Structure-from-Motion with a Non-Parametric Camera Model

Yihan Wang[1*†] Linfei Pan[2*] Marc Pollefeys[2,3] Viktor Larsson[4]
[1]EPFL [2]ETH Zurich [3]Microsoft Spatial AI Lab [4]Lund University

Abstract

In this paper, we present a new generic Structure-from-Motion pipeline, GenSfM, that uses a non-parametric camera projection model. The model is self-calibrated during the reconstruction process and can fit a wide variety of cameras, ranging from simple low-distortion pinhole cameras to more extreme optical systems such as fisheye or catadioptric cameras. The key component in our framework is an adaptive calibration procedure that can estimate partial calibrations, only modeling regions of the image where sufficient constraints are available. In experiments, we show that our method achieves comparable accuracy to traditional Structure-from-Motion pipelines in easy scenarios, and outperforms them in cases where they are unable to self-calibrate their parametric models. Code is at https://github.com/Ivonne320/GenSfM.git

1. Introduction

Structure-from-Motion (SfM) has a long history in computer vision, dating back to Ullman [32]. It tackles the problem of jointly estimating the underlying 3D structure and camera poses given an arbitrary image collection. It is for example used to build maps for localization [22, 29] or to estimate initial camera geometry and calibration for further dense or implicit reconstruction [10, 17, 25].

While there have been some recent progress in non-sequential pipelines (e.g. [19]), the incremental paradigm which alternates registering images with triangulation and bundle adjustment, has dominated the field in the last decade. Common to both approaches is to model the camera intrinsic calibration using low-dimensional parametric models, usually polynomials or rational functions with around 3 to 12 parameters. In the literature, there exists a plethora of different camera models [6, 7, 21] that fit different types of optical systems, ranging from simple pinhole cameras to complex catadioptric setups. As lens distortion is dominated by the radial component for most cameras, many models are designed to be radially symmetric.

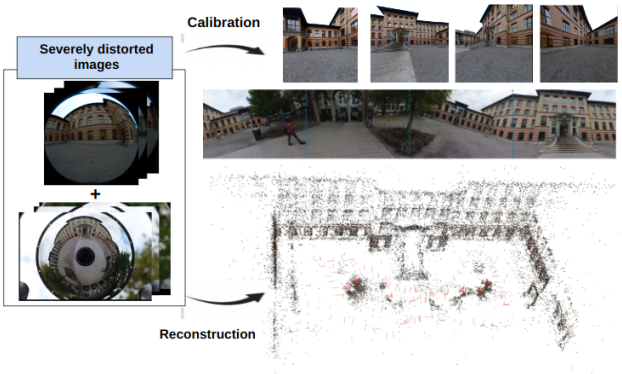

Figure 1. **Non-parametric Camera Models in Strcuture-from-Motion.** Our framework jointly performs Structure-from-Motion while calibrating a non-parametric camera model. The camera model is flexible and can fit a wide variety of cameras, including severely distorted images such as fisheye or catadioptric.

In other words, the distortion only varies with the distance to the image center, but not the angle.

In a parallel development, there are also geometric estimation methods that avoid explicitly estimating the lens distortion. Examples are methods relying on the *Radial Alignment Constraint (RAC)*, originally from Tsai [31]. These methods derive geometric constraints that are invariant to the radial distortion by ignoring the radial offset, which is the distance to the center, and only considering the radial angle towards the projection. This has been used for stratified calibration methods for radial distortion (e.g. [11, 12, 31]) and multi-view pose estimation [8]. Larsson *et al*. presented a full incremental Structure-from-Motion framework in [13] based on this projection constraint, which recovers the 3D structure and partial camera geometry while ignoring the radial distortion of the cameras. Another recent approach is to implicitly model the distortion map by regularization, for example, requiring it to be monotonic [4] or smooth [18].

In this paper, we present a novel generic incremental Structure-from-Motion pipeline, GenSfM, that instead uses a non-parametric model for the intrinsic calibration as illustrated in Fig. 1. We introduce a natural spline-based representation for the calibration map, which enables smooth,

*Equal contribution
†The work is completed during exchange study at ETH Zurich.

invertible, and generic modeling of diverse camera distortions. The proposed system adopts a stratified approach that combines the 1D radial reconstruction framework ([8, 13]) with the implicit distortion pose estimation [18] to bootstrap the reconstruction and camera poses before estimating the intrinsic calibrations. To deal with the large degrees of freedom present in the non-parametric models we propose an adaptive calibration scheme that can estimate partial calibrations for regions of images that have enough constraints. We designed a mixed triangulation method to leverage full 2D reprojection constraints for points within the calibrated regions while using the distortion-invariant radial reprojection error for points outside.

In thorough experiments, we demonstrate that our Structure-from-Motion pipeline can robustly and accurately reconstruct image collections while jointly estimating complex intrinsic calibrations. We show that our non-parametric model is flexible and can fit a wide variety of camera models, including more extreme projection mappings such as fisheye or catadioptric. We compare favorably against state-of-the-art Structure-from-Motion frameworks and surpass them in cases where they are not able to self-calibrate complex image models during reconstruction.

The proposed system complements traditional SfM pipelines by providing a bootstrapping mechanism or serving as an alternative in the presence of complex cameras or in-the-wild images with unknown distortion patterns. We believe it opens up the possibility of incorporating a much wider range of cameras into computer vision tasks.

2. Background and Related Works

Parametric and Non-Parametric Camera Models. For wide field-of-view cameras, the standard pinhole model is generally insufficient. To handle the deviation, which is also referred to as distortion, various types of non-linear distortion functions are introduced. The projection can for example then be formalized as

$$\mathbf{x} = f \cdot \mathcal{D}(\Pi(\mathbf{R}\mathbf{X} + \mathbf{t})) + \mathbf{c}, \tag{1}$$

where f is the focal length, Π the pinhole projection (dehomogenization), $\mathbf{c}$ the principal point and $\mathcal{D}$ the non-linear function for the distortion modeling. Commonly, the distortion function is radially symmetric and only depends on the radial offset from the principle point. One common choice for the distortion mapping is then

$$d(r) = 1 + k_1 r^2 + k_2 r^4 + \cdots, \tag{2}$$

which is the Brown-Conrady model [3, 5]. Such formulations fails in the case of very large field-of-view cameras (*e.g.* fisheye or catadioptric systems). To address this, [23] instead reformulate the projection as

$$\lambda \begin{bmatrix} \mathbf{x} \\ F(\|\mathbf{x}\|) \end{bmatrix} = \mathbf{R}\mathbf{X} + \mathbf{t}, \tag{3}$$

where $F : \mathbb{R}_+ \to \mathbb{R}$ represents the distortion function. Other variations exist, such as [9], which instead expresses the distortion in terms of the opening angle, i.e.

$$d(r) = 1 + k_1 \theta^2 + k_2 \theta^4 + \cdots, \tag{4}$$

where $\theta = \text{atan2}(\sqrt{x^2 + y^2}, z), (x, y, z)^\top = \mathbf{R}\mathbf{X} + \mathbf{t}$. The above formulations assume pre-determined polynomials for the camera distortion modeling, so they are referred to as *parametric camera models*.

Meanwhile, another category of camera models known as *non-parametric camera models* are introduced in the literature. Instead of relying on polynomials with fixed form, it models the camera intrinsics on a per-point basis. Since first introduced by Grossberg and Nayar [6], a series of works are proposed to improve the performances [1, 7, 20, 21, 27]. In line with this regime, [4, 18] proposes camera pose estimation algorithms based on the monotonicity of image radii concerning the image ray and on the smoothness of the focal length concerning the image radii respectively. Recently, [36] proposes to model camera intrinsics with neural networks. Due to its simplicity, differentiability, and high accuracy, we adopt [18] as the algorithm for camera pose estimation.

Multiple View Geometry of 1D Radial Cameras. The multiple view geometry of 1D radial cameras was originally studied by Thirthala and Pollefeys [30]. It is shown that the estimation of camera poses with any bi- or trifocal tensors in general position is not possible, and geometric constraints first start to appear in four views. In [30], the authors further proposed linear and minimal solvers where three views have intersecting principal axes or are planar. Larsson *et al*. [13] introduce two additional solvers with purely radial trifocal tensor and a mixed trifocal tensor with one central camera and two radial cameras in general position. Hruby *et al*. [8] take one step further and describe the minimal solutions for radial camera relative pose for quadrifocal in the uncalibrated, calibrated, and upright case for the first time. It observes that though seemingly intractable with 3584 solutions, the problem can be decomposed significantly, and can be addressed by solving a sequence of subproblems with 28, 2, and 4 solutions. In this work, we adopt the solver proposed by [8] in the initialization step.

Structure-from-Motion Pipelines. Algorithms addressing the problem of Structure-from-Motion generally fall into two categories: *incremental* method and *global* method. Though some recent progress has been made [19], the incremental pipelines have dominated the field over decades. Bundler [28], VisualSfM [35] are representative works that date back decades ago. COLMAP [24, 25] is by far the most popular Structure-from-Motion software.

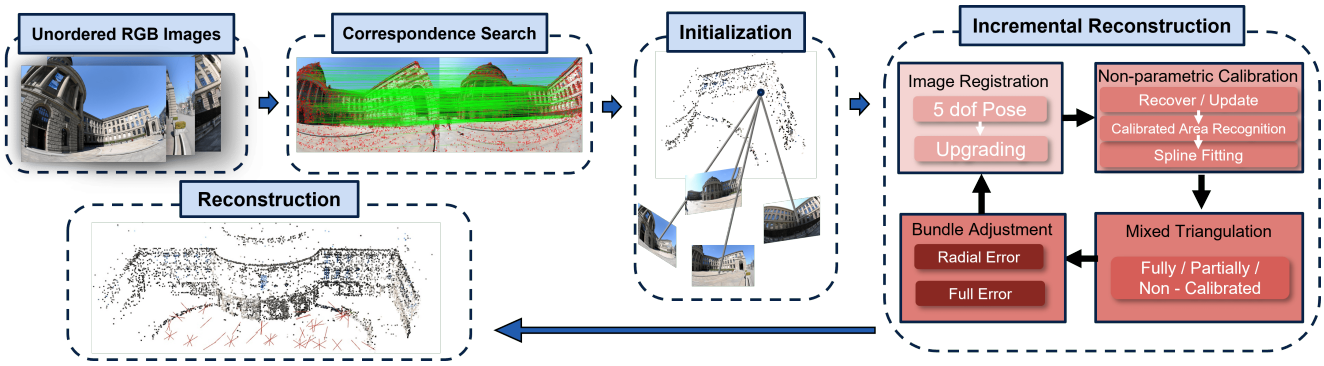

Figure 2. Pipeline for incremental SfM with non-parametric camera model. Without known calibration or specific parametric model, we collect initial 2D-3D correspondences with radial alignment constraint as in [8]. As images iteratively registered to the 3D model, we progressively calibrate the camera by fitting a non-parametric distortion map initialized with implicit distortion model [18].

More recently, PixSfM [15], VGGSfM [33] and Zhang *et al*. [37] proposes learning based methods for Structure-from-Motion task. However, the above methods share a similar drawback in that they all rely on parametric camera models and can struggle in the presence of large unknown distortion in images. In contrast, we present a non-parametric Structure-from-Motion pipeline that is applicable to various camera models, ranging from simple pinhole cameras to fisheye cameras and to catadioptric systems.

Calibration-free Structure-from-Motion. Larsson *et al*. [13] presents the first full incremental Structure-from-Motion framework, based on the radial alignment constraint (RAC) only. In [13], the authors demonstrate the possibility of reconstructing scenes in the presence of high distortion, where traditional parametric-camera model-based pipelines (*e.g*. [24]) fail. However, without camera calibration, the estimation of the camera pose is only accurate up to 5 degree-of-freedom, with the forward motion of cameras remaining ambiguous. In contrast, our framework is able to recover the full 6 degree-of-freedom camera pose by jointly self-calibrating non-parametric camera models.

3. Intrinsic Calibration Representation

In this section we present the non-parametric camera model which we use in our Structure-from-Motion pipeline. We assume that the camera is radially symmetric and that the principal point is known (taken as the image center unless otherwise stated). For ease of notation we will assume all image points are centered ($\mathbf{x} = \mathbf{x}_{\text{ori}} - \mathbf{c}$).

In this case, the intrinsics calibration of the camera can be modeled as a mapping between the opening angle θ, i.e. angle to the principal axis, and the image radius $r = \|\mathbf{x}\|$. Let $M : \mathbb{R} \to \mathbb{R}$ denote this map, i.e. $M[\theta] = r$. The 3D point $\mathbf{X}$ then projects into the image of the camera $(\mathbf{R}, \mathbf{t})$ as

$$\mathbf{x}^{\text{proj}} = M[\theta] \cdot \frac{\mathbf{R}_{12}\mathbf{X} + \mathbf{t}_{12}}{\|\mathbf{R}_{12}\mathbf{X} + \mathbf{t}_{12}\|} \tag{5}$$

where $\theta = \operatorname{atan2}\left(\|\mathbf{R}_{12}\mathbf{X} + \mathbf{t}_{12}\|, \mathbf{R}_3\mathbf{X} + \mathbf{t}_3\right)$, with $(\mathbf{R}_{12}, \mathbf{t}_{12})$ denoting the first two rows of $(\mathbf{R}, \mathbf{t})$.

In our framework, we use an adaptive spline-based representation for the map $M[\theta]$ which can model generic smooth functions. The representation consists of a collection of control points

$$(\theta_1, \ldots, \theta_K), \quad \theta_1 < \theta_2 < \cdots < \theta_K \tag{6}$$

together with corresponding image radii $(r_1, \ldots, r_K)$. The calibrated interval $[\theta_1, \theta_K]$ and the position/value $[r_1, r_K]$ of the control points are adaptively estimated and updated during the reconstruction process. This is further detailed in Section 4.3.2. The map $M[\theta]$ can then be evaluated efficiently through cubic interpolation and it is easy to compute analytic derivatives for optimization.

Note that since this mapping should be invertible (no distinct image rays should project onto the same image point), we could also have chosen to represent the intrinsic calibration in terms of the opposite map, i.e. going from radii to angles, $M^{-1}[r] = \theta$. However, since we need to optimize over the 2D reprojection error $\|\mathbf{x} - \mathbf{x}^{\text{proj}}\|$ in our framework, it is more convenient to model the $\theta - r$ mapping.

In the cases where the inverse map is needed, e.g. for triangulation, the inverse can be efficiently computed using Newton's method. Initialize θ_0 to be the corresponding value with the closest radii of the control point that $\theta_0 = \theta_k$ where $k = \arg\min_k |r_k - r|$, then we update θ with the following rule

$$\theta_{t+1} = \theta_t - \frac{M[\theta_t] - r}{M'(\theta_t)} \tag{7}$$

where $M'(\theta)$ is the first order derivative of $M(\theta)$. We proceed until convergence of $M(\theta_t) = r$.

4. Framework

The overall design of the framework follows the incremental pipeline in [13, 24], and is illustrated in Figure 2. The re-

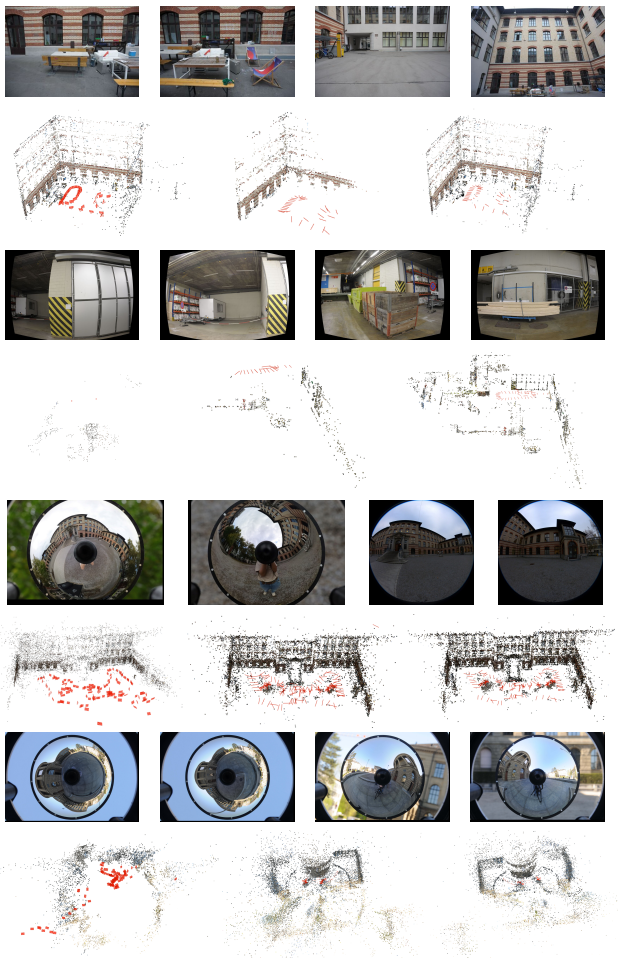

Figure 3. Reconstruction results from pinhole-like images to severely distorted images. The reconstructions in each row come from COLMAP [24], RadialSfM [13] with pose upgraded and bundle adjusted with [18] and our pipeline, from left to right.

construction is first initialized by estimating a radial quadrifocal tensor using the minimal solver from [8] and then proceeds to the iterative reconstruction stage. During this stage, images are registered with the camera pose estimation algorithm from [18], and progressively calibrated using the spline-based representation in a robust scheme. Points are triangulated with mixed constraints, using combinations of 2D and 1D radial errors depending on the availability of the calibration map. The cameras, structure and calibration are refined in interleaving local and global bundle adjustments.

4.1. Initialization

To initialize the reconstruction we select four images and estimate the radial quadrifocal tensor using the 13-point solver from [8] in a RANSAC framework. Once the initial four poses are estimated, 3D points are triangulated using the radial constraints. At this stage, all cameras are fully uncalibrated and treated as 1D radial cameras.

Note that in [13], a combination of two three-view solvers was used which make strong assumptions on the particular camera motion (intersecting principal axes). In contrast, the solver from [8] allows us to initialize with cameras in arbitrary configuration.

4.2. Image Registration

To register a new image to the reconstruction we establish tentative 2D-3D correspondences using standard 2D-2D transitive matching (following [24]). We then apply the camera pose estimation method from Pan *et al*. [18] which leverages an implict distortion model. If multiple images are taken by the same camera, we use the joint estimation scheme proposed in [18] to improve accuracy.

In the case where the image belongs to a camera that already has (partial) calibration available, it would be possible to also use the known distortion mapping together with a standard Perspective-n-Point (PnP) method. However, we found that estimating pose with [18] disentangles the steps of image registration and intrinsic estimation, making the registration process more robust to inaccurate calibrations. Also, this allows us to identify inliers in the uncalibrated region, which is the basis for growing the calibrated region.

4.3. Updating the Calibration Map

Once we have sufficient constraints on the intrinsic calibration we start estimating and updating the calibration map $M[\theta]$ for each camera. Each 2D-3D correspondence yields pairs of angles and radii (θ_i, r_i), which we de-noise using the regularization-based method from Pan *et al*. [18].

To estimate the distortion map we use a two-step process which first identifies a calibrated interval $[\theta_{min}, \theta_{max}]$ followed by a robust spline fitting scheme. The goal is to only calibrate regions of the image which have sufficient constraints for accurate calibration and postpone calibrating the remainder until more constraints become available.

4.3.1 Calibrated Area Recognition

In order to maintain robustness in the presence of poorly estimated distortion from sparse correspondences, we do not directly use all angle-radius pairs $\{(\theta_i, r_i)\}$. Instead, we split the segments into pieces based on the distance of points to their neighbors and only accept the largest piece whose observations are sufficiently close to each other. To determine the threshold, we collect the distance for every point to its previous point and denote the collection as $D = \{d_i\}$. Then, we use $\epsilon_d = \bar{D} + \sigma_d$ as the threshold. Here, $\bar{D}, \sigma_d$ stands for the mean and the standard deviation of D. When the number of observations is abundant, this threshold can be unnecessarily small. Thus, the effective threshold is chosen to be the maximum between ϵ_d and a constant c. We set c to be 0.1° in our experiments. This process ensures that

our interval $[\theta_{min}, \theta_{max}]$ contain enough points with sufficient density to constrain the calibration.

In theory, multiple disjoint segments can be maintained to further enlarge the calibrated region, however, we noticed in the experiments a single segment $[\theta_{min}, \theta_{max}]$ can cover the majority of the well-calibrated region. Also, multiple segments would introduce the problem of inconsistent endpoints from different segments. Thus, we maintain our design of using a single segment.

4.3.2 Robust Spline Fitting

To robustly estimate the calibration map, we perform RANSAC sampling of control points within the calibrated region. The two endpoints of the calibrated region are always included in the sample to ensure full coverage. The output of this step is a collection of control points $(\theta_1, \dots, \theta_K)$ together with corresponding image radii $(r_1, \dots, r_K)$, where the endpoints correspond to the interval boundaries, i.e. $\theta_1 = \theta_{min}$ and $\theta_K = \theta_{max}$. By default, we choose 10 control points in our pipeline. The impact of K is analyzed in the ablation study in Section 5.3.

4.4. Mixed Triangulation

For triangulation, we need to deal with combinations of calibrated and uncalibrated points, either coming from cameras being fully uncalibrated or points lying outside the calibrated region. For this we propose to use a mix between 2D constraints and the distortion-invariant 1D radial constraints, allowing both correspondence types to contribute.

For a 2D-3D correspondence in the uncalibrated region, it is only required to lie on the radial line, thus it yields a single constraint on the 3D point,

$$\begin{bmatrix} -y & x & 0 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{r}_1^\top & t_1 \\ \mathbf{r}_2^\top & t_2 \\ \mathbf{r}_3^\top & t_3 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{X} \\ 1 \end{bmatrix} = 0 \tag{8}$$

This can be interpreted as restricting $\mathbf{X}$ to the plane passing through the camera center and the radial camera line that $\mathbf{n}^\top \mathbf{X} + d = 0$. Here, $\mathbf{n} = x\mathbf{r}_2 - y\mathbf{r}_1$ and $d = xt_2 - yt_1$.

For correspondences in the calibrated region, we undistort the point to the normalized image coordinates $\mathbf{x}' = (x', y')$ by locally inverting $M[\theta]$ (see Section 3). Then, the normalized point (x', y') imposes two constraints:

$$\begin{bmatrix} 1 & 0 & -x' \\ 0 & 1 & -y' \end{bmatrix} \cdot \begin{bmatrix} \mathbf{r}_1^\top & t_1 \\ \mathbf{r}_2^\top & t_2 \\ \mathbf{r}_3^\top & t_3 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{X} \\ 1 \end{bmatrix} = \mathbf{0} \tag{9}$$

This is equivalent to standard triangulation. Since a 3D point has 3 degree-of-freedom thus requiring at least 3 constraints, 3 different minimal cases exist for the triangulation: 3 uncalibrated points, 1 calibrated point + 1 uncalibrated point, and 2 calibrated points. Similar to [24], we perform the triangulation process in the LO-RANSAC [14] scheme and iterate through all possible combinations. We also only accept a point if it is supported by at least 4 constraints as in [13]. For 3D points without any correspondence in the calibrated region, it is required to be seen in four views, while it is only required to be observed in three views or two views if some calibrated observations are available.

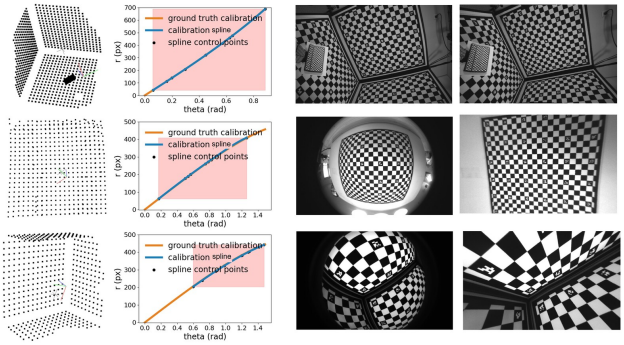

Figure 4. Qualitative result of estimated calibration map on BabelCalib [16] with our full reconstruction pipeline. The figure presents the point cloud, calibration map with highlighted calibrated area, the original images and their undistortion with the estimated calibration map.

4.5. Bundle Adjustment

The difference between bundle adjustment in the proposed pipeline and that in standard pipelines with parametric camera models is that we distinguish between the calibrated region and the uncalibrated region.

For points in the calibrated region, we minimize the full reprojection error with the estimated calibration map,

$$\varepsilon = \left\| M[\theta] \cdot \frac{\mathbf{R}_{12}\mathbf{X} + \mathbf{t}_{12}}{\|\mathbf{R}_{12}\mathbf{X} + \mathbf{t}_{12}\|} - \mathbf{x} \right\|. \tag{10}$$

For the uncalibrated region, we instead minimize the orthogonal distance to the radial line as in [13],

$$\varepsilon = \left\| \left(\frac{\mathbf{n}\mathbf{n}^\top}{\mathbf{n}^\top \mathbf{n}} - I \right) \mathbf{x} \right\|, \quad \text{where } \mathbf{n} = \mathbf{R}_{12}\mathbf{X} + \mathbf{t}_{12} \tag{11}$$

which does not depend on the radial distortion.

The calibration maps are jointly refined with the 3D points and camera poses during bundle adjustment. To facilitate the optimization process, we keep the θ values of the control points to be fixed, only optimizing r values. Notice that this does not guarantee the monotonicity of the optimization result. We perform postprocessing to restore monotonicity. This can potentially violate the geometric constraints of some points which will be removed in the filtering. However, as these correspondences are error-prone since they contribute to misleading optimization direction, the performance of the pipeline remains intact.

	[16]*	[18]*	GenSfM (Ours)†		GenSfM (Ours)	
	$\epsilon_{\text{rms}}^{BC}$	ϵ_{rms}	ϵ_{rms}	% Calib	ϵ_{rms}	% Calib
OV corner	1.52	2.09	2.11	99.8	0.89	100.0
OV cube	0.29	0.31	1.59	99.8	0.31	100.0
OV plane	0.60	0.82	0.60	99.4	0.63	97.8
Kalibr	0.21	0.30	1.30	95.4	0.25	99.9
OCamCalib	0.68	0.97	1.30	95.4	0.61	98.1
UZH DAVIS	0.41	0.42	0.71	99.5	0.42	99.9
UZH Snapdragon	0.26	0.28	0.44	100.0	0.27	99.9

Table 1. Reprojection Error (ϵ_{rms}) and ratio of calibrated points on BabelCalib [16]. Columns marked with $*$ are taken from [18]. Results with † have principal points fixed at the image center.

	AUC @ 5°				
	COLMAP (pinhole)	**COLMAP** (radial)	**COLMAP** (thin prism)	[13] + [18]	**GenSfM** (ours)
courtyard	10.8	67.2	90.0	64.2	91.1
delivery_area	9.9	3.3	93.7	26.0	90.2
electro	8.2	30.7	73.9	11.1	69.2
facade	29.3	79.7	91.7	71.1	91.0
kicker	18.1	71.8	83.0	48.4	83.6
meadow	1.3	0.9	0.9	-	-
office	15.0	36.4	0.3	28.5	45.7
pipes	18.6	20.9	1.0	-	-
playground	16.2	72.9	80.9	20.6	93.9
relief	68.6	77.4	94.0	51.0	92.7
relief_2	50.7	80.7	94.5	0.0	91.4
terrace	24.3	71.7	95.5	32.5	91.6
terrains	25.4	37.0	0.1	1.4	92.5
Average	22.8	50.0	61.5	27.3	71.8

Table 2. Camera pose evaluation on ETH3D [26]. We achieve the best or second-best performance on all the scenes and obtain the highest AUC score on average.

Alternatively, the implicit bundle adjustment process as in [18] could also be deployed. We do not adopt such a scheme because it needs to be performed in an iterative process, and it relies on comparatively weaker constraints than the calibrated bundle adjustment.

5. Experiments

In the experimental evaluation, we demonstrate the effectiveness of the proposed pipeline under various conditions. We evaluate the expressiveness of the chosen camera intrinsics representation in the controlled setting. We also evaluate the performance of the proposed self-calibrating pipeline on in-the-wild datasets. Finally, we analyze the impact of number of control points on the performance.

5.1. Controlled Checkerboard Calibration

We first evaluate the calibration map of the proposed pipeline from classical checkerboard data. In this experiment, we consider datasets BabelCalib [16]. It contains in total of 41 cameras, with a wide variety of field-of-view spanning from 88° to 187°. We obtain the calibration map $M[\theta]$ as detailed in Sec. 3 and Sec. 4.3.2, followed by bundle adjustment as described in Sec. 4.5 on the training set with fixed 3D points. On the test set, we keep the spline and 3D points fixed. We take calibration results from BabelCalib [16] and [18] as baselines.

	AUC @ 5°				
	COLMAP (pinhole)	**COLMAP** (radial)	**COLMAP** (thin prism)	[13] + [18]	**GenSfM** (ours)
courtyard	1.7	1.8	0.1	2.2	89.1
delivery_area	1.2	1.2	0.1	16.1	89.7
electro	0.8	1.0	0.1	11.0	63.0
facade	1.7	2.1	0.1	62.9	91.1
kicker	2.4	2.4	0.2	2.5	82.5
meadow	2.7	0.9	0.9	-	-
office	0.6	0.4	0.3	-	-
pipes	3.1	4.3	1.0	-	-
playground	1.1	1.2	0.1	13.0	0.8
relief	9.1	2.5	0.2	42.2	75.2
relief_2	2.9	3.1	0.4	31.4	65.2
terrace	2.1	1.9	0.4	4.5	2.1
terrains	1.5	1.7	0.1	3.0	3.6
Average	2.4	1.9	0.3	14.5	43.2

Table 3. Results on distorted ETH3D [26]. We remain comparable performance against the original dataset while other methods significantly struggle from the distortion.

Results can be found in Table 1. We report results with and without principal point refinement with [18] for the proposed method. Results show that the proposed method with optimized principal points achieves a similar level of reprojection error as BabelCalib [16]. Compared with another non-parametric model [18], our method achieves much smaller reprojection errors. We attribute this improvement to the stronger regularization from full-dimensional bundle adjustment than the smoothness constraints in [18].

Worth noting is that the reprojection error remains low with un-optimized principal points (marked as †), although principal points can deviate up to tens of pixels from optimized ones. Such observations indicate that the assumption of the principle point at image centers will not cause a significant performance drop.

Qualitative results of reconstructing the dataset in a self-calibrated manner can be found in Figure 4. The estimated spline follows closely to the parametric pseudo-ground truth and the estimated point cloud is clean.

5.2. Self-calibrating Structure-from-Motion

To validate the accuracy and robustness of the self-calibrating pipeline, we conduct end-to-end reconstruction on images with various conditions. In this section, we mainly compare with two baselines.

COLMAP [24] COLMAP is by far the most popular parametric Structure-from-Motion pipeline. It supports multiple types of camera models, ranging from pinhole cameras to camera models with larger field-of-view. We experimented with three camera models: simple pinhole model (denoted as *pinhole*) with 3 parameters (f, c_x, c_y), simple radial model [3, 5] (denoted as *radial*) with 4 parameters (f, c_x, c_y, k), and thin-prism fisheye model [34] (denoted as *thin prism*) with 12 parameters. We adopt

	AUC Completeness					AUC Accuracy					AUC F1				
	COLMAP (pinhole)	COLMAP (radial)	COLMAP (thin prism)	[13] + [18]	GenSfM (ours)	COLMAP (pinhole)	COLMAP (radial)	COLMAP (thin prism)	[13] + [18]	GenSfM (ours)	COLMAP (pinhole)	COLMAP (radial)	COLMAP (thin prism)	[13] + [18]	GenSfM (ours)
courtyard	12.7	26.6	31.9	16.9	31.3	35.7	79.0	97.7	98.4	97.9	18.6	38.6	45.6	27.9	44.9
delivery_area	5.8	2.1	28.5	8.0	25.9	9.8	2.7	92.0	86.5	88.7	6.9	2.2	40.8	14.3	37.7
electro	8.7	11.5	15.4	1.4	14.9	41.5	71.6	95.9	95.0	91.2	13.7	18.9	25.3	2.8	24.4
facade	21.0	34.4	39.0	18.3	38.4	34.1	74.9	91.2	72.5	90.9	25.8	45.4	52.5	27.7	51.9
kicker	21.8	27.4	28.0	15.8	28.5	82.2	93.8	97.8	98.0	97.0	32.6	40.3	41.2	26.3	41.6
meadow	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
office	13.7	17.6	0.0	11.8	17.0	73.1	94.3	0.0	99.1	97.5	22.5	28.2	0.0	20.2	27.4
pipes	7.4	7.4	0.0	0.0	0.0	83.9	80.7	0.0	0.0	0.0	13.3	13.3	0.0	0.0	0.0
playground	23.8	36.0	37.6	9.0	37.6	58.8	91.1	98.1	98.7	97.9	33.2	49.3	51.6	16.0	51.5
relief	20.2	20.6	26.8	12.4	25.6	80.3	98.6	99.2	99.7	99.1	31.3	32.8	40.1	21.7	38.6
relief_2	21.4	26.1	29.2	0.0	28.5	74.1	91.2	98.9	0.0	96.8	32.3	38.8	42.6	0.0	41.7
terrace	23.8	34.8	38.1	13.0	33.7	37.9	78.6	95.8	93.4	95.5	28.6	45.1	50.3	21.6	45.8
terrains	31.7	34.6	0.0	0.5	37.2	83.5	86.9	0.0	0.3	98.7	42.9	46.2	0.0	0.3	50.2
Average	16.3	21.5	21.1	8.2	24.5	53.5	72.6	66.7	64.7	80.9	23.2	30.7	30.0	13.8	35.1

Table 4. Point cloud evaluation on ETH3D [26], the AUC is calculated at tolerance thresholds at 1 cm, 2 cm, 5 cm, 10 cm, and 50 cm. We achieve comparable results as parametric camera models.

COLMAP with these three camera models as our parametric baselines.

RadialSfM [13] + Implicit Model [18] We first estimated 5 DoF camera poses with RadialSfM [13], then we upgraded them and performed bundle adjustment with the implicit camera model presented in [18]. This combination serves as a non-parametric SfM baseline.

5.2.1 Metrics

We evaluate both the quality of the estimated camera poses and point clouds. For camera poses, we report the AUC (Area Under the recall Curve) scores calculated from the maximum of relative rotation and translation error between every image pair. Such metrics are widely deployed, as seen in [19, 33]. For point cloud quality, we report the AUC scores in three aspects: completeness, accuracy, and F1, following [26], measured at distance thresholds of 1 cm to 50 cm. The completeness measures the fraction of ground truth points falling inside a certain tolerance range from the estimated points, which is equivalent to recall. The accuracy measures the fraction of estimated points within a distance threshold of the ground truth points, equivalent to precision.

5.2.2 Raw DSLR Images

We first evaluate the performance of the proposed pipelines on the ETH3D DSLR images [26]. ETH3D [26] presents a collection of high-resolution images with ground-truth pose and point cloud up to millimeter accuracy. It is a standard benchmark for evaluating the performance of Structure-from-Motion systems. Example input images and reconstructions for the *courtyard* scene can be found in the first two rows of Figure 3. We ignore the EXIF tags of the images to evaluate the performance of different systems without prior knowledge of camera intrinsics. We also share cameras for all images from the same scene.

Results on camera pose are summarized in Table 2. Entries marked with "-" indicate that the initialization fails. The proposed method largely outperforms the non-parametric Structure-from-Motion baseline ([13] + [18]). As for parametric baselines, it achieves a similar level of camera pose accuracy as COLMAP with the thin-prism fisheye model while outperforming COLMAP with the other two camera models by a large margin.

Point cloud evaluation results can be found in Table 4. Compared with [13] + [18], the proposed method achieves a significantly more complete point cloud with similar accuracy. Compared with parametric SfM pipelines, it performs similarly to COLMAP with the thin-prism fisheye camera model while obtaining more complete and more accurate results than the other two models.

The result also demonstrates how the choice of camera model largely alternates the performance of parametric pipeline. In this experiment, although images are apparently without clear distortion, fitting camera models with too few camera parameters, such as simple pinhole/radial, can still generally hamper the accuracy of the reconstruction. However, in the scenes where there are no sufficient point, complex camera models may also fail, as in the *pipes* or *office* scene. This implies the importance of a SfM pipeline with flexible camera models as the proposed one.

5.2.3 Artificially Distorted Images

While a parametric Structure-from-Motion pipeline can achieve reasonable reconstruction with sufficiently complex camera models, it starts to struggle when evident distortions are present in the image. To analyze the performance of different pipelines in such scenarios, we manually distort the ETH3D DSLR [26] images with polynomial radial distortion model implemented in OpenCV [2] and reconstruct from there. Example distorted images and reconstruction results can be found in the second two rows of Figure 3.

Quantitative results are summarized in Table 3. Results indicate that the parametric SfM pipeline COLMAP [24]

Figure 5. Undistortion results with estimated calibration map on catadioptric and fisheye images.

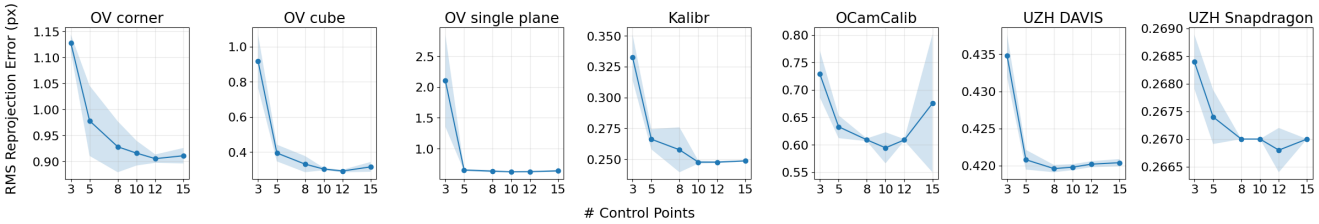

Figure 6. Impact of control point number on the calibration, evaluated on BabelCalib [16] with refined principal point

failed to obtain accurate reconstruction with all three camera models. We attribute this to failures of initializing parameters from the limited number of matches. The non-parametric baseline obtains more accurate results than COLMAP but falls short of us by a large margin. Compared across results from the original and the distorted images, it can be seen that the accuracy for pose estimation drops heavily for the non-parametric baseline while that for the proposed method is less affected. This is benefited from the calibrated region which leads to a more complete reconstructed point cloud and strong regularization.

5.2.4 Images with Large Field-of-View

To demonstrate the flexibility of the calibration representation, we reconstruct with images from more extreme optical systems. We consider fisheye datasets from [13], and datasets with catadioptric images. Example images and reconstruction can be found in the last four rows of Figure 3. Since no ground truth is available for these datasets, only qualitative results are presented. We also undistort the images with the estimated calibration map to demonstrate the quality of the estimated spline as presented in Figure 5. More qualitative results can be found in Supp. Mat.

5.3. Impact of Control Point Number

The number of control points determines the degree of freedom of the calibration map thus the representation capacity. We designed an ablation study on the number of control points for detailed analysis. We experimented on the BabelCalib [16], and used the same principal points and implicit calibration from [18] as the initialization. We tested with 3, 5, 8, 10, 12, and 15 control points and conducted 5 independent trials for each configuration. Root mean squared errors are summarized in Figure 6. From the plot, it can be seen that the performance varies largely with different numbers of control points. As the number of control points increases, the representation capacity increases, resulting in a decrease in projection errors. However, when the number of control points becomes too large, it becomes more difficult to calibrate and thus more unstable. We choose 10 as the default number in our pipeline as it strikes a balance in between. More detailed analysis can be found in Supp. Mat.

6. Conclusion

In this paper, we presented a self-calibrating Structure-from-Motion pipeline with non-parametric camera models. The highly flexible spline-based calibration map allows the pipeline to reconstruct from images ranging from simple low-distortion pinhole cameras to more extreme optical systems such as fisheye or catadioptric cameras. The key component in our framework is an adaptive calibration procedure. It only marks a region of images calibrated when sufficient constraints are available. Through extensive experiments, the proposed system demonstrates its flexibility and robustness under various conditions, both in controlled and self-calibrating scenarios. The system also successfully reconstructs with catadioptric images in a self-calibrating manner where traditionally parametric pipelines fail. We believe the proposed pipeline to be a complement to traditional SfM pipelines, either as a bootstrapping technique or as an alternative when images are highly distorted or camera calibration patterns are unknown beforehand.

Acknowledgments. The authors thank Yifan Zhou for the support of the equipment. Linfei Pan was funded by gift funding from Microsoft. Viktor Larsson was supported by ELLIIT and the Swedish Research Council (Grant No. 2023-05424).

References

[1] Johannes Beck and Christoph Stiller. Generalized b-spline camera model. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 2137–2142. IEEE, 2018. 2

[2] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 7

[3] Duane Brown. Decentering distortion of lenses. *Photogrammetric engineering*, 32(3):444–462, 1996. 2, 6

[4] Federico Camposeco, Torsten Sattler, and Marc Pollefeys. Non-parametric structure-based calibration of radially symmetric cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2192–2200, 2015. 1, 2

[5] Alexander Eugen Conrady. Decentred lens-systems. *Monthly notices of the royal astronomical society*, 79(5): 384–390, 1919. 2, 6

[6] Michael D Grossberg and Shree K Nayar. A general imaging model and a method for finding its parameters. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, pages 108–115. IEEE, 2001. 1, 2

[7] Richard Hartley and Sing Bing Kang. Parameter-free radial distortion correction with center of distortion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1309–1321, 2007. 1, 2

[8] 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. 1, 2, 3, 4

[9] Juho Kannala and Sami S Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1335–1340, 2006. 2

[10] 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. 1

[11] Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Real-time solution to the absolute pose problem with unknown radial distortion and focal length. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2816–2823, 2013. 1

[12] Viktor Larsson, Torsten Sattler, Zuzana Kukelova, and Marc Pollefeys. Revisiting radial distortion absolute pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1062–1071, 2019. 1

[13] 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. 1, 2, 3, 4, 5, 6, 7, 8

[14] Karel Lebeda, Jırı Matas, and Ondrej Chum. Fixing the locally optimized ransac–full experimental evaluation. In *British machine vision conference*. Citeseer Princeton, NJ, USA, 2012. 5

[15] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5987–5997, 2021. 3

[16] Yaroslava Lochman, Kostiantyn Liepieshov, Jianhui Chen, Michal Perdoch, Christopher Zach, and James Pritts. Babelcalib: A universal approach to calibrating central cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15253–15262, 2021. 5, 6, 8

[17] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[18] Linfei Pan, Marc Pollefeys, and Viktor Larsson. Camera pose estimation using implicit distortion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12819–12828, 2022. 1, 2, 3, 4, 6, 7, 8

[19] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 7

[20] Srikumar Ramalingam and Peter Sturm. A unifying model for camera calibration. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1309–1319, 2016. 2

[21] Dennis Rosebrock and Friedrich M Wahl. Generic camera calibration and modeling using spline surfaces. In *2012 IEEE Intelligent Vehicles Symposium*, pages 51–56. IEEE, 2012. 1, 2

[22] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8601–8610, 2018. 1

[23] Davide Scaramuzza, Agostino Martinelli, and Roland Siegwart. A toolbox for easily calibrating omnidirectional cameras. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5695–5701. IEEE, 2006. 2

[24] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2, 3, 4, 5, 6, 7

[25] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 2

[26] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 6, 7

[27] Thomas Schops, Viktor Larsson, Marc Pollefeys, and Torsten Sattler. Why having 10,000 parameters in your

camera model is better than twelve. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2535–2544, 2020. 2

[28] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 2

[29] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018. 1

[30] SriRam Thirthala and Marc Pollefeys. Radial multi-focal tensors: Applications to omnidirectional camera calibration. *International journal of computer vision*, 96:195–211, 2012. 2

[31] Roger Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal on Robotics and Automation*, 3(4):323–344, 1987. 1

[32] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. 1

[33] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21686–21697, 2024. 3, 7

[34] Juyang Weng, Paul Cohen, Marc Herniou, et al. Camera calibration with distortion models and accuracy evaluation. *IEEE Transactions on pattern analysis and machine intelligence*, 14(10):965–980, 1992. 6

[35] Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, pages 127–134. IEEE, 2013. 2

[36] Wenqi Xian, Aljaž Božič, Noah Snavely, and Christoph Lassner. Neural lens modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8435–8445, 2023. 2

[37] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024. 3