OIDC-Net: Omnidirectional Image Distortion Correction via Coarse-to-Fine Region Attention

Kang Liao, Chunyu Lin, Yao Zhao, *Senior Member, IEEE*, Moncef Gabbouj, *Fellow, IEEE*, and Yang Zheng

***Abstract*—Omnidirectional cameras have recently received significant attention in panoramic imaging systems such as virtual reality (VR) technology; however, the strong geometric distortion in omnidirectional images severely affects the object recognition and semantic understanding. In this paper, we propose an automatic omnidirectional image distortion correction approach powered by a unified learning model (OIDC-Net). This approach is applicable for almost all types of omnidirectional cameras, requiring nothing more than a distorted image. A crucial and challenging ingredient for reconstructing the real physical scene is to estimate the heterogeneous distortion coefficients in an appropriate camera model. To address this issue, we present a novel coarse-to-fine region attention mechanism to alleviate the difficulty of predicting all coefficients simultaneously. With the proposed cascade structure and deep fusion strategy, the ambiguous relationship among these heterogeneous distortion coefficients has been incrementally perceived. Our experimental results show significant improvement over the state-of-the-art methods in terms of visual appearance, while maintaining a promising quantitative performance.**

***Index Terms*—Omnidirectional image distortion correction, coarse-to-fine region attention, Incremental perception.**

I. Introduction

OMNIDIRECTIONAL cameras have been recently incorporated into the computer vision and robotics fields. The main types of omnidirectional cameras can be classified into two categories: catadioptric cameras and fisheye cameras. Compared with the view-limited conventional camera, the omnidirectional camera provides a more enhanced field of view (FOV), which has gained popularity and has been increasingly used for ego-motion estimations [1], [2], intelligent vehicles [3], [4], and panoramic displays [5], [6]. However, owing to the specific design of such structures, strong geometric distortions [7] that occur in omnidirectional images hinder the understanding of the real physical scene.

Manuscript received March 18, 2019; revised August 6, 2019; accepted November 9, 2019. Date of publication November 22, 2019; date of current version February 5, 2020. This work was supported in part by National Natural Science Foundation of China under Grant 61772066, Grant 61972323, and Grant 61972028, and in part by Fundamental Research Funds for the Central Universities under Grant 2018JBZ001. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Mai Xu. *(Corresponding author: Chunyu Lin.)*

K. Liao, C. Lin, Y. Zhao, and Y. Zheng are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China (e-mail: kang_liao@bjtu.edu.cn; cylin@bjtu.edu.cn; yzhao@bjtu.edu.cn; yang_zheng@bjtu.edu.cn).

M. Gabbouj is with the Department of Signal Processing, Tampere University of Technology, 33720 Tampere, Finland (e-mail: moncef.gabbouj@tut.fi).

Digital Object Identifier 10.1109/JSTSP.2019.2955017

Image distortion correction is generally the first step in structure from motion (SfM) and simultaneous localization and mapping (SLAM) techniques, and thus the results of the distortion correction directly determine the performance of the entire pipeline. For the purpose of recovering the real physical scene, traditional methods for the distortion correction mainly focus on hand-crafted feature detection, as well as the estimation and optimization of the distortion coefficients. In [9]–[11], and [12], the line-based algorithm is leveraged for the catadioptric camera calibration, and all of the distortion coefficients are obtained using only the line scenes. In [13]–[15], and [16], the proposed methods require multiple views of planar calibration patterns that have metric-known points, corners, or any features that could be easily detected. Other prevalent correction approaches are based on self-calibration [17]–[19], and [20], which handle a sequence of different views to calculate intrinsic parameters of the camera, without requiring knowledge of the 3D location of the feature points. However, all of the aforementioned methods demand specific physical objects or scenes so that they cannot flexibly perform for any single omnidirectional image.

Over the past few years, deep learning has begun to outperform traditional methods in computer vision, particularly in object detection [21], [22], pose estimation [23], [24], and image super-resolution [25], [26]. However, to our knowledge, little attention has been paid to correct distortion in images using convolutional neural networks (CNNs). As a pioneering study, radial distortion correction is addressed in [27], where a learning structure is trained using synthesized radial distorted images when applying the one-parameter division model proposed in [28]. Yin *et al.* [29] propose a multiple context collaborative network for fisheye image rectification. To avoid the imbalanced problem during the training process of the heterogeneous distortion coefficient estimation, the image reconstruction loss is optimized rather than the regression loss, while these intrinsic parameters are essential for the camera calibration and SfM. In addition, it is more challenging to apply this approach to all omnidirectional images, such as images obtained from catadioptric cameras.

A further major disadvantage of the aforementioned learning-based methods is that they ignore some basic but vital prior knowledge with respect to distorted images. For instance, the distortion degree exponentially increases when a pixel is farther away from the distortion center, and the distortion center shift is close to the ideal optical center within a certain range, instead of being randomly perturbed on the global distribution. This motivates an investigation into the pipeline of the distortion correction, guided by the prior knowledge of the attention mechanism.

Fig. 1. OIDC-Net helps the object detection and semantic segmentation. Top to bottom: original omnidirectional images, Mask R-CNN [8] detections and segmentations on the omnidirectional images, our corrected images, and detections and segmentations on the corrected images. Our method is able to correct the distortion in any scenario, requiring only an omnidirectional image.

In retrospect of previous works, there is still room for improvement when it comes to correct the distortion in a single omnidirectional image. This problem could be potentially mitigated if we are able to present a unified learning structure for omnidirectional cameras, both catadioptric and fisheye cameras. In this paper, we propose an omnidirectional distortion correction network (OIDC-Net) that learns the distortion information and estimates the distortion coefficients in an omnidirectional image, guided by the coarse-to-fine region attention mechanism. As shown in Fig. 1, the proposed OIDC-Net can recover the real geometric distribution and thus benefits the scene understanding tasks such as object detection and semantic segmentation. Specifically, OIDC-Net includes a distortion center network (DC-Net) and distortion parameter network (DP-Net), which predict the distortion center and distortion parameters of omnidirectional images, respectively. The overall model architecture is depicted in Fig. 2. On the one hand, DC-Net estimates the accurate location of the distortion center with the coarse region attention, which guides the network to learn location information from reasonable areas. On the other hand, DP-Net uses the attentive aggregation as input, which is generated from the fusion module, and contains the original content and geometric distortion features encoded by the prior knowledge of the omnidirectional distortion. Subsequently, a cascade structure incrementally perceives different kinds of distortion features and roughly classifies the range of each distortion parameter, with the fine region attention. Then the deep fusion strategy is utilized to hierarchically fuse the distortion features and further perceive the ambiguous relationship among the distortion parameters. Finally, DP-Net accurately predicts the deviations of all of the distortion parameters, and therefore we correct the distortion in an omnidirectional image using these estimated distortion coefficients. By separating the estimation of the heterogeneous distortion coefficients into two specific networks, our method can effectively avoid the imbalanced problem during the training process. Experimental results show that OIDC-Net significantly outperfroms the state-of-the-art methods, both the visual apperance of corrected omnidirectional images and quantitative performance.

In summary, the main contributions of this paper are three-fold:

- We present a unified and flexible learning framework for the omnidirectional distortion correction in any scenario, requiring only an image.
- We describe a novel coarse-to-fine region attention mechanism that aims to implicitly exploit the prior knowledge with regard to the distortion characteristics.
- We introduce a cascade structure and deep fusion strategy to incrementally perceive the ambiguous relationship among heterogeneous distortion coefficients.

The rest of this paper is organized as follows. We first introduce the related works in Section II. We then establish the unified omnidirectional camera model and dataset, and present the proposed OIDC-Net framework in Sections III and IV, respectively. The experiments and discussions are provided in Section V. Finally, we conclude this paper in Section VI.

II. Related Work

Distortion correction in terms of the traditional and learning-based methods is addressed in this work. Previous published works have mainly focused on the traditional and learning-based distortion corrections. We also discuss the generalized parametric model for omnidirectional cameras.

A. Traditional Distorted Images Correction

A universal distortion correction standard was provided by F. Devernay *et al.* [30], which states that a straight line must appear straight in an image. Previous classical distortion correction methods [13], [31], [32] are based on a calibration chessboard, and apply a planar calibration chessboard captured in multi-view images. Although such methods work well from sufficient priors, they are expected to perform poorly under limited conditions. Wei *et al.* [33] and Carroll *et al.* [34] required manual marking from users, such as picking out a certain number of curves that are supposed to be straight lines in the real world. However, these additional requirements make it difficult to correct a single distorted image automatically and flexibly. To extend the application scenarios, Bukhari *et al.* [35] proposed an automatic radial removal method, which works from a single distorted image and estimates circular arcs based on the one-parameter division distortion model presented by Fitzgibbon [36]. Santanacedrés *et al.* [37] improved on the work in [28], and proposed a scheme for automatic distortion correction using a two-parameter radial

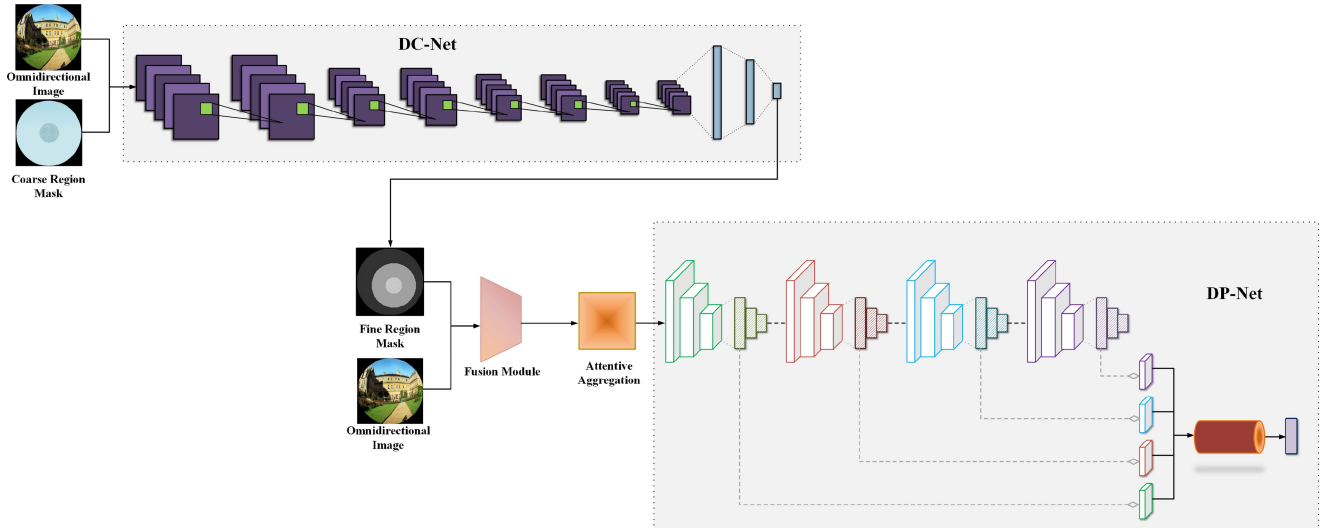

Fig. 2. Overview of the learning model architecture. OIDC-Net consists of two networks, namely, a distortion center network (DC-Net) and distortion parameter network (DP-Net). DC-Net takes an omnidirectional image and a coarse region mask as inputs, and outputs the predicted distortion center. Subsequently, the distortion center is exploited to construct a fine region mask based on the prior knowledge of the geometric distortion, which is further combined with the original omnidirectional features using a fusion module to generate an attentive aggregation. DP-Net consists of an incrementally perceived cascade structure and finally outputs the estimated distortion parameters using a deep fusion module.

distortion model with an iterative optimization algorithm. However, these three automatic methods are conventionally time-consuming due to a heavy dependence on hand-crafted feature detection and optimization, and provide inaccurate results when wrong arcs or lines are detected. All of the aforementioned approaches exploit hand-crafted features from images to correct the distortion; therefore, the process of hand-crafted feature detection and optimization is highly sensitive with regard to the final performance.

B. Learning-Based Distorted Images Correction

The learning-based method for distortion correction has recently been studied. Rong *et al.* [28] first implemented CNNs that aim to recover the real physical scene from the radial distortion. More specifically, the authors mapped a fixed range of distortion parameters into discrete integers with the intention of classifying the distorted images using neural networks. However, as the distortion model is relatively simple and only consists of one parameter, and a given assumption of the distortion center is known, this method performs poorly in certain sophisticated models, such as the omnidirectional camera model. Furthermore, CNNs occlude the classification ability using an image without any geometric and semantic features. To address these problems, Yin *et al.* [29] introduced a semantic segmentation network to guide the distortion estimation of fisheye images. They trained a multi-context collaborative deep network, namely FishEyeRecNet, using a synthetic dataset built upon the fisheye camera model. Instead of directly minmizing the regression loss, they minimized the image reconstruction loss to estimate the heterogeneous distortion coefficients. While this proposed network outperforms the state-of-the-art methods by remarkable margins, FishEyeRecNet cannot be trained without geometric and semantic information derived from undistorted and scene parsing images, respectively, thus imposing increased memory and efficiency burdens on the model. Moreover, it is more challenging to extend FishEyeRecNet to other omnidirectional cameras, such as catadioptric cameras.

C. Generalized Parametric Model for Omnidirectional Cameras

Geyer *et al.* [38] proposed a sphere model as the unified model that is applicable to any central catadioptric system. Ying *et al.* [39] directly leveraged the calibration methods for catadioptric cameras to fisheye cameras in terms of their presented unified imaging model. The general model for catadioptric cameras using a spherical projection is equivalent to pinhole-based models, as demonstrated by Courbon *et al.* [40], so that it can be directly exploited to fisheye cameras. Ramalingam *et al.* [41] employed a unified non-parametric camera model that associates one projection ray to each pixel. The imaging function of the omnidirectional camera model is described by a Taylor series expansion in Scaramuzza *et al.* [42], and the distortion coefficients were estimated using a four-step least-squares linear minimization algorithm. It is highly desirable that this method dispenses with any special models of the omnidirectional cameras and a priori knowledge of extrinsic parameters. In this work, we train the proposed network and estimate the distortion coefficients based on this unified omnidirectional camera model, by taking the generalization and flexibility of correction system into account.

III. Omnidirectional Camera Model and Image Dataset Establishment

Generally speaking, the training of a neural network with massive and standard datasets is inevitable for its learning ability. Nevertheless, a real omnidirectional image dataset requires

enormous labeling work, and the ground truth of the distortion coefficients rely heavily on the calibration process. Owing to the limitations of the utilized acquisition devices, the demand for a dataset containing a wide range of heterogeneous distortion coefficients is difficult to obtain. To this end, building a complete omnidirectional image dataset that is captured using various omnidirectional sensors and provides neural networks learning, is essentially unachievable.

Based on the previous discussion, we construct a synthesized omnidirectional image dataset with the ground truth of the distortion coefficients and matched real images with respect to the unified omnidirectional camera model.

A. Unified Omnidirectional Camera Model

We initially assume that a point $\mathbf{x} = [x, y]^{\mathrm{T}}$ in the camera plane corresponds to a scene point $\mathbf{X}$ of an incident ray $\mathbf{r}$. The relationship between $\mathbf{x}$ and $\mathbf{X}$ is given by

$$\mathbf{r} = \mathbf{h}(\mathbf{x}) = \mathrm{P}\mathbf{X}, \tag{1}$$

where $\mathbf{h}$ is the image projection non-linear function; $\mathbf{X} \in \mathbb{R}^4$ is the homogeneous coordinates in the scenes; and $\mathrm{P} \in \mathbb{R}^{3\times4}$ is a perspective projection matrix. The projection function $\mathbf{h}$ has the following expression:

$$\mathbf{h}(x, y) = (x, y, f(x, y))^{\mathrm{T}}, \tag{2}$$

where f is a Taylor series expansion that is defined as

$$f(x, y) = k_0 + k_1 r + k_2 r^2 + \cdots + k_N r^N. \tag{3}$$

Here $\{k_0, k_1, k_2, \ldots, k_N\}$ are the distortion parameters in the omnidirectional camera model and r is the distance between the point and distortion center, $\mathbf{c}$. If we assume that the coordinates of the distortion center is $[x_c, y_c]^{\mathrm{T}}$, then r can be obtained using the following expression:

$$r = \sqrt{(x - x_c)^2 + (y - y_c)^2}. \tag{4}$$

As suggested in [42], for catadioptric and fisheye cameras, the function f always satisfies the following condition:

$$\left.\frac{df}{dr}\right|_{r=0} = 0. \tag{5}$$

Thus, k_1 equals 0 and the distortion parameters are simplified to $\{k_0, k_2, \ldots, k_N\}$ in the omnidirectional camera model, and Eq. (3) can be rewritten as

$$f(x, y) = k_0 + k_2 r^2 + \cdots + k_N r^N. \tag{6}$$

Finally, the distortion coefficients $\{x_c, y_c, k_0, k_2, \ldots, k_N\}$ are used to construct the intrinsic parameters of the unified omnidirectional camera model, which should be accurately estimated for the distortion correction.

B. Omnidirectional Image Dataset Generation

We generate an omnidirectional image dataset of various distortion coefficients using the introduced unified model. The distortion coefficients consists of a series of distortion parameters and a distortion center, as described below.

Distortion Parameters Variety: The distortion parameters denote the distortion degree of the entire image. In this regard, the larger the distortion parameter values, the stronger the image distortion. As different distortion parameters have different magnitudes, the loss function unfairly treats each parameter during the training process. Therefore, we normalize the magnitude of all of the distortion parameters and exploit these values as labels of the omnidirectional images after the synthetic image generations.

Distortion Center Perturbation: In contrast, the location of the distortion center determines the distortion distribution in an omnidirectional image. The closer the pixels to the distortion center, the smaller the distortion, and vice versa. To produce different distortion centers in an omnidirectional image, we select the center of an image as the initial distortion center $\mathbf{c_o} = [x_o, y_o]^{\mathrm{T}}$, and this center is then randomly perturbed by values within the range $(-\alpha, \alpha)$. Finally, we obtain the target distortion center $\mathbf{c} = [x_o + a, y_o + b]^{\mathrm{T}}$, where a and $b \in (-\alpha, \alpha)$.

IV. Omnidirectional Distortion Correction Network (OIDC-Net)

In this section we describe the full method for learning distortion information from an omnidirectional image. We first introduce a baseline that directly estimates the heterogeneous distortion coefficients of the unified omnidirectional camera model proposed in Section III. We then present the details of DC-Net and DP-Net, which exploit the coarse-to-fine region attention mechanism with respect to the crucial prior knowledge. Finally, the loss functions of each component in OIDC-Net are proposed. The overview of OIDC-Net is illustrated in Fig. 2.

A. Baseline

First, we design a vanilla version of OIDC-Net as the baseline, which takes an omnidirectional image as the input and estimates all of the heterogeneous distortion coefficients. This structure can be divided into two sub-networks: backbone and header. Specifically, the backbone network is exploited to extract the high-level and distortion features from the input omnidirectional image. On the other hand, the header network that contains three fully connected layers is exploited to predict all of the heterogeneous distortion coefficients simultaneously. The number of units for these three layers are: 1024, 512, and n, where n is the number of distortion coefficients in the unfied omnidirectional camera model discussed in Section III-A. The activation functions for the first two fully connected layers are RELUs, while the last fully connected layer implements the linear function as the activation function.

The baseline directly estimates all of the distortion coefficients using an omnidirectional image, however a challenging problem remains due to the heterogeneous attributes of the distortion coefficients. To be more specific, the distortion coefficients contain a distortion center, $\mathbf{c}$, and a series of distortion parameters, $\{k_0, k_2, \ldots, k_N\}$, which indicate the distribution and degree of the distortion in an omnidirectional image, respectively. Moreover, the range and magnitude of these two types of distortion coefficients differ greatly. In this regard, finding a

balance between the distortion center and distortion parameters during the training process is quite challenging. Consequently, based on the divide-and-conquer algorithm, we redesign the architecture of the baseline and propose the OIDC-Net that consists of two special networks as described below.

B. Distortion Center Network

To accurately estimate the location of the distortion center in an omnidirectional image, we propose a DC-Net that utilizes the coarse region attention mechanism. We further construct a three-region geometric mask in terms of the fine region attention mechanism and generate the attentive aggregation contained in the original content and geometric distortion features.

Coarse region attention mechanism: In contrast to the general prediction task, which requires neural networks to search for an optimal solution in the global distribution, we propose a coarse region attention mechanism in the distortion center estimation task. Typically an ideal optical center locates the center of a lens. However, due to inaccurate manufacturing processes and the influence of external factors, a shift in the optical center exists in most lenses. The shift occurs around the ideal optical center within a certain range, instead of being randomly perturbed in the global distribution. This motivated the work here, to guide the neural networks to focus on a reasonable area in the omnidirectional images. More specifically, we construct a binary dense mask that contains the available region around an ideal optical center. We will compare the performance of the DC-Net both with and without the coarse region attention mechanism in Section V-B.

There are two options for the design of DC-Net. The first option is a regression-based model, which regresses the coordinate value of the distortion center in an omnidirectional image. Since the shift value is minor and the range is limited, we propose a classification-based model as the second option. As discussed in Section III-B, the shift value of distortion center $\mathbf{c}$ belongs to the range $[-a, a] \times [-b, b]$, which forms a rectangular area, $\mathbf{Q}$. We assume that the metric of the shift value is at the pixel-level so that $\mathbf{Q}$ includes $(2a+1) \times (2b+1)$ pixels in total. Finally, our goal is to classify an omnidirectional image into the category cls_k from the set $\{cls_k = (2a+1)i + j \mid 0 \leq i \leq 2a, 1 \leq j \leq (2b+1), k = (2a+1)i + j\}$. Experimental results show that the classification-based model significantly outperforms the regression-based model, and the relevant experiments and explanations will be presented in Section V-B.

DC-Net uses an omnidirectional image and a coarse region mask as the inputs, then predicts the location of the distortion center. The network uses 3×3 convolutional blocks, and all of the activation functions for each layer are ReLUs except for the last fully connected layer that leverages the softmax function (classification-based model). In more detail, we use eight convolutional layers with a maxpooling layer (2×2, stride 2) after every two convolutions. The eight convolutional layers have the following number of filters per layer: 64, 64, 128, 128, 256, 256, 256, and 256. At the end of the network, the last convolutional layer is followed by three fully connected layers, which contain the following number of units per layer: 1024, 512, and m, where m indicates the category number of the distortion center in an omnidirectional image. Fig. 2 illustrates the complete network architecture of DC-Net.

Fine region attention mechanism: As mentioned in Section III-B, the distribution of the distortion relies on the location of the distortion center in an omnidirectional image. To make full use of this prior knowledge, we structure a three-region geometric mask based on the predicted location of the distortion center. As shown in Fig. 2, different regions represent different degrees of distortion based on the characteristic of the distortion distribution. These three regions can be structured as:

$$\mathbf{v}(x,y) = \begin{cases} a_0, & x = x_c, \quad y = y_c. \\ b_0, & 0 < \mathbf{u}'(x,y) \leq r_1. \\ c_0, & r_1 < \mathbf{u}'(x,y) \leq r_2. \\ d_0, & r_2 < \mathbf{u}'(x,y), \quad \mathbf{u}''(x,y) \leq r_3. \end{cases} \tag{7}$$

where $\mathbf{v}(x,y)$ is the gray value of a pixel $\mathbf{p} = [x,y]^{\mathrm{T}}$ in the three-region geometric mask, and the size of this mask is the same as that of the original omnidirectional image. In Eq. (7), there are four hyperparameters: a_0, b_0, c_0, and d_0, and different gray values express different degrees of distortion in an omnidirectional image. Specifically, a_0 is the value of the pixel located at the distortion center $\mathbf{c} = [x_c, y_c]^{\mathrm{T}}$. The remaining hyperparameters b_0, c_0, and d_0 are the gray values of pixels in the three refine regions, where each region represent one specific degree of distortion. In addition, $\mathbf{u}' = \sqrt{(x-x_c)^2 + (y-y_c)^2}$ is the Euclidean distance between a pixel $\mathbf{p}$ and the distortion center $\mathbf{c}$. In a same manner, $\mathbf{u}'' = \sqrt{(x-x_o)^2 + (y-y_o)^2}$ is the Euclidean distance between a pixel $\mathbf{p}$ and the original center of an omnidirectional image $\mathbf{o} = (x_o, y_o)^{\mathrm{T}}$. Moreover, r_1, r_2, and r_3 denote the range of the corresponding fine regions. Details regarding the value definition of these hyperparameters will be described in Section V-B.

Attentive aggregation: To further provide more geometric features regarding the distortion for neural networks, we combine the original omnidirectional image with the three-region geometric mask using a fusion module. Specifically, this fusion module is quite lightweight, and is only comprised of one convolutional layer with a 1×1 convolutional block with a filter number of three. Compared with the DC-Net, the activation function for this convolutional layer is Tanh. As a result, the attentive aggregation includes the original content information as well as the geometric distortion information, making the DP-Net attentively perceive the effective features with regard to the distortion.

Instead of utilizing the location of the distortion center in a direct way, we implicitly structure an attentive aggregation and exploit the attention mechanism to conduct the perception of networks. As a benefit of this combination, our model makes full use of the comprehensive features, both the content and geometry, significantly outperforming methods that omit this potential prior distortion knowledge. The relevant experiments are described in Section V-B.

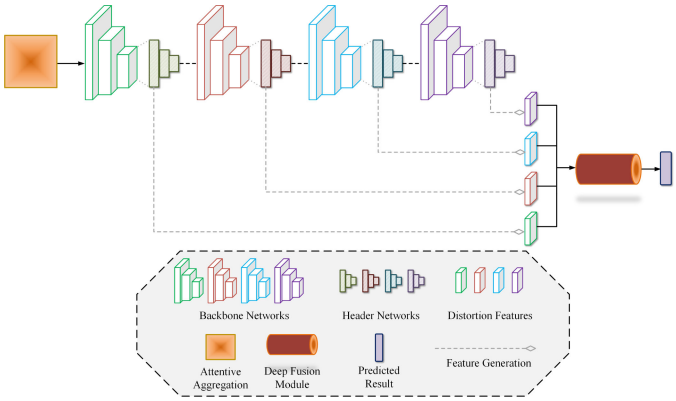

Fig. 3. Network architecture of the DP-Net. The DP-Net consists of a cascade structure that incrementally perceives the distortion features. At the end of this structure, a deep fusion module is implemented to hierarchically fuse all of distortion features.

C. Distortion Parameter Network

In the subsequent part of the OIDC-Net, the DP-Net uses the attentive aggregation as input, and estimates the distortion parameters of the omnidirectional image. The architecture of the DP-Net is illustrated in Fig. 3 and we show four parameter estimation modules.

Incrementally perceived cascade structure In analogy to the DC-Net, there are two options for the architecture of DP-Net, i.e., the regression-based and classification-based methods. However, the range of the distortion parameters is much wider than that of the distortion center, and it is tough to reasonably discrete their various values to different categories. Therefore, the classification-based method sufferes from inaccurate estimations on the large amount but slight difference data distribution. In contrast to the classification-based method, the regression-based method accurately learns a mapping relationship between the extracted features and the labels of ground truth. While this method performs poorly on a sparse and imbalanced data distribution, which provides a smooth estimation around the most frequent distortion parameters. With the accuracy and robustness in mind, we combine these two methods into a cascade way, effectively eliminating their specific shortcomings and incrementally perceiving the different distortion features.

First, the proposed structure roughly classifies the reasonable range of the distortion parameters, and then accurately predicts the deviation of each distortion parameter using the classified features. To be more specific, the range of distortion parameters is discreted to K categories with the quantization step I, so that the deviation of each distortion parameter belongs to the range of $(0, I)$. DP-Net comprises of N parameter-specific networks that estimate the corresponding distortion parameters, respectively. The structure of each network is similar to that of the baseline, and the difference between these two network versions is the number of units in the last fully connected layer, which is assigned the number of categories K. All of the backbone networks of these paramerer-specific networks share weights with each other.

Deep fusion strategy: The prediction of the distortion coefficients in an omnidirectional image is extremely challenging using neural networks, as the distortion degree of every region depends on the associated influence of all distortion coefficients. Furthermore, the ambiguous relationship among different parameters leads to difficulties in balancing the bias of each estimation of the network. To address the aforementioned problems, we further explore the mechanism of the parametric omnidirectional model. Based on previous works [43]–[45], we implement three different fusion strategies: early, late, and deep fusion.

We first denote all of the distortion features by $\{\mathbf{f}_{k_0}, \mathbf{f}_{k_2}, \ldots, \mathbf{f}_{k_N}\}$ and suppose the fusion module has M layers in which each layer is denoted by $\{\mathbf{L_i}|i = 1, 2, \ldots, M\}$. The early fusion strategy concatenates all of the distortion features at the beginning of the module, then we feed this combination into a series of abstract units to sequentially perceive the integrated features. Finally, a prediction unit, denoted by $\mathbf{P}$, simultaneously outputs the estimation of all of the distortion parameters. In brief, the early fusion strategy can be formalized as follows:

$$\mathbf{OE} = \mathbf{P}(\mathbf{L}_M(\mathbf{L}_{M-1}(\cdots\mathbf{L}_1(\mathbf{f}_{k_0} \circ \mathbf{f}_{k_2} \circ \cdots \circ \mathbf{f}_{k_N})))), \quad (8)$$

where $\circ$ is the operation of concatenation and $\mathbf{OE}$ is a N dimensional vector that includes the predicted deviation values of each distortion parameter.

The structure of the late fusion module is opposite to that of the early fusion module, where different branches independently perceive different distortion features. For a fair comparison, we suppose the number of abstract units equals that of the early fusion module in the progressive perceptions, and the dimensions of all of the units keep the same with that of the early fusion module. At the end of the abstract units, we concatenate all of the perceived features and a prediction unit $\mathbf{OL}$ outputs an estimation of the deviation vaules of distortion parameters.

$$\begin{aligned}\mathbf{OL} = \mathbf{P}(&\mathbf{L}_M^{k_0}(\mathbf{L}_{M-1}^{k_0}(\cdots\mathbf{L}_1^{k_0}(\mathbf{f}_{k_0})))\\ &\circ \mathbf{L}_M^{k_2}(\mathbf{L}_{M-1}^{k_2}(\cdots\mathbf{L}_1^{k_2}(\mathbf{f}_{k_2})))\\ &\circ \cdots\cdots\\ &\circ \mathbf{L}_M^{k_N}(\mathbf{L}_{M-1}^{k_N}(\cdots\mathbf{L}_1^{k_N}(\mathbf{f}_{k_N})))).\end{aligned} \quad (9)$$

It is noteworthy that the early and late fusions both have only one concatenation of features, which proceed at the beginning or end of a module. To benefit the neural networks and aid in the deep learning of the relationship among different distortion features, we design a multiple continuous perception in terms of the deep fusion strategy. The integrated distortion feature, denoted as $\mathbf{F}$, can be expressed by:

$$\mathbf{F} = \mathbf{f}_{k_0} \diamond \mathbf{f}_{k_2} \diamond \cdots \diamond \mathbf{f}_{k_N}, \quad (10)$$

where $\diamond$ is the operation of the element-wise mean. We also adopt the same setting for the number and dimension of the abstract units as the early fusion module, and therefore the deep fusion module is formalized as follows:

$$\mathbf{OD} = \mathbf{P}(\mathbf{F}_M), \quad (11)$$

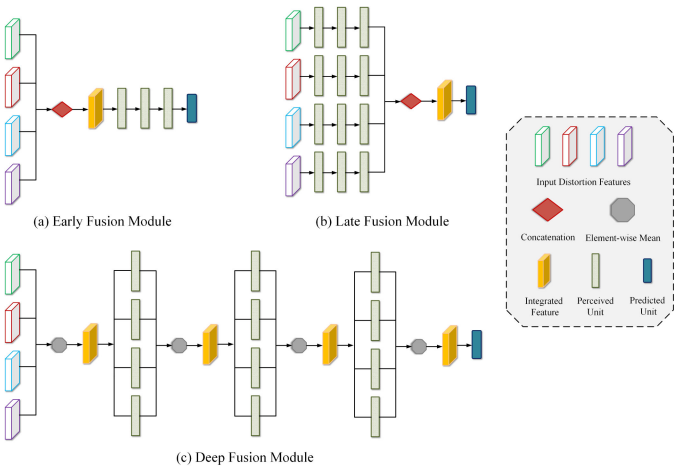

Fig. 4. Architecture of three types of fusion strategies: early, late, and deep fusion. The early and late fusions both have only one interactive operation of different features, which proceed at the beginning or end of a module. In contrast, the deep fusion has more interactive operations of different features.

where $\mathbf{OD}$ is the predicted deviations vector, and $\mathbf{F}_M$ is an integrated distortion feature derived by the M-th deep fusion layer:

$$\mathbf{F}_M = \mathbf{L}_M^{k_0}(\mathbf{F}_{M-1}) \diamond \mathbf{L}_M^{k_2}(\mathbf{F}_{M-1}) \diamond \cdots \diamond \mathbf{L}_M^{k_N}(\mathbf{F}_{M-1}). \tag{12}$$

The detailed architectures of the three fusion strategies are shown in Fig. 4, where we set M to three and the number of parameter estimation modules to four.

D. Loss Function

As discussed in Section IV-B, the DC-Net is a classification-based network that predicts the location of the distortion center from the reasonable area $\mathbf{Q}$. Based on this, the cross-entropy loss of the classification result q is used in terms of the estimated center:

$$\mathbf{L}_{\mathrm{DC}}(q, y) = \begin{cases} -\log(q), & \text{if } y = 1. \\ -\log(1-q), & \text{otherwise.} \end{cases} \tag{13}$$

Compared to the DC-Net, the DP-Net is an incrementally perceived cascade network that first classifies the general range of distortion parameters $\{k_0, k_2, \ldots, k_N\}$, and then regresses the precise deviation values of these parameters. Therefore, the loss function of the DP-Net includes two types of objectives:

$$\mathbf{L}_{DP}(\hat{\mathbf{C}}, \hat{\mathbf{R}}) = \mathbf{L}_{cls}(\mathbf{C}, \hat{\mathbf{C}}) + \lambda \mathbf{L}_{\mathrm{reg}}(\mathbf{R}, \hat{\mathbf{R}}), \tag{14}$$

where $\hat{\mathbf{C}}$ and $\hat{\mathbf{R}}$ are the estimated category and deviation value of each distortion parameter, $\mathbf{C}$ and $\mathbf{R}$ are the ground truth of these values, respectively. Besides, λ is a factor to balance the objectives of the classification and regression. Taking the stability of the traning process into account, we employ the smooth l_1 loss for the regression objective rather than the l_2 loss that may lead to the exploded gradients:

$$\mathbf{L}_{\mathrm{sm}}(x) = \begin{cases} 0.5x^2, & if \ \ |x| \leq 1. \\ |x| - 0.5, & \text{otherwise.} \end{cases} \tag{15}$$

Here, $x = p - \hat{p}$, p is the ground truth of the distortion parameters, and $\hat{p}$ is the estimated distortion parameters using the DP-Net.

V. Experiments

A. Synthetic Omnidirectional Dataset and Implementation Details

We construct a synthetic omnidirectional dataset derived from the unified model introduced in Section III-A, the sourced images are obtained from the oxford buildings dataset [46]. In particular, there are two heterogeneous distortion coefficients that need to be estimated for distortion correction, i.e., a distortion center $\mathbf{c}$ and series of distortion parameters $\{k_0, k_2, \ldots, k_N\}$. As mentioned earlier, the distortion center randomly distributes around the initial center within the range $(-\alpha, \alpha)$. Having considered the size of the synthetic omnidirectional image, we empirically choose the shift value α of the distortion center to 4 so that its category number m correspondingly becomes 81. On the other hand, Scaramuzza *et al.* [42] used a 4^{th} order polynomial to approximately fit the unified omnidirectional camera model. The method satisfactorily performed for the camera calibration, and thus the chosen number of distortion parameters is 4.

We train our learning model on a single NVIDIA GeForce GTX TITAN X GPU. Concretely, the DC-Net has been trained using Adam [47] with a base learning rate of 0.0005, which is decreased by 10 at every 5 k iterations for two steps. For the architecture of the DP-Net, we leverage InceptionV3 [48] (without fully connected layers) as the backbone network in DP-Net, which is pre-trained on ImageNet and fine-tuned on our synthetic omnidirectional dataset. The DP-Net is sequentially trained on the classification and regression tasks, with the same learning rate as that of the DC-Net. Note that the parameters of the backbone network are freezed during the training process of the regression task. Afterwards, we jointly train this incrementally perceived cascade structure with a smaller learning rate of 0.0001.

B. Ablation Study

In this part, we demonstrate the effectiveness of the proposed coarse-to-fine region attention mechanism, incrementally perceived cascade structure, and deep fuison strategy.

Coarse-to-Fine Region Attention Mechanism: We construct a coarse region mask in terms of the coarse region attention mechanism, which guides neural networks to focus on the available existing area of the distortion center, rather than searching for the optimal solution in the global distribution. The experimental results show that the coarse region attention (CRA) mechanism boosts the accuracy of the distortion center localization. A comparison of the DC-Net performances is shown in Table I. We also compared the performance of the regression-based (reg) and classification-based (cls) DC-Net, where the mean square error

TABLE I
COMPARISON OF THE DC-NET PERFORMANCE FOR DIFFERENT METHODS: THE REGRESSION-BASED (REG), CLASSIFICATION-BASED (CLS) DC-NET, AND COARSE REGION ATTENTION (CRA) MECHANISM

Methods	MSE
DC-Net (reg)	4.5447
DC-Net (cls)	2.3675
DC-Net (cls)+ CRA	**1.9175**

The DC-Net performs more accurately using both the classification strategy and CRA.

TABLE II
ABLATION STUDY OF THE DP-NET PERFORMANCE EVALUATED BY PSNR AND SSIM

Methods	PSNR	SSIM
Baseline	13.34	0.3475
Baseline + IPC	13.76	0.3622
Baseline + IPC + EF	13.94	0.3806
Baseline + IPC + LF	13.85	0.3755
Baseline + IPC + DF	14.21	0.3862
Baseline + IPC + DF + FRA	**14.63**	**0.3951**

The baseline represents the vanilla version of the DP-Net, while IPC, EF, LF, DF, and FRA indicate the incrementally perceived cascade structure, early fusion, late fusion, deep fusion, and fine region attention, respectively.

(MSE) is chosen as the evaluation metric. In the same manner, we construct a three-region geometric mask based on the mechanism of the fine region attention and the prior knowledge of the omnidirectional distortion. In contrast to the simple structure of the coarse region mask, we define two types of hyperparameters when constructing a three-region geometric mask, i.e., the gray value and range of the regions. Based on the experiments, we found that the performance of the DP-Net remains nearly stable when the gray values of the regions change, and thus we empirically choose $a_0 = 255, b_0 = 200, c_0 = 160$, and $d_0 = 50$. By comparing these values, we set $r_1 = w/8$, $r_2 = w/4$, and $r_3 = w/2$, where w is the width of the omnidirectional image.

To provide more comprehensive features regarding the original information and distortion for our learning model, we utilize a fusion module to generate the attentive aggregation contained content features, as well as the geometric distortion features. The experimental results show that this fine region attention (FRA) mechanism benefits from the performance of the DP-Net, and the evaluation of the proposal based on the peak signal to noise ratio (PSNR) and structural similarity index (SSIM) is shown in Table II. Note that we use the baseline as the evaluated DP-Net discussed in Section IV-C.

Incrementally Perceived Cascade Structure: We constructed an incrementally perceived cascade structure in the DP-Net. Besides, the deep fusion strategy is implemented to hierarchically perceive the relationships among different distortion features. To demonstrate the effects of this structure, an ablation study of the DP-Net performance using five different experiments was performed: the baseline, baseline with incrementally perceived cascade structure (IPC), baseline with IPC and early fusion (EF), baseline with IPC and late fusion (LF), and baseline with IPC and deep fusion (DF). As a benefit of the multiple continuous perception, the deep fusion strategy promotes the communications of different features in regard to the distortion parameters and adequately investigates their ambiguous relationship, achieving the better performance than the early and late fusion strategies. Table II shows that, for the DP-Net, the complete version of the proposed method achieves the best performance for both the PSNR and SSIM.

TABLE III
PERFORMANCE COMPARISON WITH OTHER METHODS FOR THE PSNR AND SSIM

Method Metric	Alemánflores [28]	Santanacedrés [37]	Rong [27]	Ours
PSNR	12.17	12.61	13.12	**14.63**
SSIM	0.2695	0.2921	0.3383	**0.3951**

C. Quantitative Measurement

The OIDC-Net based on the PSNR and SSIM was evaluated with other state-of-the-art algorithms that can automatically correct omnidirectional images without any extra conditions, including Alemánflores [28], Santanacedrés [37], and Rong [27]. All methods were evaluated using our synthesized test set.

The results of the quantitative measurement are listed in Table III. As a benefit of predicting both the distortion parameters and distortion centers based on the unified omnidirectional camera model, and guided by the coarse-to-fine region attention mechanism, our proposed OIDC-Net significantly outperformed both traditional approaches and the CNNs-based method in terms of the PSNR and SSIM. In contrast to the OIDC-Net, the state-of-the-art algorithms poorly performed on the omnidirectional image correction due to their over-simplified camera model and negligence of the prior knowledge.

D. Qualitative Results

To highlight the effect of the distortion correction, we additionally compared our method with the state-of-the-art algorithms developed by Alemánflores [28], Santanacedrés [37], and Rong [27], in terms of the visual appearance. We evaluated these methods using our synthetic omnidirectional images test set, which offers the ground truth of undistorted images. The results of the comparison are shown in Fig. 5. Intuitively, the traditional methods, which rely heavily on hand-crafted feature detection and optimization, such as Alemánflores [28] and Santanacedrés [37], consequently produced the unsatisfactory results. Rong [27] estimated the distortion parameter using the CNNs, and omitted more distortion coefficients with respect to the omnidirectional camera model, only focusing on the learning features from a single omnidirectional image without any prior knowledge. Thus this method produced inaccurate classifications on the more challenging omnidirectional distortions. In contrast, our proposed OIDC-Net taked into account the unified omnidirectional camera model and implicitly obtained the crucial prior knowledge from the coarse-to-fine attention mechanism. Therefore, our methods gains more information in terms of the omnidirectional distortions, achieving the best visual effects among all of the compared methods.

We also compared our method with the state-of-the-art approaches using real omnidirectional images from [49]. The results of the comparison are shown in Fig. 6, and as the figure

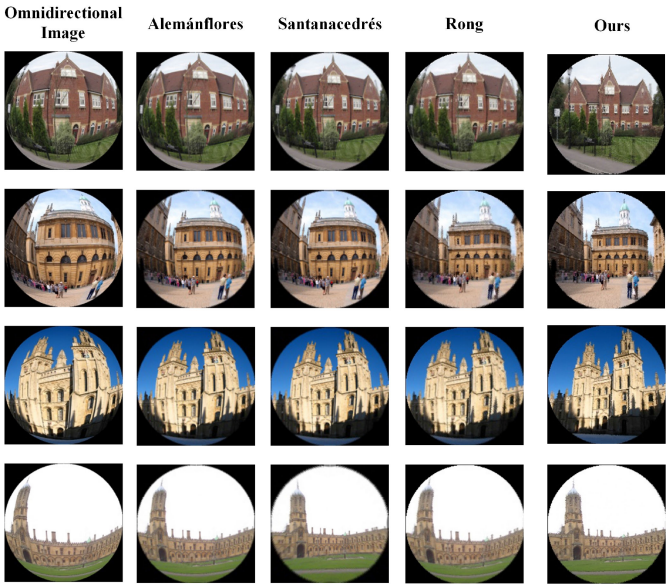

Fig. 5. Qualitative results of synthetic omnidirectional images. For each comparison, we show the omnidirectional image, results of the compared methods, namely, (Alemánflores [28], Santanacedrés [37], and Rong [27]), and the results of our proposed OIDC-Net approach, from left to right.

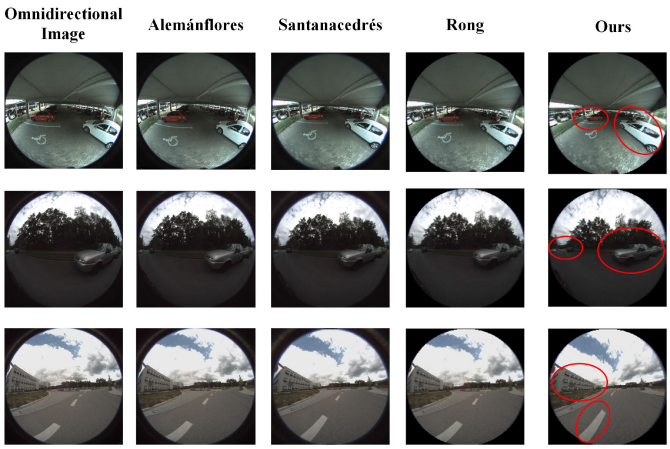

Fig. 6. Qualitative results of real omnidirectional images captured using omnidirectional cameras [49]. For each comparison, we show the omnidirectional image, results of the compared methods (Alemánflores [28], Santanacedrés [37], and Rong [27]), and results of our proposed OIDC-Net approach, from left to right.

indicates, our correction results are more close to the perspective projection especially in the boundary region, where the shapes of objects have been reasonably recovered. Therefore, our proposed OIDC-Net method outperforms the other methods in this qualitative evaluation and facilitates the scene analysis and understanding. As shown in Fig. 7, it is obvious to find that the corrected images obtained by the OIDC-Net gain more accurate object detection and semantic segmentation results, especially in the boundary areas. Our algorithm realistically recovers the real geometric distribution from the severe distortion and thus benefits other computer vision tasks.

Fig. 7. The detection and segmentation results of the original omnidirectional images (top) and corrected images using our proposed method (bottom) (best viewed in color).

VI. Conclusion and Future Work

In this paper, we present a unified and flexible learning model for omnidirectional image distortion correction. In contrast to other traditional approaches, our method avoids the constraint of having to exploit more images and a calibration pattern, and as a result outperforms the traditional methods in terms of the flexibility. Compared with methods that employ CNNs, we leverage the coarse-to-fine region attention mechanism and construct the attentive aggregation using a fusion module, which contains the original content features and geometric distortion features. Moreover, an incrementally perceived cascade structure is able to improve the accuracy of the estimation of the parameters. Finally we implement the deep fusion strategy to further explore the perception in relation to different parameters. In future work, we intend to expand the dataset with a wider range of coefficients, and aim to determine a method of eliminating the difficulty in accurately estimating the heterogeneous target values.

References

[1] B. Micusík and T. Pajdla, "Structure from motion with wide circular field of view cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 7, pp. 1135–1149, Jul. 2006.
[2] J. C. Bazin, C. Demonceaux, P. Vasseur, and I.-S. Kweon, "Motion estimation by decoupling rotation and translation in catadioptric vision," *Comput. Vis. Image Understanding*, vol. 114, pp. 254–273, 2010.
[3] S. Wang *et al.*, "Torontocity: Seeing the world with a million eyes," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 3028–3036.
[4] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, 2016.
[5] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, "Predicting head movement in panoramic video: A deep reinforcement learning approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2693–2708, Nov. 2019.
[6] H. Yang and H. Zhang, "Efficient 3D room shape recovery from a single panorama," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 5422–5430.
[7] B. Mičušík and T. Pajdla, "Estimation of omnidirectional camera model from epipolar geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2003, pp. 485–490.
[8] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[9] C. Geyer and K. Daniilidis, "Paracatadioptric camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 687–695, May 2002.
[10] X. Ying and Z. Hu, "Catadioptric camera calibration using geometric invariants," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 10, pp. 1260–1271, Oct. 2004.

[11] J. P. Barreto and H. Araújo, "Geometric properties of central catadioptric line images and their application in calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1327–1333, Aug. 2005.

[12] X. Ying and H. Zha, "Simultaneously calibrating catadioptric camera and detecting line features using Hough transform," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2005, pp. 412–417.

[13] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[14] C. Mei and P. Rives, "Single view point omnidirectional camera calibration from planar grids," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2007, pp. 3945–3950.

[15] S. Gasparini, P. F. Sturm, and J. P. Barreto, "Plane-based calibration of central catadioptric cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1195–1202.

[16] S. Shah and J. K. Aggarwal, "A simple calibration procedure for fish-eye (high-distortion) lens camera," in *Proc. IEEE Int. Conf. Robot. Autom.*, 1994, pp. 3422–3427.

[17] S. B. Kang, "Catadioptric self-calibration," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 1, Hilton Head Island, SC, USA, 2000, pp. 201–207.

[18] S. Ramalingam, P. F. Sturm, and S. K. Lodha, "Generic self-calibration of central cameras," *Comput. Vis. Image Understanding*, vol. 114, pp. 210–219, 2010.

[19] A. W. Fitzgibbon, "Simultaneous linear estimation of multiple view geometry and lens distortion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Kauai, HI, USA, 2001, pp. I–I.

[20] F. Espuny, "Generic self-calibration of central cameras from two rotational flows," *Int. J. Comput. Vis.*, vol. 91, pp. 131–145, 2007.

[21] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[22] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.

[23] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1653–1660.

[24] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1302–1310.

[25] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 1, 2016.

[26] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 105–114.

[27] J. Rong, S. Huang, Z. Shang, and X. Ying, "Radial lens distortion correction using convolutional neural networks trained with synthesized images," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 35–49.

[28] M. Alemánflores, L. Alvarez, L. Gomez, and D. Santanacedrés, "Automatic lens distortion correction using one-parameter division models," *Image Process. Line*, vol. 4, pp. 327–343, 2014.

[29] X. Yin, X. Wang, J. Yu, M. Zhang, P. Fua, and D. Tao, "FishEyeRecNet: A multi-context collaborative deep network for fisheye image rectification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 469–484.

[30] F. Ddvernay, "Straight lines have to be straight : Automatic calibration and removal of distortion from scenes of structured environments," *Mach. Vis. Appl.*, vol. 13, no. 1, pp. 14–24, 2008.

[31] R. I. Hartley and S. B. Kang, "Parameter-free radial distortion correction with centre of distortion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1309–1321, Aug. 2007.

[32] Y. Gao, C. Lin, Y. Zhao, X. Wang, S. Wei, and Q. Huang, "3-D surround view for advanced driver assistance systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 320–328, Jan. 2018.

[33] J. Wei, C. F. Li, S. M. Hu, R. R. Martin, and C. L. Tai, "Fisheye video correction," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 10, pp. 1771–1783, Oct. 2012.

[34] R. Carroll, M. Agrawal, and A. Agarwala, "Optimizing content-preserving projections for wide-angle images," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 1–9, 2009.

[35] F. Bukhari and M. N. Dailey, "Automatic radial distortion estimation from a single image," *J. Math. Imag. Vis.*, vol. 45, no. 1, pp. 31–45, 2013.

[36] A. W. Fitzgibbon, "Simultaneous linear estimation of multiple view geometry and lens distortion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Kauai, HI, USA, 2001, pp. I–I.

[37] D. Santanacedrés *et al.*, "An iterative optimization algorithm for lens distortion correction using two-parameter models," *Image Process. Line*, vol. 6, pp. 326–365, 2016.

[38] C. Geyer and K. Daniilidis, "A unifying theory for central panoramic systems and practical applications," in *Proc. Eur. Conf. Comput. Vis.*, 2000, pp. 445–461.

[39] X. Ying and Z. hu, "Can we consider central catadioptric cameras and fisheye cameras within a unified imaging model," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany, 2004, pp. 442–455.

[40] J. Courbon, Y. Mezouar, L. Eck, and P. Martinet, "A generic fisheye camera model for robotic applications," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2007, pp. 1683–1688.

[41] S. Ramalingam, P. Sturm, and S. K. Lodha, "Towards complete generic camera calibration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 1093–1098.

[42] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2006, pp. 5695–5701.

[43] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 354–370.

[44] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 826–834.

[45] J. Wang, Z. Wei, T. Zhang, and W. Zeng, "Deeply-fused nets," Preprint, 2016, *arXiv:1605.07716*.

[46] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

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

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[49] A. Eichenseer and A. Kaup, "A data set providing synthetic and real-world fisheye video sequences," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 1541–1545.