

QueryCDR: Query-based Controllable Distortion Rectification Network for Fisheye Images

Pengbo Guo^{1,2}, Chengxu Liu^{2,3}, Xingsong Hou^{2(✉)}, and
Xueming Qian^{2,3}

¹ School of Software Engineering, Xi'an Jiaotong University, Xi'an, China

² Xi'an Jiaotong University, Xi'an, China

³ Shaanxi Yulan Jiuzhou Intelligent Optoelectronic Tech. Co., Ltd, Xi'an, China
{guopengbo, chengxuliu}@stu.xjtu.edu.cn, {houxs, qianxm}@mail.xjtu.edu.cn

Abstract. Fisheye image rectification aims to correct distortions in images taken with fisheye cameras. Although current models show promising results on images with a similar degree of distortion as the training data, they will produce sub-optimal results when the degree of distortion changes and without retraining. The lack of generalization ability for dealing with varying degrees of distortion limits their practical application. In this paper, we take one step further to enable effective distortion rectification for images with varying degrees of distortion without retraining. We propose a novel Query-based Controllable Distortion Rectification network for fisheye images (QueryCDR). In particular, we first present the Distortion-aware Learnable Query Mechanism (DLQM), which defines the latent spatial relationships for different distortion degrees as a series of learnable queries. Each query can be learned to obtain position-dependent rectification control conditions, providing control over the rectification process. Then, we propose two kinds of controllable modulating blocks to enable the control conditions to guide the modulation of the distortion features better. These core components cooperate with each other to effectively boost the generalization ability of the model at varying degrees of distortion. Extensive experiments on fisheye image datasets with different distortion degrees demonstrate our approach achieves high-quality and controllable distortion rectification. Code is available at <https://github.com/PbGuo/QueryCDR>.

Keywords: Fisheye image · Distortion rectification · Controllable

1 Introduction

Benefiting from the huge field-of-view (FoV), fisheye cameras are widely utilized in various fields, including security surveillance [26, 34] and autonomous driving [11, 38]. However, the distortion brought by the fisheye lenses greatly limits the performance of downstream vision tasks [9, 21, 36, 60]. How to eliminate the distortion in fisheye images has attracted great attention in recent years.

Early methods [3, 5, 14, 18, 37, 40, 46, 59] primarily relied on identifying matching feature points or curves for automatic rectification. However, constrained by

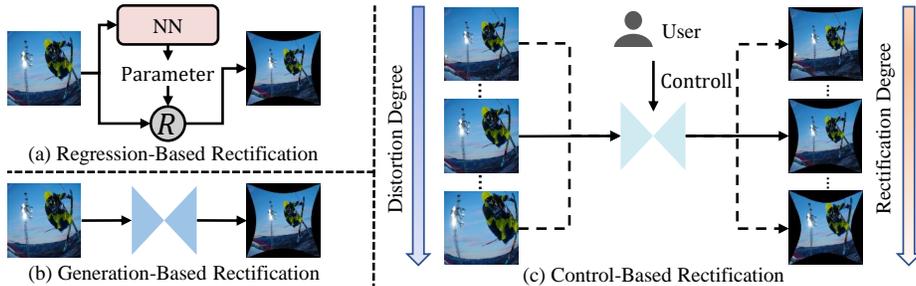


Fig. 1: Different approaches to fisheye image distortion rectification. (a) Regression-Based: Using a neural network to predict distortion-related parameters, then apply rectification algorithms R for rectification. (b) Generation-Based: Input the distorted fisheye image and directly generate the rectified image end-to-end. (c) Control-Based: Users provide control conditions to guide the rectification process, resulting in promising rectified images of various distortion degrees.

the instability of feature detection, both the generalization of the algorithm and the quality of the rectified images were unsatisfactory [10, 24, 52]. In recent years, owing to the robust learning and generalization capabilities of neural networks, deep learning-based fisheye image rectification methods have become mainstream. These methods can be divided into two categories: regression-based rectification [4, 23, 39, 50, 56] and generation-based rectification [10, 24, 25, 52, 54]. The former uses deep regression models to predict distortion parameters for image reconstruction, as shown in Fig. 1(a). The latter uses an encoder-decoder structure to generate well-rectified images directly, as shown in Fig. 1(b).

It is worth noting that these methods only achieve satisfactory results on a similar degree of distortion as the training data. This means that when handling different degrees of distortion without retraining, there will be a significant decrease in the quality of the rectified images [25]. This is because the model tends to learn fixed position mapping relationships during training. When the distortion degree changes, this relationship does not work for the new distortion distribution. Therefore, these models usually need to be retrained on images with different distortion degrees. In addition, fisheye image acquisition is difficult, and re-collecting new datasets with varying degrees of distortion would incur significant costs [52, 54]. Thus, it is essential to explore a model that can handle different degrees of distortion simultaneously.

In recent years, some methods have been proposed to achieve controllable image restoration [7, 15, 16, 33, 47, 48, 55, 61]. Typically, CFSNet [47], MM-RealSR [33], and Yao *et al.* [55] introduce scalars as control conditions to effectively restore degradation of varying degrees. It would be a promising solution to introduce an effectively controllable mechanism to deal with all degrees of distortion, as shown in Fig. 1(c). However, directly applying existing controllable mechanisms in restoration tasks to the fisheye rectification model suffers from the following challenges: 1) There is a gap between optimization objectives (*i.e.* the roles

of the controllable mechanisms). Restoration models learn the pixel-level detail restoration, whereas fisheye rectification models learn the spatial-level positional mapping relationships [10, 25]. The control mechanism lacking position information in restoration tasks is ineffective in controlling distortion rectification networks. 2) There is a gap between optimization difficulties (*i.e.* the roles of the control conditions). The distortion of a fisheye image increases gradually from the image center to boundary [51]. Therefore, it is not suitable to use a single scalar control condition in restoration tasks to deal with such spatially varying distortions. These challenges restrict the application of controllable mechanisms and control conditions in fisheye image rectification.

To address these issues, we propose the Query-based Controllable Distortion Rectification network (QueryCDR), as shown in Fig. 2. By introducing a series of learnable queries as control conditions, QueryCDR allows the users to achieve fisheye rectification with different distortion degrees. Specifically, to incorporate positional mapping relationships into control conditions, we introduce the Distortion-aware Learnable Query Mechanism (DLQM), which defines a series of queries representing different rectification control conditions. During inference, DLQM extracts position-dependent control conditions from the user-given query and feeds them into the network for controlling the rectification process. Furthermore, to enable the control conditions to guide the rectification efficiently, we propose two types of controllable modulating blocks: the Controllable Convolution Modulating Block (CCMB) based on CNN [17], and the Controllable Attention Modulating Block (CAMB) based on Transformer [45]. They are good at extracting local texture features and learning long-range distortion mapping relationships, respectively. By combining CCMB and CAMB, we construct a robust controllable rectification network. Our QueryCDR can handle various distortions without retraining, enhancing the generalization ability of the fisheye rectification model.

We summarize our contributions as follows:

- We propose QueryCDR, a Query-based Controllable Distortion Rectification network for fisheye images. Extensive experiments demonstrate that our QueryCDR can deliver superior results on a variety of distortion degrees.
- We propose the Distortion-aware Learnable Query Mechanism (DLQM), which effectively introduces the latent spatial relationships to control conditions for fisheye image rectification.
- We propose two kinds of blocks for modulating features using control conditions: the Controllable Convolution Modulating Block (CCMB) and the Controllable Attention Modulating Block (CAMB). They can effectively utilize control conditions to guide the rectification process.

2 Related Work

2.1 Traditional Fisheye Image Rectification

Traditional rectification methods can be divided into two types, multi-view-based and line-based methods. Multi-view-based methods [3, 14, 18, 20, 37, 41, 43]

calibrated fisheye images by finding corresponding feature points from multiple viewpoints. Line-based methods [2, 5, 8, 32, 40, 44, 46, 59] employed line detection to rectify the curved lines, thereby achieving distortion rectification. However, these methods require manual intervention or handcrafted feature extractors, and the rectification process is unstable, failing to achieve satisfactory results.

2.2 Deep Learning Based Fisheye Image Rectification

Rapid advances in deep learning have allowed them to shine in low-level vision tasks [6, 22, 28–31, 49, 58]. As a core task in low-level vision, distortion image rectification has received increasing attention [10, 24, 39, 52, 54, 56]. According to the network architecture, deep learning-based distortion image rectification methods can be categorized into two types, regression-based methods [4, 23, 39, 50, 56] and generation-based methods [10, 24, 25, 52, 54].

Regression-based methods [4, 23, 39, 50, 56] utilize neural networks to predict distortion-related coefficients for rectifying the image. Rong *et al.* [39] were the first to use CNNs for fisheye image rectification. They employed the network to predict across multiple distortion intervals, achieving preliminary rectification results. Yin *et al.* [56] integrated semantics as prior information to guide rectification. While these methods have shown some performance improvement, they face the challenge of non-end-to-end design, thus requiring a large number of additional labels, and increasing operational costs. Therefore, to reduce models’ complexity, Generation-based methods [10, 24, 25, 52, 54] employ generative networks to take distorted images as input and directly generate rectified images. DR-GAN [24] was the first to apply a generative adversarial network (GAN) [12] to fisheye image rectification, enabling the network to directly generate rectified images without estimating additional parameters. DDM [25] introduced distortion maps to help the model better learn the distortion distribution. PCN [52] designed a flow estimation module to predict appearance flows in fish-eye images, using it to assist the rectification progress. SimFIR [10] introduced a self-supervised rectification module, allowing the network to better learn the distortion representations at different locations. However, both paradigms require retraining when handling images with different distortion degrees, failing to address the issue of weak model generalization.

2.3 Controllable Low-level Vision

To address the issue of varying degradation levels in low-level vision tasks, an increasing number of studies [7, 15, 16, 33, 47, 48, 55, 61] propose controllable network architectures to tackle this challenge. DNI [48] interpolated all parameters of different restoration networks, which were trained with different degradation levels. By adjusting the interpolation coefficients, a smooth control of the image can be achieved. AdaFM [15] achieved better results by inserting AdaFM layers after each convolution layer to change the filters’ statistics, thus the users can interactively manipulate the restoration results by tuning a control coefficient. CFS-Net [47] introduced the tuning branch to adaptively learn the control coefficients,

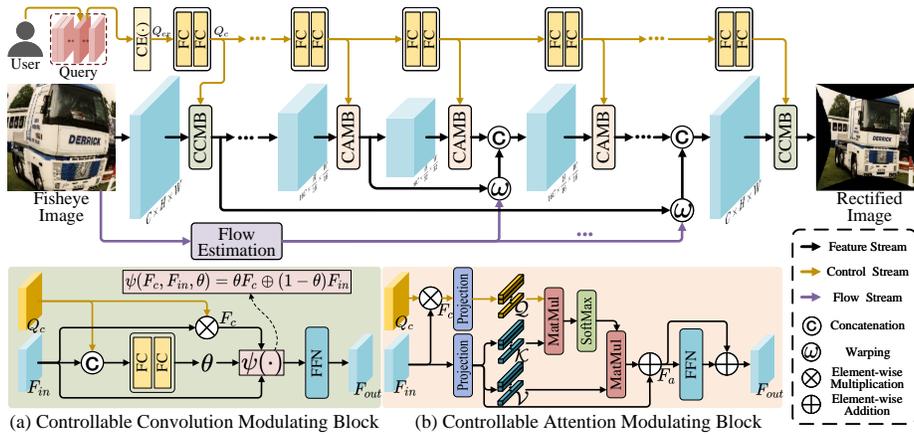


Fig. 2: Overview of our proposed Query-based Controllable Distortion Rectification network (QueryCDR). The Distortion-aware Learnable Query Mechanism (DLQM) extracts control conditions from user-given queries and feeds them layer by layer into the rectification network. The rectification network is composed of Controllable Convolution Modulating Blocks (CCMB) and Controllable Attention Modulating Blocks (CAMB), which modulate the input features F_{in} with control conditions F_c , enabling controllable rectification process.

and then use them to couple the features with the main branch. MM-RealSR [33] proposed a metric learning strategy to map unquantifiable degradation levels to a metric space as control conditions. However, due to the different optimization objectives and complexity of the control mechanisms, directly applying these methods to fisheye image distortion rectification tasks will not achieve effective control over the rectification process.

3 Methodology

3.1 Overview

The overview of our proposed Query-based Controllable Distortion Rectification network (QueryCDR) is shown in Fig. 2. First, following existing work [52], the input fisheye image is fed into the flow estimation module to obtain the appearance flow, which performs a coarse-grained rectification for image features (*i.e.*, warping $\omega(\cdot)$). Then, the Distortion-aware Learnable Query Mechanism (DLQM) (in Sec. 3.2) extracts control conditions from the user-given query and feeds them layer by layer into the rectification network. Finally, a U-shaped hierarchical network composed of several controllable modulating blocks (in Sec. 3.3) is used to rectify the distorted input image. These blocks modulate features with the control conditions given by DLQM, to get the final output.

In the following, we will provide a comprehensive description of each module and its corresponding role in the QueryCDR.

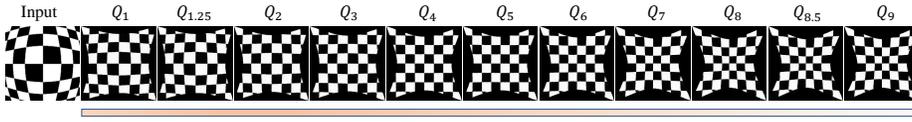


Fig. 3: Given an input image, QueryCDR can accurately produce results with different rectification degrees by feeding different queries. Moreover, by interpolating between different queries, we can achieve smooth continuous rectification for any distortion degree. For examples, $Q_{1.25} = 0.75Q_1 + 0.25Q_2$, and $Q_{8.5} = 0.5Q_8 + 0.5Q_9$.

3.2 Distortion-aware Learnable Query Mechanism (DLQM)

To achieve control over the rectification process, most existing controllable methods [33,47,55] introduce scalars as control conditions to represent the degradation degrees for image reconstruction. However, due to the peculiarities of distortion in fisheye images [10,42,53], these methods fail to achieve effective control over the rectification process. To address this issue, we propose the DLQM. It maintains a learnable query set, providing diverse effective control conditions for the rectification network. During training, DLQM projects the learned high-dimensional positional mapping relationship into a low-dimensional latent space of queries. During inference, DLQM extracts the control information from the queries and converts it into corresponding control conditions for each layer of the rectification network. Simply by providing different queries, DLQM can effectively control the rectification process, as shown in Fig. 3. At the same time, our QueryCDR can also achieve smooth continuous rectification for any distortion degrees by simply interpolating between different queries.

Specifically, we construct a query set $\mathbf{Q}_s = \{Q_i \mid Q_i \in \mathbb{R}^{C_{in} \times H_{in} \times W_{in}}, i = 1, 2, \dots, N\}$, which comprises N queries representing different distortion degrees. H_{in} , W_{in} , and C_{in} represent the query’s height, width, and channel, respectively. Each query has the same size as the input image $I_{in} \in \mathbb{R}^{C_{in} \times H_{in} \times W_{in}}$. During inference, users select a query Q_i from \mathbf{Q}_s and feed it into the DLQM. In DLQM, the control extracting part $\text{CE}(\cdot)$ is firstly used to extract the features in Q_i . The input processing part can be expressed as follows,

$$Q_{ex} = \text{CE}(Q_i), \quad (1)$$

where $Q_{ex} \in \mathbb{R}^{C_{in} \times H_{in} \times W_{in}}$ is the extracted feature, and serves as the input to the first control layer of DLQM. To minimize the computational costs and effectively extract the control information, $\text{CE}(\cdot)$ is composed of three convolution layers with a kernel size of 3×3 .

Subsequently, DLQM provides corresponding control conditions to the respective layers of the rectification network. To provide appropriate control conditions, each control layer of DLQM comprises two fully connected layers $\text{FC}_1(\cdot)$, $\text{FC}_2(\cdot)$. The l -th layer of DLQM can be represented as follows,

$$Q_c^l = \text{FC}_2^l(\text{FC}_1^l(Q_c^{l-1})), \quad (2)$$

where Q_c^{l-1} is the output control condition from the $(l-1)$ -th control layer. For $l=1$, Q_c^0 represents the extracted feature Q_{ex} . The output Q_c^l of the l -th control layer is fed into the l -th rectification layer as a control condition, and simultaneously serves as input to the $(l+1)$ -th control layer of DLQM.

Afterward, we feed the control conditions into the rectification network. Each layer’s controllable modulating block modulates the input feature with the control condition, guiding the modulated features toward the desired direction.

3.3 Controllable Modulating Block

When controlling the rectification process, to avoid obtaining results with blurred texture details or residual distortions, we introduce two types of controllable modulating blocks tailored for reconstructing local texture details and learning continuous distortion patterns. They are denoted as the Controllable Convolution Modulating Block (CCMB) based on CNN [17] (in Fig. 2(a)) and the Controllable Attention Modulating Block (CAMB) based on Transformer [45] (in Fig. 2(b)). The CCMB can adaptively fuse the original features and the controlled features, preserving more local details. The CAMB can better capture the spatial distortion information, guaranteeing the integrity of the recovered content [10,35,53]. To balance performance and computational cost, our QueryCDR constructs a U-shaped rectification network [52] composed of CCMB and CAMB. In the first three layers with larger feature maps, *i.e.*, $l = \{1, 2, 3, 9, 10, 11\}$, CCMB is employed to learn more local texture details. In the remaining layers with smaller feature maps, *i.e.*, $l = \{4, 5, 6, 7, 8\}$, CAMB is utilized to capture more global dependencies within the images.

Controllable Convolution Modulating Block (CCMB). When incorporating control conditions into the rectification network, existing methods [47,55] fail to effectively balance the fusion ratio between original features and controlled features. Directly using the controlled features or fusing them with original features at a fixed ratio will degrade the quality of the rectified images. To harmoniously incorporate control conditions into the rectification process, we design the CCMB, as shown in Fig. 2(a). CCMB can dynamically find an optimal ratio to modulate the distorted features with control conditions.

Specifically, CCMB receives an input feature $F_{in} \in \mathbb{R}^{C \times H \times W}$ and a control condition $Q_c \in \mathbb{R}^{C \times H \times W}$. Both of them are used to predict the fusion ratio. Here we omit the layer information for brevity,

$$\theta = \text{CP}(F_{in}, Q_c), \quad (3)$$

where $\text{CP}(\cdot)$ is the coefficient predictor composed of two fully connected layers, takes the concatenation of F_{in} and Q_c as input and predicts the fusion ratio θ .

Then, we perform element-wise multiplication \otimes on F_{in} and Q_c to yield the controlled features $F_c \in \mathbb{R}^{C \times H \times W}$, which can be expressed as,

$$F_c = F_{in} \otimes Q_c. \quad (4)$$

Finally, the input features F_{in} and controlled features F_c are combined in a weighted sum according to the fusion ratio θ , illustrated as $\psi(\cdot)$ in Fig. 2(a),

$$F_{out} = \psi(F_c, F_{in}, \theta) = \theta F_c \oplus (1 - \theta) F_{in}, \quad (5)$$

where \oplus represents the element-wise addition, and $F_{out} \in \mathbb{R}^{C \times H \times W}$ is the final output of CCMB.

Due to this dynamic modulation mechanism, CCMB achieves effective control over the rectification process while preserving richer texture details.

Controllable Attention Modulating Block (CAMB). Due to the lack of perception of global information, CNN-based networks struggle to learn long-range dependencies, particularly the continuous and amorphous distortions prevalent in fisheye distortions [10, 42, 53]. In contrast, Transformer-based networks effectively compensate for this, with their global attention mechanism [13, 35].

Therefore, we propose the CAMB, as illustrated in Fig. 2(b). To optimally leverage the control conditions given by DLQM, we designed the control-attention mechanism, enabling CAMB to perceive the global spatial relationships in the control conditions effectively. Specifically, we unfold and project the controlled feature F_c (in Eq. (4)) as the query $\mathcal{Q} \in \mathbb{R}^{m \times L}$, with the input feature F_{in} as the key $\mathcal{K} \in \mathbb{R}^{m \times L}$ and value $\mathcal{V} \in \mathbb{R}^{m \times L}$, where $L = H \times W$ represents the sequence length, and m denotes the dimensions of the sequences. The control-attention is described by the following,

$$\begin{aligned} \mathcal{Q} &= W_{\mathcal{Q}} F_c, \mathcal{K} = W_{\mathcal{K}} F_{in}, \mathcal{V} = W_{\mathcal{V}} F_{in}, \\ \text{CTRL-ATTN}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) &= \text{softmax}\left(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{m}}\right)\mathcal{V}, \end{aligned} \quad (6)$$

where $W_{\mathcal{Q}}, W_{\mathcal{K}}, W_{\mathcal{V}} \in \mathbb{R}^{m \times C}$ represent the projection matrices of the queries, keys, and values, respectively. CTRL-ATTN(\cdot) represents the control-attention we proposed. The overall computation of CAMB can be formulated as follows,

$$\begin{aligned} F_a &= \text{CTRL-ATTN}(\text{LN}(\mathcal{Q}, \mathcal{K}, \mathcal{V})) \oplus F_{in}, \\ F_{out} &= \text{Conv}_{1 \times 1}(\text{FFN}(\text{LN}(F_a)) \oplus F_a), \end{aligned} \quad (7)$$

where F_{in} means the input feature, $F_a \in \mathbb{R}^{m \times H \times W}$ and $F_{out} \in \mathbb{R}^{C \times H \times W}$ are the outputs of CTRL-ATTN(\cdot) and CAMB, respectively. LN(\cdot) denotes the layer normalization [1]. FFN(\cdot) stands for the feed-forward network composed of three fully connected layers, which helps CAMB to focus on the global dependencies.

CAMB can discern the global mapping relationships within fisheye images, ensuring rectification uniformity compared to the CNN-based networks.

3.4 Training Strategy

To guarantee the robust and stable controllable distortion rectification of Query-CDR, we design a two-stage training strategy that combines coarse-grained distortion pre-training and fine-grained distortion fine-tuning.

During the coarse-grained distortion pre-training phase, we choose the most commonly used dataset [10,24,52,54] that only contains one degree of distortion to train our QueryCDR. For clarity, we denote the distortion degree in this phase as d . Correspondingly, only one query, denoted as Q , is used for training. The optimization objective can be expressed as follows,

$$\mathcal{L}_{pre} = \mathcal{L}_r + \mathcal{L}_m, \quad (8)$$

where \mathcal{L}_{pre} is the overall loss function for the pre-training phase. \mathcal{L}_r denotes the reconstruction loss,

$$\mathcal{L}_r = \|I_{out}^d - I_{gt}^d\|_1, \quad (9)$$

where I_{out}^d and I_{gt}^d signify the output result and ground truth, respectively. \mathcal{L}_m denotes the multi-scale loss,

$$\mathcal{L}_m = \sum_{j=1}^{Z-1} \|S(I_{gt}^d, j) - C(F_{out}^j)\|_1, \quad (10)$$

where $S(\cdot)$ represents the operation that down-samples the input I_{gt}^d by a factor of $1/2^j$. Z represents the number of decoder’s layers, and we set Z to 6. F_{out}^j denotes the feature in j -th decoder layer. $C(\cdot)$ is 3×3 convolution for decoding the features into 3-channel RGB images. In this way, each feature map on the decoder can be effectively supervised. The I_{out}^d can be obtained as,

$$I_{out}^d = \text{QueryCDR}(I_{in}^d, Q), \quad (11)$$

where I_{in}^d represents fisheye images with distortion degree d as input. This way can effectively boost the model stability and accelerate the convergence.

During the fine-grained distortion fine-tuning phase, we use the varying distortion degrees datasets to fine-tune our QueryCDR. Before training, We replicate the weight of the query Q to the other queries to accelerate convergence on other distortion degrees. Subsequently, we fine-tune our QueryCDR using fisheye images with varying distortion degrees, and feeding corresponding queries into QueryCDR for training at the same time. This allows the query set to efficiently acquire diverse latent spatial relationships. The optimization objective can be expressed as follows,

$$\mathcal{L}_{fine}^{d_i} = \mathcal{L}_r^{d_i} + \mathcal{L}_m^{d_i}, \quad (12)$$

where $\mathcal{L}_{fine}^{d_i}$ is the overall fine-tuning loss function for distortion degree $d_i, i \in \{1, 2, \dots, 9\}$, which is similar to the \mathcal{L}_{pre} but calculated across different distortion degrees. The $I_{out}^{d_i}$ can be obtained as,

$$I_{out}^{d_i} = \text{QueryCDR}(I_{in}^{d_i}, Q_i), \quad (13)$$

where $I_{in}^{d_i}$ denotes fisheye images with a distortion degree of d_i as input, Q_i denotes the query corresponding to $d_i, i \in \{1, 2, \dots, 9\}$.

With this two-stage training strategy, our QueryCDR can effectively utilize only a small amount of varying distortion degrees data that is difficult to obtain, to rapidly finish the training of our network.

Table 1: Quantitative comparison (PSNR (dB) \uparrow , SSIM \uparrow) on COCO [27] fisheye image dataset with varying distortion degrees. **Red** indicates the best and **blue** indicates the second best performance (best viewed in color).

Method	PSNR									
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	Avg
SC [40]	10.05	11.59	11.81	11.03	11.50	10.19	11.11	10.19	9.14	10.73
DeepCalib [4]	10.00	10.69	11.01	11.19	11.28	11.46	11.45	11.13	11.11	11.04
Blind [23]	12.98	11.24	10.30	9.62	9.99	11.75	12.69	12.75	12.80	11.57
DR-GAN [24]	15.68	17.40	17.97	18.34	18.50	18.44	17.95	17.94	17.47	17.74
PCN [52]	14.93	17.43	18.43	18.86	18.86	18.88	18.74	17.35	18.26	17.97
DDA [54]	16.39	17.41	17.43	19.48	20.12	18.90	18.85	18.17	18.22	18.33
SimFIR [10]	16.57	17.88	18.43	18.97	19.31	19.28	19.19	18.65	18.48	18.53
QueryCDR	20.01	20.29	20.39	20.41	20.72	20.81	20.58	19.11	20.53	20.32
Method	SSIM									
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	Avg
SC [40]	0.101	0.113	0.149	0.182	0.283	0.175	0.141	0.126	0.093	0.151
DeepCalib [4]	0.184	0.210	0.223	0.230	0.234	0.246	0.250	0.245	0.246	0.229
Blind [23]	0.308	0.244	0.199	0.176	0.194	0.296	0.367	0.395	0.420	0.289
DR-GAN [24]	0.295	0.330	0.339	0.344	0.344	0.332	0.314	0.312	0.299	0.323
PCN [52]	0.420	0.547	0.589	0.607	0.608	0.610	0.615	0.576	0.603	0.575
DDA [54]	0.455	0.589	0.592	0.620	0.675	0.626	0.619	0.564	0.581	0.591
SimFIR [10]	0.492	0.581	0.626	0.635	0.640	0.628	0.622	0.591	0.595	0.601
QueryCDR	0.643	0.665	0.668	0.677	0.688	0.699	0.692	0.656	0.693	0.676

4 Experiment

4.1 Experimental Settings

To demonstrate the effectiveness of our proposed QueryCDR, we followed the existing works [10, 24, 52, 54], employed the four-parameter polynomial model to synthesize fisheye images. We constructed synthetic datasets based on the original images of COCO [27] and Places2 [62] datasets, respectively. Specifically, for images with distortion degree d , 40,000 images were used for the pre-training stage. And for images with varying distortion degrees $d_i, i \in \{1, 2, \dots, 9\}$, 18,000 images were used for the fine-tuning stage, and 3,600 images for testing. The images were resized to 256×256 when fed into the network. For a fair comparison, we followed existing works [10, 24, 52, 54] utilizing a batch size of 16 and the Adam [19] optimizer with a learning rate of $1e-4$.

4.2 Performance Evaluation

To evaluate the performance of our method, we retrained and validated existing fisheye image rectification methods on synthetic fisheye datasets with 9 distortion degrees. The methods can be summarized into three categories: traditional

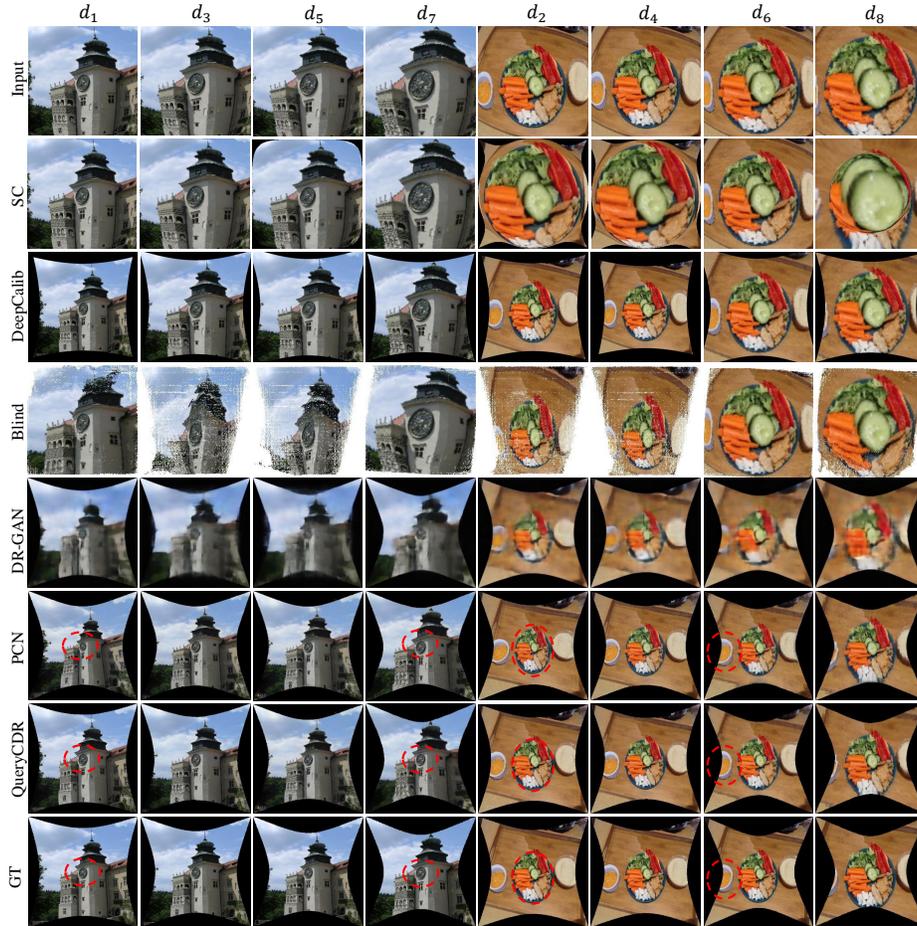


Fig. 4: Qualitative results on synthetic fisheye images.

method SC [40], regression-based methods DeepCalib [4] and Blind [23], and generation-based DR-GAN [24], PCN [52], DDA [54] and SimFIR [10]. For a fair comparison, we followed existing works [10, 24, 52, 54] employing Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) to quantify performance.

Quantitative Results. Our performance comparisons are shown in Tab. 1. Compared to existing methods, QueryCDR achieves the best performance across all distortion degrees without retraining. It is because DLQM effectively utilizes the control information in queries to guide the network in achieving varying degrees of rectification. Additionally, the capability of CCMB and CAMB allows for controlled rectification while obtaining high-quality results. Moreover, we observe that even on images with distortion degree d_5 , which the existing

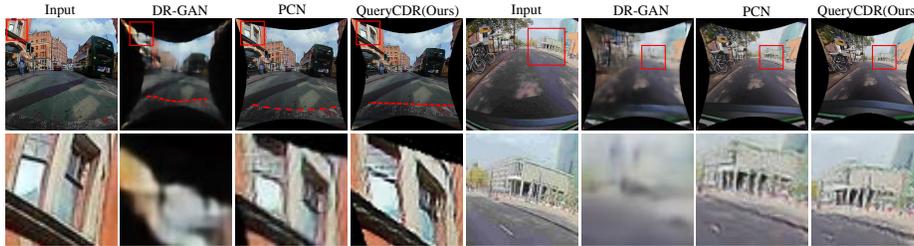


Fig. 5: Qualitative results on real-world fisheye images.

methods perform well [10, 24, 52, 54], QueryCDR still outperforms the second best method [54] by **0.60** dB of PSNR and **0.013** of SSIM. This further demonstrates the outstanding generalization ability of QueryCDR.

Qualitative Results. To further compare the visual qualities of different methods, we show the results rectified by QueryCDR and other rectification methods in Fig. 4. It can be observed intuitively that other methods fail to effectively rectify images with varying degrees of distortion. Particularly the objects in images, our QueryCDR maintains the accurate structure of objects across different distortions. Furthermore, benefiting from the modulation capability of CCMB and CAMB, QueryCDR preserves richer texture details after rectification.

In addition, to further demonstrate the generalization ability of QueryCDR, we conducted experiments on real-world fisheye image datasets [57] in Fig. 5. Despite the disparities between synthetic and real-world datasets, our QueryCDR still shows robust rectification capabilities, and preserves richer texture details. These results further validate its practicality in real-world scenarios.

4.3 Ablation Study

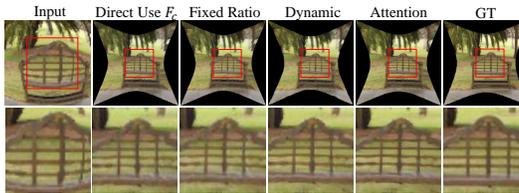
In this section, we conduct ablation for different control conditions, and study the effect of the controllable modulating blocks and the architecture setting.

Distortion-aware Learnable Query Mechanism (DLQM). To validate the effectiveness of DLQM in controlling the rectification process, we implemented different rectification networks controlled by scalar, fixed query, and our learnable query in Tab. 2. When using scalar to control, the value of the scalar increments with the distortion of the entire image. When using fixed query to control, the parameters of the query are set to incremental values from 0 to 1 with the degree of distortion from the center to the edges. Following the settings in Sec. 4.1, we use the learnable query $Q_i, i \in \{1, 2, \dots, 9\}$ to learn the distortion distribution corresponding to $d_i, i \in \{1, 2, \dots, 9\}$.

We trained different methods using the same strategy as illustrated in Sec. 3.4, the results are shown in Tab. 2 and Fig. 6. All three controllable methods outperform the uncontrollable method significantly at various degrees of distortion.

Table 2: Performance comparison of different control methods.

Method	PSNR									
	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	Avg
W/o Control	14.93	17.43	18.43	18.86	18.86	18.88	18.74	17.35	18.26	17.97
Scalar	19.52	20.25	19.97	19.43	19.89	20.07	20.03	18.76	20.15	19.78
Fixed Query	20.13	20.03	20.14	19.88	20.16	20.25	20.16	18.74	19.84	19.93
Learnable Query	20.01	20.29	20.39	20.41	20.72	20.81	20.58	19.11	20.53	20.32

**Fig. 6:** Visual results of different control mechanisms.**Fig. 7:** Visual results of different modulation methods.

This is because the controllable mechanism effectively assists the rectification network in distinguishing between different degrees of distortion, thereby enhancing the generalization capability of the rectification network. The improvement brought by scalar demonstrates the crucial role of controllable mechanisms in handling fisheye images with different degrees of distortion. The further enhancement with fixed query validates the importance of employing higher-dimensional control conditions for fisheye image rectification. Lastly, our tailored learnable query significantly outperforms all other controllable methods, verifying the unique superiority of our approach in controllable rectification.

Controllable Modulating Block. To demonstrate the capability of the CCMB and CAMB in modulating input features with control conditions, we implemented two comparison methods with different modulating approaches: one directly uses the controlled feature (in Eq. (4)), and the other adds controlled feature and original feature in a fixed 1:1 ratio. We constructed different controllable networks using each of these methods.

As shown in Tab. 3 and Fig. 7, retaining original features and fusing them with a 1:1 ratio improved by 0.03 dB, highlighting the importance of preserving original features. The dynamic modulation mechanism boosted performance by 0.06 dB, verifying its effectiveness in dynamic feature modulation. Furthermore, leveraging the control-attention mechanism led to a performance improvement of 0.12 dB, demonstrating the superiority of CAMB for perceiving global distortion distribution, and achieving better uniform global rectification.

Table 3: Ablation study of different modulation methods. Where Direct Use F_c means directly using the controlled feature F_c , Fixed Ratio means adding controlled features and original features in a 1:1 ratio, Dynamic Mechanism means the CCMB and Attention Mechanism means the CAMB.

Method	PSNR	SSIM
Direct Use F_c	20.14	0.655
Fixed Ratio	20.17	0.658
Dynamic Mechanism	20.20	0.669
Attention Mechanism	20.26	0.671

Table 4: Ablation study of different network architectures. Where $x\text{C} + (11 - x)\text{A}$ denotes that the network uses CCMB in the x layers with larger feature maps and CAMB in the remaining $11 - x$ layers with smaller feature maps.

Method	Flops(G)	Param(M)	PSNR	SSIM
PCN [52]	12.305	35.637	17.97	0.575
11C+0A	12.736	37.701	20.20	0.669
8C+3A	13.383	46.398	20.27	0.665
6C+5A	12.353	43.244	20.32	0.676
4C+7A	12.538	46.994	20.31	0.670
0C+11A	15.190	51.795	20.26	0.671

Controllable Rectification Network Architecture. As described in Sec. 3.3, rectification networks based on CCMB may ignore the long-range dependencies, while those based on CAMB may ignore the texture details and also increase computational costs. To strike a balance between these two architectures, We incorporate CCMB into the layers with larger feature map size, and integrate CAMB into the layers with smaller feature map size. To validate the effectiveness of this hybrid architecture and find an optimal combination, we conducted performance and computational cost comparisons between pure CCMB, pure CAMB, and multiple hybrid networks, as shown in Tab. 4. Pure CAMB network (*i.e.*, Row 6) outperforms pure CCMB network (*i.e.*, Row 2) in rectification performance but incurs higher computational costs. Hybrid network architectures effectively solve this problem, with higher performance and fewer parameters. After a trade-off between performance and parameters, we empirically choose the 6C + 5A hybrid architecture as our final model.

5 Conclusion

In this paper, we propose the Query-based Controllable Distortion Rectification network for fisheye images (QueryCDR), which achieves controllable rectification at different distortion degrees without retraining. In particular, we design a series of learnable queries as control conditions to guide the rectification process. Additionally, we design two different controllable modulating blocks, achieving controllable rectification while improving image quality. Extensive experiments have demonstrated the robustness and effectiveness of our QueryCDR. In the future, it is expected to further rectify distortions of different degrees by auto-controlling mechanisms to avoid human involvement. We believe our work provides an effective solution for fisheye camera applications.

Acknowledgements

This work was supported in part by the NSFC under Grant 62272376, 62272380 and 62103317, the Key Research and Development Program of Shaanxi Province under Grant 2020ZDLGY04-05, the Fundamental Research Funds for the Central Universities, China (xzy022023051), the Innovative Leading Talents Scholarship of Xi'an Jiaotong University (Corresponding author: Xingsong Hou).

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Barreto, J.P., Araujo, H.: 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)
3. Barreto, J.P., Daniilidis, K.: Fundamental matrix for cameras with radial distortion. In: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1. vol. 1, pp. 625–632. IEEE (2005)
4. Bogdan, O., Eckstein, V., Rameau, F., Bazin, J.C.: Deepcalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In: Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production. pp. 1–10 (2018)
5. Bukhari, F., Dailey, M.N.: Automatic radial distortion estimation from a single image. *Journal of mathematical imaging and vision* **45**, 31–45 (2013)
6. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: An end-to-end system for single image haze removal. *IEEE transactions on image processing* **25**(11), 5187–5198 (2016)
7. Cai, H., He, J., Qiao, Y., Dong, C.: Toward interactive modulation for photo-realistic image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 294–303 (2021)
8. Devernay, F., Faugeras, O.: Straight lines have to be straight. *Machine vision and applications* **13**, 14–24 (2001)
9. Duan, Z., Tezcan, O., Nakamura, H., Ishwar, P., Konrad, J.: Rapid: rotation-aware people detection in overhead fisheye images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 636–637 (2020)
10. Feng, H., Wang, W., Deng, J., Zhou, W., Li, L., Li, H.: Simfir: A simple framework for fisheye image rectification with self-supervised representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12418–12427 (2023)
11. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *Communications of the ACM* **63**(11), 139–144 (2020)
13. Guo, J., Han, K., Wu, H., Tang, Y., Chen, X., Wang, Y., Xu, C.: Cmt: Convolutional neural networks meet vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12175–12185 (2022)

14. Hartley, R., Kang, S.B.: Parameter-free radial distortion correction with center of distortion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**(8), 1309–1321 (2007)
15. He, J., Dong, C., Qiao, Y.: Modulating image restoration with continual levels via adaptive feature modification layers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11056–11064 (2019)
16. He, J., Dong, C., Qiao, Y.: Interactive multi-dimension modulation with dynamic controllable residual learning for image restoration. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. pp. 53–68. Springer (2020)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
18. Henrique Brito, J., Angst, R., Koser, K., Pollefeys, M.: Radial distortion self-calibration. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1368–1375 (2013)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
20. Kukulova, Z., Pajdla, T.: A minimal solution to radial distortion autocalibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**(12), 2410–2422 (2011)
21. Kumar, V.R., Klingner, M., Yogamani, S., Milz, S., Fingscheidt, T., Mader, P.: Syndistnet: Self-supervised monocular fisheye camera distance estimation synergized with semantic segmentation for autonomous driving. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. pp. 61–71 (2021)
22. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4681–4690 (2017)
23. Li, X., Zhang, B., Sander, P.V., Liao, J.: Blind geometric distortion correction on images through deep learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4855–4864 (2019)
24. Liao, K., Lin, C., Zhao, Y., Gabbouj, M.: Dr-gan: Automatic radial distortion rectification using conditional gan in real-time. *IEEE Transactions on Circuits and Systems for Video Technology* **30**(3), 725–733 (2019)
25. Liao, K., Lin, C., Zhao, Y., Xu, M.: Model-free distortion rectification framework bridged by distortion distribution map. *IEEE Transactions on Image Processing* **29**, 3707–3718 (2020)
26. Lin, L., Lu, Y., Pan, Y., Chen, X.: Integrating graph partitioning and matching for trajectory analysis in video surveillance. *IEEE transactions on Image Processing* **21**(12), 4844–4857 (2012)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. pp. 740–755. Springer (2014)
28. Liu, C., Wang, X., Li, S., Wang, Y., Qian, X.: Fsi: Frequency and spatial interactive learning for image restoration in under-display cameras. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 12537–12546 (2023)
29. Liu, C., Wang, X., Xu, X., Tian, R., Li, S., Qian, X., Yang, M.H.: Motion-adaptive separable collaborative filters for blind motion deblurring. In: *Proceedings of the*

- IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25595–25605 (2024)
30. Liu, C., Yang, H., Fu, J., Qian, X.: Learning trajectory-aware transformer for video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5687–5696 (2022)
 31. Liu, C., Yang, H., Fu, J., Qian, X.: 4d lut: learnable context-aware 4d lookup table for image enhancement. *IEEE Transactions on Image Processing* **32**, 4742–4756 (2023)
 32. Mei, C., Rives, P.: Single view point omnidirectional camera calibration from planar grids. In: Proceedings 2007 IEEE International Conference on Robotics and Automation. pp. 3945–3950. IEEE (2007)
 33. Mou, C., Wu, Y., Wang, X., Dong, C., Zhang, J., Shan, Y.: Metric learning based interactive modulation for real-world super-resolution. In: European Conference on Computer Vision. pp. 723–740. Springer (2022)
 34. Muhammad, K., Ahmad, J., Lv, Z., Bellavista, P., Yang, P., Baik, S.W.: Efficient deep cnn-based fire detection and localization in video surveillance applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **49**(7), 1419–1434 (2018)
 35. Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q.: Conformer: Local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 367–376 (2021)
 36. Plaut, E., Ben Yaacov, E., El Shlomo, B.: 3d object detection from a single fisheye image without a single fisheye training image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3659–3667 (2021)
 37. Puig, L., Bastanlar, Y., Sturm, P., Guerrero, J.J., Barreto, J.: Calibration of central catadioptric cameras using a dlt-like approach. *International Journal of Computer Vision* **93**, 101–114 (2011)
 38. Rashed, H., Mohamed, E., Sistu, G., Kumar, V.R., Eising, C., El-Sallab, A., Yogamani, S.: Generalized object detection on fisheye cameras for autonomous driving: Dataset, representations and baseline. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2272–2280 (2021)
 39. Rong, J., Huang, S., Shang, Z., Ying, X.: Radial lens distortion correction using convolutional neural networks trained with synthesized images. In: Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part III 13. pp. 35–49. Springer (2017)
 40. Santana-Cedr s, D., Gomez, L., Alem n-Flores, M., Salgado, A., Esclar n, J., Mazorra, L., Alvarez, L.: An iterative optimization algorithm for lens distortion correction using two-parameter models. *Image Processing On Line* **6**, 326–364 (2016)
 41. Scaramuzza, D., Martinelli, A., Siegwart, R.: A flexible technique for accurate omnidirectional camera calibration and structure from motion. In: Fourth IEEE International Conference on Computer Vision Systems (ICVS’06). pp. 45–45. IEEE (2006)
 42. Shen, Z., Lin, C., Liao, K., Nie, L., Zheng, Z., Zhao, Y.: Panoformer: Panorama transformer for indoor 360 depth estimation. In: European Conference on Computer Vision. pp. 195–211. Springer (2022)
 43. Sturm, P., Ramalingam, S.: A generic concept for camera calibration. In: Computer Vision–ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004. Proceedings, Part II 8. pp. 1–13. Springer (2004)
 44. Thorm hlen, T., Broszio, H., Wassermann, I.: Robust line-based calibration of lens distortion from a single view. *Mirage 2003* pp. 105–112 (2003)

45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
46. Wang, A., Qiu, T., Shao, L.: A simple method of radial distortion correction with centre of distortion estimation. *Journal of Mathematical Imaging and Vision* **35**, 165–172 (2009)
47. Wang, W., Guo, R., Tian, Y., Yang, W.: Cfsnet: Toward a controllable feature space for image restoration. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4140–4149 (2019)
48. Wang, X., Yu, K., Dong, C., Tang, X., Loy, C.C.: Deep network interpolation for continuous imagery effect transition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1692–1701 (2019)
49. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: *Proceedings of the European conference on computer vision (ECCV) workshops*. pp. 0–0 (2018)
50. Xue, Z., Xue, N., Xia, G.S., Shen, W.: Learning to calibrate straight lines for fisheye image rectification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1643–1651 (2019)
51. Yang, C.Y., Chen, H.H.: Efficient face detection in the fisheye image domain. *IEEE Transactions on Image Processing* **30**, 5641–5651 (2021)
52. Yang, S., Lin, C., Liao, K., Zhang, C., Zhao, Y.: Progressively complementary network for fisheye image rectification using appearance flow. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6348–6357 (2021)
53. Yang, S., Lin, C., Liao, K., Zhao, Y.: Fishformer: Annulus slicing-based transformer for fisheye rectification with efficacy domain exploration. *arXiv preprint arXiv:2207.01925* (2022)
54. Yang, S., Lin, C., Liao, K., Zhao, Y.: Dual diffusion architecture for fisheye image rectification: Synthetic-to-real generalization. *arXiv preprint arXiv:2301.11785* (2023)
55. Yao, M., He, D., Li, X., Li, F., Xiong, Z.: Towards interactive self-supervised denoising. *IEEE Transactions on Circuits and Systems for Video Technology* (2023)
56. Yin, X., Wang, X., Yu, J., Zhang, M., Fua, P., Tao, D.: Fisheyerecnet: A multi-context collaborative deep network for fisheye image rectification. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 469–484 (2018)
57. Yogamani, S., Hughes, C., Horgan, J., Sistu, G., Varley, P., O’Dea, D., Uricár, M., Milz, S., Simon, M., Amende, K., et al.: Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9308–9318 (2019)
58. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing* **26**(7), 3142–3155 (2017)
59. Zhang, M., Yao, J., Xia, M., Li, K., Zhang, Y., Liu, Y.: Line-based multi-label energy optimization for fisheye image rectification and calibration. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 4137–4145 (2015)
60. Zhang, Y., You, S., Gevers, T.: Automatic calibration of the fisheye camera for egocentric 3d human pose estimation from a single image. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 1772–1781 (2021)

61. Zhang, Z., Jiang, Y., Shao, W., Wang, X., Luo, P., Lin, K., Gu, J.: Real-time controllable denoising for image and video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14028–14038 (2023)
62. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* **40**(6), 1452–1464 (2017)