2022 IEEE International Conference on Robotics and Automation (ICRA)
May 23-27, 2022. Philadelphia, PA, USA

Self-Supervised Camera Self-Calibration from Video

Jiading Fang Igor Vasiljevic Vitor Guizilini Rareș Ambruș
Greg Shakhnarovich Adrien Gaidon Matthew R. Walter

***Abstract*— Camera calibration is integral to robotics and computer vision algorithms that seek to infer geometric properties of the scene from visual input streams. In practice, calibration is a laborious procedure requiring specialized data collection and careful tuning. This process must be repeated whenever the parameters of the camera change, which can be a frequent occurrence for mobile robots and autonomous vehicles. In contrast, self-supervised depth and ego-motion estimation approaches can bypass explicit calibration by inferring per-frame projection models that optimize a view-synthesis objective. In this paper, we extend this approach to explicitly calibrate a wide range of cameras from raw videos in the wild. We propose a learning algorithm to regress per-sequence calibration parameters using an efficient family of general camera models. Our procedure achieves self-calibration results with sub-pixel reprojection error, outperforming other learning-based methods. We validate our approach on a wide variety of camera geometries, including perspective, fisheye, and catadioptric. Finally, we show that our approach leads to improvements in the downstream task of depth estimation, achieving state-of-the-art results on the EuRoC dataset with greater computational efficiency than contemporary methods. The project page: https://sites.google.com/ttic.edu/self-sup-self-calib**

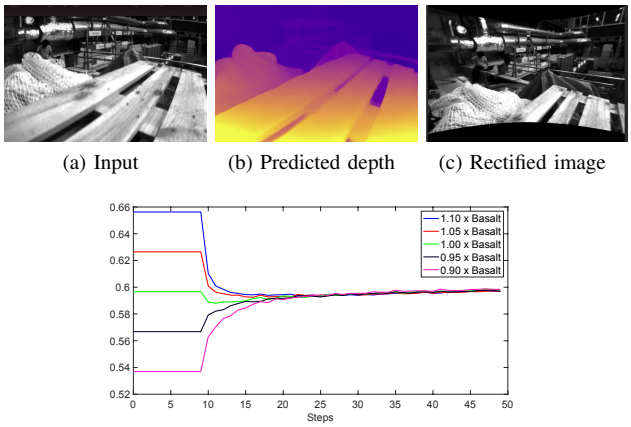
I. Introduction

Cameras provide rich information about the scene, while being small, lightweight, inexpensive, and power efficient. Despite their wide availability, camera calibration largely remains a manual, time-consuming process that typically requires collecting images of known targets (e.g., checkerboards) as they are deliberately moved in the scene [1]. While applicable to a wide range of camera models [2–4], this process is tedious and has to be repeated whenever the camera parameters change. A number of methods perform calibration "in the wild" [5–7]. However, they rely on strong assumptions about the scene structure, which cannot be met during deployment in unstructured environments. Learning-based methods relax these assumptions, and regress camera parameters directly from images, either by using labelled data for supervision [8] or by extending the framework of self-supervised depth and ego-motion estimation [9, 10] to also learn per-frame camera parameters [11, 12].

While these methods enable learning accurate depth and ego-motion without calibration, they are either over-parameterized [12] or limited to near-pinhole cameras [11].

Jiading Fang, Igor Vasiljevic, Greg Shakhnarovich, and Matthew R. Walter are with the Toyota Technological Institute at Chicago {fjd,ivas,greg,mwalter}@ttic.edu

Vitor Guizilini, Rares Ambrus, and Adrien Gaidon are with the Toyota Research Institute {vitor.guizilini, rares.ambrus, adrien.gaidon}@tri.global



(a) Input (b) Predicted depth (c) Rectified image

(d) Camera parameter re-calibration

Fig. 1: **Our self-supervised self-calibration procedure** can recover accurate parameters for a wide range of cameras using a structure-from-motion objective on raw videos (EuRoC dataset, top), enabling on-the-fly re-calibration and robustness to intrinsics perturbation (bottom).

In contrast, we propose a self-supervised camera calibration algorithm capable of learning expressive models of different camera geometries in a computationally efficient manner. In particular, our approach adopts a family of general camera models [13] that scales to higher resolutions than previously possible, while still being able to model highly complex geometries such as catadioptric lenses. Furthermore, our framework learns camera parameters per-sequence rather than per-frame, resulting in self-calibrations that are more accurate and more stable than those achieved using contemporary learning methods. We evaluate the reprojection error of our approach compared to conventional target-based calibration routines, showing comparable sub-pixel performance despite only using raw videos at training time.

Our contributions can be summarized as follows:

- We propose to self-calibrate a variety of generic camera models from raw video using self-supervised depth and pose learning as a proxy objective, providing for the first time a calibration evaluation of camera model parameters **learned purely from self-supervision**.
- We demonstrate the utility of our framework on challenging and radically different datasets, learning depth and pose on perspective, fisheye, and catadioptric images without architectural changes.
- We achieve **state-of-the-art depth evaluation results on the challenging EuRoC MAV dataset** by a large margin, using our proposed self-calibration framework.

2022 IEEE International Conference on Robotics and Automation (ICRA) | 978-1-7281-9681-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRA46639.2022.9811784

II. Related Work

Camera Calibration. Traditional calibration for a variety of camera models uses targets such as checkerboards or AprilTags to generate 2D-3D correspondences, which are then used in a bundle adjustment framework to recover relative poses as well as intrinsics [1, 14]. Targetless methods typically make strong assumptions about the scene, such as the existence of vanishing points and known (Manhattan world) scene structure [5–7]. While highly accurate, these techniques require a controlled setting and manual target image capture to re-calibrate. Several models are implemented in OpenCV [15], kalibr [16]. These methods require specialized settings to work, limiting their generalizability.

Camera Models. The pinhole camera model is ubiquitous in robotics and computer vision [17, 18] and is especially common in recent deep learning architectures for depth estimation [10]. There are two main families of models for high-distortion cameras. The first is the "high-order polynomial" distortion family that includes pinhole radial distortion [19], omnidirectional [2], and Kannala-Brandt [3]. The second is the "unified camera model" family that includes the Unified Camera Model (UCM) [20], Extended Unified Camera Model (EUCM) [21], and Double Sphere Camera Model (DS) [13]. Both families are able to achieve low reprojection errors for a variety of different camera geometries [13], however the unprojection operation of the "high-order polynomial" models requires solving for the root of a high-order polynomial, typically using iterative optimization, which is a computationally expensive operation. Further, the process of calculating gradients for these models is non-trivial. In contrast, the "unified camera model" family has an easily computed, closed-form unprojection function. While our framework is applicable to high-order polynomial models, we choose to focus on the unified camera model family in this paper.

Learning Camera Calibration. Work in learning-based camera calibration can be divided into two types: *supervised* approaches that leverage ground-truth calibration parameters or synthetic data to train single-image calibration regressors; and *self-supervised* methods that utilize only image sequences. Our proposed method falls in the latter category, and aims to self-calibrate a camera system using only image sequences. Early work on applying CNNs to camera calibration focused on regressing the focal length [22] or horizon lines [23]; synthetic data was used for distortion calibration [24] and fisheye rectification [25]. Using panorama data to generate images with a wide variety of intrinsics, Lopez et al. [26] are able to estimate both extrinsics (tilt and roll) and intrinsics (focal length and radial distortion). DeepCalib [8] takes a similar approach: given a panoramic dataset, generate projections with different focal lengths. Then, they train a CNN to regress from a set of synthetic images I to their (known) focal lengths f. Typically, training images are generated by taking crops of the desired focal lengths from 360 degree panoramas [27, 28]. While this can be done for any kind of image, and does not require image sequences, it does require access to panoramic images. Furthermore, the warped "synthetic" images are not the true 3D-2D projections. This approach has been extended to pan-tilt-zoom [29] and fisheye [25] cameras. Methods also exist for specialized problems like undistorting portraits [30], monocular 3D reconstruction [31], and rectification [32, 33].

Self-Supervised depth and ego-motion. Self-supervised learning has also been used to learn camera parameters from geometric priors. Gordon et al. [11] learn a pinhole and radial distortion model, while Vasiljevic et al. [12] learn a generalized central camera model applicable to a wider range of camera types, including catadioptric. These methods both learn calibration on a per-frame basis, and do not offer a calibration evaluation of their learned camera model. Furthermore, while Vasiljevic et al. [12] is much more general than Gordon et al. [11], it is limited to fairly low resolutions by the complex and approximate generalized projection operation. In our work, we trade some degree of generality (i.e., a global, central vs. per-pixel model) for a closed-form and efficient projection operation and ease of calibration evaluation.

III. Methodology

First, we describe the self-supervised monocular depth learning framework that we use as proxy for self-calibration. Then we describe the family of unified camera models we consider and how we learn their parameters end-to-end.

A. Self-Supervised Monocular Depth Estimation

Self-supervised depth and ego-motion architectures consist of a depth network that produces depth maps $\hat{D}_t$ for a target image I_t, as well as a pose network that predicts the relative rigid-body transformation between target t and context c frames, $\hat{\boldsymbol{X}}^{t\to c} = \begin{pmatrix} \hat{\boldsymbol{R}}^{t\to c} & \hat{\boldsymbol{t}}^{t\to c} \\ \boldsymbol{0} & \boldsymbol{1} \end{pmatrix} \in \mathrm{SE}(3)$. We train the networks jointly by minimizing the photometric reprojection error between the actual target image I_t and a synthesized image $\hat{I}_t$ generated by projecting pixels from the context image I_c (usually preceding or following I_t in a sequence) onto the target image I_t using the predicted depth map $\hat{D}_t$ and ego-motion $\hat{\boldsymbol{X}}^{t\to c}$ [10]. See Figure 2 for an overview. The general pixel-warping operation is defined as:

$$\hat{\boldsymbol{p}}^t = \pi\left(\hat{\boldsymbol{R}}^{t\to c}\phi(\boldsymbol{p}^t, \hat{d}^t, \boldsymbol{i}) + \hat{\boldsymbol{t}}^{t\to c}, \boldsymbol{i}\right), \tag{1}$$

where $\boldsymbol{i}$ are camera intrinsic parameters modeling the geometry of the camera, which is required for both projection of 3D points $\boldsymbol{P}$ onto image pixels $\boldsymbol{p}$ via $\pi(\boldsymbol{P}, \boldsymbol{i}) = \boldsymbol{p}$ and unprojection via $\phi(\boldsymbol{p}, \hat{d}, \boldsymbol{i}) = \boldsymbol{P}$ assuming an estimated pixel depth of $\hat{d}$. The camera parameters $\boldsymbol{i}$ are generally the standard pinhole model [14] defined by the 3×3 intrinsic matrix $\boldsymbol{K}$, but can include any differentiable model such as the Unified Camera Model family [13] as described next.

B. End-to-End Self-Calibration.

UCM [20] is a parametric global central camera model that uses only five parameters to represent a diverse set of camera geometries, including perspective, fisheye, and catadioptric. A 3D point is projected onto a unit sphere and then projected

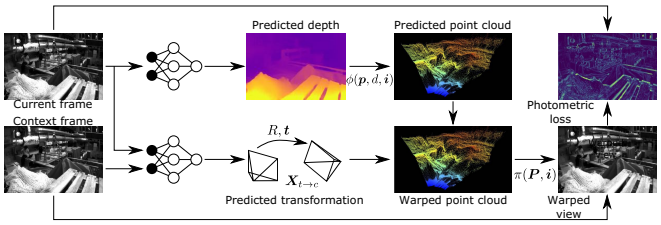


Fig. 2: **Our self-supervised self-calibration architecture.** We use gradients from the photometric loss to update the parameters of a unified camera model (Fig. 3).

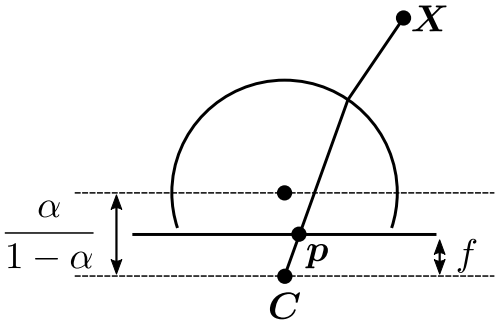


Fig. 3: **The Unified Camera Model [13] used in our self-calibration pipeline**. Points are projected onto a unit sphere before being projected onto an image plane of a standard pinhole camera offset by $\frac{\alpha}{1-\alpha}$ from the sphere center.

onto the image plane of a pinhole camera, shifted by $\frac{\alpha}{1-\alpha}$ from the center of the sphere (Fig. 3). EUCM and DS are two extensions of the UCM model. EUCM replaces the unit sphere with an ellipse as the first projection surface, and DS replaces the one unit sphere with two unit spheres in the projection process. We self-calibrate all three models (in addition to a pinhole baseline) in our experiments. For brevity, we only describe the original UCM and refer the reader to Usenko et al. [13] for details on the EUCM and DS models.

There are multiple parameterizations for UCM [20], and we use the one from Usenko et al. [13] since it has better numerical properties. UCM extends the pinhole camera model (f_x, f_y, c_x, c_y) with only one additional parameter α. The 3D-to-2D projection of $\boldsymbol{P} = (x, y, z)$ is defined as

$$\boldsymbol{\pi}(\boldsymbol{P}, \boldsymbol{i}) = \begin{bmatrix} f_x \frac{x}{\alpha d + (1-\alpha) z} \\ f_y \frac{y}{\alpha d + (1-\alpha) z} \end{bmatrix} + \begin{bmatrix} c_x \\ c_y \end{bmatrix} \tag{2}$$

where the camera parameters are $\boldsymbol{i} = (f_x, f_y, c_x, c_y, \alpha)$ and $d = \sqrt{x^2 + y^2 + z^2}$

The unprojection operation of pixel $\boldsymbol{p} = (u, v, 1)$ at estimated depth $\hat{d}$ is:

$$\phi(\boldsymbol{p}, \hat{d}, \boldsymbol{i}) = \hat{d} \frac{\xi + \sqrt{1 + (1 - \xi^2) r^2}}{1 + r^2} \begin{bmatrix} m_x \\ m_y \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ \hat{d}\zeta \end{bmatrix} \tag{3}$$

where

$$m_x = \frac{u - c_x}{f_x}(1 - \alpha) \qquad m_y = \frac{v - c_y}{f_y}(1 - \alpha) \tag{4a}$$

$$r^2 = m_x^2 + m_x^2 \qquad \zeta = \frac{\alpha}{1 - \alpha} \tag{4b}$$

As shown in Equations 2 and 3, the UCM camera model provides closed-form projection and unprojection functions that are both differentiable. Therefore, the overall architecture is end-to-end differentiable with respect to both neural network parameters (for pose and depth estimation) and camera parameters. This enables learning self-calibration end-to-end from the aforementioned view synthesis objective alone. At the start of self-supervised depth and pose training, rather than pre-calibrating the camera parameters, we initialize the camera with "default" values based on image shape only (for a detailed discussion of the initialization procedure, please see Section IV-E). Although the projection (2) and unprojection (3) are initially inaccurate, they quickly converge to highly accurate camera parameters with sub-pixel re-projection error (see Table I).

As we show in our experiments, our method combines flexibility with computational efficiency. Indeed, our approach enables learning from heterogeneous datasets with potentially vastly differing sensors for which separate parameters $\boldsymbol{i}$ are learned. As most of the parameters (in the depth and pose networks) are shared thanks to the decoupling of the projection model, this enables scaling up in-the-wild training of depth and pose networks. Furthermore, our method is efficient, with only one extra parameter relative to the pinhole model. This enables learning depth for highly-distorted catadioptric cameras at a much higher resolution than previous over-parametrized models (1024×1024 vs. 384×384 for Vasiljevic et al. [12]). Note that, in contrast to prior works [11, 12], we learn intrinsics per-sequence rather than per-frame. This increases stability compared to per-frame methods that exhibit frame-to-frame variability [12], and can be used over sequences of varying sizes.

IV. Experiments

In this section we describe two sets of experimental validations for our architecture: (i) calibration, where we find that the re-projection error of our learned camera parameters compares favorably to target-based traditional calibration toolboxes; and (ii) depth evaluation, where we achieve state-of-the-art results on the challenging EuRoC MAV dataset.

A. Datasets

Self-supervised depth and ego-motion learning uses monocular sequences [10, 11, 34, 35] or rectified stereo pairs [34, 36] from forward-facing cameras [35, 37, 38]. Given that our goal is to learn camera calibration from raw videos in challenging settings, we use the standard KITTI dataset as a baseline, and focus on the more challenging and distorted EuRoC [39] fisheye sequences.

KITTI [37] We use this dataset to show that our self-calibration procedure is able to accurately recover pinhole intrinsics alongside depth and ego-motion. Following related work [10, 11, 34, 35] we use the training protocol of [40], including filtering static images as described by Zhou et al.

Method	*Mean Reprojection Error*	
	Target-based	*Learned*
Pinhole	1.950	2.230
UCM [20]	0.145	0.249
EUCM [21]	0.144	0.245
DS [13]	0.144	0.344

TABLE I: **Mean reprojection error on EuRoC** at 256×384 resolution for UCM, EUCM and DS models using (left) AprilTag-based toolbox calibration Basalt [43] and (right) our self-supervised learned (L) calibration. Note that despite using no ground-truth calibration targets, our self-supervised procedure produces sub-pixel reprojection error.

[10]. The resulting training set contains of 39810 images, with 697 images left for evaluation.

EuRoC [39] The dataset consists of a set of indoor MAV sequences with general six-DoF motion. Consistent with recent work [11], we train using center-cropping and down-sample the images to a 384×256 resolution, while training and evaluating on the same split. For calibration evaluation, we follow Usenko et al. [13] and use the calibration sequences from the dataset. We evaluate the UCM, EUCM and DS camera models in terms of re-projection error.

OmniCam [41] A challenging outdoor catadioptric sequence, containing 12000 frames captured by an autonomous car rig. As this dataset does not provide ground-truth depth information, we only provide qualitative results.

B. Training Protocol

We implement the group of unified camera models described in [13] as differentiable PyTorch [42] operations, modifying the self-supervised depth and pose architecture of Godard et al. [34] to jointly learn depth, pose, and the unified camera model intrinsics. We use a learning rate of 2e-4 for the depth and pose network and 1e-3 for the camera parameters. We use a StepLR scheduler with $\gamma = 0.5$ and a step size of 30. All of the experiments are run for 50 epochs. The images are augmented with random vertical and horizontal flip, as well as color jittering. We train our models on a Titan X GPU with 12 GB of memory, with a batch size of 16 when training on images with a resolution of 384×256. We note that our method requires significantly less memory than that of Vasiljevic et al. [12] which learns a generalized camera model parameterized through a per-pixel ray surface.

C. Camera Self-Calibration

We evaluate the results of the proposed self-calibration method on the EuRoC dataset; detailed depth estimation evaluations are provided in Sec. IV-F. To our knowledge, ours is the first direct calibration evaluation of self-supervised intrinsics learning; although Gordon et al. [11] compare *ground-truth* calibration to their per-frame model, they do not evaluate the re-projection error for their learned parameters.

Following Usenko et al. [43], we evaluate our self-supervised calibration method on the family of unified camera models: UCM, EUCM, and DS, as well as the perspective (pinhole) model. As a lower bound, we use the Basalt [43] toolbox and compute camera calibration parameters for each unified camera model using the calibration sequences of the EuRoC dataset. We note that unlike Basalt, our method regresses the intrinsic calibration parameters directly from raw videos, without using any of the calibration sequences.

Method	f_x	f_y	c_x	c_y	α	β	ξ	w
UCM (L)	237.6	247.9	187.9	130.3	0.631	—	—	—
UCM (B)	235.4	245.1	186.5	132.6	0.650			
EUCM (L)	237.4	247.7	186.7	129.1	0.598	1.075	—	—
EUCM (B)	235.6	245.4	186.4	132.7	0.597	1.112		
DS (L)	184.8	193.3	187.8	130.2	0.561	—	-0.232	—
DS (B)	181.4	188.9	186.4	132.6	0.571		-0.230	

TABLE II: **Intrinsic calibration evaluation of different methods** on the EuRoC dataset, where B denotes intrinsics obtained from Basalt, and L denotes learned intrinsics.

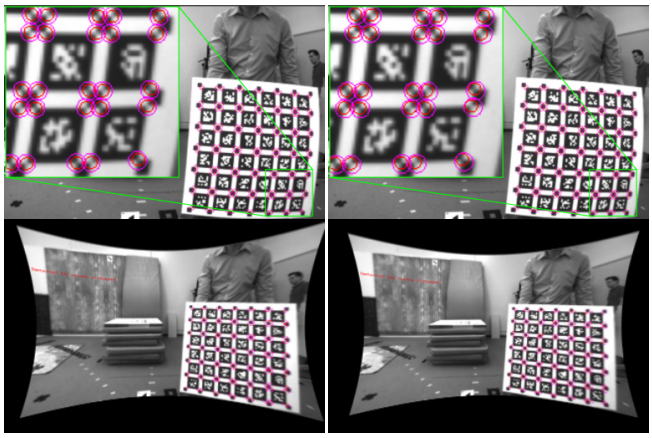


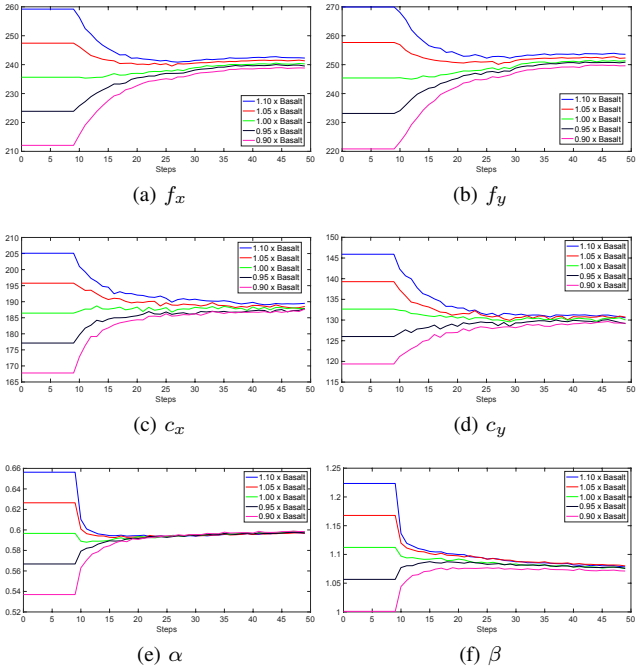
Fig. 4: **EuRoC rectification results** using images from the calibration sequences. Each column visualizes the results rendered using (left) the Basalt calibrated intrinsics and (right) our learned intrinsics. The top row shows that detected (small circles) and reprojected (big circles) corners are close using both calibration methods. The bottom row shows the same images after rectification.

Table I summarizes our re-projection error results. We use the EuRoC AprilTag [44] calibration sequences with Basalt to measure re-projection error using the full estimation procedure (Table I — *Target-based*) and learned intrinsics (Table I — *Learned*). For consistency, we optimize for both intrinsics and camera poses for the baselines and only for the camera poses for the learned intrinsics evaluation. Note that with learned intrinsics, UCM, EUCM and DS models all achieve sub-pixel mean projection error despite the camera parameters having been learned from raw video data.

Table II compares the target-based calibrated parameters to our learned parameters for different camera models trained on the *cam0* sequences of the EuRoC dataset. Though the parameter vectors were initialized with no prior knowledge of the camera model and updated purely based on gradients

Perturbation	f_x	f_y	c_x	c_y	α	β	MRE
$I_{1.10}$ init	242.3	253.6	189.5	130.7	0.5984	1.080	0.409
$I_{1.05}$ init	241.3	252.3	188.5	130.5	0.5981	1.078	0.367
I_c init	240.2	251.4	187.9	130.0	0.5971	1.076	0.348
$I_{0.95}$ init	239.5	250.9	187.8	129.2	0.5970	1.076	0.332
$I_{0.90}$ init	238.8	249.6	187.7	129.1	0.5968	1.071	0.298
I_c	235.6	245.4	186.4	132.7	0.597	1.112	0.144

TABLE III: **EUCM perturbation test results.** With perturbed initialization, all intrinsic parameters achieve subpixel convergence for mean reprojection error (**MRE**), with only a small offset to the Basalt calibration numbers.



(a) f_x (b) f_y (c) c_x (d) c_y (e) α (f) β

Fig. 5: **EuRoC perturbation test**, showing how our proposed learning-based method is able to recover from changes in camera parameters for online self-calibration.

from the reprojection error, they converge to values very close to the output of a procedure that uses bundle adjustment on calibrated image sequences.

D. Camera Rectification

Using our learned camera parameters, we rectify calibration sequences on the EuRoC dataset to demonstrate the quality of the calibration. EuRoC was captured with a fisheye camera and exhibits a high degree of radial distortion that causes the straight edges of the checkerboard grid to be curved. In Figure 4, we see that our learned parameters allow for the rectified grid to track closely to the true underlying checkerboard.

E. Camera Re-calibration: Perturbation Experiments

Thus far, we have assumed to have no prior knowledge of the camera calibration. In many real-world robotics settings, however, one may want to re-calibrate a camera based on a potentially incorrect prior calibration. Generally, this requires the capture of new calibration data. Instead, we can initialize our parameter vectors with this initial calibration (in this setting, a perturbation of Basalt calibration of the EUCM model) and see the extent to which self-supervision can nudge the parameters back to their "true value".

Given Basalt parameters $I_c = [f_x, f_y, c_x, c_y, \alpha, \beta]$, we perturb them as $I_{1.1} = 1.1 \times I_c$, $I_{1.05} = 1.05 \times I_c$, $I_{0.95} = 0.95 \times I_c$, $I_{0.9} = 0.9 \times I_c$ and initialize the camera parameters at the beginning of training with these values. All runs have warm start, i.e., freezing the gradients for the intrinsics for the first 10 epochs while we train the depth and pose networks. As Figure 5 shows, our method converges to within 3% of the Basalt estimate for each parameter. Table III provides the values of the converged parameters along with the mean projection error (MRE) for each experiment.

Method	Camera	Abs Rel↓	Sq Rel↓	RMSE↓	$\delta_{1.25}$ ↑
Gordon et al. [11]	K	0.129	0.982	5.23	0.840
Gordon et al. [11]	L(P)	0.128	0.959	5.23	0.845
Vasiljevic et al. [12]	K(NRS)	0.137	0.987	5.33	0.830
Vasiljevic et al. [12]	L(NRS)	0.134	0.952	5.26	0.832
Ours	L(P)	0.129	**0.893**	4.96	0.846
Ours	L(UCM)	**0.126**	0.951	**4.89**	**0.858**

TABLE IV: **Quantitative depth evaluation on the KITTI [39] dataset**, using the standard *Eigen* split and the *Garg* crop, for distances up to 80m (with median scaling). K and L(·) denote known and learned intrinsics, respectively. P means pinhole model.

Method	Camera	Abs Rel↓	Sq Rel↓	RMSE↓	α_1 ↑
Gordon et al. [11]	PB	0.332	0.389	0.971	0.420
Vasiljevic et al. [12]	NRS	0.303	0.056	0.154	0.556
Ours	UCM	0.282	0.048	0.141	0.591
Ours	EUCM	**0.278**	**0.047**	**0.135**	**0.598**
Ours	DS	**0.278**	0.049	0.141	0.584

TABLE V: **Quantitative depth evaluation of different methods on the EuROC [39] dataset**, using the evaluation procedure in [11] with center cropping. The training data consists of "Machine Room" sequences and the evaluation is on the "Vicon Room 201" sequence (with median scaling). PN means plum-bob model.

F. Depth Estimation

While we use depth and pose estimation as proxy tasks for camera self-calibration, the unified camera model framework allows us to achieve meaningful results compared to prior camera-learning-based approaches (see Figures 6 and 7).

KITTI results. Table IV presents the results of our method on the KITTI dataset. We note that our approach is able to model the simple pinhole setting, achieving results that are on par with approaches that are tailored specifically to this camera geometry. Interestingly, we see an increase in performance using the UCM model, which we attribute to the

Dataset	Abs Rel↓	Sq Rel↓	RMSE↓	α_1 ↑	α_2 ↑	α_3 ↑
EuRoC [11]	0.265	**0.042**	0.130	0.600	0.882	**0.966**
EuRoC+KITTI	**0.244**	0.044	**0.117**	**0.742**	**0.907**	0.961

TABLE VI: **Quantitative multi-dataset depth evaluation** on EuRoC (without cropping and with median scaling).

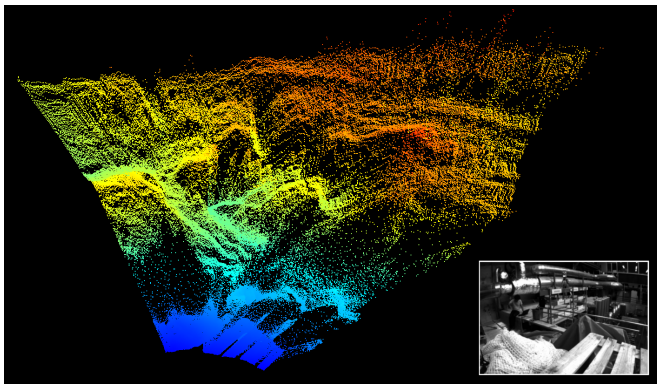


Fig. 6: **Self-supervised monocular pointcloud** for EuRoC, obtained by unprojecting predicted depth with our learned camera parameters (input image on the bottom right).

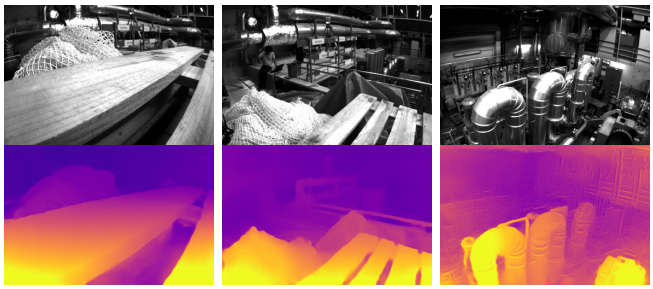
ability to further account for and correct calibration errors.

EuRoC results. Compared to KITTI, EuRoC is a significantly more challenging dataset that involves cluttered indoor sequences with six-DoF motion. Compared to the per-frame distorted camera models of Gordon et al. [11] and Vasiljevic et al. [12], we achieve significantly better absolute relative error, especially with EUCM, where the error is reduced by 16% (see Table V). We also train NRS [12] on this dataset for further comparison, using the official repository.

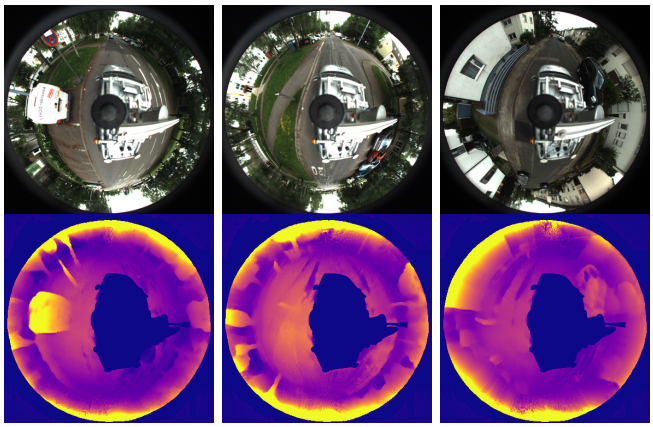
Combining heterogeneous datasets. One of the strengths of the unified camera model is that it can represent a wide variety of cameras without prior knowledge of their specific geometry. As long as we know which sequences come from which camera, we can learn separate calibration vectors that share the same depth and pose networks. This is particularly useful as a way to improve performance on smaller datasets, since it enables one to take advantage of unlabeled data from other sources. To evaluate this property, we experimented with mixing KITTI and EuRoC. In this experiment, we reshaped the KITTI images to match those in the EuRoC dataset (i.e., 384×256). As Table VI shows, our algorithm is able to take advantage of the KITTI images to improve performance on the EuRoC depth evaluation.

G. Computational Cost

Our work is closely related to the learned general camera model (NRS) of Vasiljevic et al. [12] given that in both works the parameters of a central general camera model are learned in a self-supervised way. Being a per-pixel model, NRS is more general than ours and can handle settings where there is local distortion, which a global camera necessarily cannot model. However, the computational requirements of the per-pixel NRS are significantly higher. For example, we train on



(a) EuRoC



(b) OmniCam

Fig. 7: **Qualitative depth estimation results** on non-pinhole datasets with (a) fisheye and (b) catadioptric images.

EuRoC images with a resolution of 384×256 with a batch size of 16, which consumes about 6 GB of GPU memory. Each epoch takes about 15 minutes.

On the same GPU, NRS uses 16 GB of GPU memory with a batch size of 1 to train on the same sequences, running one epoch in about 120 minutes. This is due to the high-dimensional (yet approximate) projection operation required for a generalized camera. Thus, we trade some degree of generality for significantly higher efficiency than prior work, with higher accuracy on the EuRoC dataset (see Table V).

V. Conclusion

We proposed a procedure to self-calibrate a family of general camera models using self-supervised depth and pose estimation as a proxy task. We rigorously evaluated the quality of the resulting camera models, demonstrating sub-pixel calibration accuracy comparable to manual target-based toolbox calibration approaches. Our approach generates per-sequence camera parameters, and can be integrated into any learning procedure where calibration is needed and the projection and un-projection operations are interpretable and differentiable. As shown in our experiments, our approach is particularly amenable to online re-calibration, and can be used to combine datasets of different sources, learning independent calibration parameters while sharing the same depth and pose network.

REFERENCES

[1] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000. 1, 2

[2] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A flexible technique for accurate omnidirectional camera calibration and structure from motion," in *Proceedings of the IEEE International Conference on Computer Vision Systems (ICVS)*, 2006, pp. 45–45. 1, 2

[3] J. Kannala and S. 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*, vol. 28, no. 8, pp. 1335–1340, 2006. 2

[4] M. D. Grossberg and S. K. Nayar, "A general imaging model and a method for finding its parameters," in *Proceedings of the International Conference on Computer Vision (ICCV)*, vol. 2, 2001, pp. 108–115. 1

[5] B. Caprile and V. Torre, "Using vanishing points for camera calibration," *International Journal on Computer Vision*, vol. 4, no. 2, pp. 127–139, 1990. 1, 2

[6] M. Pollefeys and L. Van Gool, "A stratified approach to metric self-calibration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1997, pp. 407–412.

[7] R. Cipolla, T. Drummond, and D. P. Robertson, "Camera calibration from vanishing points in image of architectural scenes," in *Proceedings of the British Machine Vision Conference (BMVC)*, 1999, pp. 382–391. 1, 2

[8] O. Bogdan, V. Eckstein, F. Rameau, and J.-C. Bazin, "DeepCalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras," in *Proceedings of the ACM SIGGRAPH European Conference on Visual Media Production*, 2018. 1, 2

[9] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 740–756. 1

[10] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1851–1858. 1, 2, 3, 4

[11] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019, pp. 8977–8986. 1, 2, 3, 4, 5, 6

[12] I. Vasiljevic, V. Guizilini, R. Ambrus, S. Pillai, W. Burgard, G. Shakhnarovich, and A. Gaidon, "Neural ray surfaces for self-supervised learning of depth and ego-motion," in *Proceedings of the International Conference on 3D Vision (3DV)*, 2020. 1, 2, 3, 4, 5, 6

[13] V. Usenko, N. Demmel, and D. Cremers, "The double sphere camera model," in *Proceedings of the International Conference on 3D Vision (3DV)*, 2018, pp. 552–560. 1, 2, 3, 4

[14] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, 2000. 2

[15] G. Bradski and A. Kaehler, "OpenCV," *Dr. Dobb's Journal of Software Tools*, vol. 3, 2000. 2

[16] J. Rehder, J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart, "Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 4304–4311. 2

[17] J. Leonard, J. How, S. Teller, M. Berger, S. Campbell, G. Fiore, L. Fletcher, E. Frazzoli, A. Huang, S. Karaman, O. Koch, Y. Kuwata, D. Moore, E. Olson, S. Peters, J. Teo, R. Truax, M. Walter, D. Barrett, A. Epstein, K. Maheloni, K. Moyer, T. Jones, R. Buckley, M. Antone, R. Galejs, S. Krishnamurthy, and J. Williams, "A perception-driven autonomous urban vehicle," *Journal of Field Robotics*, vol. 25, no. 10, pp. 727–774, October 2008. 2

[18] C. Urmson, J. Anhalt, D. Bagnell, C. Baker, R. Bittner, M. Clark, J. Dolan, D. Duggins, T. Galatali, C. Geyer *et al.*, "Autonomous driving in urban environments: Boss and the Urban Challenge," *Journal of Field Robotics*, vol. 25, no. 8, pp. 425–466, 2008. 2

[19] J. G. Fryer and D. C. Brown, "Lens distortion for close-range photogrammetry," *Photogrammetric Engineering and Remote Sensing*, vol. 52, pp. 51–58, 1986. 2

[20] C. Geyer and K. Daniilidis, "A unifying theory for central panoramic systems and practical implications," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2000, pp. 445–461. 2, 3, 4

[21] B. Khomutenko, G. Garcia, and P. Martinet, "An enhanced unified camera model," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 137–144, 2015. 2, 4

[22] S. Workman, C. Greenwell, M. Zhai, R. Baltenberger, and N. Jacobs, "DeepFocal: A method for direct focal length estimation," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 1369–1373. 2

[23] S. Workman, M. Zhai, and N. Jacobs, "Horizon lines in the wild," *arXiv preprint arXiv:1604.02129*, 2016. 2

[24] J. Rong, S. Huang, Z. Shang, and X. Ying, "Radial lens distortion correction using convolutional neural networks trained with synthesized images," in *Proceedings of the Asian Conference on Computer Vision*, 2016, pp. 35–49. 2

[25] X. Yin, X. Wang, J. Yu, M. Zhang, P. Fua, and D. Tao, "FishEyeRectNet: A multi-context collaborative deep network for fisheye image rectification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 469–484. 2

[26] M. Lopez, R. Mari, P. Gargallo, Y. Kuang, J. Gonzalez-Jimenez, and G. Haro, "Deep single image camera calibration with radial distortion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 11 817–11 825. 2

[27] Y. Hold-Geoffroy, K. Sunkavalli, J. Eisenmann, M. Fisher, E. Gambaretto, S. Hadap, and J.-F. Lalonde, "A perceptual measure for deep single image camera calibration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[28] R. Zhu, X. Yang, Y. Hold-Geoffroy, F. Perazzi, J. Eisenmann, K. Sunkavalli, and M. Chandraker, "Single view metrology in the wild," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 316–333. 2

[29] C. Zhang, F. Rameau, J. Kim, D. M. Argaw, J.-C. Bazin, and I. S. Kweon, "DeepPTZ: Deep self-calibration for PTZ cameras," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1041–1049. 2

[30] Y. Zhao, Z. Huang, T. Li, W. Chen, C. LeGendre, X. Ren, A. Shapiro, and H. Li, "Learning perspective undistortion of portraits," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019, pp. 7849–7859. 2

[31] W. Yin, J. Zhang, O. Wang, S. Niklaus, L. Mai, S. Chen, and C. Shen, "Learning to recover 3D scene shape from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 204–213. 2

[32] S. Yang, C. Lin, K. Liao, C. Zhang, and Y. Zhao, "Progressively complementary network for fisheye image rectification using appearance flow," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6348–6357. 2

[33] K. Liao, C. Lin, and Y. Zhao, "A deep ordinal distortion estimation approach for distortion rectification," *IEEE Trans-*

actions on Image Processing, vol. 30, pp. 3362–3375, 2021. 2

[34] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019, pp. 3828–3838. 3, 4

[35] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3D packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[36] S. Pillai, R. Ambrus, and A. Gaidon, "SuperDepth: Self-supervised, super-resolved monocular depth estimation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2019. 3

[37] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361. 3

[38] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 621–11 631. 3

[39] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016. 3, 4, 5

[40] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *arXiv preprint arXiv:1406.2283*, 2014. 3

[41] M. Schönbein, T. Strauß, and A. Geiger, "Calibrating and centering quasi-central catadioptric cameras," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 4443–4450. 4

[42] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NeurIPS Workshop*, 2017. 4

[43] V. Usenko, N. Demmel, D. Schubert, J. Stueckler, and D. Cremers, "Visual-inertial mapping with non-linear factor recovery," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 422–429, 2020. 4

[44] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 3400–3407. 4