# Rethinking Generic Camera Models for Deep Single Image Camera Calibration to Recover Rotation and Fisheye Distortion

Nobuhiko Wakai[1], Satoshi Sato[1], Yasunori Ishii[1], and Takayoshi Yamashita[2]

[1] Panasonic Holdings {*lastname.firstname*}@jp.panasonic.com
[2] Chubu University takayoshi@isc.chubu.ac.jp

**Abstract.** Although recent learning-based calibration methods can predict extrinsic and intrinsic camera parameters from a single image, the accuracy of these methods is degraded in fisheye images. This degradation is caused by mismatching between the actual projection and expected projection. To address this problem, we propose a generic camera model that has the potential to address various types of distortion. Our generic camera model is utilized for learning-based methods through a closed-form numerical calculation of the camera projection. Simultaneously to recover rotation and fisheye distortion, we propose a learning-based calibration method that uses the camera model. Furthermore, we propose a loss function that alleviates the bias of the magnitude of errors for four extrinsic and intrinsic camera parameters. Extensive experiments demonstrated that our proposed method outperformed conventional methods on two large-scale datasets and images captured by off-the-shelf fisheye cameras. Moreover, we are the first researchers to analyze the performance of learning-based methods using various types of projection for off-the-shelf cameras.

**Keywords:** camera calibration, fisheye camera, rectification

## 1 Introduction

Learning-based perception methods are widely used for surveillance, cars, drones, and robots. These methods are well established for many computer vision tasks. Most computer vision tasks require undistorted images; however, fisheye images have the superiority of a large field of view (FOV) in visual surveillance [18], object detection [51], pose estimation [11], and semantic segmentation [41]. To use fisheye cameras by removing distortion, camera calibration is a desirable step before perception. Camera calibration is a long-studied topic in areas of computer vision, such as image undistortion [37,66], image remapping [60], virtual object insertion [24], augmented reality [3], and stereo measurement [43]. In camera calibration, we cannot escape the trade-off between accuracy and usability that we need a calibration object; hence, tackling the trade-off has been an open challenge, which we explain further in the following.

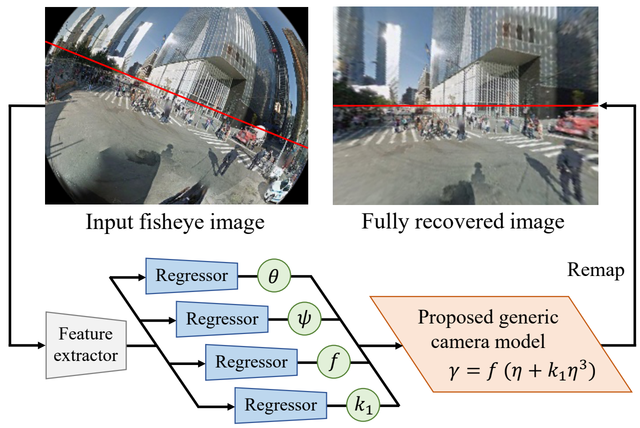



**Fig. 1:** Concept illustrations of our work. Our network predicts parameters in our proposed generic camera model to obtain fully recovered images using remapping. Red lines indicate horizontal lines in each of the images, for which we used [46].

Calibration methods are classified into two categories: geometric-based and learning-based methods. Geometric-based calibration methods achieve high accuracy but require a calibration object, such as a cube [59] or a plane [68], to obtain a strong geometric constraint. By contrast, learning-based methods can calibrate cameras without a calibration object from a general scene image [8,13,34,37,45,60,66], which is called deep single image camera calibration. Although learning-based methods do not require a calibration object, the accuracy of these methods is degraded for fisheye images because of the mismatch between the actual projection and expected projection in conventional methods. In particular, calibration methods [45,60] that predict both camera rotation and distortion have much room for improvement regarding addressing complex fisheye distortion. López-Antequera's method [45] was designed for non-fisheye cameras with radial distortion and cannot process fisheye distortion. Although four standard camera models are used for fisheye cameras, Wakai's method [60] supports only one fisheye camera model.

Based on the observations above, we propose a new generic camera model for various fisheye cameras. The proposed generic camera model has the potential to address various types of distortion. For the generic camera model, we propose a learning-based calibration method that predicts extrinsic parameters (tilt and roll angles), focal length, and a distortion coefficient simultaneously from a single image, as shown in Figure 1. Our camera model is utilized for learning-based methods through a closed-form numerical calculation of camera projection. To improve the prediction accuracy, we use a joint loss function composed of each loss for the four camera parameters. Unlike heuristic approaches in conventional methods, our loss function makes considerable progress; that is, we can determine the optimal joint weights based on the magnitude of errors for these camera parameters instead of the heuristic approaches.

To evaluate the proposed method, we conducted extensive experiments on two large-scale datasets [12,46] and images captured by off-the-shelf fisheye cameras. This evaluation demonstrated that our method meaningfully outperformed nine conventional geometric-based [2,55] and learning-based [8,13,34,37,45,60,66] methods. The major contributions of our study are summarized as follows:

- We propose a learning-based calibration method for recovering camera rotation and fisheye distortion using the proposed generic camera model that has an adaptive ability for off-the-shelf fisheye cameras. To the best of our knowledge, we are the first researchers to calibrate extrinsic and intrinsic parameters of generic camera models addressing various types of projection in off-the-shelf fisheye cameras from a single image.
- We propose a new loss function that alleviates the bias of the magnitude of errors between the ground-truth and predicted camera parameters for four extrinsic and intrinsic parameters to obtain accurate camera parameters.
- We first analyze the performance of learning-based methods using off-the-shelf fisheye cameras consisting of four types of fisheye projection: stereographic projection, equidistance projection, equisolid angle projection, and orthogonal projection.

## 2 Related work

**Camera calibration:** Camera calibration estimates parameters composed of extrinsic parameters (rotation and translation) and intrinsic parameters (image sensor and distortion parameters). Geometric-based calibration methods have been developed using a strong constraint based on a calibration object [59,68], line detection [2,7,10,19,55,67], or vanishing points [44,49]. This constraint explicitly represents the relation between world coordinates and image coordinates to achieve stable calibration optimization. By contrast, learning-based methods using convolutional neural networks calibrate cameras from a single image in the wild. In this study, we focus on learning-based calibration methods and describe them below.

Calibration methods for only extrinsic parameters have been proposed that are aimed at narrow view cameras [27,48,56,57,62,63] and panoramic 360° images [16]. These methods cannot calibrate intrinsic parameters; that is, they cannot remove distortion. For extrinsic parameters and focal length, narrow-view camera calibration was developed with depth estimation [14,22] and room layout [52]. These methods are not suitable for fisheye cameras with over 180° FOV because fisheye distortion is not negligible.

Calibration methods for only undistortion have been proposed using regressors or generators. These regressors predicted camera parameters of polynomial models [35], division distortion models [25,39,53], unified spherical models [8], or generic camera models using fisheye projection [36,64,66]. These camera models have room for improvement in learning-based methods because the models were originally designed for geometric-based calibration methods. By contrast, the generators predicted undistorted images using multi-scale information [65] or discriminators [13,38] in generative adversarial networks (GAN) [20]. Only undistortion methods cannot recover camera rotation.

To calibrate both extrinsic and intrinsic parameters, López-Antequera *et al*. [45] proposed a pioneering method for non-fisheye cameras. This method estimated distortion using a polynomial function model of perspective projection similar to Brown's quartic polynomial model [9]. This polynomial function against the distance from a principal point has two coefficients for the second- and fourth-order terms. The method is only trainable for the second-order coefficient, and the fourth-order coefficient is calculated using a quadratic function of the second-order one. This method does not calibrate fisheye cameras effectively because the camera model does not represent fisheye camera projection. Additionally, Wakai *et al*. [60] proposed a calibration method for extrinsic parameters and focal length in fisheye cameras. Although four types of standard fisheye projection are used for camera models, for example, equisolid angle projection, Wakai's method [60] only expects equisolid angle projection. Li *et al*. [34] proposed image transformation for rotation and distortion. Images with rotation and distortion degrade accuracy because this method needs to employ rotation and distortion transformation separately. As discussed above, conventional learning-based calibration methods do not fully calibrate extrinsic and intrinsic parameters of generic camera models from a single image.

**Exploring loss landscapes:** To optimize networks effectively, loss landscapes have been explored after training [15,21,33] and during training [23]. In learning-based calibration methods, we have the problem that joint weights are difficult to determine before training. The joint loss function was often defined to stabilize training or to merge heterogeneous loss components [37,45,60,66]. However, these joint weights were defined using experiments or the same values, that is, unweighted joints. These joint weights are hyperparameters that depend on networks and datasets. A hyperparameter search method was proposed by Akiba *et al*. [1]. However, hyperparameter search tools require high computational costs because they execute various conditions. Additionally, to analyze optimizers, Goodfellow *et al*. [21] proposed an examination method for loss landscapes using linear interpolation from the initial network weights to the final weights. To overcome the saddle points of loss landscapes, Dauphin *et al*. [15] proposed an optimization method based on Newton's method. Furthermore, Li *et al*. [33] developed an approach for visualizing loss landscapes. Although these methods can explore high-order loss landscapes, the optimal values of joint loss weights have not been determined in learning-based calibration methods. Moreover, the aforementioned methods cannot explore loss landscapes before training because they require training results.

## 3 Proposed method

First, we describe our proposed camera model based on a closed-form solution for various fisheye cameras. Second, we explain our learning-based calibration method for recovering rotation and fisheye distortion. Finally, we introduce a new loss function, with its notation and mechanism.

### 3.1 Generic camera model

Camera models are composed of extrinsic parameters $[\ \mathbf{R} \mid \mathbf{t}\ ]$ and intrinsic parameters, and these camera models represent the mapping from world coordinates $\tilde{\mathbf{p}}$ to image coordinates $\tilde{\mathbf{u}}$ in homogeneous coordinates. This projection can be expressed for radial distortion of perspective projection [6,9,17,50] and fisheye projection [5,30,58] as

$$\tilde{\mathbf{u}} = \begin{bmatrix} \gamma/d_u & 0 & c_u \\ 0 & \gamma/d_v & c_v \\ 0 & 0 & 1 \end{bmatrix} [\ \mathbf{R} \mid \mathbf{t}\ ]\,\tilde{\mathbf{p}}, \tag{1}$$

where $\gamma$ is distortion, $(d_u, d_v)$ is an image sensor pitch, $(c_u, c_v)$ is a principal point, $\mathbf{R}$ is a rotation matrix, and $\mathbf{t}$ is a translation vector. The subscripts of $u$ and $v$ denote the horizontal and vertical direction, respectively.

The generic camera model including fisheye lenses [30] is defined as

$$\gamma = \tilde{k_1}\eta + \tilde{k_2}\eta^3 + \cdots\ , \tag{2}$$

where $\tilde{k_1}$, $\tilde{k_2}, \ldots$ are distortion coefficients and $\eta$ is an incident angle. Note that the focal length is not defined explicitly; that is, the focal length is set to 1 mm, and the distortion coefficients represent distortion and implicit focal length.

### 3.2 Proposed camera model

Many off-the-shelf fisheye cameras have 180° FOV and more. To calibrate them, camera models need to support over 180° FOV; that is, the models require fisheye projection defined by a distortion function against $\eta$. In accordance with the FOV range, we select fisheye projection rather than perspective projection for our camera model. A generic camera model with high order has the potential to achieve high calibration accuracy. However, this high-order function leads to unstable optimization, particularly for learning-based methods. Considering this problem, we propose a generic camera model for learning-based fisheye calibration using explicit focal length, given by

$$\gamma = f(\eta + k_1\eta^3), \tag{3}$$

where $f$ is the focal length and $k_1$ is a distortion coefficient.

**Evaluating our camera model:** Our generic camera model is a third-order polynomial function corresponding to the Taylor series expansion of the trigonometric function in fisheye cameras, that is, stereographic projection, equidistance projection, equisolid angle projection, and orthogonal projection. In the following, we show that our model can express trigonometric function models with slight errors.

**Table 1:** Comparison of absolute errors in fisheye camera models

| Reference model[1] | Mean absolute error [pixel] STG | EQD | ESA | ORT |
|---|---|---|---|---|
| Stereographic (STG) | – | 9.33 | 13.12 | 93.75 |
| Equidistance (EQD) | 9.33 | – | 3.79 | 23.58 |
| Equisolid angle (ESA) | 13.12 | 3.79 | – | 14.25 |
| Orthogonal (ORT) | 93.75 | 23.58 | 14.25 | – |
| Unified spherical model [5] | 0.71 | 0.19 | 0.00 | 0.51 |
| Proposed generic model | **0.54** | **0.00** | **0.02** | **0.35** |

[1] Each reference model is compared with other fisheye models

To evaluate camera models using fisheye projection and a few parameters, we compared the projection function, $\gamma = g(\eta)$, of the four trigonometric function models, the unified spherical model [5], and our generic camera model, as shown in Table 1. All the camera models have two or fewer intrinsic camera parameters for fisheye projection. In this comparison, we calculated the mean absolute errors $\epsilon$ between pairs of the projection function $g_1$ and $g_2$. We defined the errors as $\epsilon = 1/(\pi/2) \int_0^{\pi/2} \mid g_1(\eta) - g_2(\eta) \mid d\eta$. These mean absolute errors simply represent mean distance errors in image coordinates. Overall, our model is useful for various fisheye models because our model had small mean absolute errors, as shown in Table 1.

**Calculation easiness:** For our generic camera model, it is easy to calculate back-projection, which converts image coordinates to corresponding incident angles. When using back-projection, we must solve the generic camera model against incident angles $\eta$ in Equation (3). Practically, we can solve equations on the basis of iterations or closed forms. Non-fisheye cameras often use the iteration approaches [4]. By contrast, we cannot use the iteration approaches for fisheye cameras because large distortion prevents us from obtaining the initial values close to solutions. We, therefore, use a closed-form approach because the Abel-Ruffini theorem [70] shows that fourth-order or less algebraic equations are solvable. Refer to the supplementary material for the details of our model.

### 3.3 Proposed calibration method

To calibrate various fisheye cameras, we propose a learning-based calibration method that uses our generic camera model. We use DenseNet-161 [26] pre-trained on ImageNet [54] to extract image features and details as follows: First, we convert the image features using global average pooling [40] for regressors. Second, four individual regressors predict the normalized parameters (from 0 to 1) of a tilt angle $\theta$, a roll angle $\psi$, focal length $f$, and a distortion coefficient $k_1$. Each regressor consists of a 2208-channel fully connected (FC) layer with Mish activation [47] and a 256-channel FC layer with sigmoid activation. Batch normalization [29] uses these FC layers. Finally, we predict a camera model by recovering the ranges of the normalized camera parameters to their original ranges. Following conventional studies [24,45,60], we scale the input images to $224 \times 224$ pixels.

### 3.4 Harmonic non-grid bearing loss

Unlike a loss function based on image reconstruction, Wakai *et al.* proposed the non-grid bearing loss function $L$ [60] based on projecting image coordinates to world coordinates as

$$L_\alpha = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Huber}(||\mathbf{p}_{\alpha i} - \hat{\mathbf{p}}_i||_2), \tag{4}$$

where $n$ is the number of sampling points; $\alpha$ is a parameter, $\alpha = \{\theta, \psi, f, k_1\}$; $\mathbf{p}_\alpha$ is a projected world coordinate using a predicted parameter $\alpha$ and ground-truth values for the remaining parameters; and $\hat{\mathbf{p}}$ is the ground-truth value of world coordinates $\mathbf{p}_\alpha$. The Huber $(\bullet)$ denotes the Huber loss function with $\delta = 1$ [28]. The loss function $L_\theta$ uses a predicted $\theta$ and ground-truth parameters for $\psi$, $f$, and $k_1$. Additionally, $L_\psi$, $L_f$, and $L_{k_1}$ are determined in the same manner. We obtain the world coordinates $\mathbf{p}_\alpha$ from the image coordinates in sampled points. The sampled points are projected from a unit sphere. For sampling on the unit sphere, we use uniform distribution within valid incident angles that depend on $k_1$. The loss function achieved stable optimization using the unit sphere. The joint loss is defined as

$$L = w_\theta L_\theta + w_\psi L_\psi + w_f L_f + w_{k_1} L_{k_1}, \tag{5}$$

where $w_\theta$, $w_\psi$, $w_f$, and $w_{k_1}$ are the joint weights of $\theta$, $\psi$, $f$, and $k_1$, respectively. Although this loss function can effectively train networks, we need to determine the joint weights for each camera parameter. Wakai *et al.* [60] and López-Antequera *et al.* [45] used joint weights set to the same values. To determine the optimal joint weights, they needed to repeat training and validation. However, they did not search for the optimal joint weights because of high computational costs.

To address this problem, we surprisingly found that numerical simulations instead of training can analyze loss landscapes. This loss function can be divided into two steps: predicting camera parameters from an image and projecting sampled points using camera parameters. The latter step requires only the sampled points and camera parameters. Therefore, we focused on the latter step independent of input images. Figure 2 (a) shows the loss landscapes for camera parameters along normalized camera parameters. The landscapes express that the magnitude of loss values of the focal length is the smallest among $\theta$, $\psi$, $f$, and $k_1$, and the focal length is relatively hard to train. Our investigation suggests that the optimal joint loss weights $w$ are estimated as follows: We calculate areas $S$ under the loss function $L$ for $\theta$, $\psi$, $f$, and $k_1$. Assuming practical conditions, we set the ground-truth values to 0.5, which means that the center of the normalized parameter ranges from 0 to 1 in Figure 2 (a). This area $S$ is calculated using the integral of $L$ from 0 to 1, as illustrated in Figure 2 (b) and is given by

$$S_\alpha = \int_0^1 L_\alpha d\alpha = \int_0^1 \frac{1}{n} \sum_{i=1}^{n} \mathrm{Huber}(||\mathbf{p}_{\alpha i} - \hat{\mathbf{p}}_i||_2)\ d\alpha, \tag{6}$$

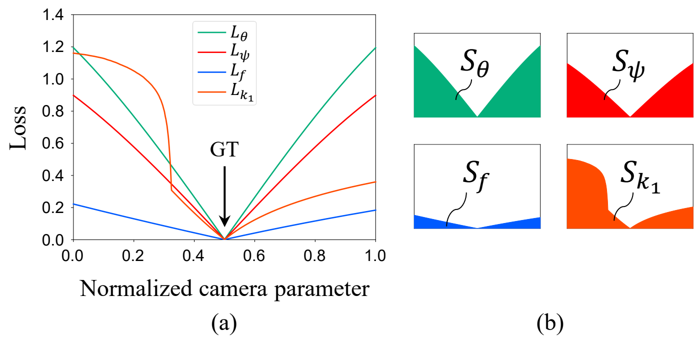



**Fig. 2:** Difference between the non-grid bearing loss functions [60] for the camera parameters. (a) Each loss landscape along the normalized camera parameters using a predicted camera parameter with a subscript parameter and ground-truth parameters for the remaining parameters, and the ground-truth values are set to 0.5. (b) Areas $S$ calculated using the integral of $L$ with respect to $\theta$, $\psi$, $f$, and $k_1$ from 0 to 1.

These areas $S$ represent the magnitude of each loss for $\theta$, $\psi$, $f$, and $k_1$. Therefore, we define the joint weights $w$ in Equation (5) using normalization as follows:

$$w_\alpha = \tilde{w}_\alpha \ / \ W, \tag{7}$$

where $\tilde{w}_\alpha = 1/S_\alpha$ and $W = \sum_\alpha \tilde{w}_\alpha$. We call a loss function using the weights in Equation (7) "harmonic non-grid bearing loss (HNGBL)." As stated above, our joint weights can alleviate the bias of the magnitude of the loss for camera parameters. Remarkably, we determine these weights before training.

## 4 Experiments

To validate the adaptiveness of our method to various types of fisheye cameras, we conducted massive experiments using large-scale synthetic images and off-the-shelf fisheye cameras.

### 4.1 Datasets

We used two large-scale datasets of outdoor panoramas called the StreetLearn dataset (Manhattan 2019 subset) [46] and the SP360 dataset [12]. First, we divided each dataset into train and test sets following in [60]: $55,599$ train and 161 test images for StreetLearn, and $19,038$ train and 55 test images for SP360. Second, we generated image patches, with 224-pixel image height ($H_{img}$) and image width ($W_{img} = H_{img} \cdot A$), where $A$ is the image aspect ratio, from panorama images: $555,990$ train and $16,100$ test image patches for StreetLearn, and $571,140$ train and $16,500$ test image patches for SP360. Table 2 shows the random distribution of the train set when we generated image patches using camera models with the maximum incident angle $\eta_{\max}$. The test set was generated using the

**Table 2:** Distribution of the camera parameters for our train set

| Parameters | Distribution | Range or values[1] |
|---|---|---|
| Pan $\phi$ | Uniform | $[0, 360)$ |
| Tilt $\theta$ | Mix<br>Normal<br>Uniform | Normal 70%, Uniform 30%<br>$\mu = 0, \sigma = 15$<br>$[-90, 90]$ |
| Roll $\psi$ | Mix<br>Normal<br>Uniform | Normal 70%, Uniform 30%<br>$\mu = 0, \sigma = 15$<br>$[-90, 90]$ |
| Aspect ratio | Varying | {1/1 9%, 5/4 1%, 4/3 66%, 3/2 20%, 16/9 4%} |
| Focal length $f$ | Uniform | $[6, 15]$ |
| Distortion $k_1$ | Uniform | $[-1/6, 1/3]$ |
| Max angle $\eta_{\max}$ | Uniform | $[84, 96]$ |

[1] Units: $\phi$, $\theta$, $\psi$, and $\eta_{\max}$ [deg]; $f$ [mm]; $k_1$ [dimensionless]

**Table 3:** Off-the-shelf fisheye cameras with experimental IDs

| ID | Camera body | Camera lens |
|---|---|---|
| 1 | Canon EOS 6D | Canon EF8-15mm F4L Fisheye USM |
| 2 | Canon EOS 6D | Canon EF15mm F2.8 Fisheye |
| 3 | Panasonic LUMIX GM1 | Panasonic LUMIX G FISHEYE 8mm F3.5 |
| 4 | FLIR BFLY-U3-23S6C | FIT FI-40 |
| 5 | FLIR FL3-U3-88S2 | FUJIFILM FE185C057HA-1 |
| 6 | KanDao QooCam8K | Built-in |

uniform distribution instead of the mixed and varying distribution applied to the train set. During the generation step, we set the minimum image circle diameter to the image height, assuming practical conditions. Note that each generated camera parameter means the ground-truth parameter. Refer to the supplementary material for the details of the camera parameter ranges.

### 4.2 Off-the-shelf fisheye cameras

We evaluated off-the-shelf fisheye cameras because fisheye cameras have complex lens distortion, unlike narrow-view cameras. Table 3 shows various fisheye cameras that we used for evaluation. Note that we only used the front camera in the QooCam8K camera, which has front and rear cameras. Using the off-the-shelf cameras, we captured outdoor fisheye images in Kyoto, Japan.

### 4.3 Parameter and network settings

To simplify the camera model, we fixed $d_u = d_v$ and the principal point $(c_u, c_v)$ as the image center following in [45,60]. Because the scale factor depends on the focal length and the image sensor size, which is arbitrary for undistortion, we assumed that the image sensor height was 24 mm, which corresponds to a full-size image sensor. We ignored the arbitrary translation vector $\mathbf{t}$. Because the origin of the pan angle is arbitrary, we provided the pan angle for training and evaluation. Therefore, we focused on four trainable parameters, that is, a tilt angle $\theta$, a roll angle $\psi$, focal length $f$, and a distortion coefficient $k_1$, in our method. Note that we considered camera rotation based on the horizontal line, unlike calibration methods [31,32] under the Manhattan world assumption.

We optimized our network for a 32 mini-batch size using a rectified Adam optimizer [42], whose weight decay was 0.01. We set the initial learning rate to $1 \times 10^{-4}$ and multiplied the learning rate by 0.1 at the 50th epoch. Additionally, we set the joint weights in Equation (5) using $w_\theta = 0.103$, $w_\psi = 0.135$, $w_f = 0.626$, and $w_{k_1} = 0.136$.

**Table 4:** Feature summarization of the conventional methods and our method

| Method | DL[1] | Rot[1] | Dist[1] | >180° FOV[1] | Projection | Network |
|---|---|---|---|---|---|---|
| Alemán-Flores [2] | | | ✓ | | Perspective | – |
| Santana-Cedrés [55] | | | ✓ | | Perspective | – |
| Liao [37] | ✓ | | ✓ | | Perspective | Regressor |
| Yin [66] | ✓ | | ✓ | ✓ | Generic camera [30] | Regressor |
| Chao [13] | ✓ | | ✓ | – | – | Generator (GAN) |
| Bogdan [8] | ✓ | | ✓ | ✓ | Unified spherical model [5] | Regressor |
| Li (GeoNetS-$\mathcal{B}$) [34] | ✓ | | ✓ | – | – | Generator |
| López-Antequera [45] | ✓ | ✓ | ✓ | | Perspective | Regressor |
| Wakai [60] | ✓ | ✓ | ✓ | ✓ | Equisolid angle | Regressor |
| Ours | ✓ | ✓ | ✓ | ✓ | Proposed generic camera | Regressor |

[1] DL denotes learning-based method; Rot denotes rotation; Dist denotes distortion; ">180° FOV" denotes supporting over 180° FOV

**Table 5:** Comparison of the absolute parameter errors and reprojection errors on the test set for our generic camera model

| Method | StreetLearn | | | | | SP360 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean absolute error ↓ | | | | REPE ↓ | Mean absolute error ↓ | | | | REPE ↓ |
| | Tilt $\theta$ [deg] | Roll $\psi$ [deg] | $f$ [mm] | $k_1$ | [pixel] | Tilt $\theta$ [mm] | Roll $\psi$ [deg] | $f$ [mm] | $k_1$ | [pixel] |
| López-Antequera [45] | 27.60 | 44.90 | 2.32 | – | 81.99 | 28.66 | 44.45 | 3.26 | – | 84.56 |
| Wakai [60] | 10.70 | 14.97 | 2.73 | – | 30.02 | 11.12 | 17.70 | 2.67 | – | 32.01 |
| Ours w/o HNGBL[1] | 7.23 | 7.73 | 0.48 | 0.025 | 12.65 | 6.91 | 8.61 | 0.49 | 0.030 | 12.57 |
| Ours | **4.13** | **5.21** | **0.34** | **0.021** | **7.39** | **3.75** | **5.19** | **0.39** | **0.023** | **7.39** |

[1] "Ours w/o HNGBL" refers to replacing HNGBL with non-grid bearing loss [60]

### 4.4 Experimental results

In Table 4, we summarize the features of the conventional methods. We implemented the methods according to the corresponding papers, except that StreetLearn [46] and SP360 [12] were used for training. Note that we trained Yin's method [66] using ADE20K [69] following Yin's implementation because the method requires semantic segmentation data. In Li's methods [34], we selected GeoNetS-$\mathcal{B}$, which is the single-model distortion network to remove barrel distortion. For Alemán-Flores's [2] and Santana-Cedrés's [55] methods, we excluded test images with few lines because they require many lines for calibration.

**Parameter and reprojection errors.** To validate the accuracy of the predicted camera parameters, we compared methods that can predict rotation and distortion parameters. We evaluated the mean absolute errors of the camera parameters and the mean reprojection errors (REPE) on the test set for our generic camera model. We did not compare the focal length in Bogdan's method [8] because the unified spherical model [5] has ambiguity between the focal length and the distortion parameter [8]. Table 5 shows that our method achieved the lowest mean absolute errors and REPE among all methods. This REPE reflected the errors of both extrinsic and intrinsic parameters. To calculate the REPE, we generated $32,400$ uniform world coordinates on a unit sphere within less than $90°$ incident angles because of the lack of calibration points for image-based calibration methods. López-Antequera's method [45] did not seem to work well because it expects non-fisheye input images. Our method substantially reduced focal length errors and camera rotation errors (tilt and roll angles) by 86%

**Table 6:** Comparison of mean PSNR and SSIM on the test set for our generic camera model

| Method | StreetLearn | | | | | | SP360 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | | | SSIM ↑ | | | PSNR ↑ | | | SSIM ↑ | | |
| | Diag[1] | Circ[1] | All | Diag | Circ | All | Diag | Circ | All | Diag | Circ | All |
| Alemán-Flores [2] | 14.79 | 11.70 | 13.25 | 0.354 | 0.271 | 0.313 | 14.57 | 11.03 | 12.82 | 0.408 | 0.311 | 0.360 |
| Santana-Cedrés [55] | 16.27 | 13.17 | 14.65 | 0.384 | 0.306 | 0.341 | 16.06 | 12.38 | 14.26 | 0.438 | 0.343 | 0.390 |
| Liao [37] | 13.92 | 13.48 | 13.71 | 0.355 | 0.369 | 0.362 | 14.08 | 13.61 | 13.85 | 0.401 | 0.408 | 0.404 |
| Yin [66] | 14.24 | 13.57 | 13.91 | 0.344 | 0.354 | 0.349 | 14.37 | 13.68 | 14.03 | 0.389 | 0.391 | 0.390 |
| Chao [13] | 17.36 | 14.89 | 16.13 | 0.439 | 0.378 | 0.409 | 17.23 | 14.86 | 15.88 | 0.480 | 0.417 | 0.449 |
| Bogdan [8] | 14.81 | 14.32 | 14.57 | 0.360 | 0.353 | 0.356 | 17.82 | 16.20 | 17.02 | 0.517 | 0.459 | 0.488 |
| Li (GeoNetS-$\mathcal{B}$) [34] | 18.77 | 15.15 | 16.98 | 0.529 | 0.410 | 0.470 | 18.76 | 15.13 | 16.97 | 0.572 | 0.452 | 0.513 |
| López-Antequera [45] | 19.17 | 16.58 | 17.88 | 0.547 | 0.449 | 0.499 | 17.72 | 14.73 | 16.24 | 0.542 | 0.429 | 0.486 |
| Wakai [60] | 21.12 | 22.04 | 21.57 | 0.604 | 0.640 | 0.622 | 21.03 | 20.93 | 20.98 | 0.640 | 0.637 | 0.639 |
| Ours w/o HNGBL[2] | 27.12 | 27.70 | 27.41 | 0.801 | 0.801 | 0.801 | 25.93 | 27.07 | 26.49 | 0.790 | 0.812 | 0.801 |
| Ours | **28.39** | **29.63** | **29.01** | **0.828** | **0.847** | **0.838** | **27.19** | **29.03** | **28.10** | **0.819** | **0.852** | **0.835** |

[1] Diag denotes evaluation using only diagonal fisheye images; Circ denotes evaluation using only circumferential fisheye images
[2] "Ours w/o HNGBL" refers to replacing HNGBL with non-grid bearing loss [60]

and 66%, respectively, on average for the two datasets compared with Wakai's method [60]. Furthermore, our method reduced the REPE by 76% on average for the two datasets compared with Wakai's method [60]. Therefore, our method predicted accurate extrinsic and intrinsic camera parameters.

We also evaluated our method, referred to as "Ours w/o HNGBL," replacing our loss function with non-grid bearing loss [60] to analyze the performance of our loss function, as shown in Table 5. This result demonstrates that our loss function effectively reduced the rotation errors in the tilt and roll angles by 3.05° on average for the two datasets compared with the "Ours w/o HNGBL" case. In addition to rotation errors, the REPE for our method with HNGBL was 5.22 pixels on average for the two datasets smaller than that for "Ours w/o HNGBL." These results suggest that our loss function enabled networks to accurately predict not only focal length but also other camera parameters.

**Comparison using PSNR and SSIM.** To demonstrate validity and effectiveness in images, we used the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) [61] for intrinsic parameters. When performing undistortion, extrinsic camera parameters are arbitrary because we consider only intrinsic camera parameters, image coordinates, and incident angles. Table 6 shows the performance of undistortion on the test set for our generic camera model. We verified that circumferential fisheye images did not degrade the accuracy of our method because our camera model supports over 180° FOV. By contrast, the circumferential fisheye images degraded the performance of methods using perspective projection in Alemán-Flores [2], Santana-Cedrés [55], and López-Antequera [45]. Note that the comparison of camera projection is shown in Table 4. Our method notably improved the image quality of undistortion by 7.28 for the PSNR and 0.206 for the SSIM on average for the two datasets compared with Wakai's method [60], as shown in Table 6 (*All*). Therefore, our

**Table 7:** Comparison of mean PSNR on the test set for the trigonometric function models

| Method | StreetLearn | | | | | SP360 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Stereo-graphic | Equi-distance | Equisolid angle | Ortho-gonal | All | Stereo-graphic | Equi-distance | Equisolid angle | Ortho-gonal | All |
| Alemán-Flores [2] | 13.23 | 12.25 | 11.70 | 9.72 | 11.72 | 12.89 | 11.69 | 10.99 | 8.53 | 11.03 |
| Santana-Cedrés [55] | 14.68 | 13.20 | 12.49 | 10.29 | 12.66 | 14.25 | 12.57 | 11.77 | 9.34 | 11.98 |
| Liao [37] | 13.63 | 13.53 | 13.52 | 13.74 | 13.60 | 13.76 | 13.66 | 13.67 | 13.92 | 13.75 |
| Yin [66] | 13.81 | 13.62 | 13.59 | 13.77 | 13.70 | 13.92 | 13.74 | 13.72 | 13.94 | 13.83 |
| Chao [13] | 15.86 | 15.12 | 14.87 | 14.52 | 15.09 | 15.60 | 15.02 | 14.83 | 14.69 | 15.03 |
| Bogdan [8] | 14.55 | 14.43 | 14.46 | 14.71 | 14.54 | 16.92 | 16.34 | 16.14 | 15.65 | 16.26 |
| Li (GeoNetS-$\mathcal{B}$) [34] | 16.37 | 15.41 | 15.07 | 14.58 | 15.36 | 16.22 | 15.33 | 15.04 | 14.72 | 15.33 |
| López-Antequera [45] | 17.84 | 16.84 | 16.43 | 15.15 | 16.57 | 15.72 | 14.94 | 14.68 | 14.52 | 14.97 |
| Wakai [60] | 22.39 | 23.62 | 22.91 | 17.79 | 21.68 | 22.29 | 22.65 | 21.79 | 17.54 | 21.07 |
| Ours w/o HNGBL[1] | 26.49 | 29.08 | 28.56 | **23.97** | 27.02 | 25.35 | 28.53 | 28.26 | 23.85 | 26.50 |
| Ours | **26.84** | **30.10** | **29.69** | 23.70 | **27.58** | **25.74** | **29.28** | **28.95** | **23.93** | **26.98** |

[1] "Ours w/o HNGBL" refers to replacing HNGBL with non-grid bearing loss [60]

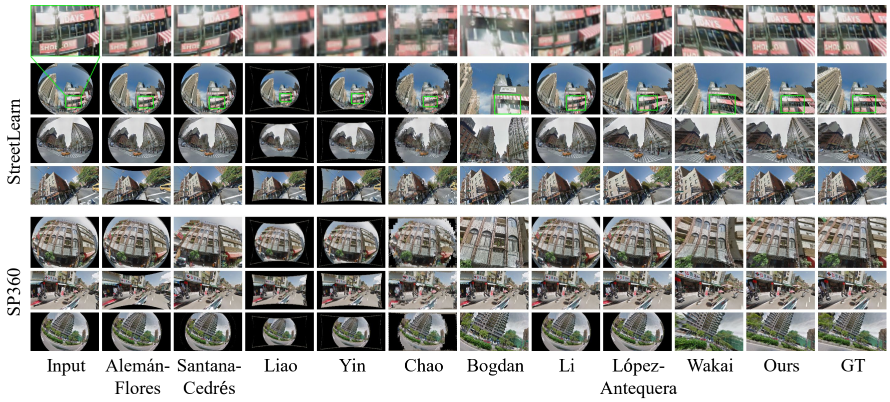
method outperformed conventional methods on both diagonal and circumferential fisheye images.

To validate the dependency of the four types of fisheye camera models, we also evaluated the performance on the trigonometric function models in Table 7. Although orthogonal projection decreased the PSNR, our method addressed all the trigonometric function models; hence, our method had the highest PSNR in all cases. This suggests that our generic camera model precisely behaved like a trigonometric function model. Therefore, our method has the potential to calibrate images from various fisheye cameras.

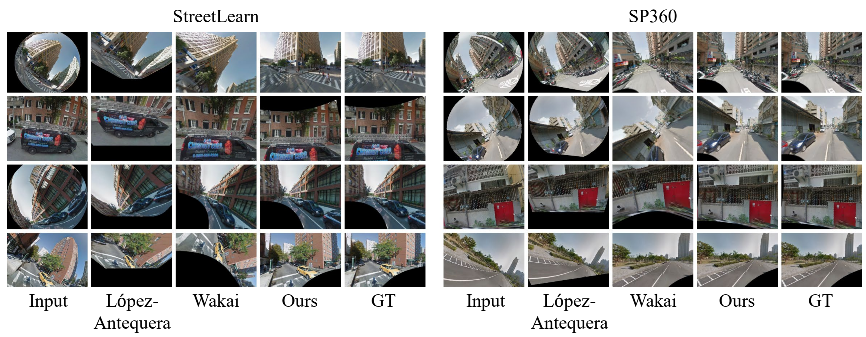
**Qualitative evaluation.** We evaluated the performance of undistortion and full recovery for not only synthetic images but also off-the-shelf cameras to describe the image quality after calibration.

**Synthetic images:** Figure 3 shows the qualitative results on the test set for our generic camera model. Our results are the most similar to the ground-truth images in terms of undistortion and fully recovering rotation and fisheye distortion. Our method worked well for various types of distortion and scaling. By contrast, it was difficult to calibrate circumferential fisheye images with large distortion using Alemán-Flores's [2], Santana-Cedrés's [55], Liao's [37], Yin's [66], Chao's [13], and Li's (GeoNetS-$\mathcal{B}$) [34] method. Furthermore, López-Antequera's [45] and Wakai's [60] methods did not remove distortion, although the scale was close to the ground truth. When fully recovering rotation and distortion, López-Antequera's [45] and Wakai's [60] methods tended to predict camera rotation with large errors in the tilt and roll angles. As shown in Figure 3, our synthetic images consisted of zoom-in images of parts of buildings and zoom-out images of skyscrapers. Our method processed both types of images; that is, it demonstrated scale robustness.

**Off-the-shelf cameras:** We also validated calibration methods using off-the-shelf fisheye cameras to analyze the performance of actual complex fisheye distortion. Figure 4 shows the qualitative results of fully recovering rotation and




(a) Only undistortion




(b) Fully recovered rotation and distortion

**Fig. 3:** Qualitative results on the test images for our generic camera model. (a) Undistortion results shown in the input image, results of the compared methods (Alemán-Flores [2], Santana-Cedrés [55], Liao [37], Yin [66], Chao [13], Bogdan [8], Li (GeoNetS-$\mathcal{B}$) [34], López-Antequera [45], and Wakai [60]), our method, and the ground-truth image from left to right. (b) Fully recovered rotation and distortion shown in the input image, results of the compared methods (López-Antequera [45] and Wakai [60]), our method, and the ground-truth image from left to right.

fisheye distortion for methods that can predict extrinsic and intrinsic camera parameters. These methods were trained using the StreetLearn [46] or SP360 [12] datasets. The results for López-Antequera's method had rotation and/or distortion errors. Our method outperformed Wakai's method [60], which often recovered only distortion for all our cameras. Our fully recovered images demonstrated the effectiveness of our method for off-the-shelf fisheye cameras with four types of projection: stereographic projection, equidistance projection, equisolid angle projection, and orthogonal projection.

In all the calibration methods, images captured by off-the-shelf cameras seemingly degraded the overall performance in the qualitative results compared with synthetic images. This degradation probably occurred because of the complex

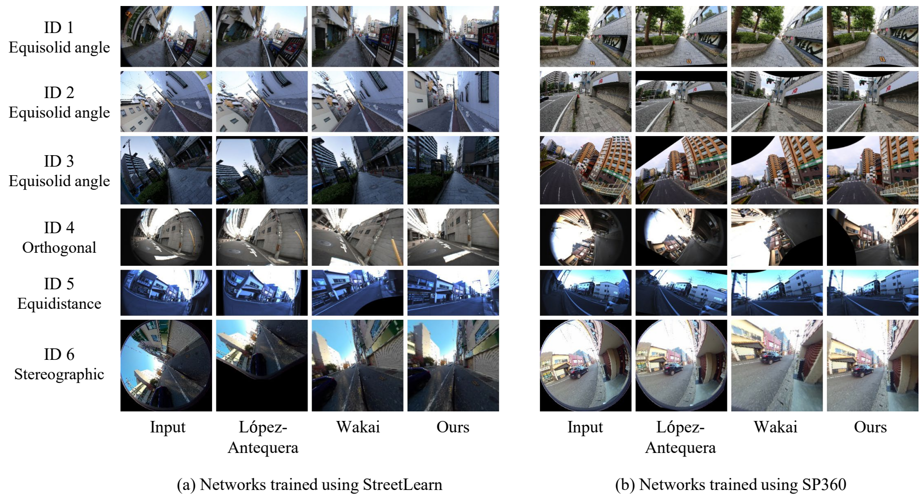



**Fig. 4:** Qualitative results of fully recovering rotation and fisheye distortion for the off-the-shelf cameras shown in the input image, results of the compared methods (López-Antequera [45] and Wakai [60]), and our method from left to right for each image. The IDs correspond to IDs in Table 3, and the projection names are attached to the IDs from specifications (ID: 3–5) and our estimation (ID: 1, 2, and 6). Qualitative results of the methods trained using StreetLearn [46] and SP360 [12] in (a) and (b), respectively.

distortion of off-the-shelf fisheye cameras and the dataset domain mismatch between the two panorama datasets and our captured images. Overall, our method outperformed the conventional methods in the qualitative evaluation of off-the-shelf cameras. As described above, our method precisely recovered both rotation and fisheye distortion using our generic camera model. Refer to the supplementary material for error distribution and additional calibration results.

## 5 Conclusion

We proposed a learning-based calibration method using a new generic camera model to address various types of camera projection. Additionally, we introduced a novel loss function that has optimal joint weights determined before training. These weights can alleviate the bias of the magnitude of each loss for four camera parameters. As a result, we enabled networks to precisely predict both extrinsic and intrinsic camera parameters. Extensive experiments demonstrated that our method substantially outperformed conventional geometric-based and learning-based methods on two large-scale datasets. Moreover, we demonstrated that our method fully recovered rotation and distortion using off-the-shelf fisheye cameras consisting of stereographic projection, equidistance projection, equisolid angle projection, and orthogonal projection. To improve the calibration performance in off-the-shelf cameras, in future work, we will study the dataset domain mismatch.

## References


1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of International Conference on Knowledge Discovery and Data Mining (KDD). pp. 2623–2631 (2019). https://doi.org/10.1145/3292500.3330701
2. Alemán-Flores, M., Alvarez, L., Gomez, L., Santana-Cedrés, D.: Automatic lens distortion correction using one-parameter division models. Image Processing On Line (IPOL) **4**, 327–343 (2014). https://doi.org/10.5201/ipol.2014.106
3. Alhaija, H.A., Mustikovela, S.K., Mescheder, L., Geiger, A.: Augmented reality meets computer vision: Efficient data generation for urban driving scenes. International Journal of Computer Vision (IJCV) **126**(9), 961–972 (2018). https://doi.org/10.1007/s11263-018-1070-x
4. Alvarez, L., Gómez, L., Sendra, J.: An algebraic approach to lens distortion by line rectification. Journal of Mathematical Imaging and Vision (JMIV) **35**(1), 36–50 (2009). https://doi.org/10.1007/s10851-009-0153-2
5. Barreto, J.: A unifying geometric representation for central projection systems. Computer Vision and Image Understanding (CVIU) **103**(3), 208–217 (2006). https://doi.org/10.1016/j.cviu.2006.06.003
6. Basu, A., Licardie, S.: Modeling fish-eye lenses. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). vol. 3, pp. 1822–1828 (1993). https://doi.org/10.1109/IROS.1993.583883
7. Benligiray, B., Topal, C.: Blind rectification of radial distortion by line straightness. In: Proceedings of European Signal Processing Conference (EUSIPCO). pp. 938–942 (2016). https://doi.org/10.1109/EUSIPCO.2016.7760386
8. Bogdan, O., Eckstein, V., Rameau, F., Bazin, J.C.: DeepCalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In: Proceedings of SIGGRAPH European Conference on Visual Media Production (CVMP) (2018). https://doi.org/10.1145/3278471.3278479
9. Brown, D.: Close-range camera calibration. Photogrammetric Engineering **37**(8), 855–866 (1971)
10. Bukhari, F., Dailey, M.: Automatic radial distortion estimation from a single image. Journal of Mathematical Imaging and Vision (JMIV) **45**, 31–45 (2013). https://doi.org/10.1007/s10851-012-0342-2
11. Cao, Z., Simon, T., Wei, S., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1302–1310 (2017). https://doi.org/10.1109/CVPR.2017.143
12. Chang, S., Chiu, C., Chang, C., Chen, K., Yao, C., Lee, R., Chu, H.: Generating 360 outdoor panorama dataset with reliable sun position estimation. In: Proceedings of SIGGRAPH Asia. pp. 1–2 (2018). https://doi.org/10.1145/3283289.3283348
13. Chao, C., Hsu, P., Lee, H., Wang, Y.: Self-supervised deep learning for fisheye image rectification. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 2248–2252 (2020). https://doi.org/10.1109/ICASSP40776.2020.9054191
14. Chen, Y., Schmid, C., Sminchisescu, C.: Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In: Proceedings of IEEE International Conference on Computer Vision (ICCV). pp. 7062–7071 (2019). https://doi.org/10.1109/ICCV.2019.00716

15. Dauphin, Y.N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., Bengio, Y.: Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). vol. 2, pp. 2933–2941 (2018). https://doi.org/10.48550/arXiv.1406.2572
16. Davidson, B., Alvi, M., Henriques, J.F.: 360° Camera alignment via segmentation. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 579–595 (2020). https://doi.org/10.1007/978-3-030-58604-1_35
17. Fitzgibbon, A.: Simultaneous linear estimation of multiple view geometry and lens distortion. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. I–I (2001). https://doi.org/10.1109/CVPR.2001.990465
18. Fu, Z., Liu, Q., Fu, Z., Wang, Y.: Template-free visual tracking with space-time memory networks. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13774–13783 (2021). https://doi.org/10.1109/CVPR46437.2021.01356
19. González-Aguilera, D., Gómez-Lahoz, J., Rodríguez-Gonzálvez, P.: An automatic approach for radial lens distortion correction from a single image. IEEE Sensors Journal (JSEN) **11**(4), 956–965 (2011). https://doi.org/10.1109/JSEN.2010.2076403
20. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). vol. 2, pp. 2672–2680 (2014). https://doi.org/10.48550/arXiv.1406.2661
21. Goodfellow, I.J., Vinyals, O., Saxe, A.M.: Qualitatively characterizing neural network. In: Proceedings of International Conference on Learning Representations (ICLR). pp. 1–20 (2015). https://doi.org/10.48550/arXiv.1412.6544
22. Gordon, A., Li, H., Jonschkowski, R., Angelova, A.: Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8976–8985 (2019). https://doi.org/10.1109/ICCV.2019.00907
23. Groenendijk, R., Karaoglu, S., Gevers, T., Mensink, T.: Multi-loss weighting with coefficient of variations. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1468–1477 (2021). https://doi.org/10.1109/WACV48630.2021.00151
24. Hold-Geoffroy, Y., Sunkavalli, K., Eisenmann, J., Fisher, M., Gambaretto, E., Hadap, S., Lalonde, J.: A perceptual measure for deep single image camera calibration. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2354–2363 (2018). https://doi.org/10.1109/CVPR.2018.00250
25. Hosono, M., Simo-Serra, E., Sonoda, T.: Self-supervised deep fisheye image rectification approach using coordinate relations. In: Proceedings of IEEE International Conference on Machine Vision and Applications (MVA). pp. 1–5 (2021). https://doi.org/10.23919/MVA51890.2021.9511349
26. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2261–2269 (2017). https://doi.org/10.1109/CVPR.2017.243
27. Huang, Z., Xu, Y., Shi, J., Zhou, X., Bao, H., Zhang, G.: Prior guided dropout for robust visual localization in dynamic environments. In: Proceedings of IEEE/CVF

International Conference on Computer Vision (ICCV). pp. 2791–2800 (2019). https://doi.org/10.1109/ICCV.2019.00288
28. Huber, P.J.: Robust estimation of a location parameter. Annals of Mathematical Statistics **35**(1), 73–101 (1964). https://doi.org/10.1214/aoms/1177703732
29. Ioffe, S., Szegedy, C.: Batch Normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of International Conference on Machine Learning (ICML). vol. 37, pp. 448–456 (2015). https://doi.org/10.48550/arXiv.1502.03167
30. Kannala, J., Brandt, S.S.: A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **28**(8), 1335–1340 (2006). https://doi.org/10.1109/TPAMI.2006.153
31. Lee, J., Go, H., Lee, H., Cho, S., Sung, M., Kim, J.: CTRL-C: Camera calibration TRansformer with line-classification. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16228–16237 (2021). https://doi.org/10.1109/ICCV48922.2021.01592
32. Lee, J., Sung, M., Lee, H., Kim, J.: Neural geometric parser for single image camera calibration. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 541–557 (2020). https://doi.org/10.1007/978-3-030-58610-2_32
33. Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T.: Visualizing the loss landscape of neural nets. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). pp. 6391–6401 (2018). https://doi.org/10.48550/arXiv.1712.09913
34. Li, X., Zhang, B., Sander, P.V., Liao, J.: Blind geometric distortion correction on images through deep learning. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4855–4864 (2019). https://doi.org/10.1109/CVPR.2019.00499
35. Liao, K., Chunyu, L., Liao, L., Zhao, Y., Lin, W.: Multi level curriculum for training a distortion-aware barrel distortion rectification model. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4369–4378 (2021). https://doi.org/10.1109/ICCV48922.2021.00435
36. Liao, K., Lin, C., Wei, Y., Li, F., Yang, S., Zhao, Y.: Towards complete scene and regular shape for distortion rectification. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14549–14558 (2021). https://doi.org/10.1109/ICCV48922.2021.01430
37. Liao, K., Lin, C., Zhao, Y.: A deep ordinal distortion estimation approach for distortion rectification. IEEE Transactions on Image Processing (TIP) **30**, 3362–3375 (2021). https://doi.org/10.1109/TIP.2021.3061283
38. Liao, K., Lin, C., Zhao, Y., Gabbouj, M.: DR-GAN: Automatic radial distortion rectification using conditional GAN in real-time. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) **30**(3), 725–733 (2020). https://doi.org/10.1109/TCSVT.2019.2897984
39. Liao, K., Lin, C., Zhao, Y., Xu, M.: Model-free distortion rectification framework bridged by distortion distribution map. IEEE Transactions on Image Processing (TIP) **29**, 3707–3718 (2020). https://doi.org/10.1109/TIP.2020.2964523
40. Lin, M., Chen, Q., Yan, S.: Network in network. In: Proceedings of International Conference on Learning Representations (ICLR). pp. 1–10 (2014). https://doi.org/10.48550/arXiv.1312.4400
41. Liu, C., Chen, L., Schroff, F., Adam, H., Hua, W., Yuille, A.L., Fei-Fei, L.: Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation.

In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 82–92 (2019). https://doi.org/10.1109/CVPR.2019.00017
42. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. In: Proceedings of International Conference on Learning Representations (ICLR). pp. 1–14 (2020). https://doi.org/10.48550/arXiv.1908.03265
43. Locher, A., Perdoch, M., Gool, L.V.: Progressive prioritized multi-view stereo. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3244–3252 (2016). https://doi.org/10.1109/CVPR.2016.353
44. Lochman, Y., Dobosevych, O., Hryniv, R., Pritts, J.: Minimal solvers for single-view lens-distorted camera auto-calibration. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 2886–2895 (2021). https://doi.org/10.1109/WACV48630.2021.00293
45. López-Antequera, M., Marí, R., Gargallo, P., Kuang, Y., Gonzalez-Jimenez, J., Haro, G.: Deep single image camera calibration with radial distortion. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11809–11817 (2019). https://doi.org/10.1109/CVPR.2019.01209
46. Mirowski, P., Banki-Horvath, A., Anderson, K., Teplyashin, D., Hermann, K.M., Malinowski, M., Grimes, M.K., Simonyan, K., Kavukcuoglu, K., Zisserman, A., Hadsell, R.: The StreetLearn environment and dataset. arXiv preprint arXiv:1903.01292 (2019). https://doi.org/10.48550/arXiv.1903.01292
47. Misra, D.: Mish: A self regularized non-monotonic neural activation function. In: Proceedings of British Machine Vision Conference (BMVC). pp. 1–14 (2020). https://doi.org/10.48550/arXiv.1908.08681
48. Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., Zhang, J.J.: Total3DUnderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 52–61 (2020). https://doi.org/10.1109/CVPR42600.2020.00013
49. Pritts, J., Kukelova, Z., Larsson, V., Chum, O.: Radially-distorted conjugate translations. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1993–2001 (2018). https://doi.org/10.1109/CVPR.2018.00213
50. Puskorius, G., Feldkamp, L.: Camera calibration methodology based on a linear perspective transformation error model. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA). vol. 3, pp. 1858–1860 (1988). https://doi.org/10.1109/ROBOT.1988.12340
51. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, real-time object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91
52. Ren, L., Song, Y., Lu, J., Zhou, J.: Spatial geometric reasoning for room layout estimation via deep reinforcement learning. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 550–565 (2020). https://doi.org/10.1007/978-3-030-58583-9_33
53. Rong, J., Huang, S., Shang, Z., Ying, X.: Radial lens distortion correction using convolutional neural networks trained with synthesized images. In: Proceedings of Asian Conference on Computer Vision (ACCV). vol. 10113, pp. 35–49 (2017). https://doi.org/10.1007/978-3-319-54187-7_3

54. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
55. Santana-Cedrés, D., Gomez, L., Alemán-Flores, M., Salgado, A., Esclarín, J., Mazorra, L., Alvarez, L.: An iterative optimization algorithm for lens distortion correction using two-parameter models. Image Processing On Line (IPOL) **6**, 326–364 (2016). https://doi.org/10.5201/ipol.2016.130
56. Saputra, M.R.U., Gusmao, P., Almalioglu, Y., Markham, A., Trigoni, N.: Distilling knowledge from a deep pose regressor network. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). pp. 263–272 (2019). https://doi.org/10.1109/ICCV.2019.00035
57. Sha, L., Hobbs, J., Felsen, P., Wei, X., Lucey, P., Ganguly, S.: End-to-End camera calibration for broadcast videos. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13624–13633 (2020). https://doi.org/10.1109/CVPR42600.2020.01364
58. Shah, S., Aggarwal, J.K.: A simple calibration procedure for fish-eye (high distortion) lens camera. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA). vol. 4, pp. 3422–3427 (1994). https://doi.org/10.1109/ROBOT.1994.351044
59. Tsai, R.Y.: A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. IEEE Journal of Robotics and Automation (JRA) **3**(4), 323–344 (1987). https://doi.org/10.1109/JRA.1987.1087109
60. Wakai, N., Yamashita, T.: Deep single fisheye image camera calibration for over 180-degree projection of field of view. In: Proceedings of IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 1174–1183 (2021). https://doi.org/10.1109/ICCVW54120.2021.00137
61. Wang, Z., Bovik, A.C.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing (TIP) **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
62. Xian, W., Li, Z., Snavely, N., Fisher, M., Eisenman, J., Shechtman, E.: UprightNet: Geometry-aware camera orientation estimation from single images. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9973–9982 (2019). https://doi.org/10.1109/ICCV.2019.01007
63. Xue, F., Wang, X., Yan, Z., Wang, Q., Wang, J., Zha, H.: Local supports global: Deep camera relocalization with sequence enhancement. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2841–2850 (2019). https://doi.org/10.1109/ICCV.2019.00293
64. Xue, Z., Xue, N., Xia, G., Shen, W.: Learning to calibrate straight lines for fisheye image rectification. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1643–1651 (2019). https://doi.org/10.1109/CVPR.2019.00174
65. Yang, S., Lin, C., Liao, K., Zhang, C., Zhao, Y.: Progressively complementary network for fisheye image rectification using appearance flow. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6344–6353 (2021). https://doi.org/10.1109/CVPR46437.2021.00628
66. Yin, X., Wang, X., Yu, J., Zhang, M., Fua, P., Tao, D.: FishEyeRecNet: A multi-context collaborative deep network for fisheye image rectification. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 475–490 (2018). https://doi.org/10.48550/arXiv.1804.04784

67. Zhang, M., Yao, J., Xia, M., Li, K., Zhang, Y., Liu, Y.: Line-based multi-label energy optimization for fisheye image rectification and calibration. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4137–4145 (2015). https://doi.org/10.1109/CVPR.2015.7299041
68. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **22**(11), 1330–1334 (2000). https://doi.org/10.1109/34.888718
69. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5122–5130 (2017). https://doi.org/10.1109/CVPR.2017.544
70. Żołądek, H.: The topological proof of Abel–Ruffini theorem. Topological Methods in Nonlinear Analysis **16**, 253–265 (2000)