17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

# Self-Supervised Deep Fisheye Image Rectification Approach using Coordinate Relations

Masaki Hosono
Waseda University, Utagoe Inc.
Tokyo, Japan
masaki_5292@fuji.waseda.jp

Edgar Simo-Serra
Waseda University
Tokyo, Japan
ess@waseda.jp

Tomonari Sonoda
Utagoe Inc.
Tokyo, Japan
sonoda@utagoe.com

## Abstract

*With the ascent of wearable camera, dashcam, and autonomous vehicle technology, fisheye lens cameras are becoming more widespread. Unlike regular cameras, the videos and images taken with fisheye lens suffer from significant lens distortion, thus having detrimental effects on image processing algorithms. When the camera parameters are known, it is straight-forward to correct the distortion, however, without known camera parameters, distortion correction becomes a non-trivial task. While learning-based approaches exist, they rely on complex datasets and have limited generalization. In this work, we propose a CNN-based approach that can be trained with readily available data. We exploit the fact that relationships between pixel coordinates remain stable after homogeneous distortions to design an efficient rectification model. Experiments performed on the cityscapes dataset show the effectiveness of our approach. Our code is available at GitHub*[1].

## 1 Introduction

Fisheye lens cameras have a much wider field of view than conventional cameras, making them suitable for autonomous vehicle and robotic tasks among others. Although a single fisheye camera can capture the same area as the combination of many conventional cameras, this comes at the cost of heavy image distortion, which limits the applicability of traditional computer vision algorithms and off-the-shelf machine learning algorithms. Thus it becomes fundamental to rectify the images to gain access to a diversity of existing algorithms.

When the camera parameters such as focal length and viewing angle are known, distortions can be corrected through simple calculations. However, such situations are rare, and many existing algorithms focus on the case when parameters are not known. Recently, learning-based approaches have been employed in the fisheye correction task using supervised pairs of fisheye and corrected images [1], annotations of expected camera parameters [2, 3], etc. Since the data has a high impact on the generalization results, much care has to be taken when assembling the data, and even then, generalization is not guaranteed.

In this paper, we introduce a learning-based image rectification model that is trained using existing data only, without any specific annotations. Our approach is based on the simple idea that pixels on straight lines should be aligned to straight lines again after distorting and correcting them. Our contributions are:

- A dataset construction approach based on distortion-free images.
- A Line Reconstruction error loss based on the fact that relationships between pixel coordinates remain stable after homogeneous distortions
- Significant impromevents over existing state-of-the-art.

## 2 Background

### 2.1 Distortion Parameters

Distortions on the images are caused by the optic characteristics of lens. Fitzgibbon [4] represents the transformation relationship of pixel coordinates between the normal images and the distorted images with a model. Considering the pixel coordinates as the points on a normalized coordinate system, they can be written as $p_i = (x_i, y_i)$ where $x_i, y_i \in [-1, 1]$. When a radius of the pixels belonging to the undistorted and distorted images that respects to their center coordinates are written as $r_u = \sqrt{x_u^2 + y_u^2}$ and $r_d = \sqrt{x_d^2 + y_d^2}$, respectively, the relationship between them can be approximated as follows:

$$r_u = f(\, r_d \mid k_i,\ 0 < i \leq n) = \frac{r_d}{1 - \sum_{i=1}^{n} k_i r_d^{2i}} \quad (1)$$

An amount of the distortion is related to $k_i$ only, thus they are called as *distortion parameters*. In this paper, we chose $n = 1$ to simplify the problem. Considering the affinity matrix $A_{ij}$ representing the transformation of distorting corrected, distortion-free, images,

[1] https://github.com/MasakiHosono/SelfSupervisedFisheyeRectification

pixel $I_i$ in an input will be mapped to the pixel $\hat{I}_j$ in a distorted image as follows:

$$\hat{I}_j = \sum_{i=1}^{N} A_{ij} I_i$$
$$where\ A_{ij} = \begin{cases} 1 & (x_i, y_i) = f(\hat{x_j}, \hat{y_j} \mid k) \\ 0 & otherwise \end{cases} \quad (2)$$

### 2.2 Curriculum Learning

Curriculum learning is a fundamental technique to improve training of difficult tasks [5]. The technique is based on the idea; humans and animals in the nature learn better when the examples are given in a meaningful order. Training starts with easy tasks such as small vocabularies for natural language processing and simple object images for image classification, and gradually the tasks are made more complex as training progresses. Models trained with this method performed better than models trained on difficult tasks from the start. In our research, we use this technique to train appropriate distortion parameters step by step. More concretely, the model should identify two significantly different distortion parameters, then gradually increase the number of patterns of the parameters for every switch epochs.

### 2.3 Deep Fisheye Rectification Approaches

While there are many non-learning based approaches, most modern techniques are based on learning from data. They can be divided into two categories: (1) models learned with supervision; and (2) generative models learned with adversarial networks.

#### 2.3.1 Supervised Learning Approaches

Supervised learning approaches are generally based on a CNN trained with supervision to predict distortion parameters. Zhucun Xue *et al.* [2, 3] used data which contained the distorted images $I$, the ground truth distortion parameters, ground truth distorted line maps, the ground truth distortion-free line maps, and the corresponding (rectified) line segments $L$ of image $I$. Manuel López *et al.* [6] introduced a model which could recover not only distortion parameters and focal length but also tilt and roll of the camera of the given image. They used SUN360 panorama dataset [7] for training their model; pan, offset, roll, aspect ratio, focal length, and the distortion parameters were sampled respectively. Jinlong Fan *et al.* [8] proposed a self-supervised image rectification method, but it needs a set of images of the exactly same scene from multiple different lenses. There is also an approach which needs a pair of distorted image and a rectification ground truth proposed by Xiaoqing Yin *et al.* [1]. The latest

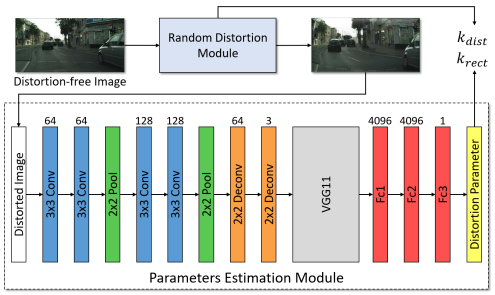



Figure 1. Our proposed network architecture.

approach which has proposed by Kang Liao *et al.* [9] in 2021 also used a ground truth ordinal distortions to calculate the loss functions.

Most of these approaches need a somewhat complex dataset. However, it is difficult to prepare such a dataset of a new domain, such as a real street environment in Japan (*e.g.* we can never take an exactly same shots using multiple cameras). To tackle this problem, we propose using only distortion-free images during training.

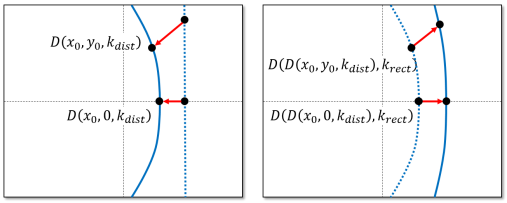
#### 2.3.2 GAN-based Approaches

The other type of approaches are based on generative-adversarial networks. One of the latest approach is Fisheye GAN (FE-GAN) [10]. It consists of two major components: a generator $G$ which aims at recovering the distortion-free images from distorted inputs, and a discriminator $D$ which aims to distinguish between real distortion-free images and rectified images. The generator takes an input distorted image $x_i \in M_{H \times W \times 3}$ and outputs a pixel warping flow map $f_i \in M_{H \times W \times 2}$. It may appear to be working correctly, but it can potentially learn to behave physically unnatural ways. Thus, we opt to predict the distortion parameters, then rectify using approximated optical formulations.

## 3 Proposed Approach

### 3.1 Network Architecture

The architecture of our network is shown in Figure 1. We base our model on the parameters estimation module proposed by Shangrong Yang *et al.* [11]. Our network uses an encoder-decoder structure jointly with a VGG11 model. It takes an image with distortion as an input and predicts the distortion parameters. The encoder-decoder structure aims at extracting low-level information and high-level information of the image at same time. Our encoder uses four convolutional layers in combination with two deconvolutional layers to extract features which are then fed into a VGG11 architecture. Finally, the output is connected to three fully-connected layers, and the distortion parameter is estimated. We use the Adam optimizer to train our model.




(a) (b)

Figure 2. An overview of LRE: (a) coordinate mappings during distorting images; (b) those mappings during restoring images. Difference between x-coordinates of these points should be minimized.

The dataset which we use to train the model contains the set of distortion-free images only. Therefore, they were distorted using randomly generated distortion parameter $k_{dist} \in [0, 1]$ (determined as uniform distribution), model had been trained to extract appropriate parameter $k_{rect} \in [-1, 0]$ which rectifies the images. After random distortions, distorted images are cropped from the distortion center as large as possible so that distorted edges between the image and the margin won't affect to the model's prediction.

### 3.2 Line Reconstruction Error

Since we won't use a complex dataset with a variety of supervision for training, our loss function relies on the fact that relationships between pixel coordinates remain stable after homogeneous distortions. More concretely, pixels on the arbitrary straight line will return to the same line again after distorting and correcting operations. We call our loss function *Line Reconstruction Error* (LRE) loss.

Figure 2 shows an overview of our loss. Consider the vertical straight line on a distortion-free image which can be represented as $x = x_0$ where $x$ is in a normalized coordination. When the image is once distorted and correctly rectified again, all the points that were on the line will become vertical straight line again. Projected coordinates after distorting with $k$ can be calculated as follows.

$$(\hat{x}, \hat{y}) = D(x, y, k) = \left(\frac{\sqrt{4k(x^2 + y^2) + 1} - 1}{2k(x + \frac{y^2}{x})}, \frac{y}{x}\hat{x}\right) \quad (3)$$

When $(x_0, 0)$ and $(x_0, y_0)$ are mapped correctly to the distorted image and the rectified image in turn, they must have the same x-coordinates. Therefore, we set this constraint in the LRE $E_{LR}$ and minimized it during training so that the model could predict $k_{rect}$.

$$\begin{aligned} E_{LR} = M \parallel\ & D_x(D(x_0, 0, k_{dist}), k_{rect}) \\ & - D_x(D(x_0, y_0, k_{dist}), k_{rect}) \parallel^2 \end{aligned} \quad (4)$$

where $M$ defines a magnitude of the error and we set it to 100 in our experiments. Our hypothesis holds for any straight line on a distortion-free image. Thus we assumed a perpendicular straight line for simplify the calculation. $(x_0, y_0)$ was defined heuristically so that a line do not stick out of the image even if $k_{dist} = 1$.

Table 1. Hyper-parameters settings for our model.

| Name | Value |
|---|---|
| Resized input height | 256 |
| Resized input width | 512 |
| Max epoch | 1000 |
| Batch size | 16 |
| Learning rate | 0.0001 |
| Switch epoch | every 30 epochs |

## 4 Experiments

First, we examine the effects of curriculum learning on the fisheye distortion rectification task. After that, we compare our approach with the method of Faisal Bukhari *et al.* [12] and FE-GAN [10]. Hyper-parameters we used while training our model are listed in the Table 1.

### 4.1 Dataset

Since our approach requires only distortion-free images for training, we can use readily available datasets. We chose the *Cityscapes dataset* [13] for experiments, since its data is similar to the frames that taken by a dashcam. Cityscapes is a large-scale dataset that usually used for training semantic segmentation tasks. It contains two different types of the images: frames from a diverse set of street scene videos, and multi-quality pixel-level annotations. However, we only use the street scene images: 2975 for training, 500 for validation, and 1525 for testing.

Images in the cityscapes dataset contains arcs belonging to the cars. In addition, natural objects such as trees can be a source of noise. These factors will make distortion rectification with this dataset difficult.

### 4.2 Results

#### 4.2.1 Effectiveness of Curriculum Learning

We train our model in two ways: using curriculum learning techniques and simple random sampling, then compare mean SSIM for each epochs. According to the Figure 4.a, the mean SSIM remained stable with random sampling, it indicates that the quality of the rectified images did not improved during training. On the other hand, the mean SSIM becomes larger after training when we apply our curriculum leaning techniques. Figure 4.b represents the outputs of our model. It also shows how effectiveness of the curriculum learning techniques, because model trained with random sampling predicts almost same values for all input images distorted with every $k_{dist}$.

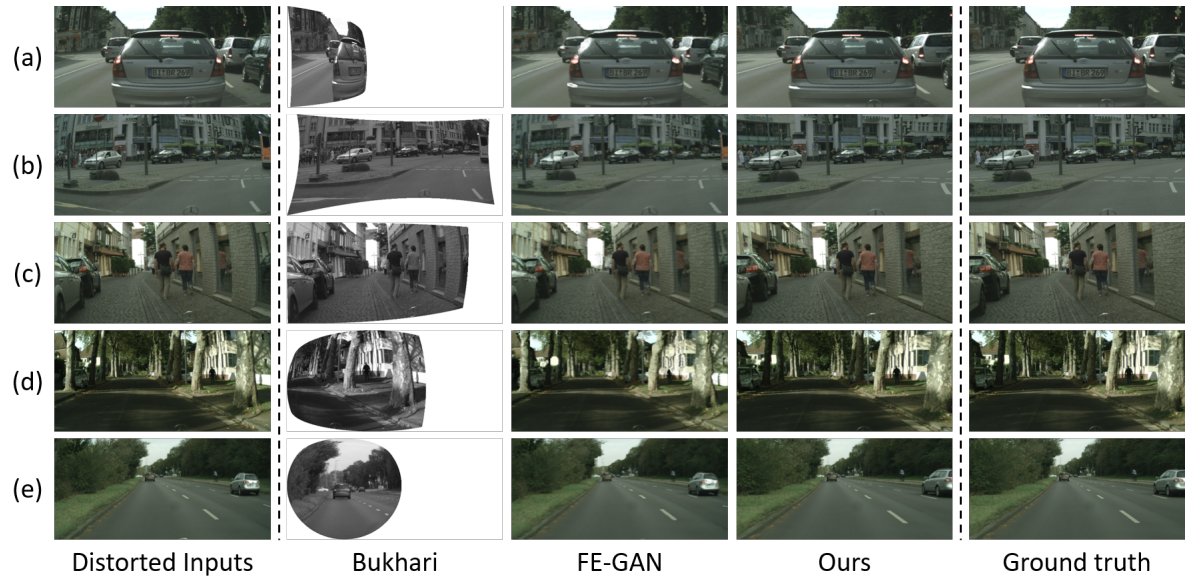



Figure 3. Rectification results of: Bukhari *et al.* [12], FE-GAN [10], and our approach.

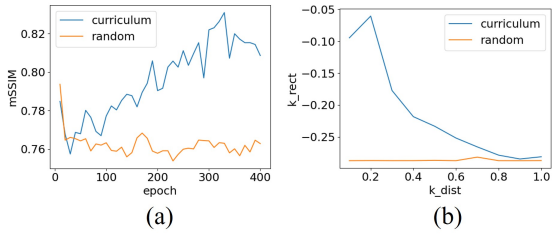



Figure 4. Effectiveness of curriculum learning: (a) mSSIM while training; (b) predicted distortion parameters for each input distorted images.

Table 2. Quantitative comparison between our method and FE-GAN

| | PSNR↑ | | SSIM↑ | |
|---|---|---|---|---|
| Image | Ours | FE-GAN | Ours | FE-GAN |
| a | **24.206** | 17.460 | **0.8459** | 0.6725 |
| b | **25.137** | 20.278 | **0.8240** | 0.6498 |
| c | **27.654** | 18.710 | **0.8160** | 0.5647 |
| d | **19.677** | 14.347 | **0.6732** | 0.5220 |
| e | **28.785** | 21.298 | **0.8852** | 0.7811 |

#### 4.2.2 Comparisons with Other Methods

We chose five different scenes for comparison: the large shot of rear of the car, relatively wide street, back alley with people walking around, residential area with trees, and the expressway. An experimental results for each of the three methods is shown in the Figure 3.

When the method of [12] was applied to the images, results changed with each run. Their method estimates distortion parameters based on the arcs found with RANSAC like algorithm. Arcs in the distorted images had been assumed to have been straight lines in the ground truth. However, a lot of the arcs in the images from cityscapes dataset are also arcs in the ground truth. Thus, their approach didn't work well in this experiment.

Both FE-GAN and our method are learning-based approaches, but they are quite different in qualitative results. According to the paper of FE-GAN [10], its generator $G$ outputs the flow map $f$, then a discriminator $D$ identifies the fake undistorted images from true distortion free images. However, the flow map may not always be the optically correct result. On the other hand, our method predicts a distortion parameter which is used to calculate the flow with optically approximated formulations. This difference significantly affects the quality of rectified images.

Table 2 shows the results of quantitative comparisons between our method and FE-GAN. As listed in the table, we compute the PSNR and SSIM between the results and the ground truth images. They shows that our method performs significantly better than FE-GAN for distorted cityscapes dataset even though ground truth images were only used for training.

## 5 Conclusions

We proposed a self-supervised fisheye image distortion rectification network which can be trained using distortion-free images only. Our model was trained using a Line Reconstruction Error loss on synthetic data and results on the cityscapes dataset show significant improvement over existing approaches. Currently, we approximate the image distortions to a single parameter division model [12]. As future work, a multi-parameter division model should be used in order to improve the quality of the approximation.

## References


[1] Xiaoqing Yin, Xinchao Wang, Jun Yu, Maojun Zhang, Pascal Fua, and Dacheng Tao. Fisheyerecnet: A multi-context collaborative deep network for fisheye image rectification. In *Proceedings of the European Conference on Computer Vision*, September 2018.

[2] Zhu-Cun Xue, Nan Xue, and Gui-Song Xia. Fisheye distortion rectification from deep straight lines. *arXiv preprint arXiv:2003.11386*, 2020.

[3] Zhucun Xue, Nan Xue, Gui-Song Xia, and Weiming Shen. Learning to calibrate straight lines for fisheye image rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1651, 2019.

[4] Andrew W Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2001.

[5] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[6] Manuel Lopez, Roger Mari, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, and Gloria Haro. Deep single image camera calibration with radial distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11817–11825, 2019.

[7] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702. IEEE, 2012.

[8] Jinlong Fan, Jing Zhang, and Dacheng Tao. Sir: Self-supervised image rectification via seeing the same scene from multiple different lenses. *arXiv:2011.14611*, 2020.

[9] Kang Liao, Chunyu Lin, and Yao Zhao. A deep ordinal distortion estimation approach for distortion rectification. *IEEE Transactions on Image Processing*, 30:3362–3375, 2021.

[10] Chun-Hao Chao, Pin-Lun Hsu, Hung-Yi Lee, and Yu-Chiang Frank Wang. Self-supervised deep learning for fisheye image rectification. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2248–2252. IEEE, 2020.

[11] Shangrong Yang, Chunyu Lin, Kang Liao, Yao Zhao, and Meiqin Liu. Unsupervised fisheye image correction through bidirectional loss with geometric prior. *Journal of Visual Communication and Image Representation*, 66:102692, 2020.

[12] Faisal Bukhari and Matthew N. Dailey. Automatic radial distortion estimation from a single image. *Journal of Mathematical Imaging and Vision*, 2012.

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.