SO(3)}(\mathsf{R}_1 \mathsf{R}_2^T)||^2)$ and $\rho_{\mathsf{R}}(\cdot)$ is a robust loss function.

We search for an optimal setting for the focal length by selecting random samples for $f \in [f_{min}, f_{max}]$ and evaluating the cost function C. For a given setting of f, to evaluate C we need to initialize the rotations according to the relative rotations. If the images were captured sequentially in a video, we use the temporal sequence of relative rotations to initialize the poses; i.e., we set $\mathsf{R}^1 = \mathsf{I}$ and $\mathsf{R}^i = \mathsf{R}^{(i-1)i} \mathsf{R}^{i-1}$ for all $i > 1$. Otherwise, for an unordered image set, we apply hybrid rotation averaging [10] to initialize the rotations.

In the SM we analyze the behavior of the cost function (12) on the datasets used in our real data experiments.

After finding the best focal length setting by random search, we refine the best estimate for f and the rotations $\mathsf{R}^1, \ldots, \mathsf{R}^n$ using iterative non-linear optimization on C.

4.3. Point triangulation and bundle adjustment

Now that we have initialized and refined the rotations and focal length estimate, we proceed to triangulate 3D points and apply bundle adjustment. After collecting point observations into feature tracks, we use a robust point triangulation procedure similar to COLMAP [29] by wrapping the triangulation procedure in a LO-RANSAC loop [11, 22]. We then apply robust bundle adjustment (BA) with a spherical motion constraint by fixing the translation vector of each camera to $[0\ 0\ -1]^T$. After BA we re-triangulate the points and apply BA a second time. Re-triangulating allows for points that were previously considered outliers to be re-considered as the camera pose estimates become more accurate.

Finally we remove the spherical motion constraint by including the translation vectors in the bundle adjustment, and run two more rounds of re-triangulation and BA. For these final refinement steps we use COLMAP's functions for triangulation and BA, for fair comparison against COLMAP's camera pose initialization procedure.

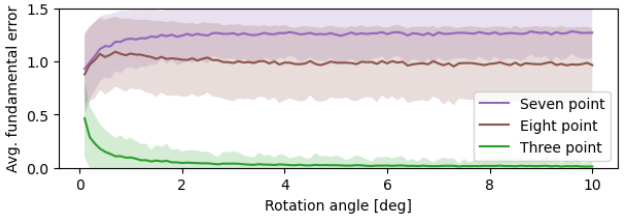


Figure 3. Comparison of error in fundamental matrix estimation under spherical motion with increasing rotation angle and 1 px std. dev. observation noise. The solid line shows the average error and the shaded area shows one std. dev. above and below the mean.

5. Experiments

5.1. Synthetic data experiment

We performed a synthetic experiment data as an initial investigation of why traditional structure-from-motion pipelines tend to fail on spherical motion sequences. Since COLMAP [29] and GLOMAP [26] both rely on fundamental matrix estimation as the basis for uncalibrated SFM, we compared the accuracy of the traditional eight-point linear and seven-point non-linear solvers [13] versus the three-point solver [37] for fundamental matrix estimation. See the SM for more detail.

Figure 3 plots the Frobenius norm of the error in the estimated fundamental matrix as the rotation angle increases. The seven- and eight-point solvers are unstable on this particular class of relative pose problem, while the three-point solver is far more reliable. This motivates our use of the three-point solver in our proposed uncalibrated spherical SFM pipeline.

5.2. Real data experiments

5.2.1. Methods

We tested the following uncalibrated SFM systems in our real data experiments: COLMAP [29], GLOMAP [26], RadialSFM [14, 21] and Ours. All methods used the same features [24] and features matches extracted by COLMAP. We were unable to test the method of Sweeney et al. [36] as there is no public implementation available.

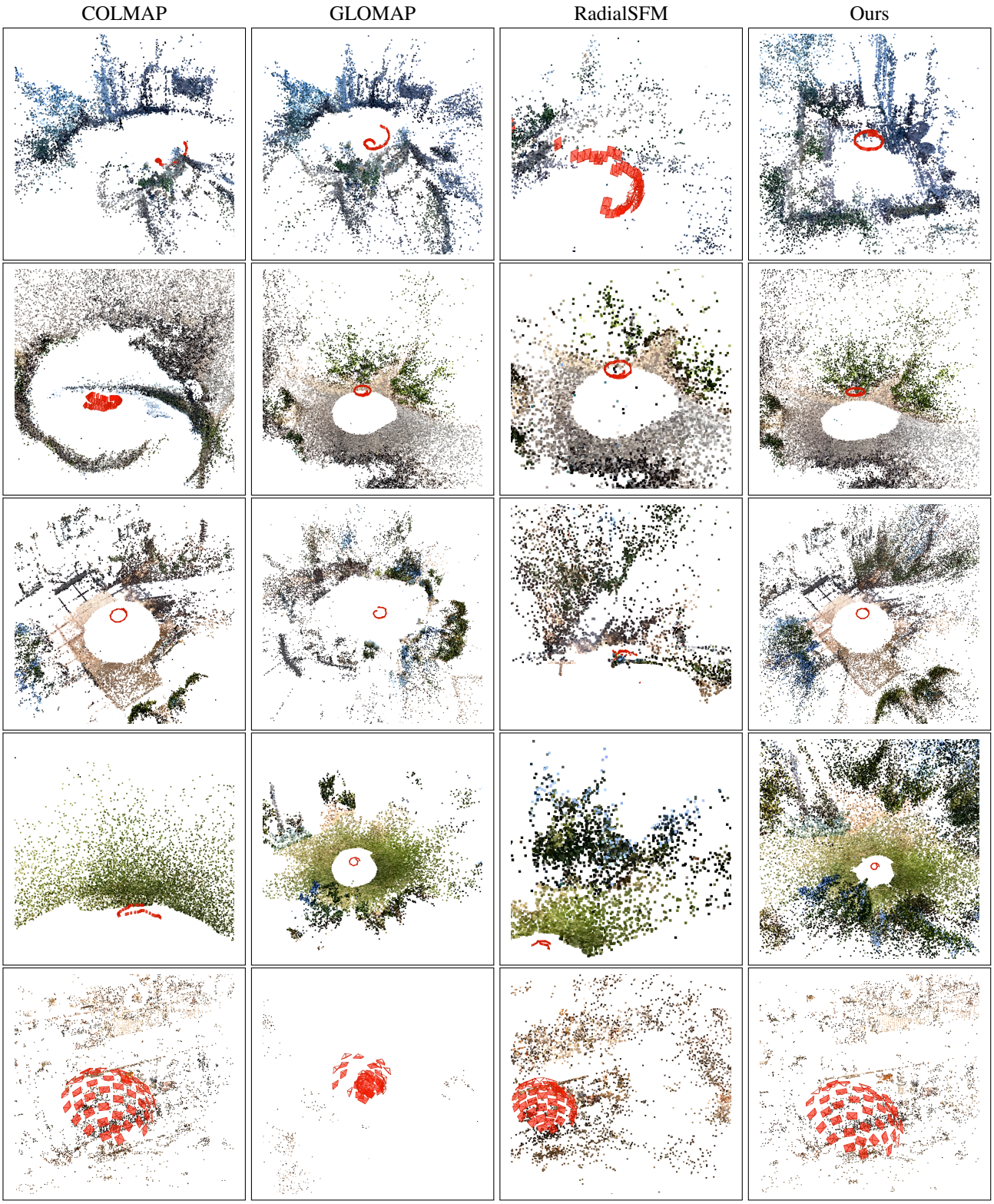


Figure 4. Example reconstruction results. COLMAP, GLOMAP and RadialSFM exhibit many failed reconstructions, where often the camera motion is incorrectly interpreted as inward-facing instead of outward-facing. In contrast, our method reliably reconstructs these scenes. From top to bottom: iPhone 13 scenes *deck* and *ocean* and Nexus 5X scenes *arboretum7* and *education1* from the PhoneSweep dataset; scene *01_Welder* from the Deep View Video dataset.

Table 1. Results for PhoneSweep dataset. Our method is consistently the best performer across all metrics.

Device	Method	RRA↑ @5	RRA↑ @15	RRA↑ @30	RTA↑ @5	RTA↑ @15	RTA↑ @30	AUC@30↑	AFE↓
iPhone 13	COLMAP	75.50	83.80	96.02	59.10	81.64	90.59	70.99	2.81
	GLOMAP	65.55	77.22	86.59	57.76	77.84	86.81	66.93	4.50
	RadialSFM	89.13	91.13	92.47	59.48	83.77	89.85	74.52	-
	Ours	100.00	100.00	100.00	86.65	98.77	99.77	91.45	0.25
Nexus 5X	COLMAP	86.91	98.80	100.00	68.42	85.68	91.78	78.33	3.21
	GLOMAP	40.91	67.22	89.66	29.12	57.68	86.49	47.49	4.70
	RadialSFM	70.89	81.83	85.85	33.40	58.67	69.55	48.98	–
	Ours	100.00	100.00	100.00	83.48	96.56	98.83	90.43	0.97
Combined	COLMAP	78.89	91.91	97.62	56.18	75.47	83.54	67.04	2.95
	GLOMAP	51.43	71.49	88.35	41.34	66.28	86.63	55.79	4.59
	RadialSFM	79.05	85.99	88.81	45.07	69.90	78.63	60.41	–
	Ours	100.00	100.00	100.00	84.83	97.50	99.23	90.87	0.58
	COLMAP (calib.)	92.00	92.25	92.25	76.85	90.88	94.73	83.16	0.00

Table 2. Results for Deep View Video dataset. Our method consistently the best performer across all metrics. We observed the best performance from our method when skipping the final general BA steps.

Method	Recall↑	RRA↑ @5	RRA↑ @15	RRA↑ @30	RTA↑ @5	RTA↑ @15	RTA↑ @30	AUC@30↑	AFE↓
COLMAP	58.15	78.09	95.40	100.00	18.76	38.55	59.07	36.44	4.01
GLOMAP	15.80	22.35	65.29	89.78	6.71	25.38	50.20	21.03	13.33
RadialSFM	27.50	72.61	80.32	85.01	17.02	33.45	46.12	30.04	–
Ours (with general BA)	87.74	99.71	100.00	100.00	74.53	91.56	96.42	83.87	0.70
Ours (without general BA)	93.12	99.56	100.00	100.00	93.63	99.78	99.97	91.99	0.59
COLMAP (calib.)	91.18	100.00	100.00	100.00	74.97	94.02	98.86	86.17	0.00

Our method was implemented in C++ using the Ceres library [1] for non-linear optimization. For all experiments we used the following settings: $\tau_{inlier} = 2$ px, $\tau_{num} = 100$, $f_{init} = (W + H)/2$ where W and H are the width and height of the image, respectively, $f_{min} = f_{init}/4$, and $f_{max} = 2f_{init}$. For the robust cost function in (12) we used the "Soft L1 Loss" from Ceres: $\rho_R(s) = 2a^2(\sqrt{1 + s/a^2} - 1)$ with $a = 0.03$. For robust BA we use the Cauchy loss $\rho(s) = \log(1 + s)$. All experiments were run on a Linux server with a 2.0 GHz 32-core CPU and 256 GB of RAM.

5.2.2. Datasets

PhoneSweep We introduce the PhoneSweep dataset which consist of thirteen scenes captured with both spherical camera motion. We captured seven scenes with an iPhone 13 camera and six scenes with a Nexus 5X camera. The iPhone 13 camera has a field-of-view (FOV) of $120°$ and the Nexus 5X FOV is $94°$. We captured the scenes with the phone in portrait orientation to maximize the vertical FOV.

In each scene we recorded a video with outward-facing spherical motion by holding the phone in an outstretched hand and slowly turning in a circle. The scenes include a variety of indoor and outdoor locations in both urban and natural settings. We sub-sampled the 30 frames per second (FPS) videos to either 6 or 3 FPS, depending on the speed of motion in the video and the distance to the scene, to ensure sufficient feature matches between frames.

Because we recorded these sequences outside of a laboratory setting, we could not capture precise ground truth for the camera motion using an external tracking system. Instead, we opted to use SFM techniques to produce pseudo-ground truth. However, traditional SFM techniques perform unreliably on spherical motion sequences – hence the motivation for this work. So, in each scene we captured a second video with general motion. We first reconstructed each general motion video in COLMAP and then localized the spherical motion video to the COLMAP reconstruction. The estimated camera poses for the spherical motion video are used as pseudo-ground truth in the subsequent evaluation of each method.

Deep View Video The Deep View Video dataset [5] was captured with a hemi-spherical rig of 16 Yi cameras and provides metric ground truth for the camera extrinsics and intrinsics obtained via external calibration. We used the

Table 3. Comparison of reconstruction speed, measured as time per image reconstructed. On the the PhoneSweep dataset, GLOMAP is the fastest on our method is second fastest. One Deep View Video, our method (without the general BA steps) is the fastest.

		COLMAP		GLOMAP		RadialSFM		Ours	
Dataset	N	N_r	T / N_r	N_r	T / N_r	N_r	T / N_r	N_r	T / N_r
PhoneSweep	1109	975	8.97	1109	2.56	1019	11.53	1109	3.97
Deep View Video	669	573	4.96	667	1.73	596	14.16	624	1.13

provided ground truth intrinsics to remove the distortion in the images, resulting in perspective images with a FOV of about 108°. This dataset includes 15 multi-view videos of both indoor and outdoor scenes captured in a variety of lighting conditions. We used the first timestamp in each video for our experiments.

This dataset was created for the purpose of evaluating view synthesis techniques, rather than SFM. Accordingly, some of the images are less than ideal for feature matching, and for example capture a view of the sky or clouds.

5.2.3. Metrics

To evaluate camera pose accuracy, we compute the Relative Rotation Accuracy (RRA), Relative Translation Accuracy (RTA), and Area Under Curve (AUC) metrics [38]. These metrics are invariant to global scale and thus are appropriate for evaluating monocular SfM reconstruction.

For the Deep View Video Dataset, which has metric ground truth, we also robustly align the predicted camera centers to the ground truth and compute the Recall metric as the percentage of cameras with an error below 10 cm.

To evaluate camera calibration accuracy, we compute the Absolute Focal Error (AFE) as $AFE = |f_{pred} - f_{gt}|/f_{gt}$.

To evaluate reconstruction speed, we compute the time per reconstructed image as T/N_r where T is the processing time in seconds and N_r is the number of images reconstructed by the method out of all N images. We did not include feature extraction and matching in the timing since these steps are common to all of the methods.

5.2.4. Results

Tables 1 and 2 present accuracy metrics for the PhoneSweep and Deep View Video datasets. Our method consistently outperforms the other methods across all metrics.

On the Deep View Video dataset we found that while our method accurately estimated the camera rotations, in some cases there was error in the camera positions (mostly along the optical axis). This indicates that in some scenes there may be insufficient feature matches to fully constrain the camera translation. We tested our method without the general BA step, so that the cameras are restricted to a perfect spherical configuration, and saw a significant improvement in the translation accuracy (Table 2, second-to-last row).

The last row of Tables 1 and 2, indicated by COLMAP (calib.), provide accuracy metrics for COLMAP when provided with the ground truth focal length and the intrinsics fixed during optimization. Even when provided the correct focal length, COLMAP still produces less accurate results than our method.

Table 3 compares the methods in terms of reconstruction speed. Our method was the second fastest after GLOMAP on the PhoneSweep dataset, with a runtime of about 4 seconds per image, and was the fastest on the Deep View Dataset, with a runtime of about 1 second per image.

Figures 1 and 4 compare example reconstructions on several scenes. COLMAP in some cases produces incomplete and inaccurate reconstructions, while GLOMAP and RadialSFM tend to reconstruct more cameras but in some cases produces inaccurate camera poses. All three competing methods have a tendency to confuse the motion as inward-facing instead of outward-facing.

In the SM we highlight view synthesis and dense depth results based on the camera poses and 3D points estimated by our method.

6. Conclusions and Future Work

Our work provides a thorough analysis of uncalibrated spherical SFM in the case of a single unknown focal length. We elucidate the relationship between focal length and rotation in two-view spherical motion, and prove that self-calibration is possible from three or more views. We develop and validate a global SFM approach based on the tools of uncalibrated spherical SFM that can accurately handle deviations from perfect spherical motion. In our comparative evaluation, our method outperformed general SFM methods on data from handheld cameras and a hemispherical camera rig. Our method enables a casual user to easily make a dense 3D reconstruction of a scene by simply holding their phone and spinning in a circle.

Future work includes alternate optimization strategies and extending the method to handle radial distortion, as well as exploring learning-based approaches as a complementary direction [6, 39, 40].

Acknowledgments

This work was supported by the National Science Foundation under Award No. 2144822, the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, the strategic research project ELLIIT, and the Swedish Research Council (grant no. 2023-05424).

References

[1] Sameer Agarwal, Keir Mierle, and The Ceres Solver Team. Ceres Solver, 2023. 7

[2] Lewis Baker, Steven Mills, Stefanie Zollmann, and Jonathan Ventura. CasualStereo: Casual capture of stereo panoramas with spherical structure-from-motion. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 782–790. IEEE, 2020. 1, 2

[3] Lewis Baker, Jonathan Ventura, Stefanie Zollmann, Steven Mills, and Tobias Langlotz. SPLAT: Spherical localization and tracking in large spaces. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 809–817. IEEE, 2020. 2

[4] Lewis Baker, Jonathan Ventura, Tobias Langlotz, Shazia Gul, Steven Mills, and Stefanie Zollmann. Localization and tracking of stationary users for augmented reality. *The Visual Computer*, 40(1):227–244, 2024. 1, 2

[5] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew Duvall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics (TOG)*, 39(4):86–1, 2020. 1, 7

[6] Lucas Brynte, José Pedro Iglesias, Carl Olsson, and Fredrik Kahl. Learning structure-from-motion with graph attention networks. In *Conference on Computer Vision and Pattern Recognition*, 2024. 8

[7] Jay Busch, Peter Hedman, Matthew DuVall, Matt Whalen, Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, and Paul Debevec. Brutus: A mid-range multi-camera array for immersive light field video capture. In *CVPR Workshop on Computational Cameras and Displays*, 2021. 1

[8] Martin Byröd, Matthew Brown, and Karl Åström. Minimal solutions for panoramic stitching with radial distortion. In *The 20th British Machine Vision Conference*. British Machine Vision Association (BMVA), 2009. 2

[9] Avishek Chatterjee and Venu Madhav Govindu. Efficient and robust large-scale rotation averaging. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 521–528, 2013. 2

[10] Yu Chen, Ji Zhao, and Laurent Kneip. Hybrid rotation averaging: A fast and robust rotation averaging approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10358–10367, 2021. 5

[11] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized RANSAC. In *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25*, pages 236–243. Springer, 2003. 5

[12] Anders Eriksson, Carl Olsson, Fredrik Kahl, and Tat-Jun Chin. Rotation averaging and strong duality. In *Conference on Computer Vision and Pattern Recognition*, 2018. 2

[13] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. Second Edition. 5

[14] Petr Hruby, Viktor Korotynskiy, Timothy Duff, Luke Oeding, Marc Pollefeys, Tomas Pajdla, and Viktor Larsson. Four-view geometry with unknown radial distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8990–9000, 2023. 2, 5

[15] Kyungdon Joo, Hongdong Li, Tae-Hyun Oh, Yunsu Bok, and In So Kweon. Globally optimal relative pose estimation for camera on a selfie stick. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4983–4989. IEEE, 2020. 2

[16] Kyungdon Joo, Hongdong Li, Tae-Hyun Oh, and In So Kweon. Robust and efficient estimation of relative pose for cameras on selfie sticks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5460–5471, 2021. 2

[17] F. Kahl. Critical motions and ambiguous Euclidean reconstructions in auto-calibration. In *International Conference on Computer Vision*, 1999. 1

[18] Fredrik Kahl and Bill Triggs. Critical motions in Euclidean structure from motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1999. 3

[19] Fredrik Kahl, Bill Triggs, and Kalle Åström. Critical motions for auto-calibration when some intrinsic parameters can vary. *Journal of Mathematical Imaging and Vision*, 13(2): 131–146, 2000. 1, 4

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[21] Viktor Larsson, Nicolas Zobernig, Kasim Taskin, and Marc Pollefeys. Calibration-free structure-from-motion with calibrated radial trifocal tensors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 382–399. Springer, 2020. 2, 5

[22] Karel Lebeda, Jirí Matas, and Ondrej Chum. Fixing the locally optimized RANSAC. In *Procedings of the British Machine Vision Conference 2012*. British Machine Vision Association, 2012. 5

[23] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981. 1

[24] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 5

[25] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion. *Acta Numerica*, 26:305–364, 2017. 1

[26] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes Lutz Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 5

[27] Shmuel Peleg and Moshe Ben-Ezra. Stereo panorama with a single camera. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, pages 395–401. IEEE, 1999. 1

[28] Christian Richardt, Yael Pritch, Henning Zimmer, and Alexander Sorkine-Hornung. Megastereo: Constructing high-resolution stereo panoramas. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1256–1263, 2013. 1, 2

[29] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 5

[30] J.G. Semple and G.T. Kneebone. *Algebraic projective geometry*. Clarendon Press, 1979. 4

[31] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *SIGGRAPH06: Special Interest Group on Computer Graphics and Interactive Techniques Conference*, pages 835–846, 2006. 2

[32] Henrik Stewénius, Christopher Engels, and David Nistér. Recent developments on direct relative orientation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60(4):284–294, 2006. 2

[33] Peter Sturm. Critical motion sequences for monocular self-calibration and uncalibrated euclidean reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1997. 3

[34] Peter Sturm. Critical motion sequences for the self-calibration of cameras and stereo systems with variable focal length. *Image and Vision Computing*, 20(5):415–426, 2002. 1, 4

[35] Chris Sweeney, Torsten Sattler, Tobias Hollerer, Matthew Turk, and Marc Pollefeys. Optimizing the viewing graph for structure-from-motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 801–809, 2015. 2

[36] Chris Sweeney, Aleksander Holynski, Brian Curless, and Steve M Seitz. Structure from motion for panorama-style videos. *arXiv preprint arXiv:1906.03539*, 2019. 1, 2, 5

[37] Jonathan Ventura. Structure from motion on a sphere. In *European Conference on Computer Vision*, 2016. 1, 2, 4, 5

[38] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 8

[39] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Conference on Computer Vision and Pattern Recognition*, 2025. 8

[40] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Conference on Computer Vision and Pattern Recognition*, 2024. 8

[41] Kyle Wilson and Noah Snavely. Robust global translations with 1dsfm. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 61–75. Springer, 2014. 2

[42] Congrong Xu, Justin Kerr, and Angjoo Kanazawa. Splatfacto-W: A Nerfstudio implementation of Gaussian Splatting for unconstrained photo collections. *arXiv preprint arXiv:2407.12306*, 2024. 2, 3

[43] Yongcong Zhang, Yifei Xue, Ming Liao, Huiqing Zhang, and Yizhen Lao. DFR: Depth from rotation by uncalibrated image rectification with latitudinal motion assumption. In *International Conference on Media and Expo (ICME)*, 2023. 2