A Deep Ordinal Distortion Estimation Approach for Distortion Rectification

Kang Liao, *Graduate Student Member, IEEE*, Chunyu Lin, *Member, IEEE*, and Yao Zhao, *Senior Member, IEEE*

***Abstract*— Radial distortion has widely existed in the images captured by popular wide-angle cameras and fisheye cameras. Despite the long history of distortion rectification, accurately estimating the distortion parameters from a single distorted image is still challenging. The main reason is that these parameters are implicit to image features, influencing the networks to learn the distortion information fully. In this work, we propose a novel distortion rectification approach that can obtain more accurate parameters with higher efficiency. Our key insight is that distortion rectification can be cast as a problem of learning an *ordinal distortion* from a single distorted image. To solve this problem, we design a local-global associated estimation network that learns the ordinal distortion to approximate the realistic distortion distribution. In contrast to the implicit distortion parameters, the proposed ordinal distortion has a more explicit relationship with image features, and significantly boosts the distortion perception of neural networks. Considering the redundancy of distortion information, our approach only uses a patch of the distorted image for the ordinal distortion estimation, showing promising applications in efficient distortion rectification. In the distortion rectification field, we are the first to unify the heterogeneous distortion parameters into a learning-friendly intermediate representation through ordinal distortion, bridging the gap between image feature and distortion rectification. The experimental results demonstrate that our approach outperforms the state-of-the-art methods by a significant margin, with approximately 23% improvement on the quantitative evaluation while displaying the best performance on visual appearance.**

***Index Terms*— Distortion rectification, neural networks, learning representation, ordinal distortion.**

I. Introduction

IMAGES captured by wide-angle camera usually suffer from a strong distortion, which influences the important scene perception tasks such as the object detection and recognition [1]–[3], semantic segmentation [4], [5], and image denoising [6], [7]. The distortion rectification tries to recover the real geometric attributes from distorted scenes. It is a fundamental and indispensable part of image processing, which has a long research history extending back 60 years. In recent, distortion rectification through deep learning has attracted increasing attention [8]–[14].

Accurately estimating the distortion parameters derived from a specific camera, is a crucial step in distortion rectification. However, two main limitations that make the distortion parameters learning challenging. (i) The distortion parameters are not observable and hard to learn from a single distorted image, such as the principal point and distortion coefficients. Compared with the intuitive targets, such as the object classification and bounding box detection studied in other research regions, the distortion parameters have a more complicated and implicit relationship with image features. As a result, the neural networks obtain an ambiguous and insufficient distortion perception, which leads to inaccurate estimation and poor rectification performance. (ii) The different components of distortion parameters have different magnitudes and ranges of values, showing various effects on an image's global distortion distribution. Such a heterogeneous representation confuses the distortion cognition of neural networks and causes a heavy imbalance problem during the training process.

To overcome the above limitations, previous methods exploit more guided features such as the semantic information and distorted lines [9], [10], or introduce the pixel-wise reconstruction loss [11]–[13]. However, the extra features and supervisions impose increased memory/computation cost. In this work, we would like to draw attention from the traditional calibration objective to a learning-friendly perceptual target. The target is to unify the implicit and heterogeneous parameters into an intermediate representation, thus bridging the gap between image feature and distortion estimation in the field of distortion rectification.

In particular, we redesign the whole pipeline of deep distortion rectification and present an intermediate representation based on the distortion parameters. The comparison of the previous methods and the proposed approach is illustrated in Fig. 1. Our key insight is that distortion rectification can be cast as a problem of learning an *ordinal distortion* from a distorted image. The ordinal distortion indicates the distortion levels of a series of pixels, which extend outward from the principal point. To predict the ordinal distortion, we design a local-global associated estimation network optimized with an ordinal distortion loss function. A distortion-aware perception

Manuscript received July 20, 2020; revised January 3, 2021 and February 18, 2021; accepted February 18, 2021. Date of publication March 1, 2021; date of current version March 5, 2021. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2020YJS024 and in part by the National Natural Science Foundation of China under Grant 61772066 and Grant 61972028. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Abdesselam S. Bouzerdoum. *(Corresponding author: Chunyu Lin.)*

The authors are with the Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China, and also with the Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China (e-mail: kang_liao@bjtu.edu.cn; cylin@bjtu.edu.cn; yzhao@bjtu.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3061283

layer is exploited to boost the feature extraction of different degrees of distortion.

The proposed learning representation offers three unique advantages. First, the ordinal distortion is directly perceivable from a distorted image, and it solves a more straightforward estimation problem than the implicit metric regression. As we can observe, the farther the pixel is away from the principal point, the larger the distortion degree is, and vice versa. This prior knowledge enables the neural networks to build a clear cognition with respect to the distortion distribution. Thus, the learning model gains a sufficient distortion perception of image features and shows faster convergence, without any extra feature guidances and pixel-wise supervisions.

Second, the ordinal distortion is homogeneous as all its elements share a similar magnitude and description. Therefore, the imbalanced optimization problem no longer exists during the training process, and we do not need to focus on the cumbersome factor-balancing task anymore. Compared to the distortion parameters with different types of components, our learning model only needs to consider one optimization objective, thus achieving more accurate estimation and more realistic rectification results.

Third, the ordinal distortion can be estimated using only a part of a distorted image. Unlike the semantic information, the distortion information is redundant in images, showing the central symmetry and mirror symmetry to the principal point. Consequently, the efficiency of rectification algorithms can be significantly improved when taking the ordinal distortion estimation as a learning target. More importantly, the ordinal relationships are invariant to monotonic transformations of distorted images, thereby increasing the robustness of the rectification algorithm.

With lots of experiments, we verify that the proposed ordinal distortion is more suitable than the distortion parameters as a learning representation for deep distortion rectification. The experimental results also show that our approach outperforms the state-of-the-art methods with a large margin, approximately 23% improvement on the quantitative evaluation while using fewer input images, demonstrating its efficiency on distortion rectification.

The rest of this paper is organized as follows. We first introduce the related work in Section II. We then present our approach in Section III. The experiments are provided in Section IV. Finally, we conclude this paper in Section V.

II. Related Work

In this section, we briefly review the previous distortion rectification methods and classify them into two groups: the traditional vision-based one and the deep learning one.

A. Traditional Distortion Rectification

There is a rich history of exploration in the field of distortion rectification. The most common method is based on a specific physical model. [15]–[17] utilized a camera to capture several views of a 2D calibration pattern that covered points, corners, or other features, and then computed the distortion parameters of the camera. However, these methods cannot handle images captured by other cameras and thus are restricted to the application scenario. Self-calibration was leveraged for distortion parameter estimation in [18]–[20]; however, the authors failed in the geometry recovery using only a single image. To overcome the above limitations and achieve automatic distortion rectification, Bukhari *et al.* [21] employed a one-parameter camera model [22] and estimated distortion parameters using the detected circular arcs. Similarly, [23], [24] also utilized the simplified camera model to correct the radial distortion in images. However, these methods perform poorly on scenes that are lacking enough hand-crafted features. Thus, the above traditional methods are difficult to handle on the single distorted image rectification in various scenes.

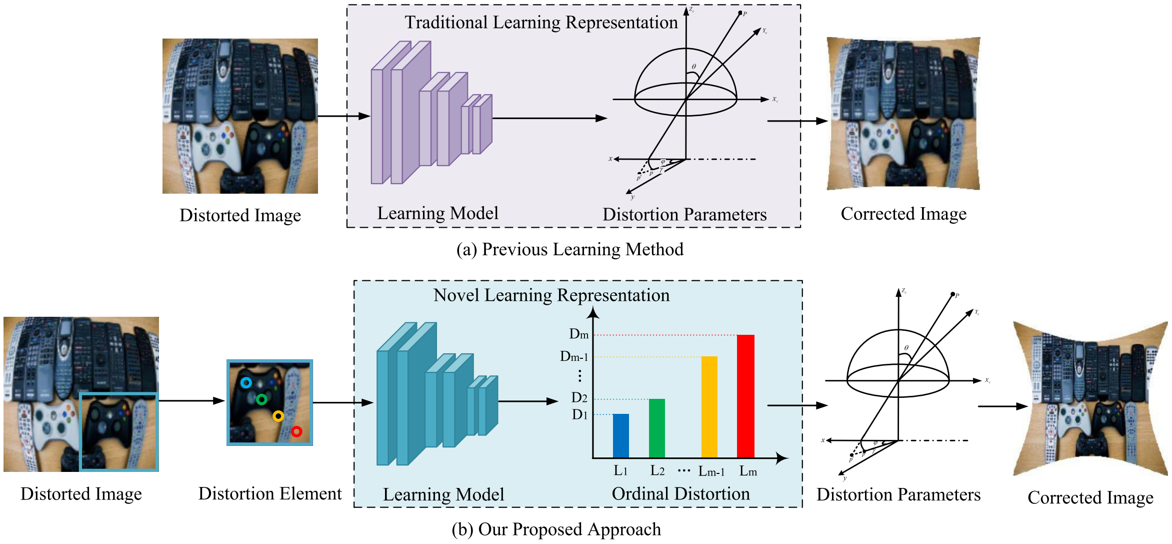
B. Deep Distortion Rectification

In contrast to the long history of traditional distortion rectification, learning methods began to study distortion rectification in the last few years. Rong *et al.* [8] quantized the values of the distortion parameter to 401 categories based on the one-parameter camera model [22] and then trained a network to classify the distorted image. This method achieved the deep distortion rectification for the first time, while the coarse values of parameters and the simplified camera model severely influenced its generalization ability. To expand the application, Yin *et al.* [9] rectified the distortion in terms of the fisheye camera model using a multi-context collaborative deep network. However, their correction results heavily rely on the semantic segmentation results, leading to a strong cascading effect. Xue *et al.* [10] improved the performance of distortion parameter estimation by distorted lines. In analogy to traditional methods [21], [23], [24], the extra introduced hand-crafted features limit the robustness of this algorithm and decrease the efficiency of the rectification. Note that the above methods directly estimates distortion parameters from a single distorted image, such an implicit and heterogeneous calibration objective hinders sufficient learning concerning the distortion information. To solve the imbalance problem in the distortion parameter estimation, recent works [11]–[13] optimized the image reconstruction loss rather than the parameters regression loss for rectification. However, their models are based on the parameter-free mechanism and cannot estimate the distortion parameters, which are important for the structure from motion and camera calibration. Manuel *et al.* [14] proposed a parameterization scheme for the extrinsic and intrinsic camera parameters, but they only considered one distortion coefficient for the rectification and cannot apply the algorithm to more complicated camera models.

With the proposed intermediate representation, i.e., ordinal distortion, our approach can boost the efficient learning of neural networks and eliminate the imbalance problem, obtaining accurate parameters for better rectification performance.

III. Approach

In this section, we describe how to learn the ordinal distortion given a single distorted image. We first define the proposed objective in Section III-A. Next, we introduce the



(a) Previous Learning Method

(b) Our Proposed Approach

Fig. 1. Method Comparisons. (a) Previous learning methods, (b) Our proposed approach. We aim to transfer the traditional calibration objective into a learning-friendly representation. Previous methods roughly feed the whole distorted image into their learning models and directly estimate the implicit and heterogeneous distortion parameters. In contrast, our proposed approach only requires a part of a distorted image (distortion element) and estimates the ordinal distortion. Due to its explicit description and homogeneity, we can obtain more accurate distortion estimation and achieve better corrected results.

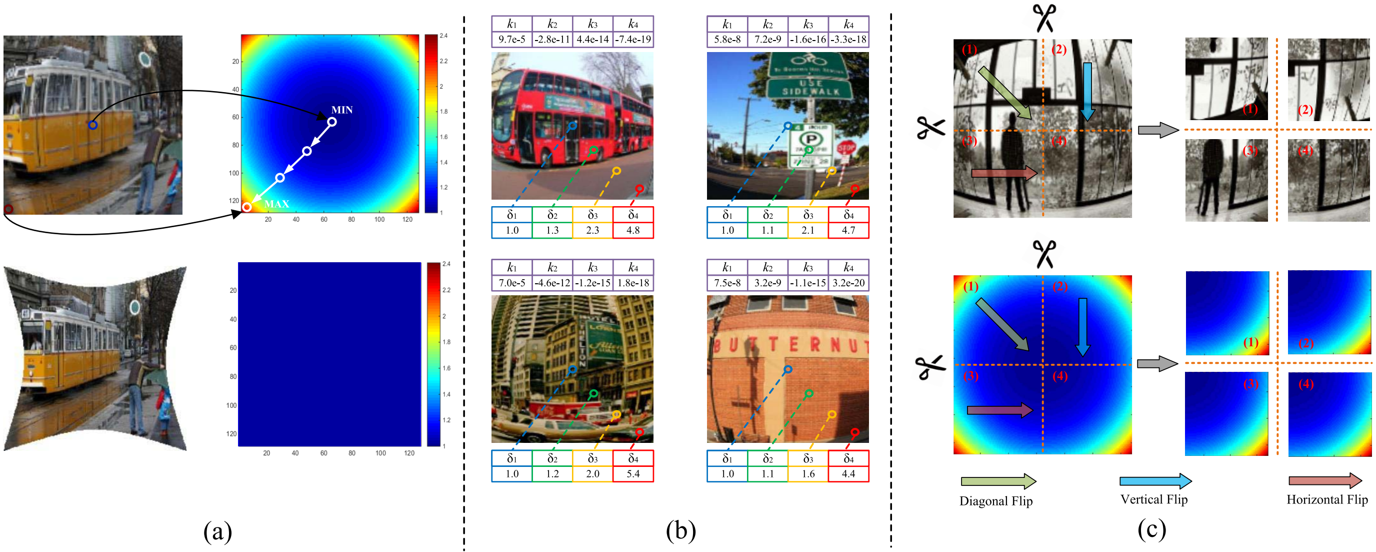


Fig. 2. Attributes of the proposed ordinal distortion. (a) Explicitness. The ordinal distortion is observable in an image and explicit to image features, which describes a series of distortion levels from small to large (top); the ordinal distortion always equals one in an undistorted image (bottom). (b) Homogeneity. Compared with the heterogeneous distortion parameters $\mathcal{K} = [k_1 \;\; k_2 \;\; k_3 \;\; k_4]$, the ordinal distortion $\mathcal{D} = [\delta_1 \;\; \delta_2 \;\; \delta_3 \;\; \delta_4]$ is homogeneous, representing the same concept of distortion distribution. (c) Redundancy. After different flip operations, although the semantic features of four patches have not any relevance (top), the ordinal distortion of four patches keeps the same in distribution with each other (bottom).

network architecture and training loss in Section III-B. Finally, Section III-C describes the transformation between the ordinal distortion and distortion parameter.

A. Problem Definition

1) Parameterized Camera Model: We assume that a point in the distorted image is expressed as $\mathbf{P} = [x, y]^{\mathrm{T}} \in \mathbb{R}^2$ and a corresponding point in the corrected image is expressed as $\mathbf{P}' = [x', y']^{\mathrm{T}} \in \mathbb{R}^2$. The polynomial camera model can be described as

$$\begin{aligned} x' &= x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6 + k_4 r^8 + \cdots) \\ y' &= y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6 + k_4 r^8 + \cdots), \end{aligned} \qquad (1)$$

where $[k_1 \;\; k_2 \;\; k_3 \;\; k_4 \;\; \cdots]$ are the distortion coefficients, r is the Euclidean distance between the point $\mathbf{P}$ and the principal point $\mathbf{C} = [x_c, y_c]^{\mathrm{T}}$ in the distorted image, which can be expressed as

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

This polynomial camera model fits well for small distortions but requires more distortion parameters for severe distortions. As an alternative camera model, the division model is formed by:

$$x' = \frac{x}{1 + k_1 r^2 + k_2 r^4 + k_3 r^6 + k_4 r^8 + \cdots}$$

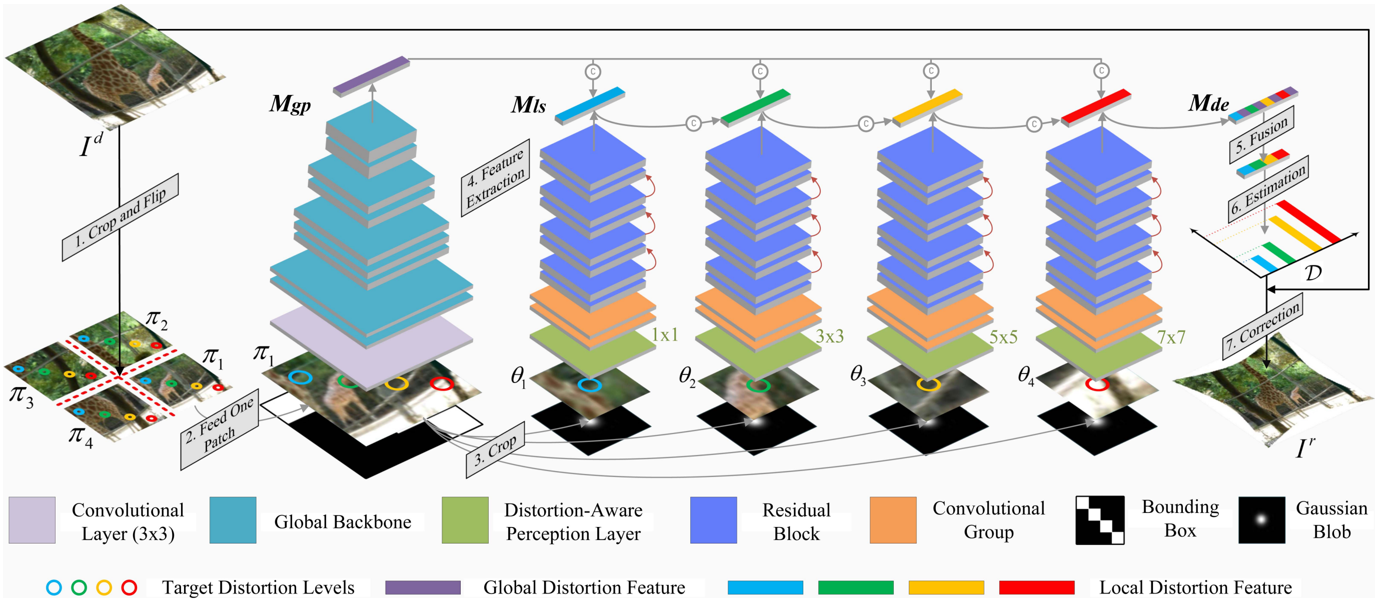


Fig. 3. Architecture of the proposed network. This network consists of a global perception module M_{gp}, local Siamese module M_{ls}, and distortion estimation module M_{de}. During the network's training process, we use four parts, i.e., distortion elements: $\Pi = [\pi_1 \ \ \pi_2 \ \ \pi_3 \ \ \pi_4]$ of the distorted image I^d to train its ability of distortion perception sequentially, in which the distortion blocks: $\Theta = [\theta_1 \ \ \theta_2 \ \ \theta_3 \ \ \theta_4]$ derived from a distortion element provide the local distortion information. In the test or application stage, we only need one part of the input distorted image to estimate the ordinal distortion $\mathcal{D}$. Finally, the rectified image I^r can be generated using the estimated ordinal distortion and the input distorted image.

$$y' = \frac{y}{1 + k_1 r^2 + k_2 r^4 + k_3 r^6 + k_4 r^8 + \cdots}. \qquad (3)$$

Compared with the polynomial camera model, the division model requires fewer parameters in terms of the strong distortion and is thus more suitable for approximating wide-angle cameras [25].

2) Ordinal Distortion: As mentioned above, most previous learning methods correct the distorted image based on the distortion parameters estimation. However, due to the implicit and heterogeneous representation, the neural network suffers from the insufficient learning problem and imbalance regression problem. These problems seriously limit the learning ability of neural networks and cause inferior distortion rectification results. To address the above problems, we propose a fully novel concept, i.e., ordinal distortion. Fig. 2 illustrates the attributes of the proposed ordinal distortion.

The ordinal distortion represents the image feature in terms of the distortion distribution, which is jointly determined by the distortion parameters and location information. We assume that the camera model is the division model, and the ordinal distortion $\mathcal{D}$ can be defined as

$$\mathcal{D} = [\delta(r_1) \ \ \delta(r_2) \ \ \delta(r_3) \ \ \cdots \ \ \delta(r_n)], \\ 0 \leq r_1 < r_2 < r_3 < \cdots < r_n \leq R, \qquad (4)$$

where R is the maximum distance between a point and the principal point, $\delta(\cdot)$ indicates the distortion level of a point P_i in the distorted image:

$$\delta(r_i) = \frac{x_i}{x_i'} = \frac{y_i}{y_i'} = 1 + k_1 {r_i}^2 + k_2 {r_i}^4 + k_3 {r_i}^6 + k_4 {r_i}^8 + \cdots. \qquad (5)$$

Intuitively, the distortion level expresses the ratio between the coordinates of $\mathbf{P}$ and $\mathbf{P}'$. The larger the distortion level is, the stronger the distortion of a pixel is, and vice versa. For an undistorted or ideally rectified image, $\delta(\cdot)$ always equals 1. Therefore, the ordinal distortion represents the distortion levels of pixels in a distorted image, which increases outward from the principal point sequentially.

We assume the width and height of a distorted image are W and H, respectively. Then, the distortion level satisfies the following equation:

$$\begin{aligned} \delta(x_i, y_i) &= \delta(\frac{W}{2} - x_i + x_c, y_i) = \delta(x_i, \frac{H}{2} - y_i + y_c) \\ &= \delta(\frac{W}{2} - x_i + x_c, \frac{H}{2} - y_i + y_c). \end{aligned} \qquad (6)$$

Thus, the ordinal distortion displays the mirror symmetry and central symmetry to the principal point in a distorted image. This prior knowledge ensures less data required in the ordinal distortion estimation process.

B. Network

Our network consists of three main modules: global perception module M_{gp}, local Siamese module M_{ls}, and distortion estimation module M_{de} as shown in Fig. 3. The first module extracts the global distortion features from a patch of the input distorted image. The second module extracts the local distortion features from a series of distortion blocks, corresponding to the different distortion levels. The final one fuses the extracted global and local distortion features and estimates the proposed ordinal distortion.

1) Network Input: The network input includes two parts. The first is the global distortion context, which provides a distortion element $\pi_i \in \Pi$ with the overall distortion information. The second is the local distortion context, which

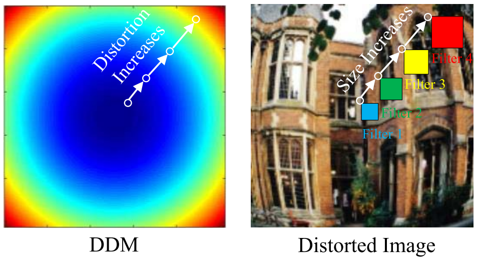


Fig. 4. Motivation of the designed distortion-aware perception layer. Left: the distortion distribution map (DDM) that describes the degree of distortion for each pixel. Right: the corresponding distorted image. Particularly, we use the filters with increasing sizes to perceive the increasing degrees of distortions along the extended path (the white arrows).

provides the distortion blocks $\Theta = [\theta_1 \quad \theta_2 \quad \theta_3 \quad \cdots \quad \theta_n]$ with the detailed distortion levels. Considering the principal point is slightly disturbed in the image center, we first cut the distorted image into four patches along the center of the image, and dub these patches as distortion elements $\Pi = [\pi_1 \quad \pi_2 \quad \pi_3 \quad \pi_4]$ with size of $\frac{W}{2} \times \frac{H}{2} \times 3$. Although most distortion information covers in one patch, the distortion distribution of each patch is spatially different. To normalize this diversity, we flip three of the four elements to keep a similar distortion distribution with that of the selected one. As shown in Fig. 2 (c), the top left, top right, and bottom left distortion parts are handled with the diagonal, vertical, and horizontal flip operations, respectively.

We further crop a distortion element into the distortion blocks $\Theta = [\theta_1 \quad \theta_2 \quad \theta_3 \quad \cdots \quad \theta_n]$, in which a block θ_i is leveraged to provide the local distortion feature and predict the distortion level δ_i. To boost neural networks to learn the distortion features, we construct the masks consisting of the bounding boxes $\mathcal{M}_B \in \mathbb{R}^{w_b \times h_b \times 1}$ and Gaussian blobs $\mathcal{M}_G \in \mathbb{R}^{w_g \times h_g \times 1}$ of the distortion blocks, where $w_b = \frac{W}{2}, h_b = \frac{H}{2}, w_g = \frac{W}{2n}, h_g = \frac{H}{2n}$. Concretely, the mask represents the Region of Interest (RoI) of input data, which offers the ranges of global and local distortion information for $\mathcal{M}_B$ and $\mathcal{M}_G$:

$$\mathcal{M}_B(p, q) = \begin{cases} 255, & \mathit{if}\ (p, q) \in \Omega. \\ 0, & \mathit{otherwise}. \end{cases} \tag{7}$$

where $\Omega = \{(p, q) | (t-1)w_g \leq p \leq tw_g, (t-1)h_g \leq q \leq th_g\}$ is the region of distortion blocks and $t \in \mathbb{Z}$ derives from the range of $[1, n]$. For $\mathcal{M}_G$, the mask value can be weighted by the Gaussian Distribution $X \sim \mathcal{N}(\mu, \sigma^2)$:

$$\mathcal{M}_G(d) = 255 \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(d-\mu)^2}{2\sigma^2}}, \tag{8}$$

where d indicates the Euclidean distance between a pixel and the center of the distortion block. In our implementation, μ and σ are set to 0 and 1 respectively.

2) Network Architecture:

a) Global perception module: For the global perception module, its architecture can be divided into two sub-networks, a backbone network, and a header network. Specifically, the general representation of the global distortion context is extracted using the backbone network composed of convolutional layers. This representation indicates the high-level information including the semantic features. Any prevalent networks such as VGG16 [26], ResNet [27], and InceptionV3 [28] (without fully connected layers) can be plugged into the backbone network. We pretrain the backbone network on ImageNet [29] and fine-tune on our synthesized distorted image dataset. The header network contains three fully connected layers. It aggregates the input's general representation and further abstracts the high-level information in the form of a feature vector, which . The numbers of units for these layers are 4096, 2048, and 1024. The activation functions for all of the fully connected layers are ReLUs. The extracted features of the global distortion context, dubbed as $\mathcal{F}_g$, are combined with the features of the local distortion context, derived from the local Siamese module.

b) Local siamese module: The local Siamese module consists of n components, each component also can be divided into a backbone network and a header network. In detail, we first use two convolutional layers to extract the low-level features from the input local distortion context. Then, we feed the feature maps into a pyramid residual module consisting of five residual blocks and get the high-level features. The pyramid residual module shares the weights in each component. Subsequently, a header network with three fully connected layers aggregates the general representation of the local distortion features: $[\mathcal{F}_l^{(1)} \quad \mathcal{F}_l^{(2)} \quad \cdots \quad \mathcal{F}_l^{(n)}]$, which are one-to-one correspondence to the estimated ordinal distortion $\hat{\mathcal{D}} = [\hat{\delta}(r_1) \quad \hat{\delta}(r_2) \quad \cdots \quad \hat{\delta}(r_n)]$.

Having observed that the distortion degree increases with the distance of a pixel to the principal point, we design a distortion-aware perception layer to extract the different distortion features. The motivation is illustrated in Fig. 4. In general, the filter size indicates the size of the receptive field, which determines the context of reasoning features. Therefore, it is more reasonable to grasp the prior knowledge of the variable distortion distribution using filters with different sizes instead of the same size. In our implementation, for different distortion blocks of a patch along the extended path (the white arrows in Fig. 4), we use convolutional filters with increasing sizes to extract the distortion features. Concretely, the distortion-aware perception layer is applied before feeding the input contexts to the network. For the local distortion context, the distortion blocks $\Theta = [\theta_1 \quad \theta_2 \quad \cdots \quad \theta_n]$ are processed using the filters with sizes of $W_{l1} \times H_{l1}, W_{l2} \times H_{l2}, \cdots, W_{ln} \times H_{ln}$, from small to large. Namely, all sizes of filters satisfies the following relationship: $W_{l1} \times H_{l1} < W_{l2} \times H_{l2} < \cdots < W_{ln} \times H_{ln}$. As a result, our learning model can explicitly perceive the different degrees of distortions in a distorted image, thus achieving a better approximation of ordinal distortion. The relevant experimental results will be described in Section IV-C.

c) Distortion estimation module: To comprehensively reason the distortion information, we combine each local distortion feature with the global distortion feature and fuse these features using two fully connected layers f, which constructs a hybrid feature vector $\mathcal{F}_h$:

$$\mathcal{F}_h = f([\mathcal{F}_l^{(1)} \oplus \mathcal{F}_g \quad \mathcal{F}_l^{(2)} \oplus \mathcal{F}_g \quad \cdots \quad \mathcal{F}_l^{(n)} \oplus \mathcal{F}_g]), \tag{9}$$

where $\oplus$ represents the concatenation operation. Finally, a fully connected layer F with the unit number of n and linear activation function takes the $\mathcal{F}_h$ as input, estimating the ordinal distortion $\hat{\mathcal{D}}$ of a distorted image by

$$\hat{\mathcal{D}} = [\hat{\delta}(r_1)\quad \hat{\delta}(r_2)\quad \cdots\quad \hat{\delta}(r_n)] = F(\mathcal{F}_h). \tag{10}$$

3) Training Loss: After predicting the distortion labels of a distorted image, it is direct to use the distance metric loss such as $\mathcal{L}_1$ loss or $\mathcal{L}_2$ loss to learn the network parameters. Nevertheless, these loss functions cannot measure the ordered relationship in the distortion labels, while the proposed ordinal distortion possesses a strong ordinal correlation in terms of the distortion distribution. To this end, we regard the distortion estimation problem as an ordinal distortion regression problem and design an ordinal distortion loss to train our learning model.

Suppose that the ground truth ordinal distortion $\mathcal{D} = [\delta(r_1)\quad \delta(r_2)\quad \cdots\quad \delta(r_n)]$ is an increasing vector, which means $\delta(r_1) < \delta(r_2) < \cdots < \delta(r_n)$. Due to the available distortion parameters in dataset, we can easily get the ground truth of ordinal distortion of any single image based on Eq. 5. Recall that $\mathcal{F}_g$ indicates the global distortion feature which is extracted by the global perception module M_{gp}; $[\mathcal{F}_l^{(1)}\quad \mathcal{F}_l^{(2)}\quad \cdots\quad \mathcal{F}_l^{(n)}]$ indicate the local distortion features which are extracted by the local Siamese module M_{ls}. Subsequently, a distortion estimation module M_{de} fuses the global feature and local features into a hybrid feature vector $\mathcal{F}_h$, that is used to predict the target ordinal distortion $\hat{\mathcal{D}} = [\hat{\delta}(r_1)\quad \hat{\delta}(r_2)\quad \cdots\quad \hat{\delta}(r_n)]$. Let ξ contains the weights of the final fully connected layer F, and then the ordinal distortion loss $\mathcal{L}(\mathcal{F}_h, \xi)$ can be described by the following formulation over the entire sequence:

$$\mathcal{L}(\mathcal{F}_h, \xi) = \frac{1}{n}\sum_{i=1}^{n}(1 + \mathcal{C}_o)\mathcal{L}_d(i, \mathcal{F}_h, \xi). \tag{11}$$

The term $\mathcal{C}_o$ is to weight the loss function and measures the ordinal correlation in $\hat{\mathcal{D}}$:

$$\mathcal{C}_o = \sum_{k=1}^{i}\log(\mathcal{P}_i^k) + \sum_{k=i+1}^{n}\log(1 - \mathcal{P}_i^k), \tag{12}$$

where $\mathcal{P}_i^k = P(\hat{\delta}(r_i) > \hat{\delta}(r_k))$ indicates the probability that $\hat{\delta}(r_i)$ is larger than $\hat{\delta}(r_k)$. $\mathcal{L}_d(i, \mathcal{F}_h, \xi)$ minimizes the difference between the $\hat{\mathcal{D}}$ and the ground truth $\mathcal{D}$ based on the smooth $\mathcal{L}_1$ measurement [30]:

$$\mathcal{L}_d(i, \mathcal{F}_h, \xi) = \begin{cases} 0.5\Phi_i^2, & \textit{if}\ \ |\Phi_i| \leq 1. \\ |\Phi_i| - 0.5, & \textit{otherwise}, \end{cases} \tag{13}$$

where $\Phi_i = \hat{\delta}(r_i) - \delta(r_i)$. The $\mathcal{L}_d$ with smooth $\mathcal{L}_1$ measurement can be cast as the composition of the $\mathcal{L}_1$ and $\mathcal{L}_2$ losses, which can eliminate the exploding gradient problem during the training process. Therefore, our ordinal distortion loss function reasons both the increasing ordinal correlation in the predicted elements and the accurate distortion levels.

Algorithm 1 Training Process of the Proposed Network

```
Input: Distorted Image $I^d$
Output: Ordinal Distortion $\hat{\mathcal{D}} = [\delta(\hat{r_1})\ \ \delta(\hat{r_2})\ \ \cdots\ \ \delta(\hat{r_n})]$
Crop and flip $I^d$ into four distortion elements $\Pi = [\pi_1\ \ \pi_2\ \ \pi_3\ \ \pi_4]$
repeat
    for all $\pi_i \in \Pi$ do
        Generate the feature vector of overall distortion $\mathcal{F}_g^{(i)}$ using the global perception module $M_{gp}$: $\mathcal{F}_g^{(i)} \leftarrow M_{gp}(\pi_i)$
        Crop $\pi_i$ into same size of distortion blocks $\Theta^{(i)} = [\theta_1^{(i)}\ \ \theta_2^{(i)}\ \ \cdots\ \ \theta_n^{(i)}]$
        for all $\theta_j^{(i)} \in \Theta^{(i)}$ do
            Generate the feature vector of detailed distortion $\mathcal{F}_l^{(i,j)}$ using the local Siamese module $M_{ls}$: $\mathcal{F}_l^{(i,j)} \leftarrow M_{ls}(\theta_j^{(i)})$
        end for
        Generate the hybrid feature vector $\mathcal{F}_h^{(i)}$ by fusing $\mathcal{F}_g^{(i)}$ and $[\mathcal{F}_l^{(i,1)}\ \ \mathcal{F}_l^{(i,2)}\ \ \cdots\ \ \mathcal{F}_l^{(i,n)}]$ based on Eq. 9
        Estimate the $\hat{\mathcal{D}}$ using $\mathcal{F}_h^{(i)}$ based on Eq. 10
        Update the parameters of neural network by optimizing the difference of $\hat{\mathcal{D}}$ and the ground truth $\mathcal{D}$ based on Eq. 11
    end for
until Convergence
```

C. Ordinal Distortion to Distortion Parameter

Once the ordinal distortion is estimated by neural networks, the distortion coefficients $\mathcal{K} = [k_1\quad k_2\quad \cdots\quad k_n]$ of a distorted image can be easily obtained by

$$\begin{bmatrix} k_1 & k_2 & \cdots & k_n \end{bmatrix} = \begin{bmatrix} \delta(r_1) - 1 \\ \delta(r_2) - 1 \\ \vdots \\ \delta(r_n) - 1 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} r_1^2 & r_2^2 & \cdots & r_n^2 \\ r_1^4 & r_2^4 & \cdots & r_n^4 \\ \vdots & \vdots & \ddots & \vdots \\ r_1^{2n} & r_2^{2n} & \cdots & r_n^{2n} \end{bmatrix}^{-1}. \tag{14}$$

For clarity, we rewrite Eq. 14 as follows:

$$\mathcal{K} = \mathcal{D}^* \cdot \mathcal{R}^{-1}, \tag{15}$$

where $\mathcal{D}^* = \hat{\mathcal{D}} - [\underbrace{1\quad 1\quad \cdots\quad 1}_{n}]$ and $\hat{\mathcal{D}}$ expresses the estimated ordinal distortion, and the location information with different powers is included in $\mathcal{R}$.

Finally, the rectified image can be warped by each pixel of the distorted image using the computed distortion parameters based on Eq. 1 or Eq. 3.

In summary, we argue that by presenting our distortion rectification framework, we can have the following advantages.

1. The proposed ordinal distortion is a learning-friendly representation for neural networks, which is explicit and homogeneous compared with the implicit and heterogeneous distortion parameters. Thus, our learning model gains sufficient distortion perception of features and shows faster convergence. Moreover, this representation enables more efficient learning with less data required.

2. The local-global associate ordinal distortion estimation network considers different scales of distortion features, jointly

Algorithm 2 Test Process of the Proposed Network

```
Input: Distorted Image $I^d$
Output: Rectified Image $I^r$
Crop and flip $I^d$ into four distortion elements $\Pi = [\pi_1 \; \pi_2 \; \pi_3 \; \pi_4]$, randomly feed one element $\pi_t$ into the trained network
Generate the feature vector of overall distortion $\mathcal{F}_g^{(t)}$ using the global perception module $M_{gp}$: $\mathcal{F}_g^{(t)} \leftarrow M_{gp}(\pi_t)$
Crop $\pi_t$ into same size of distortion blocks $\Theta^{(t)} = [\theta_1^{(t)} \; \theta_2^{(t)} \; \cdots \; \theta_n^{(t)}]$
for all $\theta_j^{(t)} \in \Theta^{(t)}$ do
    Generate the feature vector of detailed distortion $\mathcal{F}_l^{(t,j)}$ using the local Siamese module $M_{ls}$: $\mathcal{F}_l^{(t,j)} \leftarrow M_{ls}(\theta_j^{(t)})$
end for
Generate the hybrid feature vector $\mathcal{F}_h^{(t)}$ by fusing $\mathcal{F}_g^{(t)}$ and $[\mathcal{F}_l^{(t,1)} \; \mathcal{F}_l^{(t,2)} \; \cdots \; \mathcal{F}_l^{(t,n)}]$ based on Eq. 9
Estimate the $\hat{\mathcal{D}}$ using $\mathcal{F}_h^{(t)}$ based on Eq. 10
Compute the distortion coefficients $\hat{\mathcal{K}}$ using the $\hat{\mathcal{D}}$ based on Eq. 14
Warp each pixel of $I^d$ using $\hat{\mathcal{K}}$ based on Eq. 3, to obtain $I^r$
```

reasoning the local distortion context and global distortion context. Also, the devised distortion-aware perception layer boosts the feature extraction of different degrees of distortion.

3. Our ordinal distortion loss fully measures the strong ordinal correlation in the proposed representation, facilitating the accurate approximation of distortion distribution.

4. We can easily calculate the distortion parameters with the estimated ordinal distortion. In contrast to previous methods, our method can handle various camera models and different distortion types due to the unified learning representation.

IV. Experiments

In this section, we first state the details of the synthetic distorted image dataset and the training process of our learning model. Subsequently, we analyze the learning representation for distortion estimation. To demonstrate the effectiveness of each module in our framework, we conduct an ablation study to show the different performances. Additionally, the experimental results of our approach compared with the state-of-the-art methods are exhibited, in both quantitative measurement and visual qualitative appearance. Finally, we discuss two main limitations of our approach and present the possible solutions for future work.

A. Implementation Settings

1) Dataset: We construct a standard synthetic distorted image dataset in terms of the division model discussed in Section III-A. The original images are collected from the MS-COCO dataset [31]. Following the implementations of previous literature [9], [13], [32], we also use a 4^{th} order polynomial based on Eq. 3, which is able to approximate most projection models with high accuracy. Additionally, all of the distortion coefficients are randomly generated from their corresponding ranges: $k_1 \in [-e^{-3}, -e^{-8}]$, $k_2 \in [-e^{-7}, -e^{-12}]$ or $[e^{-12}, e^{-7}]$, $k_3 \in [-e^{-11}, -e^{-16}]$ or $[e^{-16}, e^{-11}]$, and $k_4 \in [-e^{-15}, -e^{-20}]$ or $[e^{-20}, e^{-15}]$. Our synthetic dataset contains 20,000 training images, 2,000 test images, and 2,000 validation images.

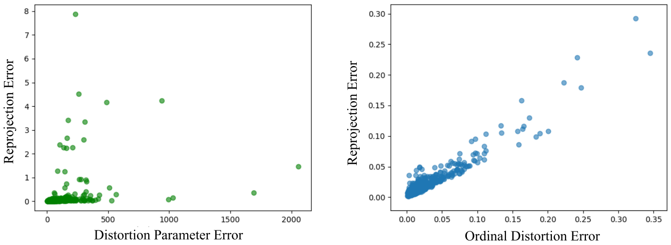


Fig. 5. Comparison of two learning representations for distortion estimation, distortion parameter (left) and ordinal distortion (right). In contrast to the ambiguous relationship between the distortion distribution and distortion parameter, the proposed ordinal distortion displays an evident positive correlation to the distortion reprojection error.

2) Training/Testing Setting: We train our learning model on a NVIDIA RTX 2080 Ti GPU for 300 epochs, and the mini-batch size is 128. The backbone network of the global perception module is pre-trained on the ImageNet [29], and we fine-tune the learning model using the constructed synthetic distorted image dataset with a relatively small learning rate 5×10^{-4}, following the principle of transfer learning. The Adam [31] is chosen as the optimizer with the parameters $\beta_1 = 0.5$ and $\beta_2 = 0.9$.

In the training stage, we crop each distorted image into four distortion elements and learn the parameters of the neural network using all data. Note that this training process is data-independent, where each part of the entire image is fed into the network one by one without the data correlation. In the test stage, we only need one distortion element, i.e., 1/4 of an image, to estimate the ordinal distortion. For a clear exhibition of our approach, we draw the detailed algorithm schemes of the training process and test process as listed in Algorithm 1 and Algorithm 2, respectively.

3) Evaluation Metrics: Crucially, evaluating the performance of different methods with reasonable metrics benefits experimental comparisons. In the distortion rectification problem, the corrected image can be evaluated with the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM). For the evaluation of the estimated distortion label, it is straightforward to employ the root mean square error (RMSE) between the estimated coefficients $\hat{\mathcal{K}}$ and ground truth coefficients $\mathcal{K}$:

$$RMSE = \frac{1}{N} \sum_{i=1}^{N} \sqrt{(\hat{\mathcal{K}}_i - \mathcal{K}_i)^2}, \tag{16}$$

where N is the number of estimated distortion coefficients. However, we found that different groups of distortion coefficients may display similar distortion distributions in images. To more reasonably evaluate the estimated distortion labels,

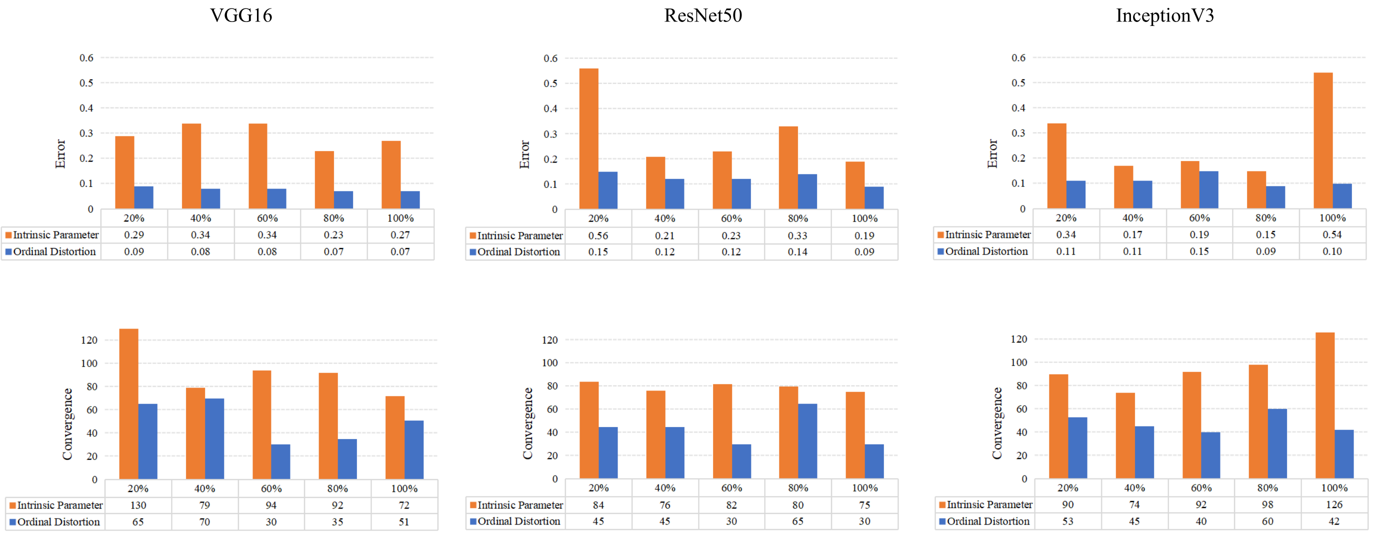


Fig. 6. Analysis of two learning representations in terms of the error and convergence. We show the the histogram of error (top) and convergence (bottom) of two learning representations using three backbone networks, VGG16, ResNet50, and InceptionV3. Compared with the distortion estimation task, our proposed ordinal distortion estimation task achieves lower errors and faster convergence on all backbone networks.

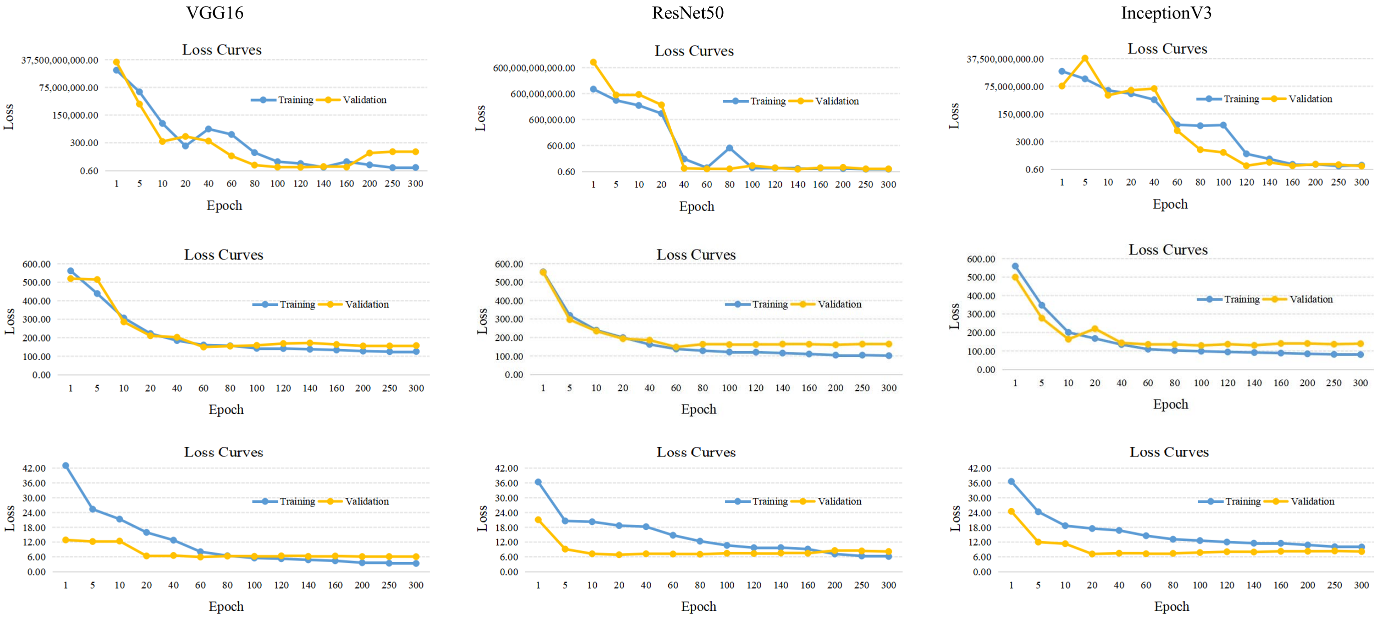


Fig. 7. Analysis of two learning representation in terms of the training and validation loss curves. We show the learning performance of the distortion parameter estimation without (top) and with (middle) the normalization of magnitude, and the ordinal distortion estimation (bottom). Our proposed ordinal distortion estimation task displays the fast convergence and stable trend on both training and validation sets.

we propose a metric based on the reprojection error, mean distortion level deviation (MDLD):

$$MDLD = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}|\hat{\delta}(i,j) - \delta(i,j)|, \tag{17}$$

where W and H are the width and height of a distorted image, respectively. The ground truth distortion level $\delta(i, j)$ of each pixel can be obtained using Eq. 5.

In contrast to RMSE, MDLD is more suitable for parameter evaluation due to the uniqueness of the distortion distribution. Moreover, RMSE fails to evaluate the different numbers and attributes of estimated parameters for different camera models. Thanks to the objective description of the distortion, MDLD is capable of evaluating different distortion estimation methods using different camera models.

B. Analysis of Learning Representation

Previous learning methods directly regress the distortion parameters from a distorted image. However, such an implicit and heterogeneous representation confuses the distortion learning of neural networks and causes the insufficient distortion perception. To bridge the gap between image feature and

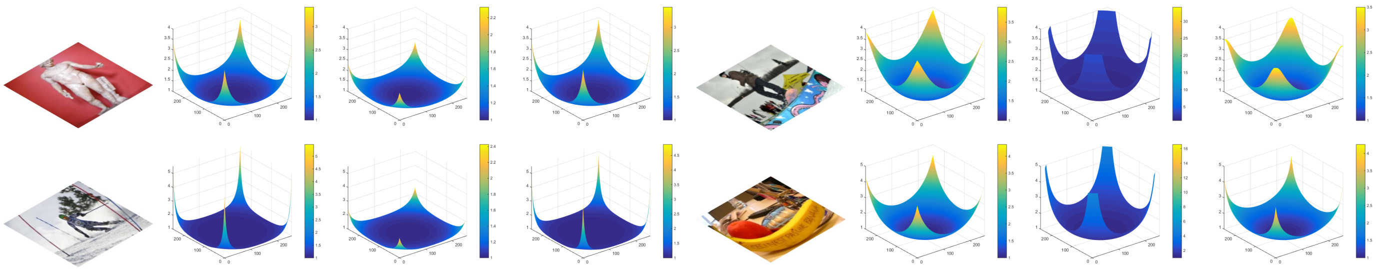
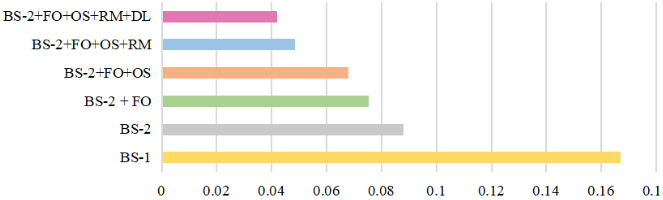


Fig. 8. Qualitative comparison of two learning representations. For each comparison, we show the distorted image, the ground truth 3D DDM, the 3D DDM constructed by the estimated distortion parameter, and ordinal distortion, from left to right.

TABLE I

THE LEARNING-FRIENDLY RATES OF TWO LEARNING REPRESENTATION EVALUATED WITH THREE BACKBONE NETWORKS

Learning Representation	VGG16	ResNet50	InceptionV3
Distortion Parameter	0.50	0.60	0.59
Ordinal Distortion	**2.23**	1.43	1.50



	BS-1	BS-2	BS-2 + FO	BS-2+FO+OS	BS-2+FO+OS+ RM	BS-2+FO+OS+ RM+DL
■ Error	0.1671	0.0879	0.0752	0.0679	0.0486	0.0421

Fig. 9. Ablation study of the proposed ordinal distortion estimation approach.

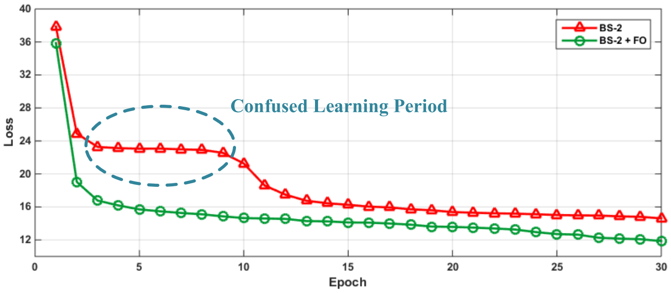


Fig. 10. Training loss of first 30 epochs derived from the BS-2 and BS-2 + FO. The flip operation that normalizes the distortion distribution of inputs is able to significantly accelerate the convergence of the learning process.

calibration objective, we present a novel intermediate representation, i.e., ordinal distortion, which displays a learning-friendly attribute for learning models. For an intuitive and comprehensive analysis, we compare these two representations from the following three aspects.

1) Relationship to Distortion Distribution: We first emphasize the relationship between two learning representations and the realistic distortion distribution of a distorted image. In detail, we train a learning model to estimate the distortion parameters and the ordinal distortions separately, and the errors of estimated results are built in the relationship to the distortion reprojection error. As shown in Fig. 5, we visualize the scatter diagram of two learning representations using 1,000 test distorted images. For the distortion parameter, its relationship to the distortion distribution is ambiguous and the similar parameter errors are related to quite different reprojection errors, which indicates that optimizing the parameter error would confuse the learning of neural networks. In contrast, the ordinal distortion error displays an evident positive correlation to the distortion distribution error, and thus the learning model gains intuitive distortion perception. Therefore, the proposed representation helps to decrease the error of distortion estimation.

2) Distortion Learning Evaluation: Then, we introduce three key elements for evaluating the learning representation: training data, convergence, and error. Supposed that the settings such as the network architecture and optimizer are the same, a better learning representation can be described from the less the training data is, the faster convergence and the lower error are. For example, a student is able to achieve the highest test grade (the lowest error) with the fastest learning speed and the least homework, meaning that he grasps the best learning strategy compared with other students. In terms of the above description, we evaluate the distortion parameter and ordinal distortion as shown in Fig. 6 and Fig. 7.

To exhibit the performance fairly, we employ three common network architectures VGG16, ResNet50, and InceptionV3 as the backbone networks of the learning model. The proposed MDLD metric is used to express the distortion estimation error due to its unique and fair measurement for distortion distribution. To be specific, we visualize the error and convergence epoch when estimating two representations under the same number of training data in Fig. 6, which is sampled with 20%, 40%, 60%, 80%, and 100% from the entire training data. Besides, the training and validation loss curves of two learning representations are shown in Fig. 7, in which the distortion parameters are processed without (top) and with (middle) the normalization of magnitude. From these learning evaluations, we can observe:

(1) Overall, the ordinal distortion estimation significantly outperforms the distortion parameter estimation in both convergence and accuracy, even if the amount of training data is 20% of that used to train the learning model. Note that we only use 1/4 distorted image to predict the ordinal distortion. As we pointed out earlier, the proposed ordinal distortion is explicit

to the image feature and is observable from a distorted image; thus it boosts the neural networks' learning ability. On the other hand, the performance of the distortion parameter estimation drops as the amount of training data decreases. In contrast, our ordinal distortion estimation performs more consistently due to the homogeneity of the learning representation.

(2) For each backbone network, the layer depths of VGG16, InceptionV3, and ResNet50 are 23, 159, and 168, respectively. These architectures represent the different extraction abilities of image features. As illustrated in Fig. 6, the distortion parameter estimation achieves the lowest error (0.15) using InceptionV3 as the backbone under 80% training data, which indicates its performance requires more complicated and high-level features extracted by deep networks. With the explicit relationship to image features, the ordinal distortion estimation achieves the lowest error (0.07) using the VGG16 as the backbone under 100% training data. This promising performance indicates the ordinal distortion is a learning-friendly representation, which is easy to learn even using a very shallow network.

(3) From the loss curves in Fig. 7, the ordinal distortion estimation achieves the fastest convergence and best performance on the validation dataset. It is also worth to note that the ordinal distortion estimation already performs well on the validation at the first twenty epochs, which verifies that this learning representation yields a favorable generalization for neural networks. In contrast, suffering from the heterogeneous representation, the learning process of distortion parameter estimation displays a slower convergence. Moreover, the training and validation loss curves show unstable descend processes when the distortion parameters are handled without the normalization of magnitude, demonstrating the distortion parameter estimation is very sensitive to the label balancing.

We further present a *learning-friendly rate* (Γ_{lr}) to evaluate the effectiveness of learning representation or strategy quantitatively. To our knowledge, this is the first evaluation metric to describe the effectiveness of learning representation for neural networks. As mentioned above, the required training data, convergence, and error can jointly describe a learning representation, and thus we formulate the learning-friendly rate as follows

$$\Gamma_{lr} = \frac{1}{M}\sum_{i=1}^{N}\frac{T_i}{T}\left(\frac{1}{E_i}\log(2-\frac{C_i}{C})\right), \tag{18}$$

where M is the number of split groups, E_i, T_i, and C_i indicate the error, number of training data, the epoch of convergence of the i-th group, respectively. T and C indicate the total number of training data and total training epochs for the learning model. We compute the learning-friendly rates of two learning representations and list the quantitative results in Table I. The results show that our scheme outperforms the distortion parameter estimation on all backbone settings, and thus the proposed ordinal distortion is much suitable for the neural network as a learning representation.

3) Qualitative Comparison: To qualitatively show the performance of different learning representations, we visualize the 3D distortion distribution maps (3D DDM) derived from the ground truth and these two schemes in Fig. 8, in which each pixel value of the distortion distribution map represents the distortion level. Since the ordinal distortion estimation pays more attention to the realistic distortion perception and reasonable learning strategy, our scheme achieves results much closer to the ground truth 3D DDM. Due to implicit learning, the distortion parameter estimation generates inferior reconstructed results, such as the under-fitting (left) and over-fitting (right) on the global distribution approximation as shown in Fig. 8.

C. Ablation Study

To validate the effectiveness of each component in our approach, we conduct an ablation study to evaluate the error of distortion estimation, as shown in Fig. 9. Concretely, we first use the VGG16 network without the fully connected layers as the backbone of the ordinal distortion estimation network, based on the analysis of the learning representation in Section IV-B. Subsequently, we implement the learning model without the flip operation (FO) on global distortion context, ordinal supervision (OS), region of interest mask (RM), and distortion-aware perception layer (DL) as the baseline (BS), and then gradually add these removed components to show the different estimation performance. In addition, we perform two loss functions: $\mathcal{L}_2$ and $\mathcal{L}_{sm}$ to optimize the baseline model, in which $\mathcal{L}_{sm}$ is the smooth $\mathcal{L}_1$ loss function [30] that combines the attributes of $\mathcal{L}_1$ and $\mathcal{L}_2$. We name these two types of baseline models as BS-1 and BS-2. During the training process, we crop four patches from the distorted image and shuffle the orders of all input patches. Subsequently, the patches are fed into the learning model. In the test stage, we only use one patch of a distorted image to evaluate the model.

Overall, the completed framework achieves the lowest error of distortion estimation as shown in Fig. 9, verifying the effectiveness of our proposed approach. For the optimization strategy, the BS-2 used $\mathcal{L}_{sm}$ performs much better than BS-1 used $\mathcal{L}_2$ since the $\mathcal{L}_{sm}$ loss function boosts a more stable training process. Due to the effective normalization of distortion distribution, the network gains explicit spatial guidance with the flip operation on the global distortion context. We also show the training loss of the first 30 epochs derived from the BS-2 and BS-2 + FO in Fig. 10, where we can observe that the distribution normalization can significantly accelerate the convergence of the training process. On contrary, the BS-2 without flip operation suffers from a *confused learning period* especially in the first 10 epochs, which indicates that the neural network is unsure how to find a direct optimization way from the distribution difference. Moreover, the ordinal supervision fully measures the strong ordinal correlation in the proposed representation, and thus facilitates the accurate approximation of distortion distribution. With the special attention mechanism and distortion feature extraction, our learning model gains further improvements using the region of interest mask and distortion-aware perception layer.

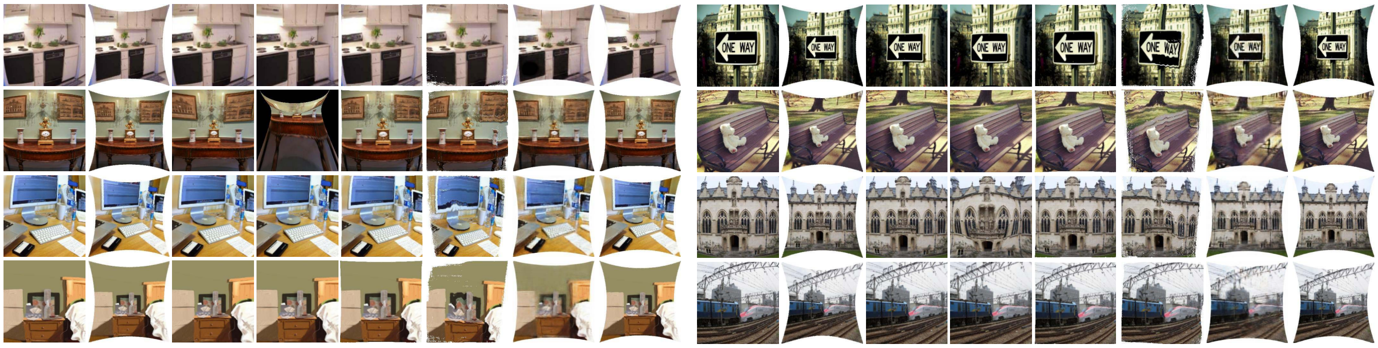


Fig. 11. Qualitative evaluations of the rectified distorted images on indoor (left) and outdoor (right) scenes. For each evaluation, we show the distorted image, ground truth, and corrected results of the compared methods: Alemán-Flores [23], Santana-Cedrés [24], Rong [8], Li [11], and Liao [12], and rectified results of our proposed approach, from left to right.

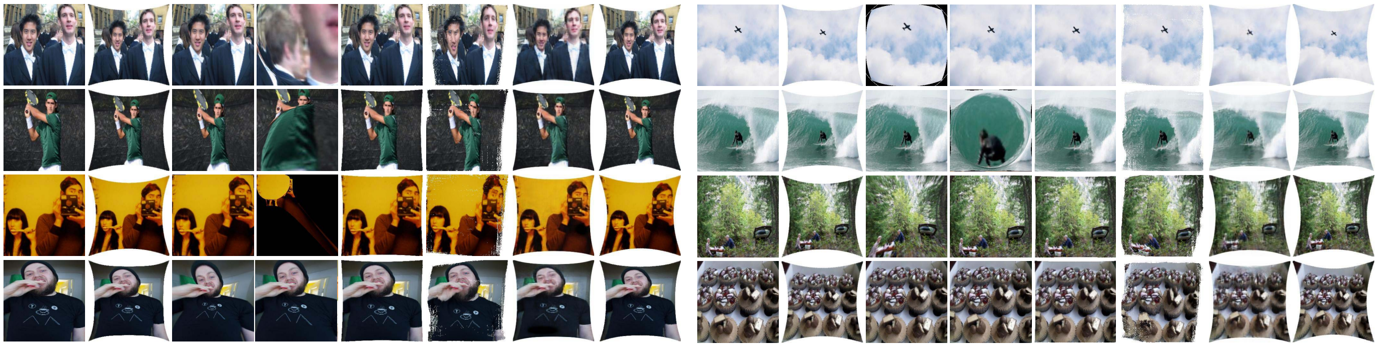


Fig. 12. Qualitative evaluations of the rectified distorted images on people (left) and challenging (right) scenes. For each evaluation, we show the distorted image, ground truth, and corrected results of the compared methods: Alemán-Flores [23], Santana-Cedrés [24], Rong [8], Li [11], and Liao [12], and rectified results of our proposed approach, from left to right.

D. Comparison Results

In this part, we compare our approach with the state-of-the-art methods in both quantitative and qualitative evaluations, in which the compared methods can be classified into traditional methods [23], [24] and learning methods [8], [11] [12]. Note that our approach only requires a patch of the input distorted image to estimate the ordinal distortion.

1) Quantitative Evaluation: To demonstrate a quantitative comparison with the state-of-the-art approaches, we evaluate the rectified images based on the PSNR (peak signal-to-noise ratio), SSIM (structural similarity index), and the proposed MDLD (mean distortion level deviation). All the comparison methods are used to conduct the distortion rectification on the test dataset including 2,000 distorted images. For the PSNR and SSIM, we compute these two metrics using the pixel difference between each rectified image and the ground truth image. For the MDLD, we first exploit the estimated distortion parameters to obtain all distortion levels of the test distorted image based on Eq. 5. Then, the value of MDLD can be calculated by the difference between estimated distortion levels and the ground truth distortion levels based on Eq. 17. Note that the generated-based methods such as Li *et al.* [11] and Liao *et al.* [12] directly learn the transformation manner of the pixel mapping instead of estimating the distortion parameters, so we only evaluate these two methods in terms of the PSNR and SSIM.

As listed in Table II, our approach significantly outperforms the compared approaches in all metrics, including the highest metrics on PSNR and SSIM, as well as the lowest metric on MDLD. Specifically, compared with the traditional methods [23], [24] based on the hand-crafted features, our approach overcomes the scene limitation and simple camera model assumption, showing more promising generality and flexibility. Compared with the learning distortion rectification methods [8], [11] [12], which omit the prior knowledge of the distortion, our approach transfers the heterogeneous estimation problem into a homogeneous one, eliminating the implicit relationship between image features and predicted values in a more explicit expression. As benefits of the effective ordinal supervision and guidance of distortion information during the learning process, our approach outperforms Liao *et al.* [12] by a significant margin, with approximately 23% improvement on PSNR and 17% improvement on SSIM. Besides the high quality of the rectified image, our approach can obtain the accurate distortion parameters of a distorted image, which is crucial for the subsequent tasks such as the camera calibration. However, the generation-based methods [11], [12] mainly focus on the pixel reconstruction of a rectified image and ignore the parameter estimation.

2) Qualitative Evaluation: We visually compare the corrected results from our approach with state-of-the-art methods using our synthetic test set and the real distorted images.

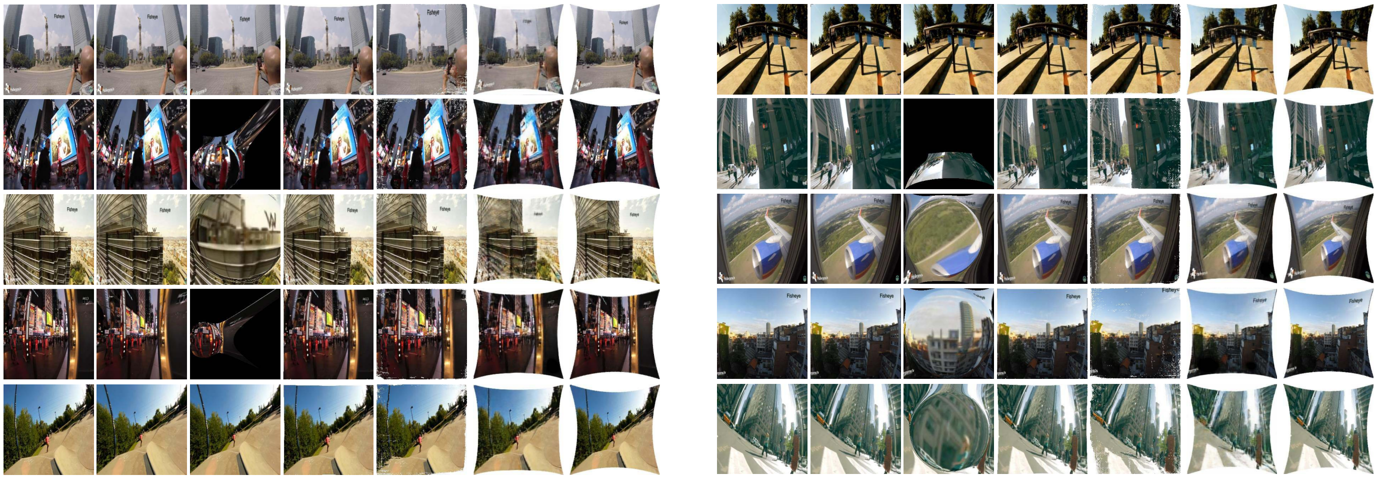


Fig. 13. Qualitative evaluations of the rectified distorted images on real-world scenes. For each evaluation, we show the distorted image and corrected results of the compared methods: Alemán-Flores [23], Santana-Cedrés [24], Rong [8], Li [11], and Liao [12], and rectified results of our proposed approach, from left to right.

TABLE II

QUANTITATIVE EVALUATION OF THE RECTIFIED RESULTS OBTAINED BY DIFFERENT METHODS

Comparison Methods	PSNR ↑	SSIM ↑	MDLD ↓
Traditional Methods			
Alemán-Flores [23]	9.47	0.31	0.26
Santana-Cedrés [24]	7.90	0.25	1.18
Learning Methods			
Rong [8]	10.37	0.29	0.23
Li [11]	13.87	0.64	-
Liao [12]	20.28	0.72	-
Ours	**24.82**	**0.84**	**0.04**

To show the comprehensive rectification performance under different scenes, we split the test set into four types of scenes: indoor, outdoor, people, and challenging scenes. The indoor and outdoor scenes are shown in Fig. 11, and the people and challenging scenes are shown in Fig. 12. Our approach performs well on all scenes, while the traditional methods [23], [24] show inferior corrected results under the scene that lacks sufficient hand-crafted features, especially in the people and challenging scenes. On the other hand, the learning methods [8], [11], [12] lag behind in the sufficient distortion perception and cannot easily adapt to scenes with strong geometric distortion. For example, the results obtained by Rong *et al.* [8] show coarse rectified structures, which are induced by the implicit learning of distortion and simple model assumption. Li *et al.* [11] leveraged the estimated distortion flow to generate the rectified images. However, the accuracy of the pixel-wise reconstruction heavily relies on the performance of scene analysis, leading to some stronger distortion results under complex scenes. Although Liao *et al.* [12] generated better rectified images than the above learning methods in terms of global distribution; the results display unpleasant blur local appearances due to the used adversarial learning manner. In contrast, our results achieve the best performance on global distribution and local appearance, which benefit from the proposed learning-friendly representation and the effective learning model.

The comparison results of the real distorted image are shown in Fig. 13. We collect the real distorted images from the videos on YouTube, captured by popular fisheye lenses, such as the SAMSUNG 10mm F3, Rokinon 8mm Cine Lens, Opteka 6.5mm Lens, and GoPro. As illustrated in Fig. 13, our approach generates the best rectification results compared with the state-of-the-art methods, showing the appealing generalization ability for blind distortion rectification. To be specific, the salient objects such as buildings, streetlights, and roads are recovered into their original straight structures by our approach, which exhibits a more realistic geometric appearance than the results of other methods. Since our approach mainly focuses on the design of learning representation for distortion estimation, the neural networks gain more powerful learning ability to the distortion perception and achieve more accurate estimation results.

E. Limitation Discussion

In this work, we presented a new learning representation for the deep distortion rectification and implemented a standard and widely-used camera model to validate its effectiveness. The rectification results on the synthesized and real-world scenarios also demonstrated our approach's superiority compared with the state-of-the-art methods. Like most of the assumptions in the other works [8], [11], [12], [14], [21], [23], our approach has two main limitations to extend to more complicated applications.

The first limitation is that the principal point needs to be at the center of the image. Observing that the principal point is slightly disturbed around the center of the image, we mainly consider the estimation of distortion coefficients using the proposed ordinal distortion in our work. Nevertheless, our method can be easily extended to more scenarios when the network predicts more target labels. For example, suppose we wish to estimate a principal point (x_c, y_c) and four distortion coefficients (k_1, k_2, k_3, k_4) of a distorted image (six variables in total). In that case, we only need to predict the ordinal distortion $\mathcal{D} = [\delta(r_1) \quad \delta(r_2) \quad \delta(r_3) \quad \delta(r_4) \quad \delta(r_5) \quad \delta(r_6)]$ with two extra distortion levels $\delta(r_5)$ and $\delta(r_6)$ than the original scheme, namely, building simultaneous equations for solving six variables based on Eq. 5. Moreover, in our previous work [33], we developed a VGG-like network to regress the principal point given a distorted image, and then other distortion parameters are estimated accordingly. Thus, this sequential estimation solution also could be used in more complicated cases.

The second limitation is that the distortion needs to be radially symmetric. This problem may be addressed by the grid optimization technique in computer graphics, and we can teach the network to learn an asymmetric grid to warp each pixel of the distorted image. Based on the above limitations and the presented solutions, we plan to achieve a more comprehensive and robust distortion rectification framework in future work.

V. Conclusion

In this paper, we present a learning-friendly representation for the deep distortion rectification, bridging the gap between image feature and calibration objective. Compared with the implicit and heterogeneous distortion parameters, the proposed ordinal distortion offers three unique advantages: explicitness, homogeneity, and redundancy, enabling a sufficient and efficient learning on the distortion. To learn this representation, we design a local-global associated estimation network optimized with an ordinal distortion loss function, and a distortion-aware perception layer is used to boost the features extraction of different degrees of distortion. As the benefit of the proposed learning representation and learning model, our approach outperforms the state-of-the-art methods by a remarkable margin while only leveraging a part of data for distortion estimation.

Acknowledgment

The authors would like to thank the editor and reviewers for their valuable suggestions and meaningful discussions.

References

[1] G. Li, Y. Gan, H. Wu, N. Xiao, and L. Lin, "Cross-modal attentional context learning for RGB-D object detection," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1591–1601, Apr. 2019.

[2] P. Zhang, W. Liu, H. Lu, and C. Shen, "Salient object detection with lossless feature reflection and weighted structural loss," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3048–3060, Jun. 2019.

[3] D. Tao, Y. Guo, Y. Li, and X. Gao, "Tensor rank preserving discriminant analysis for facial recognition," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 325–334, Jan. 2018.

[4] B. Kang and T. Q. Nguyen, "Random forest with learned representations for semantic segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3542–3555, Jul. 2019.

[5] C. Redondo-Cabrera, M. Baptista-Ríos, and R. J. López-Sastre, "Learning to exploit the prior network knowledge for weakly supervised semantic segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3649–3661, Jul. 2019.

[6] H. Li, X. He, D. Tao, Y. Tang, and R. Wang, "Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning," *Pattern Recognit.*, vol. 79, pp. 130–146, Jul. 2018.

[7] Y. Hou *et al.*, "NLH: A blind pixel-level non-local method for real-world image denoising," *IEEE Trans. Image Process.*, vol. 29, pp. 5121–5135, 2020.

[8] 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.

[9] 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.

[10] Z. Xue, N. Xue, G.-S. Xia, and W. Shen, "Learning to calibrate straight lines for fisheye image rectification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1643–1651.

[11] X. Li, B. Zhang, P. V. Sander, and J. Liao, "Blind geometric distortion correction on images through deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4855–4864.

[12] K. Liao, C. Lin, Y. Zhao, and M. Xu, "Model-free distortion rectification framework bridged by distortion distribution map," *IEEE Trans. Image Process.*, vol. 29, pp. 3707–3718, 2020.

[13] K. Liao, C. Lin, Y. Zhao, and M. Gabbouj, "DR-GAN: Automatic radial distortion rectification using conditional GAN in real-time," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 725–733, Mar. 2020.

[14] M. Lopez, R. Mari, P. Gargallo, Y. Kuang, J. Gonzalez-Jimenez, and G. Haro, "Deep single image camera calibration with radial distortion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11817–11825.

[15] Q. Zhang, C. Zhang, J. Ling, Q. Wang, and J. Yu, "A generic multi-projection-center model and calibration method for light field cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2539–2552, Nov. 2019.

[16] X. Chen and Y.-H. Yang, "A closed-form solution to single underwater camera calibration using triple wavelength dispersion and its application to single camera 3D reconstruction," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4553–4561, Sep. 2017.

[17] Y. Bok, H.-G. Jeon, and I. S. Kweon, "Geometric calibration of micro-lens-based light field cameras using line features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 287–300, Feb. 2017.

[18] S. B. Kang, "Catadioptric self-calibration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2000, pp. 201–207.

[19] S. Ramalingam, P. Sturm, and S. K. Lodha, "Generic self-calibration of central cameras," *Comput. Vis. Image Understand.*, vol. 114, no. 2, pp. 210–219, Feb. 2010.

[20] F. Espuny and J. I. B. Gil, "Generic self-calibration of central cameras from two rotational flows," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 131–145, Jan. 2011.

[21] 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, Jan. 2013.

[22] A. W. Fitzgibbon, "Simultaneous linear estimation of multiple view geometry and lens distortion," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2001, p. 1.

[23] 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, Nov. 2014.

[24] D. Santana-Cedrés *et al.*, "An iterative optimization algorithm for lens distortion correction using two-parameter models," *Image Process. Line*, vol. 5, pp. 326–364, Dec. 2016.

[25] Z. Tang, R. Grompone von Gioi, P. Monasse, and J.-M. Morel, "A precision analysis of camera distortion models," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2694–2704, Jun. 2017.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, pp. 1–14, Sep. 2015.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[28] 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. (CVPR)*, Jun. 2016, pp. 2818–2826.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[31] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

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

[33] K. Liao, C. Lin, Y. Zhao, M. Gabbouj, and Y. Zheng, "OIDC-Net: Omnidirectional image distortion correction via coarse-to-fine region attention," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 1, pp. 222–231, Jan. 2020.

Kang Liao (Graduate Student Member, IEEE) received the B.S. degree in software engineering from Shaanxi Normal University, Xi'an, Shaanxi, China, in 2017. He is currently pursuing the Ph.D. degree in signal and information processing with the Institute of Information Science, Beijing Jiaotong University, Beijing, China.

His current research interests include image and video processing, 3-D scene understanding, and adversarial learning.

Chunyu Lin (Member, IEEE) received the doctor's degree from Beijing Jiaotong University (BJTU), Beijing, China, in 2011.

From 2009 to 2010, he was a Visiting Researcher with the ICT Group, Delft University of Technology, The Netherlands. From 2011 to 2012, he was a Post-Doctoral Researcher with the Multimedia Laboratory, Gent University, Belgium. He is currently a Full Professor with BJTU. His research interests include image/video compression and robust transmission, 3-D video coding, virtual reality video processing, and ADAS.

Yao Zhao (Senior Member, IEEE) received the B.S. degree from the Radio Engineering Department, Fuzhou University, Fuzhou, China, in 1989, the M.E. degree from the Radio Engineering Department, Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He became an Associate Professor at BJTU in 1998 and became a Professor in 2001. From 2001 to 2002, he was a Senior Research Fellow with the Information and Communication Theory Group, Faculty of Information Technology and Systems, Delft University of Technology, Delft, The Netherlands. He is currently the Director of the Institute of Information Science, BJTU. His current research interests include image/video coding, digital watermarking and forensics, and video analysis and understanding. He is also a Fellow of IET. He also serves on the Editorial Boards for several international journals, including as an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS, a Senior Associate Editor for the IEEE SIGNAL PROCESSING LETTERS, and an Area Editor for *Signal Processing: Image Communication*. He was named a Distinguished Young Scholar by the National Science Foundation of China in 2010, and was elected as a Chang Jiang Scholar of Ministry of Education of China in 2013.