Multi-Level Curriculum for Training A Distortion-Aware Barrel Distortion Rectification Model

Kang Liao[1,2] Chunyu Lin[1,2*] Lixin Liao[1,2] Yao Zhao[1,2] Weiyao Lin[3]
[1]Institute of Information Science, Beijing Jiaotong University
[2]Beijing Key Laboratory of Advanced Information Science and Network
[3]Department of Electronic Engineering, Shanghai Jiaotong University
{kang_liao, cylin, 16112056, yzhao}@bjtu.edu.cn, wylin@sjtu.edu.cn

Abstract

Barrel distortion rectification aims at removing the radial distortion in a distorted image captured by a wide-angle lens. Previous deep learning methods mainly solve this problem by learning the implicit distortion parameters or the nonlinear rectified mapping function in a direct manner. However, this type of manner results in an indistinct learning process of rectification and thus limits the deep perception of distortion. In this paper, inspired by the curriculum learning, we analyze the barrel distortion rectification task in a progressive and meaningful manner. By considering the relationship among different construction levels in an image, we design a multi-level curriculum that disassembles the rectification task into three levels, structure recovery, semantics embedding, and texture rendering. With the guidance of the curriculum that corresponds to the construction of images, the proposed hierarchical architecture enables a progressive rectification and achieves more accurate results. Moreover, we present a novel distortion-aware pre-training strategy to facilitate the initial learning of neural networks, promoting the model to converge faster and better. Experimental results on the synthesized and real-world distorted image datasets show that the proposed approach significantly outperforms other learning methods, both qualitatively and quantitatively.

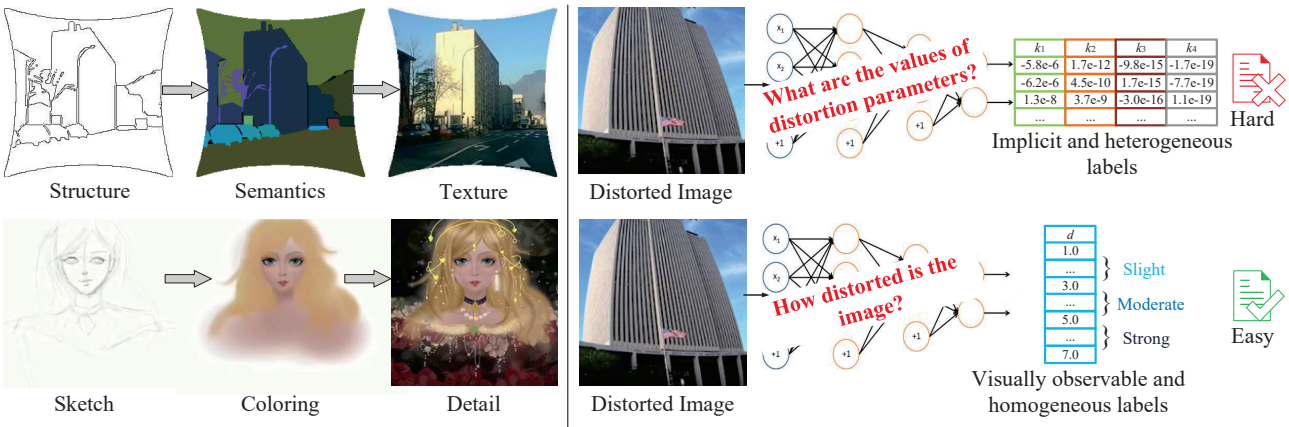
1. Introduction

Rectifying the distorted images is an indispensable preprocessing step for most computer vision tasks since the geometric distortion changes the original scene distribution. Recent works [25][32][31][16][19][17][5] learn the barrel distortion rectification model in a direct manner, which feeds distorted images into networks and only supervises the final outputs. Despite the end-to-end architecture, directly learning such a complex nonlinear mapping function between different domains (from the distortion domain into the alignment domain) is challenging. The pixel-level supervision on final output cannot fully guide the rectification of geometry distribution. Moreover, this process cannot explicitly reason different construction levels of a distorted image, limiting the models' learning of the distortion features in rectification task. Thus, previous direct learning manners hinder the performance improvement of the rectification algorithm. In this paper, inspired by curriculum learning [4], we consider improving the barrel distortion rectification in a progressive and meaningful manner.

Curriculum learning, proposed by Bengio [4], is one general paradigm that introduces a guided and meaningful strategy to train a machine learning model. By imitating the learning process of humans, the model can converge faster by learning different knowledge at different learning stages based on a curriculum. Inspired by this process, we construct a multi-level curriculum to train the deep barrel distortion rectification model. As illustrated in Fig. 1 (a), similar to human painting from sketch, coloring to details, the procedure of our curriculum displays a simple-to-complex order from the structure, semantics to texture.

Additionally, to facilitate the initial learning of the rectification model, we propose a distortion-aware pre-training strategy. Pre-training on ImageNet [14] is a widely used strategy in computer vision. Nevertheless, He et al. [9] verified that it helps less if the target task is more sensitive to localization. Thus, it is not suitable for the rectification task requiring a precise coordinate transformation. Pre-training on ImageNet was demonstrated that it counts against the distortion estimation task in [18] as ImageNet does not contain any distorted images. Since it is difficult to make the model learn the implicit distortion parameters, our distortion-aware pre-training strategy permits the model a better network initialization and helps to perceive how dis-

*Corresponding author



(a) Multi-level curriculum and the human painting procedure

(b) Distortion-aware pre-training focuses on the general concept rather than obscure details

Figure 1. Motivation of the proposed method. (a) Similar to the human painting procedure (the bottom), the constructed curriculum (the top) addresses the barrel distortion rectification task into three levels, structure, semantics, and texture. (b) Instead of the implicit and detailed distortion parameters (top), the distortion-aware pre-training strategy focuses on the explicit and general distortion level (bottom).

torted the image is, as shown in Fig. 1 (b).

In particular, we propose a multi-level curriculum with a distortion-aware pre-training strategy for training the deep barrel distortion rectification model. First, we construct a curriculum with three levels, structure, semantics, and texture. The curriculum is also related to the construction of an image, as Marr [22] emphasized understanding an image is a multiple stages procedure, where different components at different construction levels of an image are strongly linked. Subsequently, we develop a distortion-aware pre-training strategy to enhance the distortion perception of the model, teaching it to grasp the general prior knowledge of distortion rather than obscure details. To gradually learn the image rectification, we design a hierarchical framework consisting of three modules, structure recovery, semantics embedding, and texture rendering. Such an architecture enables the progressive rectification from the low-level features to high-level features. Compared with previous methods, the proposed rectification process can tackle the barrel distortion by supervising the intermediate product of each module. Experimental results on the synthesized and real-world datasets demonstrate our approach outperforms the state-of-the-art methods with a large margin.

In general, our contributions are summarized as follows:

- We propose a curriculum to train the deep barrel distortion rectification model in a progressive and meaningful manner.
- A distortion-aware pre-training strategy is proposed to enhance the initialization of the learning model.
- To learn the proposed multi-level curriculum, we design an effective hierarchical rectification framework.

2. Related Work

2.1. Barrel Distortion Rectification

Traditional methods mainly rely on the detection of hand-crafted features [3][23][10][13][1][28]. However, these methods usually performed poorly due to specific constraint such as the plumb-line and curve, leading to a poor generalization ability to other scenes. Recently, the accuracy of rectification has been improved using deep learning [25][32][31][16][19][17][5]. Rong et al. [25] first used convolutional neural networks (CNNs) to estimate the distortion parameters. However, the simple camera model and AlexNet architecture limit its general application. To rectify more complicated distortions, DR-GAN [17] and FishEyeRecNet [32] trained their models based on adversarial learning [8] and multi-task learning [6]. But these two methods cannot guide the networks to explicitly learn distortion features due to the one-pass direct training manner. Xue et al. [31] improved the performance through the curve guidance, but this method suffered from inferior robustness when facing the scene containing fewer hand-crafted features. Liao et al. [19] and Li et al. [16] proposed to unify different types of distortions using the distortion distribution map and distortion flow, respectively. Nevertheless, their designed learning models are still hard to correct the distortion in a progressive and meaningful way.

2.2. Curriculum Learning

The conception of curriculum learning can be found in Elman et al. [7], which highlights the importance of starting small and then gradually process the more challenging levels. This work displayed a learning process as human

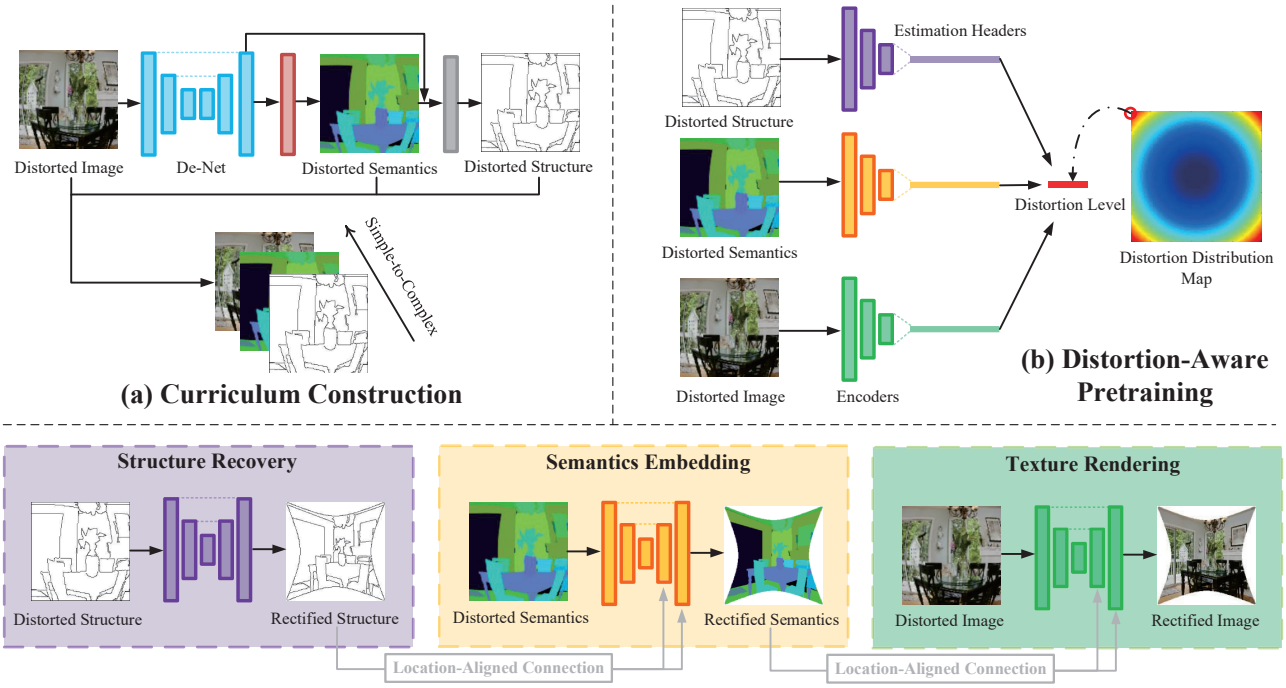


Figure 2. The overview of our proposed approach. (a) We first decompose a distorted image into different levels and construct a simple-to-complex curriculum. (b) We develop a distortion-aware pre-training strategy for the general cognition of distortion. (c) A hierarchical rectification framework is presented to progressively correct the distortion from low-level to high-level features.

infants do, inspiring a more efficient strategy for machine learning. Like methods in robotics [27], Krueger et al. [15] exploited a shaping scheme to speed up the convergence of the learning process. Bengio et al. [4] thoroughly described the concept, details, and experiment of curriculum learning. This strategy improved the performances of many challenging tasks by addressing them in a simple-to-complex order.

3. Method

In this section, we describe the proposed approach in detail. The overview of our approach is illustrated in Fig. 2.

3.1. Parametric Camera Model

Assumed that a point $\mathbf{P}_w$ in the world coordinate projects to a point $\mathbf{P}_c$ in the camera plane. Then, the relationship between $\mathbf{P}_w$ and $\mathbf{P}_c$ can be given by:

$$\mathbf{P}_c = M\mathbf{P}_w, \tag{1}$$

where $M \in \mathbb{R}^{3\times4}$ is a perspective projection matrix, $\mathbf{P}_c = (x, y, 1)^{\mathrm{T}} \in \mathbb{R}^{3\times1}$ and $\mathbf{P}_w \in \mathbb{R}^{4\times1}$ represent the homogeneous coordinates in the camera and world coordinate system, respectively. For the barrel distortion, there will be non-linear mapping introduced. The projection $h(\cdot)$ is such a non-linear function to describe the radial distortion:

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

The wide-angle lens such as fisheye lens violates the perspective projection mode, and then $f(\cdot)$ can be approximated by a Taylor series expansion as follows:

$$f(x, y) = 1 + k_1 r + k_2 r^2 + k_3 r^3 + \cdots + k_N r^N, \tag{3}$$

where $k_1, k_2, \cdots$ are the distortion parameters. r indicates the Euclidean distance between the distortion center $\mathbf{P}_d = [x_{dc}, y_{dc}]^{\mathrm{T}} \in \mathbb{R}^2$ and point in image.

3.2. Construction of Multi-level Curriculum

We first present a decomposition network (De-Net) to decompose an image into structure, semantics, and texture levels, constructing a simple-to-complex curriculum as shown in Fig. 2 (a). To be more specific, De-Net takes a distorted image $I^d_{tex} \in \mathbb{R}^{h\times w\times3}$ as input and gradually outputs the distorted semantics $I^d_{sem} \in \mathbb{R}^{h\times w\times c}$ and distorted structure $I^d_{str} \in \mathbb{R}^{h\times w\times1}$, where h and w denote the height and width of distorted image, and c is the number of object categories. The backbone of De-Net is designed based on a U-Net [26] style, in which an encoder-decoder network with skip connections gets a final feature map $I^d_{fea} \in \mathbb{R}^{h\times w\times64}$.

Then, two convolutional groups with 1×1 kernels activated by softmax function outputs I^d_{sem} and I^d_{str}, respectively.

For the texture level, we keep the original appearance of a distorted image due to its rich RGB information. DeNet is trained using the label of semantic segmentation in ADE20K dataset [34], which covers most of the scenes in life. Besides, some objects such as the rainbow and arch have their curve structure, which should not be detected as the distorted line. Thus, we apply the contour of the semantic segmentation map as our structure level, which contains more general information than distorted lines do.

3.3. Distortion-Aware Pre-training Strategy

To enhance the distortion perception of the model, we present a distortion-aware pre-training strategy. A standard pre-training needs to meet two requirements: easy to learn and helpful for the subsequent task. Although pre-training on ImageNet [14] is widely used in computer vision, it cannot facilitate the rectification task since the dataset contains no distorted images. Moreover, the original classification task on ImageNet is hard to inspire the rectification task, which requires more accurate localization in terms of the coordinate transformation. In this work, we initialize our model based on the distortion level estimation [19] in the proposed distortion-aware pre-training strategy.

The distortion level indicates the degree of distortion for a pixel in the image, which is visually observable and explicit to the image features. All distortion levels constitute a distortion distribution map. Compared with the distortion parameter, the distortion level is a more general description. Thus, it is easier to teach network to learn how distorted is an image than what are the specific values of parameters in an image. Specifically, we select the maximum distortion level $\mathcal{D}_{max}$ as the learning label given by:

$$\mathcal{D}_{max} = \frac{1}{1 + k_1 r_{max} + k_2 r^2_{max} + k_3 r^3_{max} + \cdots}, \quad (4)$$

where r_{max} indicates the farthest Euclidean distance between the distortion center and pixel in a distorted image. As shown in Fig. 2 (b), three encoder networks extract the feature of corresponding construction levels, and the estimation headers are used to estimate $\mathcal{D}_{max}$. The estimation header consists of three fully connected layers with the following number of units: 512, 256, and 1. During the pre-training process, encoder networks pay more attention to extracting the geometric distortion feature constrained by $\mathcal{D}_{max}$. Thus, the neural network gains a significant improvement in the extraction ability of the distortion feature, accelerating the convergence of rectification tasks.

3.4. Hierarchical Rectification Framework

As illustrated in Fig. 2 (c), the hierarchical rectification framework consists of structure recovery, semantics embedding, and texture rendering modules. In particular, the structure recovery module aims to recover the realistic structure from the distortion distribution. This module is a fully convolutional neural network including an encoder and a decoder network, with skip connections between the encoder and decoder features at the same spatial resolution. There are 5 hierarchies progressively extracting the structure feature in the encoder, where each hierarchy has a convolutional layer with 3×3 kernels and 2 strides. Unlike the encoder, at the beginning of each hierarchy in the decoder, a bilinear upsampling layer is implemented to increase the spatial dimension by a factor of 2. Observing that the input and output domain differs greatly in geometric distribution, we employ Coordinate Convolution [21] in the structure recovery module since it can facilitate the generalization ability of coordinate transformation in the network.

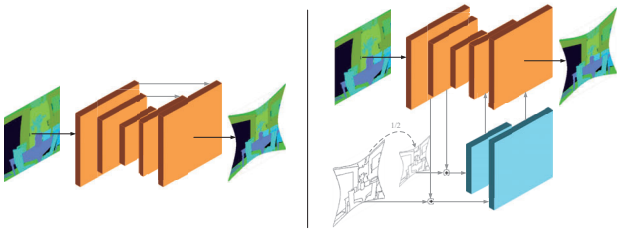


Figure 3. Comparison of the original skip connection operation (left) and our location-aligned connection mechanism that aims to revise the distortion information from encoder to decoder (right).

After the structure recovery module, the distorted semantics can be corrected by embedding them into the rectified structure. The architecture of the semantics embedding module is similar to that of the structure recovery module. To be noted, the skip connection from shallow layers would introduce the distortion information to deep layers in the encoder-decoder network. To revise this distortion information, we present a location-aligned connection mechanism as illustrated in Fig. 3. In the implementation, we first downsample the resolution of rectified structure to match each feature map in the encoder, which is regarded as the aligned target. Then, the feature maps in the encoder that would introduce the distortion information, are concatenated with the corresponding aligned targets. Finally, we leverage a convolutional layer as a revised layer to align the spatial distributions of feature maps, providing the low-level and undistorted features to the decoder network.

Given a rectified semantics, the texture rendering module performs the rectification in the final construction level of a distorted image. The architecture of this module is similar to that of the semantics embedding module, except for two special designs. The first design is that we use both the rectified structure and rectified semantics to improve the final rectification result, which shows more coherent details in the boundary of the scene. The second one is that we lever-

age the Instance-Normalization layer [30] to replace the Batch-Normalization layer since it it can reduce the number of artifacts in the generated images. Also, we use the location-aligned connection mechanism to replace the skip connection in the texture rendering module. Therefore, we design a complete hierarchical framework to perform the distortion rectification in a simple-to-complex order.

We argue that by presenting the hierarchical rectification framework, we can have the following advantages.

1. In contrast to the direct manner, our framework can be trained in a progressive and meaningful manner, promoting the learning of the complex nonlinear mapping function between the distortion domain and the alignment domain.

2. With the distortion-aware pre-training strategy, the learning model obtains a general and clear prior knowledge of distortion. In addition, this pre-training strategy can boost convergence in the rectification training process.

3. Our framework fully considers different levels of features in the image, such as the low-level structure and high-level semantics. Thus, we gain more robust rectification performance than other methods considering only one level.

3.5. Training Loss Functions

For the multi-level curriculum construction, we train the De-Net by optimizing a hybrid per-pixel cross-entropy loss:

$$\mathcal{L}_{De} = \lambda \mathcal{L}_{De}^{sem} + \mathcal{L}_{De}^{str}, \tag{5}$$

where $\mathcal{L}_{De}^{sem}$ and $\mathcal{L}_{De}^{str}$ express the cross-entropy loss for semantic and structure segmentation, respectively.

To enable the distortion-aware pre-training strategy, the difference of the estimated distortion level $\hat{\mathcal{D}}$ and ground truth $\mathcal{D}$ is measured by the smooth $\mathcal{L}_1$ loss [24]:

$$\mathcal{L}_{Pre} = \begin{cases} 0.5t^2, & if \ |t| \leq 1. \\ |t| - 0.5, & otherwise, \end{cases} \tag{6}$$

where $t = \mathcal{D} - \hat{\mathcal{D}}$. $\mathcal{L}_{Pre}$ can be interpreted as the composition of $\mathcal{L}_1$ and $\mathcal{L}_2$ loss, which alleviates the problem of explosive gradient in the training process.

Based on the multi-level curriculum, the final training loss of hierarchical rectification framework is given by:

$$\mathcal{L}_{Rec} = \alpha \mathcal{L}_{Rec}^{str} + \beta \mathcal{L}_{Rec}^{sem} + \gamma \mathcal{L}_{Rec}^{tex}, \tag{7}$$

where α, β, and γ are the weights to balance the losses of structure recovery, semantics embedding, and texture rendering modules. Concretely, we formulate $\mathcal{L}_{Rec}^{str}$ as follows:

$$\mathcal{L}_{Rec}^{str} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} ||\hat{S}_{x,y} - S_{x,y}||_1 + \mathcal{L}_w, \tag{8}$$

where W and H are the width and height of distorted image, $\hat{S}$ and S are the rectified structure and ground truth, respectively. $\mathcal{L}_w$ indicates the loss of the wasserstein-GAN (WGAN) [2] , it can improve the stability of adversarial training and the quality of generated images.

Table 1. Quantitative evaluation of the rectified results obtained by different methods.

Comparison Methods	PSNR ↑	SSIM ↑
Traditional Methods		
Alemán-Flores [1]	8.42	0.13
Santana-Cedrés [28]	9.22	0.18
Learning Methods		
Rong (ACCV'16) [25]	12.98	0.37
DR-GAN (TCSVT'19) [17]	16.43	0.56
Li (CVPR'19) [16]	17.19	0.63
DeepCalib (CVMP'18) [5]	18.43	0.67
Liao (TIP'20) [19]	23.02	0.71
Ours	**26.71**	**0.88**

Like the structure loss function, we minimize the $\mathcal{L}_1$ loss between rectified semantics $\hat{M}$ and ground truth M by:

$$\mathcal{L}_{Rec}^{sem} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} ||\hat{M}_{x,y} - M_{x,y}||_1 + \mathcal{L}_w. \tag{9}$$

Finally, we implement the perceptual loss [11] and wasserstein loss to train the texture rendering module:

$$\mathcal{L}_{Rec}^{tex} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} ||\phi_{i,j}(\hat{I})_{x,y} - \phi_{i,j}(I)_{x,y}||_2 + \mathcal{L}_w, \tag{10}$$

where the difference of rectified texture $\hat{I}$ and ground truth I are minimized on the feature map $\phi_{i,j}$, which is obtained from the j-th convolution (after activation) before the i-th max-pooling layer in the VGG19 network [29].

4. Experiments

4.1. Experimental Settings

Dataset: To train and evaluate the proposed rectification model, we build a comprehensive synthetic image dataset. To be specific, we first select images and segmentation maps in the ADE20K dataset [34] as the source data. The distorted images and distorted segmentation maps are then generated based on the parametric camera model in Section 3.1. For the structure level, we use the contour of the segmentation map due to its general representation of the semantics level. To achieve the distortion-aware pre-training strategy, we provide the label of maximum distortion level for each image. In total, this dataset contains 20,210 training, 1,000 test, and 1,000 validation image sequences.

Implementation Details: The training of our learning model is divided into three parts following the procedure in Fig. 2. We first train De-Net to construct a multi-level curriculum, optimized using SGD with the learning rate of 0.02. Then, we train the encoder networks and estimation headers using Adam [12] with the learning rate of 1×10^{-3}, to conduct the distortion-aware pre-training. For the hierarchical rectification framework, the pre-trained weights are loaded in each encoder, and then the structure recovery, semantics embedding, and texture rendering modules are fine-tuned based on our multi-level curriculum using Adam with the learning rate of 1×10^{-4}. All the networks are trained on NVIDIA GeForce RTX 2080 Ti GPUs.

4.2. Comparison Results

Quantitative Evaluation: We compare our approach with previous rectification methods including the traditional methods: Alemán-Flores [1], Santana-Cedrés [28] and learning methods: Rong [25], DR-GAN [17], DeepCalib [5], Li [16], Liao [19]. The corrected images generated by the state-of-the-art approaches are evaluated based on the PSNR (peak signal-to-noise ratio) and SSIM (structural similarity index). All the methods are leveraged to perform the distortion rectification on the test dataset, including 1,000 images. Then, we compute these two metrics through the pixel difference between each rectified image and the ground truth image. As shown in Table 1, the rectified images are evaluated with ground truth on PSNR and SSIM. Owing to the strong dependence on hand-crafted features, traditional methods [1][28] show poor performance and are hard to apply for the scene-agnostic barrel rectification task. Learning methods [16][19][17][25][5] outperform traditional methods due to the deep perception of semantic features, but the direct training manner limits the comprehensive understanding of distortion rectification.

Quantitative results demonstrate that our approach is superior to other methods in both pixel correction and structure maintenance, achieving the best performance in quantitative evaluation. There are three reasons: (1) the proposed multi-level curriculum guides the learning of the rectification model in a progressive and meaningful manner. (2) the distortion-aware pre-training strategy enhances the distortion perception of deep neural networks. (3) the hierarchical rectification framework reasons the different features of different levels in a distorted image, and thus we gain more robust rectification performance than other methods.

Qualitative Evaluation: To display an intuitive comparison, we visualize the rectified images of different methods using our synthetic dataset in this part. As shown in Fig. 4, the rectified image derived from Santana-Cedrés et al. [28] displays a more severe distortion effect (the first and fourth row). The main reason is that they heavily rely on detecting distorted lines and the optimization of distortion parameters, thus performing poorly in the scene where the hand-crafted features are hard to distinguish. As a benefit of the global semantic features provided by neural networks, the learning methods [16][19][17][25][5] achieve better rectification performance in terms of the visual appearance. Nevertheless, these methods are hard to recover the accurate distribution from the severely distorted scene, influenced by the insufficient and plain learning manner. In contrast, our approach obtains the best rectification performance and leads most compared methods in the qualitative evaluation.

To evaluate the generalization ability of algorithms, we compare our approach with the state-of-the-art methods on the real-world images captured by various wide-angle lenses, as illustrated in Fig. 5. For this evaluation, we collect the real-world barrel distorted images from the videos on YouTube, captured by widely used wide-angle lenses such as the SAMSUNG 10mm F3, Rokinon 8mm Cine Lens, Opteka 6.5mm Lens, and GoPro. From Fig. 5, we can observe that our approach well rectifies the distorted objects such as the buildings and roads, outperforming other methods in terms of global scene distribution and local visual appearance. These results demonstrate that our approach is more competent in practical barrel distortion rectification. More qualitative comparison results can be found in the supplementary material.

4.3. Exploring the Learning Strategy

To validate the effectiveness of the curriculum learning and pre-training, we visualize the training loss curves and the rectified images of different learning schemes: the direct learning without pre-training (DL), pre-training strategy based on the estimation of the distortion parameters (DL + DP-1), pre-training strategy based on the estimation of the maximum distortion level $\mathcal{D}_{max}$ (DL + DP-2, also named the distortion-aware pre-training), and the designed curriculum learning with the distortion-aware pre-training strategy (Ours) as shown in Fig. 6.

Overall, our curriculum learning with the distortion-aware pre-training strategy achieves the best performance on the convergence of training loss and visual rectification result. Specifically, the proposed distortion-aware pre-training strategy is based on the homogeneous and explicit distortion level, enabling a proper initialization of the neural network with the general prior knowledge of distortion. Thus, DL + DP-2 performs much better than DL and DL + DP-1, demonstrating that our pre-training strategy is more suitable for the distortion rectification task. We also present a multi-level curriculum for training the distortion rectification model. Such a learning manner enables a progressive rectification process and weakens the difficulty of one-pass rectification. Therefore, Ours obtains the fastest and best learning process, which generates the rectified image with a visually pleasing appearance.

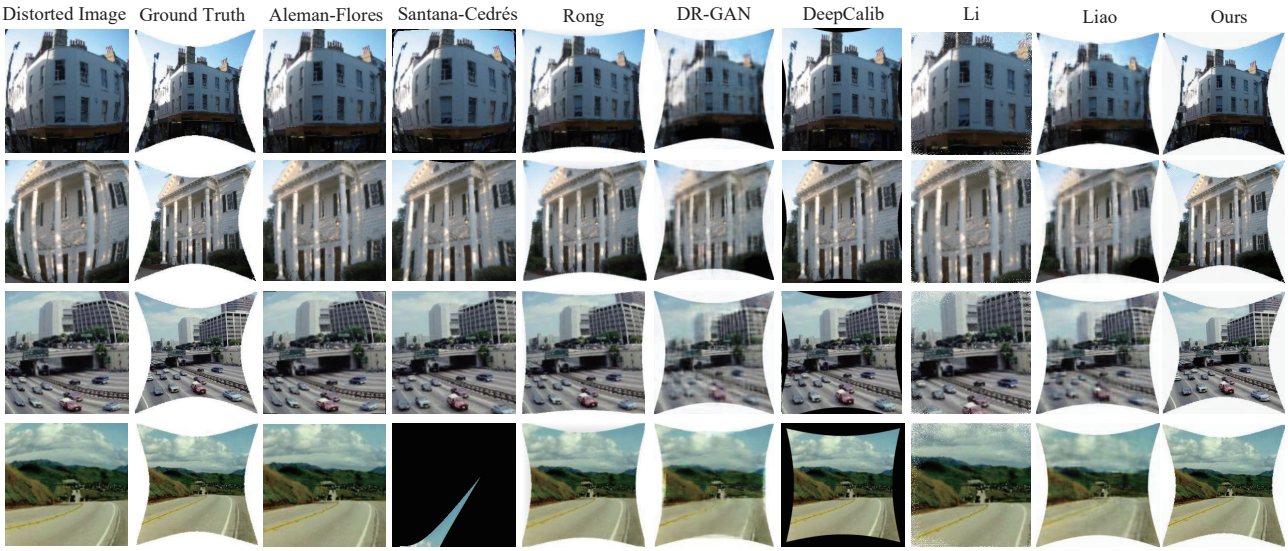


Figure 4. Qualitative results on our synthesized datasets. For each comparison, we show the distorted image, ground truth, and the rectified results of the compared methods: Alemán-Flores [1], Santana-Cedrés [28], Rong [25], DR-GAN [17], DeepCalib [5], Li [16], Liao [19], and our proposed approach, from left to right.

Figure 5. Qualitative results on the real-world distorted images. For each comparison, we show the distorted image and the rectified results from methods: Santana-Cedrés [28], Rong [25], DR-GAN [17], Li [16], Liao [19], and our proposed approach, from left to right.

Table 2. Ablation study of the proposed rectification framework, where HRF = TR + SR + SE + LAC.

Modules					Metrics	
TR	TR+SR	TR+SR+SE	HRF	HRF+DPS	PSNR ↑	SSIM ↑
✓	✗	✗	✗	✗	18.23	0.65
✓	✓	✗	✗	✗	20.12	0.69
✓	✓	✓	✗	✗	23.87	0.75
✓	✓	✓	✓	✗	25.27	0.80
✓	✓	✓	✓	✓	**26.71**	**0.88**

4.4. Ablation Study

We also investigate an ablation study to evaluate each component in the proposed approach as shown in Table 2. We mainly consider the crucial parts in hierarchical rectification framework (HRF) as follows: texture rendering module (TR), structure recovery module (SR), semantics embedding module (SE), location-aligned connection (LAC), and distortion-aware pre-training strategy (DPS). From Table 2, we can observe: (1) TR + SR + SE achieves better performance than only considering one or two construction components, indicating that the proposed multi-level curriculum is beneficial and robust for the distortion rectification. (2) HRF (TR + SR + SE + LAC) obtains higher values on both PSNR and SSIM than TR + SR + SE. Our location-aligned connection can revise the distortion distribution of feature maps from shallow layers and thus boost more useful information messaging in the network. (3) The complete version (HRF + DPS) gains the most considerable improvement over other baselines. As discussed in Section 4.3, the proposed scheme initializes the neural network with the general prior knowledge of distortion, which meets two key requirements of a standard pre-training strategy: easy to learn and helpful for the subsequent task.

4.5. Cross-Domain Evaluation

We further examine the generalization ability of the proposed learning model across different domains, especially

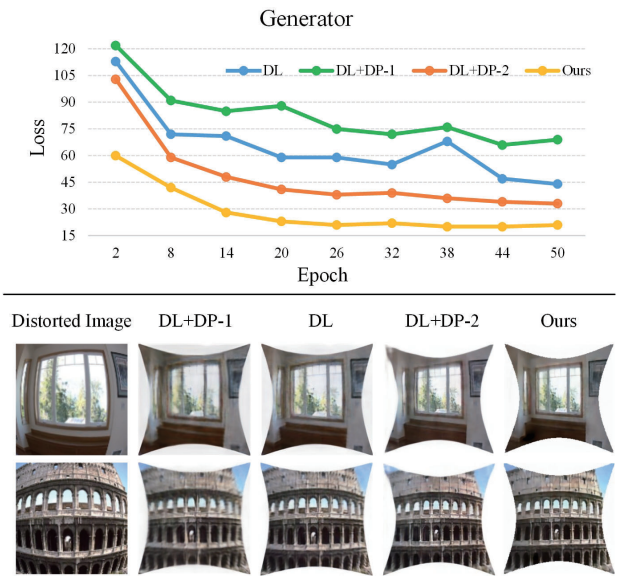


Figure 6. The comparisons of different learning schemes, which are evaluated by the training loss curves (top) and the rectified results (bottom).

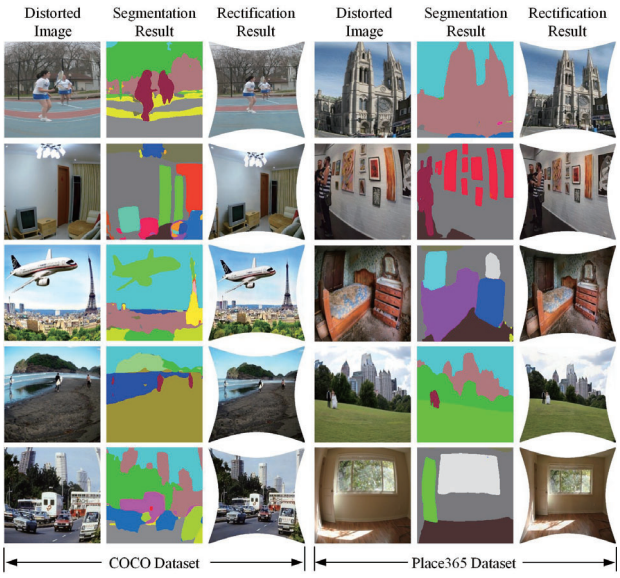


Figure 7. Cross-domain qualitative evaluation on COCO dataset [20] and Place365 dataset [33].

in two prevalent large scale datasets: COCO dataset [20] and Place365 dataset [33]. In the implementation, we leverage the DeNet trained on the ADE20K dataset [34] to perform the semantic segmentation on the distorted images derived from COCO and Place365 datasets. Then, the segmentation results are fed into our distortion rectification module. The experimental results are shown in Fig. 7. As we know, the ADE20K dataset is the largest open-source

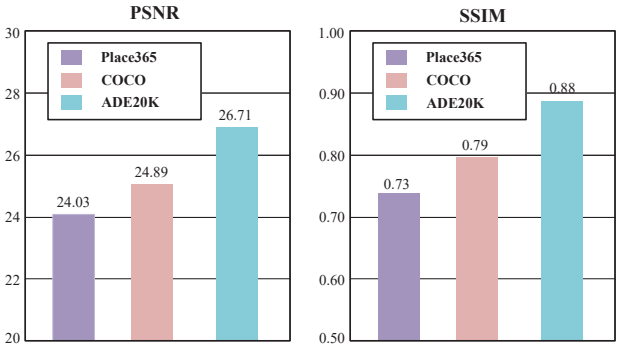


Figure 8. Cross-domain quantitative evaluation on ADE20K dataset [34] COCO dataset [20] and Place365 dataset [33].

dataset for semantic segmentation, covering most of the scenes in life (150 categories in total). Therefore, most segmentation results tested on the COCO dataset and Place365 dataset look plausible and coherent, constructing reasonable rectification results. As shown in Fig. 8, the quantitative evaluations on ADE20K, COCO, and Place365 are reported. Although the qualitative results on COCO dataset and Place365 dataset look plausible and visually pleasing, there are performance decreases compared to ADE20K shown in the quantitative evaluation. The main reason is that the domain difference influences the recognition ability of neural networks. Nevertheless, our learning model still outperforms the methods training on COCO dataset, such as DR-GAN [19] and Liao [17].

5. Conclusions

In this paper, we revisit the challenging barrel distortion rectification task and present a multi-level curriculum with a distortion-aware pre-training strategy for training the deep rectification model. We first propose a simple-to-complex curriculum, following the construction levels of an image. Then, we develop a distortion-aware pre-training strategy to enhance the distortion perception of the deep rectification model. By breaking down the rectification learning process, we design a hierarchical rectification framework consisting of structure recovery, semantics embedding, and texture rendering modules. With the multi-level curriculum and the distortion-aware pre-training strategy, the model learns to rectify distorted images progressively and converges fast. Experimental results demonstrate that our approach outperforms state-of-the-art methods, both quantitatively and qualitatively.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (No.62172032, No.61772066). We thank Yuying Shi from POP MART for her help providing the paintings for Fig. 1(a) (bottom).

References

[1] Miguel Alemán-Flores, Luis Alvarez, Luis Gomez, and Daniel Santana-Cedrés. Automatic lens distortion correction using one-parameter division models. *Image Processing On Line*, 4:327–343, 2014. 2, 5, 6, 7

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017. 5

[3] J. P. Barreto and H. Araujo. Geometric properties of central catadioptric line images and their application in calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1327–1333, 2005. 2

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of International Conference on Machine Learning*, pages 41–48, 2009. 1, 3

[5] Oleksandr Bogdan, Viktor Eckstein, Francois Rameau, and Jean-Charles Bazin. Deepcalib: a deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*, pages 1–10, 2018. 1, 2, 5, 6, 7

[6] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997. 2

[7] Jeffrey L. Elman. Learning and development in neural networks: the importance of starting small. *Cognition*, 48:71–99, 1993. 2

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 2

[9] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4918–4927, 2019. 1

[10] M. Hu, M. Chang, J. Wu, and L. Chi. Robust camera calibration and player tracking in broadcast basketball video. *IEEE Transactions on Multimedia*, 13(2):266–279, 2011. 2

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711, 2016. 5

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[13] H. I. Koo. Segmentation and rectification of pictures in the camera-captured images of printed documents. *IEEE Transactions on Multimedia*, 15(3):647–660, 2013. 2

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 1, 4

[15] Kai A Krueger and Peter Dayan. Flexible shaping: How learning in small steps helps. *Cognition*, 110(3):380–394, 2009. 3

[16] Xiaoyu Li, Bo Zhang, Pedro V Sander, and Jing Liao. Blind geometric distortion correction on images through deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4855–4864, 2019. 1, 2, 5, 6, 7

[17] Kang Liao, Chunyu Lin, Yao Zhao, and Moncef Gabbouj. DR-GAN: Automatic radial distortion rectification using conditional GAN in real-time. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(3):725–733, 2019. 1, 2, 5, 6, 7, 8

[18] K. Liao, C. Lin, Y. Zhao, and M. Gabbouj. Distortion rectification from static to dynamic: A distortion sequence construction perspective. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):3870–3882, 2020. 1

[19] Kang Liao, Chunyu Lin, Yao Zhao, and Mai Xu. Model-free distortion rectification framework bridged by distortion distribution map. *IEEE Transactions on Image Processing*, 29:3707–3718, 2020. 1, 2, 4, 5, 6, 7, 8

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 8

[21] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *Advances in Neural Information Processing Systems*, pages 9605–9616, 2018. 4

[22] David Marr. Vision: A computational investigation into the human representation and processing of visual information. *Inc., New York, NY*, 2(4.2), 1982. 2

[23] Rui Melo, Michel Antunes, João Pedro Barreto, Gabriel Falcão, and Nuno Gonçalves. Unsupervised intrinsic calibration from a single frame using a plumb-line approach. In *IEEE International Conference on Computer Vision*, pages 537–544, 2013. 2

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 5

[25] Jiangpeng Rong, Shiyao Huang, Zeyu Shang, and Xianghua Ying. Radial lens distortion correction using convolutional neural networks trained with synthesized images. In *Asian Conference on Computer Vision*, pages 35–49, 2016. 1, 2, 5, 6, 7

[26] O. Ronneberger, P.Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, volume 9351, pages 234–241, 2015. 3

[27] Terence D Sanger. Neural network learning control of robot manipulators using gradually increasing task difficulty. *IEEE Transactions on Robotics and Automation*, 10(3):323–333, 1994. 3

[28] Daniel Santana-Cedrés, Luis Gomez, Miguel Alemán-Flores, Agust¨ªn Salgado, Julio Esclarín, Luis Mazorra, and Luis Alvarez. An iterative optimization algorithm for lens distortion correction using two-parameter models. *Image Processing on Line*, 6:326–365, 2016. 2, 5, 6, 7

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[30] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017. 5

[31] Zhucun Xue, Nan Xue, Gui-Song Xia, and Weiming Shen. Learning to calibrate straight lines for fisheye image rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1643–1651, 2019. 1, 2

[32] Xiaoqing Yin, Xinchao Wang, Jun Yu, Maojun Zhang, Pascal Fua, and Dacheng Tao. FishEyeRecNet: A multi-context collaborative deep network for fisheye image rectification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 469–484, 2018. 1, 2

[33] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 8

[34] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 4, 5, 8