Degeneracy in Self-Calibration Revisited and a Deep Learning Solution for Uncalibrated SLAM

Bingbing Zhuang[1] Quoc-Huy Tran[2] Gim Hee Lee[1] Loong Fah Cheong[1] Manmohan Chandraker[2,3]

Abstract— Self-calibration of camera intrinsics and radial distortion has a long history of research in the computer vision community. However, it remains rare to see real applications of such techniques to modern Simultaneous Localization And Mapping (SLAM) systems, especially in driving scenarios. In this paper, we revisit the geometric approach to this problem, and provide a theoretical proof that explicitly shows the ambiguity between radial distortion and scene depth when two-view geometry is used to self-calibrate the radial distortion. In view of such geometric degeneracy, we propose a learning approach that trains a convolutional neural network (CNN) on a large amount of synthetic data. We demonstrate the utility of our proposed method by applying it as a checkerboard-free calibration tool for SLAM, achieving comparable or superior performance to previous learning and hand-crafted methods.

I. INTRODUCTION

With the research efforts in the last decade, SLAM has achieved significant progress and maturity. It has reached the stage that one can expect satisfactory results to be obtained by the various state-of-the-art SLAM systems [1], [2], [3] under well-controlled environments. However, as pointed out in [4], many challenges still exist under more unconstrained conditions. These typically include image degradation, e.g., motion blurs [5], [6] and rolling shutter effects [7], [8], and unknown camera calibration including radial distortion. In this paper, we work towards a robust camera calibration system that enables SLAM on a video input without knowing the camera calibration a priori. Specifically, we revisit the ill-posedness of the self-calibration problem as set forth in a conventional geometric approach, and propose an effective data-driven approach.

Radial distortion estimation has been a long-standing problem in geometric vision. Many minimal solvers (e.g, [9], [10]) utilize two-view epipolar constraints to estimate radial distortion parameters. However, such methods have not been integrated into state-of-the-art SLAM systems [1], [2], [3]. Therefore, its robustness in practical scenarios, e.g. driving scenes [11], [12], [13], has not been tested as extensively as one would have liked. In this paper, we provide a further theoretical and numerical study on such geometric approaches. Specifically, although it has been mentioned in existing works [14], [15] that forward camera motion constitutes a degenerate case for two-view
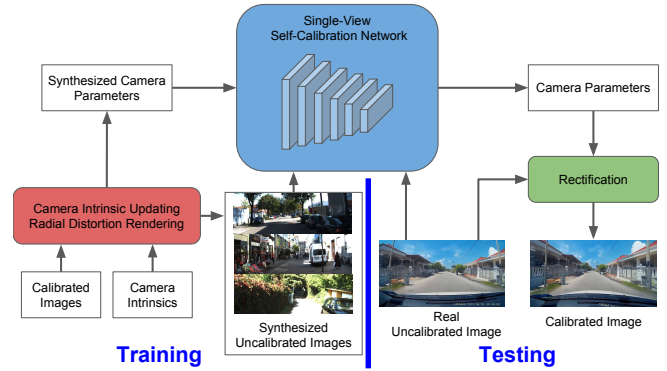

Fig. 1. Overview of our approach to single-view camera self-calibration. From a set of calibrated images with known camera intrinsics and no radial distortion, we generate synthetic uncalibrated images for training our network by randomly adjusting camera intrinsics and randomly adding radial distortion. At testing, given a single real uncalibrated image, our network predicts accurate camera parameters, which are used for producing the calibrated image.

radial distortion self-calibration, we give a more general and formal proof of this observation that explicitly describes the ambiguity between radial distortion and scene depth in two-view geometry. We further show that such degeneracy causes the solution to be very unstable when applied to driving scenes under (near-)forward motion. Moreover, it is well known from previous works [16], [17] that focal length estimation from two views also suffers from degeneracy under various configurations, e.g., when the optical axes of the two views intersect at a single point.

To mitigate such geometric degeneracy, we propose to exploit a data-driven approach that learns camera self-calibration from a single image (see Fig. 1). This allows us to leverage any regularity in the scenes so that the intrinsic degeneracies of the self-calibration problem can be overcome. Specifically, we target at driving scenes and synthesize a large amount of images with different values of radial distortion, focal length, and principal point. We show that a network trained on such synthetic data is able to generalize well to unseen real data. Furthermore, in this work we exploit the recent success of CNNs with deep supervision [18], [19], [20], [21], [22], which suggests that providing explicit supervisions to intermediate layers contributes to improved regularization. We observe that it performs better than a typical multi-task network [23] in terms of radial distortion and focal length estimation. Lastly, we empirically show that it is feasible to perform SLAM on uncalibrated videos from KITTI Raw [11] and YouTube with the self-calibration by our network.
In summary, our contributions include:

[1]B. Zhuang, G.H. Lee, and L. F. Cheong are with National University of Singapore zhuang.bingbing@u.nus.edu, dcslgh@nus.edu.sg, eleclf@nus.edu.sg.
[2]Q.-H. Tran and M. Chandraker are with NEC Laboratories America, Inc. qhtran@nec-labs.com, manu@nec-labs.com.
[3]M. Chandraker is also with University of California, San Diego mkchandraker@eng.ucsd.edu.

- Using a general distortion model and a discrete Structure from Motion (SfM) formulation, we prove that two-view self-calibration of radial distortion is degenerate under forward motion. This is important for understanding intrinsic and algorithm-independent properties of radially-distorted two-view geometry estimation.
- We propose a CNN-based approach which leverages deep supervision for learning both radial distortion and camera intrinsics from a single image. We demonstrate the application of our single-view camera self-calibration network to SLAM operating on uncalibrated videos from KITTI Raw and YouTube.

II. RELATED WORK

Geometric Degeneracy With Unknown Radial Distortion: It has been shown that self-calibration of unknown radial distortion parameters using two-view geometry suffers from degeneracy under forward motion. Specifically, [14] shows that two-view radially-distorted geometry under forward motion is degenerate with the division distortion model [24] and a discrete SfM formulation. More recently, critical configurations of radially-distorted cameras under infinitesimal motions are identified and proved in [15] for a general distortion model, but with an approximate differential SfM formulation [25]. In this work, we prove that radially-distorted two-view geometry under forward motion is degenerate using a more general setting than [14], [15], i.e. a general distortion model and a discrete SfM formulation.

Multiple-View Camera Self-Calibration: Multiple-view methods use multiple-view geometry (usually two-view epipolar constraints) in radially-distorted images to correct radial distortion [9], [26], [10] as well as estimate camera intrinsics [27], via keypoint correspondences between the images. Despite being able to handle different camera motions or scene structures, they need two input images and often require solving complex polynomial systems.

Single-View Camera Self-Calibration: Single-view methods rely on extracted line/curve features in the input image to remove radial distortion [28], [29] and/or compute camera intrinsics [30], [31], [32], [33]. Moreover, some of them assume special scene structures, e.g., Manhattan world [33]. One problem with these hand-crafted methods is that they cannot work well when the scenes contain very few line/curve features or the underlying assumptions on the scene structures are violated. Recently, single-view CNN-based methods have been proposed for radial distortion correction [34] and/or camera intrinsic calibration [35], [36]. Our method also uses powerful CNN-extracted features, however, unlike [35], [36] which use multi-task supervision, we apply deep supervision. More importantly, we demonstrate the practical utility of the deep learning method for uncalibrated SLAM.

III. DEGENERACY IN TWO-VIEW RADIAL DISTORTION SELF-CALIBRATION UNDER FORWARD MOTION

We denote the 2D coordinates of a distorted point on a normalized image plane as $s_d = [x_d, y_d]^\top$ and the
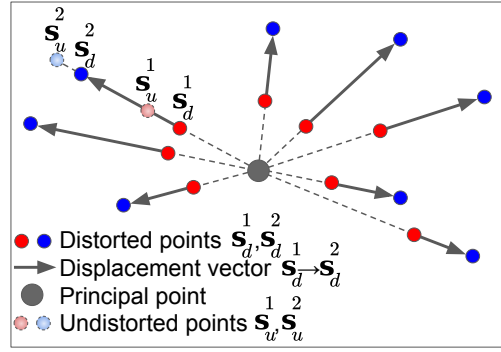

Fig. 2. Degeneracy in two-view radial distortion self-calibration under forward motion. There are infinite number of valid combinations of radial distortion and scene structure, including the special case with zero radial distortion.

corresponding undistorted point as $s_u = [x_u, y_u]^\top = f(s_d; \boldsymbol{\theta})s_d$. $\boldsymbol{\theta}$ is the radial distortion parameters and $f(s_d; \boldsymbol{\theta})$ is the undistortion function which scales s_d to s_u. The specific form of $f(s_d; \boldsymbol{\theta})$ depends on the radial distortion model being used. For instance, we have $f(s_d; \lambda) = \frac{1}{1+\lambda r^2}$ for the division model [24] with one parameter, or we have $f(s_d; \lambda) = 1 + \lambda r^2$ for the polynomial model [37] with one parameter. In both models, λ is the 1D radial distortion parameter and $r = \sqrt{x_d^2 + y_d^2}$ is the distance from the principal point. We use the general form $f(s_d; \boldsymbol{\theta})$ for the analysis below.

We formulate the two-view geometric relationship under forward motion, i.e., how a pure translational camera motion along the optical axis is related to the 2D correspondences and their depths. Let us consider a 3D point S, expressed as $S_1 = [X_1, Y_1, Z_1]^\top$ and $S_2 = [X_2, Y_2, Z_2]^\top$, respectively, in the two camera coordinates. Under forward motion, we have $S_2 = S_1 - T$ with $T = [0, 0, t_z]^\top$. Without loss of generality, we fix $t_z = 1$ to remove the global scale ambiguity. Projecting the above relationship onto the image planes, we obtain $s_u^2 = \frac{Z_1}{Z_1-1}s_u^1$, where s_u^1 and s_u^2 are the 2D projections of S_1 and S_2, respectively (i.e., $\{s_u^1, s_u^2\}$ is a 2D correspondence). Expressing the above in terms of the observed distorted points s_d^1 and s_d^2 yields:

$$f(s_d^2; \boldsymbol{\theta}_2)s_d^2 = \frac{Z_1}{Z_1 - 1}f(s_d^1; \boldsymbol{\theta}_1)s_d^1, \qquad (1)$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ represent radial distortion parameters in the two images respectively (note that $\boldsymbol{\theta}_1$ may differ from $\boldsymbol{\theta}_2$).

Eq. 1 represents all the information available for estimating the radial distortion and the scene structure. However, the correct radial distortion and point depth cannot be determined from the above equation. We can replace the ground truth radial distortion denoted by $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ with a fake radial distortion $\{\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'\}$ and the ground truth point depth Z_1 for each 2D correspondence with the following fake depth Z_1' such that Eq. 1 still holds:

$$Z_1' = \frac{\alpha Z_1}{(\alpha - 1)Z_1 + 1}, \quad \alpha = \frac{f(s_d^2; \boldsymbol{\theta}_2')f(s_d^1; \boldsymbol{\theta}_1)}{f(s_d^1; \boldsymbol{\theta}_1')f(s_d^2; \boldsymbol{\theta}_2)}. \qquad (2)$$

In particular, we can possibly set $\forall s_d^1 : f(s_d^1; \boldsymbol{\theta}_1') = 1, \forall s_d^2 :$

3767

$f(\boldsymbol{s}_d^2; \boldsymbol{\theta}_2') = 1$ as the fake radial distortion, and use the corrupted depth Z_1' computed according to Eq. 2 so that Eq. 1 still holds. This special solution corresponds to the pinhole camera model, i.e., $\boldsymbol{s}_u^1 = \boldsymbol{s}_d^1$ and $\boldsymbol{s}_u^2 = \boldsymbol{s}_d^2$. In fact, this special case can be inferred more intuitively. Eq. 1 indicates that all 2D points move along 2D lines radiating from the principal point, as illustrated in Fig. 2. This pattern is exactly the same as in the pinhole camera model, and is the sole cue to recognize the forward motion.

Intuitively, the 2D point movements induced by radial distortion alone, e.g., between $\boldsymbol{s}_u^1$ and $\boldsymbol{s}_d^1$, or between $\boldsymbol{s}_u^2$ and $\boldsymbol{s}_d^2$, are along the same direction as the 2D point movements induced by forward motion alone, e.g., between $\boldsymbol{s}_u^1$ and $\boldsymbol{s}_u^2$ (see Fig. 2). Hence, radial distortion only affects the magnitudes of 2D point displacements but not their directions in cases of forward motion. Furthermore, such radial distortion can be compensated with an appropriate corruption in the depths so that a corrupted scene structure that explains the image observations, i.e., 2D correspondences, exactly in terms of reprojection errors can still be recovered.

We arrive at the following proposition:

Proposition: *Two-view radial distortion self-calibration is degenerate for the case of pure forward motion. In particular, there are infinite number of valid combinations of radial distortion and scene structure, including the special case of zero radial distortion.*

IV. Learning Radial Distortion and Camera Intrinsics from a Single Image

In this section, we present the details of our learning approach for single-view camera self-calibration. Fig. 1 shows an overview of our approach.

A. Network Architecture

We adopt the ResNet-34 architecture [38] as our base architecture and make the following changes. We first remove the last average pooling layer which was designed for ImageNet classification. Next, we add a set of 3×3 convolutional layers (each followed by a BatchNorm layer and a ReLU activation layer) for extracting features, and then a few 1×1 convolutional layers (with no bias) for regressing the output parameters, including radial distortion λ (defined on the normalized plane), focal length f, and principal point $\{c_x, c_y\}$. Here, we adopt the widely used division model [24] with one parameter for radial distortion (the corresponding undistortion only requires to solve a simple quadratic equation).

To apply deep supervision [21] on our problem, we need to select a dependence order between the predicted parameters. Knowing that: (1) a known principal point is clearly a prerequisite for estimating radial distortion, and (2) image appearance is affected by the composite effect of radial distortion and focal length, we predict the parameters in the following order: (1) principal point in the first branch, and (2) both focal length and radial distortion in the second branch. Fig. 3 shows our network architecture. We have also tried separating the prediction of λ and f in two branches but
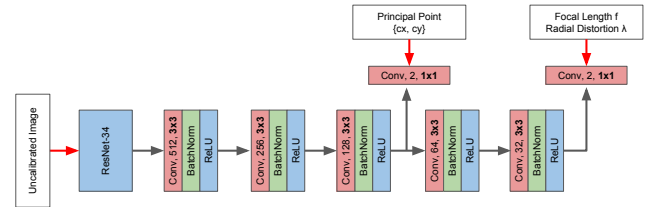

Fig. 3. Detailed architecture of our single-view camera self-calibration network. White boxes denote input data (left) or supervision signals (top).

it does not perform as well. We train our network by using an L_1-norm regression loss for each predicted parameter (L_1-norm is preferred since it is more robust than L_2-norm).

B. Training Data Generation

Since there is no large-scale dataset of real uncalibrated images with ground truth camera parameters available, we generate synthetic images with ground truth parameters based on the KITTI Raw dataset [11] to train our network. We use all sequences in the 'City' and 'Residential' categories, except for 5 sequences that contain mostly static scenes (i.e., sequences '0017', '0018', '0057', '0060' in '2011_09_26', and sequence '0001' in '2011_09_28') and 2 sequences that are used for validation (i.e., sequence '0071' in '2011_09_29') and evaluation in Sec. V (i.e., sequence '0027' in '2011_10_03'). In total, we use 42 sequences with around 30K frames to generate the training data.

For each calibrated image with known camera intrinsics and no radial distortion, we randomly vary the camera intrinsics ($\{c_x, c_y\} \in [0.45, 0.55] \times [0.45, 0.55]$, $f \in [0.6, 1.8]$) and randomly add a radial distortion component ($\lambda \in [-1.0, 0]$) — note that f and c_x are normalized by image width, while c_y is normalized by image height. We find the above parameter ranges to be sufficient for the videos we test as shown in Sec. V. The modified parameters are then used to generate a new uncalibrated image accordingly by generating radial distortion effect, image cropping, and image resizing. For each calibrated image, we repeat the above procedure 10 times to generate 10 uncalibrated images, resulting in a total of around 300K synthetic uncalibrated images with known ground truth camera parameters for training our network.

C. Training Details

We set the weight of each regression loss to 1.0 and use the ADAM optimization [39] with learning rate of 10^{-3} to train our network. We use the pre-trained weights of ResNet-34 on ImageNet classification to initialize the common layers in our network, and the newly added layers are randomly initialized by using [40]. The input image size is 320×960 pixels. We set the batch size to 40 images and train our network for 6 epochs (chosen based on validation performance). Our network is implemented in PyTorch [41].

V. Experiments

We first analyze the performance of traditional geometric method and our learning method for radial distortion self-calibration in near-degenerate cases in Sec. V-A. Next, we

3768

compare our method against state-of-the-art methods for single-view camera self-calibration in Sec. V-B. Lastly, Sec. V-C demonstrates the application of our method to uncalibrated SLAM. Note that for our method, we use our model trained solely on the synthetic data from Sec. IV-B without any fine-tuning with real data.

Competing Methods: We benchmark our method ('DeepSup') against state-of-the-art methods [30], [32], [36] for single-view camera self-calibration. Inspired by [36], we add a multi-task baseline ('MultiTask'), which has similar network architecture as ours but employing multi-task supervision, i.e., predicting principal point, focal length, and radial distortion all at the final layer. We train both 'MultiTask' and 'DeepSup' on the same set of synthetic data in Sec. IV-B. In addition, we evaluate the performance of two hand-crafted methods: (1) 'HandCrafted', which uses lines or curves and Hough Transformation to estimate radial distortion and principal point [32] and (2) 'HandCrafted+OVP', which uses orthogonal vanishing points to estimate radial distortion and focal length [30]. For 'HandCrafted+OVP', we use the binary code provided by the authors. For 'HandCrafted', we need to manually upload each image separately to the web interface[1] to obtain the results, thus we only show a few qualitative comparisons with this method in the following experiments.

Test Data: Our test data consists of 4 real uncalibrated sequences. In particular, 1 sequence is originally provided by KITTI Raw [11] (i.e., sequence '0027' in '2011_10_03' — with over 4.5K frames in total and no overlap with the sequences used for generating the training data in Sec. IV-B) and 3 sequences are extracted from YouTube videos (with 1.2K, 2.4K, and 1.6K frames respectively). Ground truth camera parameters and camera poses are available for the KITTI Raw test sequence (we convert the distortion model used in KITTI Raw, i.e., OpenCV calibration toolbox, to the division model [24] with one parameter by regressing the new model parameter). Ground truth for YouTube test sequences is not available. We also note that the aspect ratio of the YouTube images is different from that of the training data. Therefore, we first crop the YouTube images to match the aspect ratio of the training data before feeding them to the network, and then convert the predicted camera parameters according to the aspect ratio of the original YouTube images.

A. Geometric Approach vs. Learning Approach

We have shown in Sec. III that estimating radial distortion from radially-distorted two-view geometry is degenerate when the camera undergoes ideal forward motion. Here, we investigate the behavior of traditional geometric approach in a practical scenario, i.e., the KITTI Raw test sequence, where the camera motion is near-forward motion. To this end, we implement a minimal solver (9-point algorithm) to solve for both radial distortion λ (i.e., division model [24] with one parameter) and relative pose by using the following radial
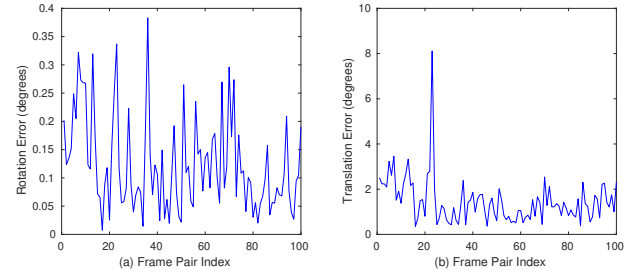
Fig. 4. Relative pose estimation by geometric approach. (a) Rotation errors. (b) Translation errors.
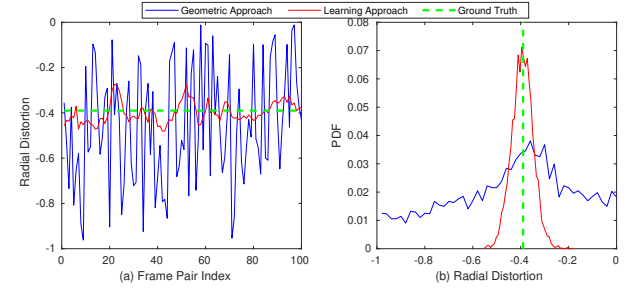

Fig. 5. Radial distortion estimation by geometric and learning approaches. (a) Radial distortion estimates. (b) PDF of radial distortion estimates.
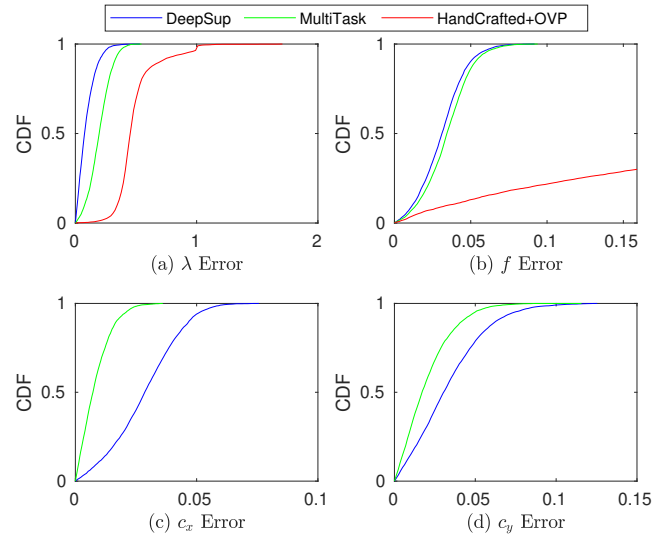

Fig. 6. Quantitative results of single-view camera self-calibration on the KITTI Raw test sequence. (a) λ errors. (b) f errors. (c) c_x errors. (d) c_y errors. In (b), we only show the CDF corresponding to the smallest 30% f errors for visibility. 'HandCrafted+OVP' assumes known principal point and hence is excluded in (c)-(d).

distortion aware epipolar constraint:

$$[x_d^2, \ y_d^2, \ 1+\lambda(x_d^2+y_d^2)] \ \boldsymbol{E} \ [x_d^2, \ y_d^2, \ 1+\lambda(x_d^2+y_d^2)]^\top = 0,$$
(3)

where $[x_d^{\{1,2\}}, \ y_d^{\{1,2\}}]^\top$ is the observed 2D correspondence in the two images and $\boldsymbol{E}$ is the essential matrix where the relative pose can be extracted. We apply the minimal solver within RANSAC [42], [43] to the first 100 consecutive frame pairs of the KITTI Raw test sequence. We first plot the errors of relative pose estimation by the above geometric approach in Fig. 4, which shows that both rotation and translation are
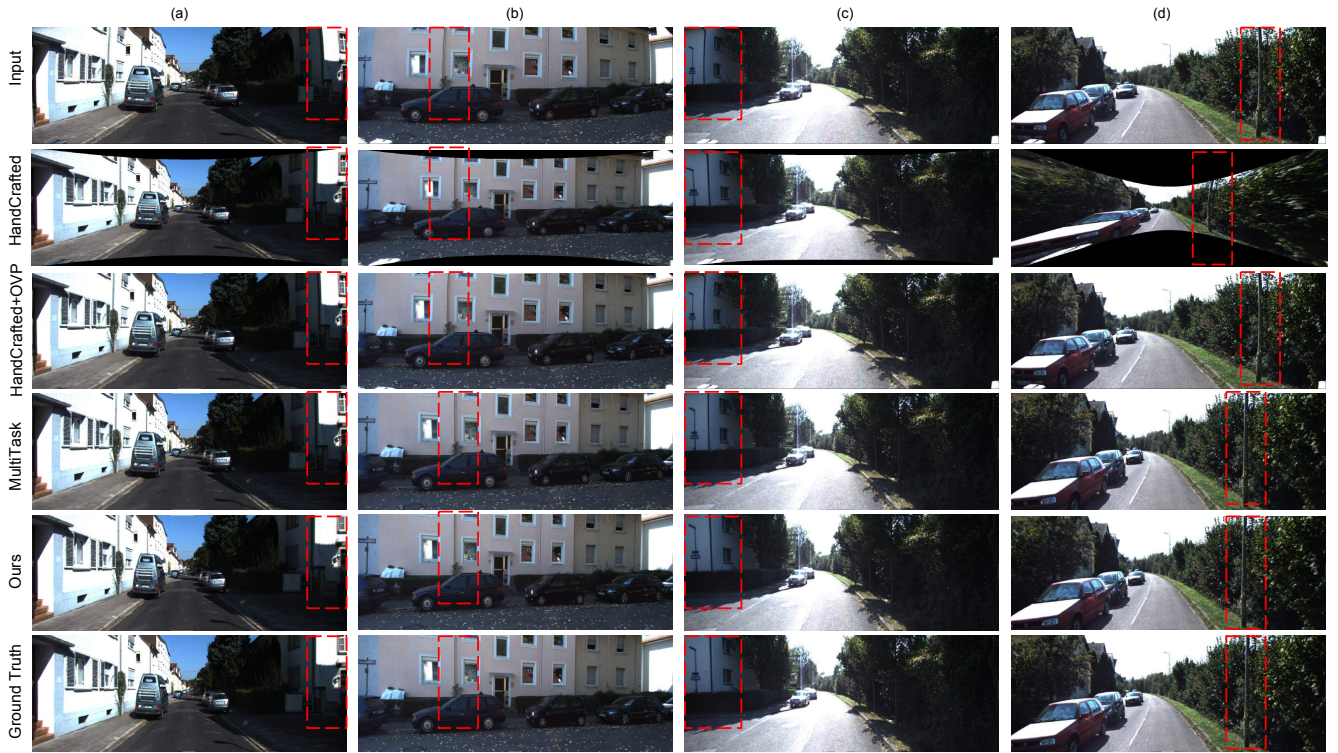
Fig. 7. Qualitative results of single-view camera self-calibration on the KITTI Raw test sequence.

estimated with relatively small errors. Next, we plot the results of radial distortion estimation by the geometric approach in Fig. 5, including the radial distortion estimates in (a) and the probability density function (PDF) of the estimates in (b). In addition, we include the results of our learning approach in Fig. 5 for comparison. Note that for a fair comparison, in the geometric approach we exclude any minimal sample yielding radial distortion estimate outside $[-1.0, 0]$ (which is the radial distortion range of the training data for our learning approach). From the results, it is evident that compared to the geometric approach, the radial distortions estimated from our learning approach are more concentrated around the ground truth ($\lambda = -0.39$), indicating a more robust performance in the (near-)degenerate setting.

B. Single-View Camera Self-Calibration

KITTI Raw: We first quantitatively evaluate the estimation accuracy of radial distortion λ, focal length f, and principal point c_x, c_y against their ground truth values of $\bar{\lambda}$, $\bar{f}$, $\bar{c}_x$, $\bar{c}_y$. We compute the relative errors $\frac{|\lambda - \bar{\lambda}|}{|\bar{\lambda}|}$, $\frac{|f - \bar{f}|}{|\bar{f}|}$, $\frac{|c_x - \bar{c}_x|}{|\bar{c}_x|}$, $\frac{|c_y - \bar{c}_y|}{|\bar{c}_y|}$ and plot the cumulative distribution function (CDF) of the errors from all test images in Fig. 6. It is clear that CNN-based methods (i.e., 'DeepSup' and 'MultiTask') outperform the hand-crafted method [30] (i.e., 'HandCrafted+OVP') by a significant margin. In particular, we find that the focal lengths estimated by 'HandCrafted+OVP' are evidently unstable and suffer from large errors. We also observe that 'DeepSup' predicts λ and f more accurately than 'MultiTask', although it has lower accuracy on c_x and c_y estimation than 'MultiTask'.

Next, we show a few qualitative results in Fig. 7. We can see in (a)-(b) that 'HandCrafted+OVP' and 'HandCrafted' are able to produce reasonable results because the scenes in (a)-(b) contain many structures with line/curve features. However, when the scenes become more cluttered or structure-less as in (c)-(d), such traditional methods cannot work well, especially, the hand-crafted method [32] (i.e., 'HandCrafted') produces significantly distorted result in (d). In contrast, 'DeepSup' and 'MultiTask' achieve good performance, even in the challenging cases of cluttered or structure-less scenes in (c)-(d). Compared to 'MultiTask', 'DeepSup' produces results that are visually closer to ground truth. For instance, the line is more straight in (a) and the perspective effect is kept better in (c) for 'DeepSup'.

YouTube: Due to lack of ground truth, we only show a few qualitative results for YouTube test sequences in Fig. 8. From the results, 'DeepSup' is able to perform relatively well for all 3 sequences. It is worth noting that the cameras in Sequences 2 and 3 are placed inside the car, thus the hood of the car always appears in the images. Although the hood of the car is not in our synthetic training data in Sec. IV-B, our network is able to deal with it and extract useful information from other regions. 'MultiTask' produces similar results, which are not shown in this figure. In addition, we observe that although 'HandCrafted+OVP' works reasonably for Sequence 1, which mainly consists of well-structured scenes, it fails to calibrate Sequences 2 and 3, where the scenes are more cluttered or structure-less. Similarly, 'HandCrafted' performs reasonably for Sequences 1 and 3, but fails for Sequence 2.

3770

Fig. 8. Qualitative results of single-view camera self-calibration on YouTube test sequences. (a)-(b) Sequence 1. (c)-(d) Sequence 2. (e)-(f) Sequence 3.
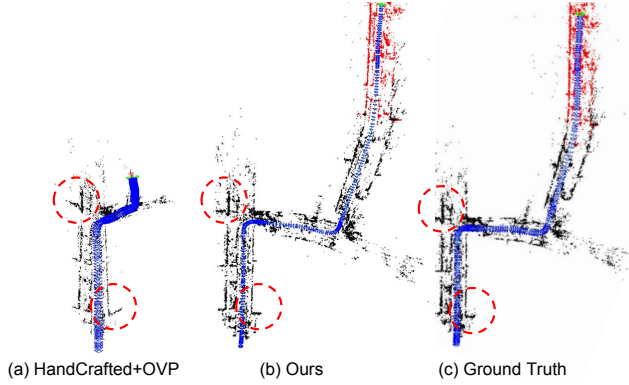

Fig. 9. Qualitative results of uncalibrated SLAM on the first 400 frames of the KITTI Raw test sequence. Red points denote the latest local map.
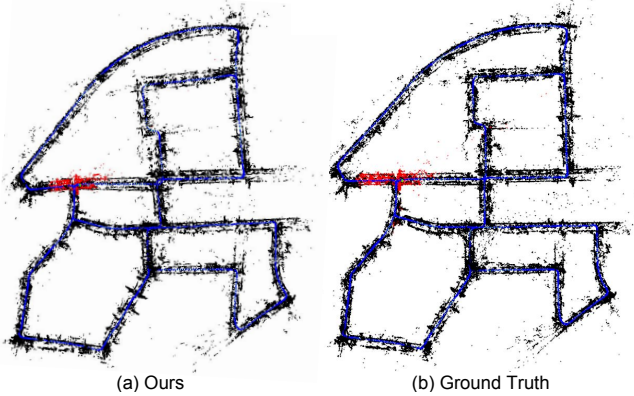

Fig. 10. Qualitative results of uncalibrated SLAM on the entire KITTI Raw test sequence. Red points denote the latest local map.

Methods	RMSE (m)
HandCrafted+OVP	X
Ours	8.04
Ground Truth	7.29

TABLE I

QUANTITATIVE COMPARISONS IN TERMS OF MEDIAN RMSE OF KEYFRAME TRAJECTORY OVER 5 RUNS ON THE ENTIRE KITTI RAW TEST SEQUENCE. X DENOTES THAT 'HANDCRAFTED+OVP' BREAKS ORB-SLAM IN ALL 5 RUNS.

C. Uncalibrated SLAM

We now demonstrate the application of our single-view camera self-calibration method to SLAM with uncalibrated videos, where calibration with a checkerboard pattern is not available or prohibited. For this purpose, we introduce a two-step approach. We first apply our method on the first 100 frames of the test sequence and take the median values of the per-frame outputs as the estimated parameters, which are then used to calibrate the entire test sequence. Next, we employ state-of-the-art SLAM systems designed for calibrated environments (here we use ORB-SLAM [2]) on the calibrated images. We compare our two-step approach against similar two-step approaches which use 'HandCrafted+OVP' or ground truth parameters (termed as 'Ground Truth' — only available for the KITTI Raw test sequence) for calibration.

KITTI Raw: We first evaluate the performance of the above approaches on a subset of the KITTI Raw test sequence (i.e., the first 400 frames). Fig. 9 presents qualitative results. From the results, although 'HandCrafted+OVP' does not break ORB-SLAM in the initial forward motion stage, the reconstructed scene structure and camera trajectory are distorted evidently (see red circles). Subsequently, ORB-SLAM breaks down when the car turns. In contrast, the scene structure and camera trajectory estimated by our method are much more closer to the 'Ground Truth'. More importantly, our method allows ORB-SLAM to run successfully on the entire KITTI Raw test sequence without significant drifts from the result of 'Ground Truth' — see Fig. 10 for typical reconstructed scene structure and camera trajectory. In addition, we quantitatively evaluate the camera trajectory computed by our method and 'Ground Truth' on the entire KITTI Raw test sequence using
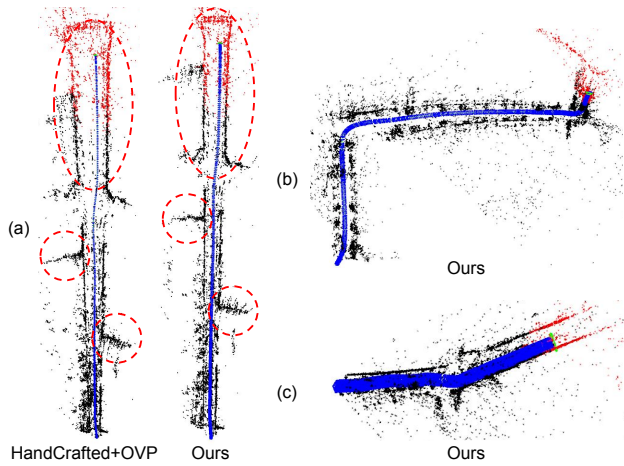
Fig. 11. Qualitative results of uncalibrated SLAM on YouTube test sequences. (a) Sequence 1. (b) Sequence 2. (c) Sequence 3. Red points denote the latest local map.

the public EVO toolbox [2]. Tab. I presents the median root mean square error (RMSE) of the keyframe trajectory over 5 runs by our method and 'Ground Truth'. It can be seen that the error gap between our method and 'Ground Truth' is marginal, while our method does not require calibration with a checkerboard pattern.

YouTube: Fig. 11 illustrates qualitative results of our method and 'HandCrafted+OVP' on YouTube test sequences. 'Ground Truth' is not included since ground truth parameters are not available for these sequences. While our method enables ORB-SLAM to run successfully for all 3 sequences, 'Hand-Crafted+OVP' breaks ORB-SLAM for the last 2 sequences (i.e., ORB-SLAM cannot initialize and hence completely fails). In addition, although ORB-SLAM does not fail in the first sequence for 'HandCrafted+OVP', the reconstructed scene structure is more distorted compared to our method. For instance, the parallelism of the two sides of the road is preserved better by our method (see red ellipses). Please see the supplementary video[3] for more details.

VI. CONCLUSION

In this paper, we revisit the camera self-calibration problem which still remains an important open problem in the computer vision community. We first present a theoretical study of the degeneracy in two-view geometric approach for radial distortion self-calibration. A deep learning solution is then proposed and contributes towards the application of SLAM to uncalibrated videos. Our future work includes gaining a better understanding of the network, e.g., using visualization tools such as Grad-CAM [44]. In addition, we want to work towards a more complete system that allows SLAM to succeed in videos with more challenges, e.g. motion blurs and rolling shutter distortion, as well as explore SfM [45], 3D object localization [46] and top-view mapping [13] in unconstrained scenarios.

[2]https://michaelgrupp.github.io/evo/

[3]https://youtu.be/cfWq9uz2Zac

REFERENCES

[1] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[3] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2018.

[4] N. Yang, R. Wang, X. Gao, and D. Cremers, "Challenges in monocular visual odometry: Photometric calibration, motion bias, and rolling shutter effect," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 2878–2885, 2018.

[5] H. S. Lee, J. Kwon, and K. M. Lee, "Simultaneous localization, mapping and deblurring," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1203–1210.

[6] H. Park and K. Mu Lee, "Joint estimation of camera pose, depth, deblurring, and super-resolution from a blurred image sequence," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4613–4621.

[7] B. Zhuang, L.-F. Cheong, and G. Hee Lee, "Rolling-shutter-aware differential sfm and image rectification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 948–956.

[8] B. Zhuang, Q.-H. Tran, P. Ji, L.-F. Cheong, and M. Chandraker, "Learning structure-and-motion-aware rolling shutter correction," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] Z. Kukelova and T. Pajdla, "A minimal solution to the autocalibration of radial distortion," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–7.

[10] Z. Kukelova, J. Heller, M. Bujnak, A. Fitzgibbon, and T. Pajdla, "Efficient solution to the epipolar geometry for radially distorted cameras," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2309–2317.

[11] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012.

[12] V. Dhiman, Q.-H. Tran, J. J. Corso, and M. Chandraker, "A continuous occlusion model for road scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4331–4339.

[13] Z. Wang, B. Liu, S. Schulter, and M. Chandraker, "A parametric top-view representation of complex road scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[14] C. Steger, "Estimating the fundamental matrix under pure translation and radial distortion," *ISPRS journal of photogrammetry and remote sensing*, vol. 74, pp. 202–217, 2012.

[15] C. Wu, "Critical configurations for radial distortion self-calibration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 25–32.

[16] R. Hartley, "Extraction of focal lengths from the fundamental matrix."

[17] P. Sturm, "On focal length calibration from two views," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 2. IEEE, 2001, pp. II–II.

[18] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015, pp. 562–570.

[19] C. Li, M. Z. Zia, Q.-H. Tran, X. Yu, G. D. Hager, and M. Chandraker, "Deep supervision with shape concepts for occlusion-aware 3d object parsing," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 388–397.

[20] C.-Y. Lee, V. Badrinarayanan, T. Malisiewicz, and A. Rabinovich, "Roomnet: End-to-end room layout estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4865–4874.

[21] C. Li, M. Z. Zia, Q. Tran, X. Yu, G. D. Hager, and M. Chandraker, "Deep supervision with intermediate concepts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

[22] M. E. Fathy, Q.-H. Tran, M. Zeeshan Zia, P. Vernaza, and M. Chandraker, "Hierarchical metric learning and matching for 2d and 3d geometric correspondences," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 803–819.

[23] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[24] A. W. Fitzgibbon, "Simultaneous linear estimation of multiple view geometry and lens distortion," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[25] B. K. Horn, "Motion fields are hardly ever ambiguous," *International Journal of Computer Vision*, vol. 1, no. 3, pp. 259–274, 1988.

[26] M. Byrod, Z. Kukelova, K. Josephson, T. Pajdla, and K. Astrom, "Fast and robust numerical solutions to minimal problems for cameras with radial distortion," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[27] F. Jiang, Y. Kuang, J. E. Solem, and K. Åström, "A minimal solution to relative pose with unknown focal length and radial distortion," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 443–456.

[28] F. Bukhari and M. N. Dailey, "Automatic radial distortion estimation from a single image," *Journal of mathematical imaging and vision*, vol. 45, no. 1, pp. 31–45, 2013.

[29] X. Mei, S. Yang, J. Rong, X. Ying, S. Huang, and H. Zha, "Radial lens distortion correction using cascaded one-parameter division model," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3615–3619.

[30] H. Wildenauer and B. Micusik, "Closed form solution for radial distortion estimation from a single vanishing point."

[31] M. Zhang, J. Yao, M. Xia, K. Li, Y. Zhang, and Y. Liu, "Line-based multi-label energy optimization for fisheye image rectification and calibration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4137–4145.

[32] D. Santana-Cedrés, L. Gomez, M. Alemán-Flores, A. Salgado, J. Esclarín, L. Mazorra, and L. Alvarez, "Automatic correction of perspective and optical distortions," *Computer Vision and Image Understanding*, vol. 161, pp. 1–10, 2017.

[33] M. Antunes, J. P. Barreto, D. Aouada, and B. Ottersten, "Unsupervised vanishing point detection and camera calibration from a single manhattan image with radial distortion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4288–4296.

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

[35] X. Yin, X. Wang, J. Yu, M. Zhang, P. Fua, and D. Tao, "Fisheyerecnet: A multi-context collaborative deep network for fisheye image rectification," *arXiv preprint arXiv:1804.04784*, 2018.

[36] 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 15th ACM SIGGRAPH European Conference on Visual Media Production*. ACM, 2018, p. 6.

[37] J. Wang, F. Shi, J. Zhang, and Y. Liu, "A new calibration model of camera lens distortion," *Pattern Recognition*, vol. 41, no. 2, pp. 607–615, 2008.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[40] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[41] 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 *NIPS-W*, 2017.

[42] M. A. Fischler and R. C. Bolles, "A paradigm for model fitting with applications to image analysis and automated cartography (reprinted in readings in computer vision, ed. ma fischler,"," *Comm. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[43] Q. H. Tran, "Robust parameter estimation in computer vision: geometric fitting and deformable registration." Ph.D. dissertation, 2014.

[44] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[45] B. Zhuang, L.-F. Cheong, and G. Hee Lee, "Baseline desensitizing in translation averaging," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4539–4547.

[46] S. Srivastava, F. Jurie, and G. Sharma, "Learning 2d to 3d lifting for object detection in 3d for autonomous vehicles," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019.