

# Minimalist and High-Quality Panoramic Imaging with PSF-aware Transformers

Qi Jiang\*, Shaohua Gao\*, Yao Gao, Kailun Yang<sup>†</sup>, Zhonghua Yi, Hao Shi, Lei Sun, and Kaiwei Wang<sup>†</sup>

**Abstract**—High-quality panoramic images with a Field of View (FoV) of  $360^\circ$  are essential for contemporary panoramic computer vision tasks. However, conventional imaging systems come with sophisticated lens designs and heavy optical components. This disqualifies their usage in many mobile and wearable applications where thin and portable, minimalist imaging systems are desired. In this paper, we propose a Panoramic Computational Imaging Engine (PCIE) to address minimalist and high-quality panoramic imaging. With less than three spherical lenses, a Minimalist Panoramic Imaging Prototype (MPIP) is constructed based on the design of the Panoramic Annular Lens (PAL), but with low-quality imaging results due to aberrations and small image plane size. We propose two pipelines, *i.e.* Aberration Correction (AC) and Super-Resolution and Aberration Correction (SR&AC), to solve the image quality problems of MPIP, with imaging sensors of small and large pixel size, respectively. To provide a universal network for the two pipelines, we leverage the information from the Point Spread Function (PSF) of the optical system and design a PSF-aware Aberration-image Recovery Transformer (PART), in which the self-attention calculation and feature extraction are guided via PSF-aware mechanisms. We train PART on synthetic image pairs from simulation and put forward the PALHQ dataset to fill the gap of real-world high-quality PAL images for low-level vision. A comprehensive variety of experiments on synthetic and real-world benchmarks demonstrates the impressive imaging results of PCIE and the effectiveness of plug-and-play PSF-aware mechanisms. We further deliver heuristic experimental findings for minimalist and high-quality panoramic imaging, in terms of the choices of prototype and pipeline, network architecture, training strategies, and dataset construction. Our dataset and code will be available at <https://github.com/zju-jiangqi/PCIE-PART>.

**Index Terms**—Panoramic imaging, minimalist optical systems, computational imaging, vision transformer, point spread function.

## I. INTRODUCTION

**I**MAGE processing of panoramic images with an ultra-wide Field of View (FoV) of  $360^\circ$  is growing popular for achieving a holistic understanding of the entire surrounding scene [1], [2], [3], [4]. To produce high-quality panoramas for photography and down-stream tasks, a variety of low-level

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 12174341 and in part by Hangzhou SurImage Technology Company Ltd.

Q. Jiang, S. Gao, Y. Gao, Z. Yi, H. Shi, L. Sun, and K. Wang are with the State Key Laboratory of Modern Optical Instrumentation and the National Engineering Research Center of Optical Instrumentation, Zhejiang University, Hangzhou 310027, China.

K. Yang is with the School of Robotics and the National Engineering Research Center of Robot Visual Perception and Control Technology, Hunan University, Changsha 410082, China.

\*Equal contribution.

<sup>†</sup>Corresponding authors: Kaiwei Wang and Kailun Yang. (E-mail: wangkaiwei@zju.edu.cn, kailun.yang@hnu.edu.cn.)

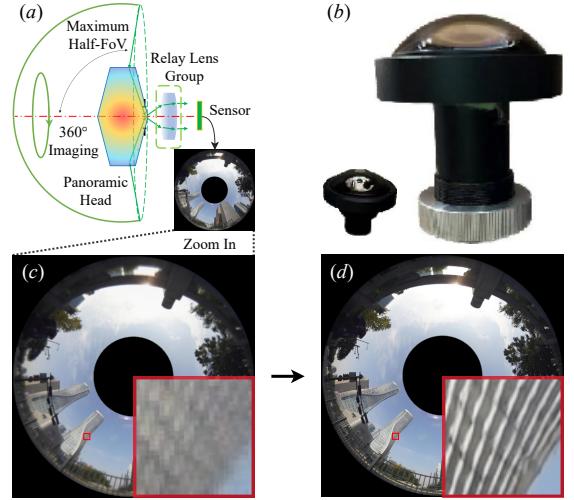


Fig. 1. Illustration of proposed MPIP and its key issue of low image quality, which is properly addressed with PSF-aware transformer: PART. (a) Minimalist Panoramic Imaging Prototype (MPIP); (b) Comparison between real products of conventional panoramic imaging systems and PAL-based MPIP; (c) Low-quality image captured by MPIP. (d) High-quality image recovered by PART. In this way, we realize minimalist and high-quality panoramic imaging with PSF-aware transformers.

vision works is conducted to deal with the image quality of  $360^\circ$  panoramas, *e.g.*, low angular resolutions [5], [6], [7] and severe image distortions [8], [9], [10].

Most contemporary works on panoramic images are based on the common sense that the optical system is aberration-free where the imaging result is clear and sharp without any blur. While widely applied, conventional panoramic optical systems, come with notoriously sophisticated lens designs, composed of multiple sets of lenses with complex surface types [11], [12], [13], to reach high imaging quality. However, this is not often the case as the demand for thin, portable imaging systems, *i.e.*, Minimalist Optical Systems (MOS), grows stronger in mobile and wearable applications [14]. Without sufficient lens groups for aberration correction, the aberration-induced image blur is inevitable for MOS. In this case, the imaging quality drops significantly and often catastrophically, and the unsatisfactory imaging performance disqualifies its potential usage in upper-level applications. This leads to an appealing issue and we ask if we may strike a fine balance between high-quality panoramic imaging and minimalist panoramic optical systems.

With the rapid development of digital image processing, Computational Imaging (CI) methods for MOS [15], [16], [17] appear as a preferred solution to this issue. These methods often propose optical designs with few necessary optical components to meet the basic demands of specific

applications, *e.g.*, the FoV, depth of field, and focal length, then apply an image recovery pipeline to remove the aberration-induced image blur. Recent research works [18], [19], [20] further design end-to-end deep learning frameworks for joint optimization of optical systems and image recovery networks. In this paper, based on the idea of computational imaging, we propose *Panoramic Computational Imaging Engine (PCIE)*, a framework for minimalist and high-quality panoramic imaging, to solve the trade-off between high-quality panoramic imaging and minimalist panoramic optical systems as a whole, without sitting on only one of its sides.

Motivated by modern panoramic lens designs [12], [21], [22], PCIE builds on a Minimalist Panoramic Imaging Prototype (MPIP) shown in Fig. 1(a), which is composed of an essential panoramic head for 360° panoramic imaging and a relay lens group for aberration correction. In specific, we select the structure of Panoramic Annular Lens (PAL) [23], [24]: a more compact solution to 360° panoramas, as an example for MPIP, where a catadioptric PAL head is equipped to replace the complex lens group [11] in the conventional fisheye lens [25], [26]. To achieve a minimalist design, the proposed MPIP is composed of 1 spherical lens for the PAL head, and 1 or 2 simple spherical lenses for the relay lens group, which can image over 360° FoV with only 40% of the numbers of lenses and 60% of the volumes of conventional panoramic imaging systems, as shown in Fig. 1(b). However, as illustrated in Fig. 1(c), it brings a key issue of corrupted image quality caused by the uncorrected optical aberrations and the limited image plane size.

To address the aberration-induced spatially-variant image blur and low imaging resolution of MPIP, engaged with the information of Point Spread Function (PSF) from optical design, we propose *PSF-aware Aberration-image Recovery Transformer (PART)*: a transformer-based low-quality image recovery paradigm for MPIP. Different from previous transformer baselines, *e.g.*, SwinIR [27], PART exploits the PSF, the forward function characterizing the aberration-induced blur, to attain better results. Inspired by the spatially-variant aberration distributions of MPIP, we first design the PSF-aware Feature Modulator (PFM), where pixel-adaptive convolution kernels are learned from the compressed PSF map of the applied optical system, to modulate the feature map during representation learning. PFM is a plug-and-play PSF-aware mechanism that can be inserted into other recovery models.

In addition, PSF-aware Mix-Attention Block (PMAB) is proposed as the basic unit of PART, which comprises: (1) Vanilla window attention of SwinIR [27] for capturing long-range dependency; (2) PSF-aware Varied-Size Attention (P-VSA), where diverse windows of varied sizes and locations are learned from PSFs to provide dynamic receptive fields; (3) PFM of small kernel size for enhancing the feature extraction of local details. With PART, the low-quality image captured by MPIP can be smoothly recovered (see Fig. 1(d)) for minimalist and high-quality panoramic imaging.

To facilitate the training of PART, wave-based imaging simulation with random perturbation [28] is utilized for generating clear-blur image pairs. To fill the gap of ground-truth images of PAL, we record a high-quality PAL images

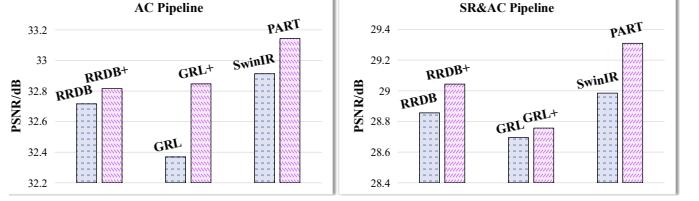


Fig. 2. The proposed plug-and-play PSF-aware mechanism, PFM, consistently and significantly improves the performance of several baseline models in two pipelines. “+” means the model inserted the PFM in the same way as PART.

dataset named *PALHQ* through a well-designed PAL in varied scenes. Based on PALHQ, we set up two tasks to formalize the key issue of low-quality MPIP images: (1) *Aberration Correction (AC)* of high-resolution images taken by sensors with small pixel size, and (2) *Super-Resolution and Aberration Correction (SR&AC)* of low-resolution images from sensors with large pixel size. Then, representative models of image super-resolution (SR) [27], [29], [30], [31], [32], [33], [34], image deblurring (Deblur) [35], [36], [37], [38], and image restoration with PSF-aware mechanisms (PSF-aware) [28] are evaluated, where PCIE enables all models to produce impressive panoramic imaging results.

Furthermore, we manufacture an MPIP sample with better image quality and capture real-world low-quality image datasets *RealMPAL* to evaluate models on real-world data. Experimental results reveal that PFM enhances the performance of SR models (see Fig. 2) and PART sets the state of the art on both synthetic and real-world benchmarks. We also conduct extensive experiments to investigate the potential of GAN-based training strategies and the effectiveness of PALHQ in PCIE. The generative model appears to be competitive for generating more realistic details if the artifacts can be well suppressed. Additionally, PALHQ serves as the cornerstone of PCIE for training a robust model for annular images. At a glance, we deliver the following contributions:

- We propose the Panoramic Computational Imaging Engine (PCIE), a novel framework for minimalist and high-quality panoramic imaging, as shown in Fig. 3.
- A Minimalist Panoramic Imaging Prototype (MPIP) is designed for 360° panoramic imaging with an essential panoramic head and simple relay lens group.
- We introduce the PSF-aware Aberration-image Recovery Transformer (PART) to process the low-quality images of MPIP, where the newly designed PSF-aware Feature Modulator (PFM) and PSF-aware Mix-Attention Block (PMAB) are embedded in the architecture to exploit PSF information for enhancing the recovery performance.
- We raise two pipelines of low-quality MPIP image processing: Aberration Correction (AC) and Super-Resolution and Aberration Correction (SR&AC).
- We record the real-world panoramic image datasets PALHQ and RealMPAL of high-quality and low-quality respectively for benchmarking our two pipelines, which are the first real-world PAL datasets for low-level vision.
- A comprehensive variety of experiments is designed to evaluate the effectiveness of PCIE, which provides heuristic findings in terms of optical design, network architecture, training strategies, and dataset construction.

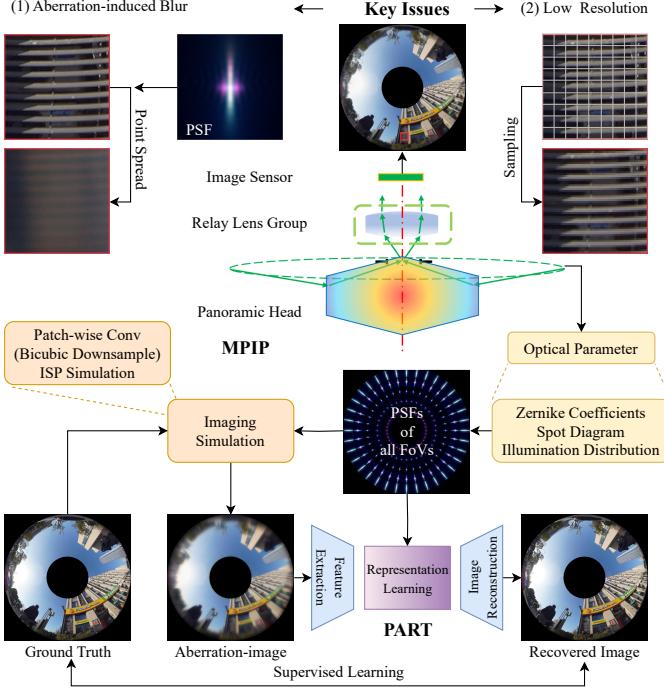


Fig. 3. Overview of the proposed Panoramic Computational Imaging Engine (PCIE) for minimalist and high-quality panoramic imaging. The designed MPIP brings two key issues of low image quality: (1) aberration-induced blur due to lack of enough lenses for aberration correction and (2) low resolution caused by limited image plane size. We introduce the PART, which is trained on synthetic data pairs generated by imaging simulation, to recover the low-quality aberration image with the guidance of PSFs information of the MPIP.

## II. RELATED WORK

### A. Image Processing of Panoramic Images

Recent research interest in panoramic images is booming for immersive visual experiences [11], [39]. Semantic segmentation [1], [40], depth estimation [2], [3], and visual SLAM [41], [42] are widely explored on panoramic images for a holistic understanding of the surrounding scene. To this intent, high-quality panoramic images are urgently required for robust performance. A considerable amount of work is conducted to improve the image quality of panoramic images, such as super-resolution [5], [6], [7] and distortion correction [8], [9], [10].

However, the above image processing of panoramic images is based on the aberration-free images captured by the conventional panoramic lens, where multiple sets of lenses with complex surface types [11], [12], [13] are applied for high-quality imaging. This work focuses on capturing panoramic images with a Minimalist Optical System (MOS) composed of much fewer lenses for volume-limited applications, where we process the aberration images via computational methods.

### B. Computational Imaging for Minimalist Optical System

The aberration-induced image blur is inevitable for MOS due to insufficient lens groups for aberration correction. Recently, computational imaging methods for MOS appear as a preferred solution to this issue, where optical designs with few necessary optical components are equipped with image recovery pipelines for both minimalist and aberration-free

imaging [16], [17], [43]. Some research works [18], [44], [45] even design end-to-end deep learning frameworks for joint optimization of MOS and post-processing networks to achieve the best match between MOS and recovery models to further improve the final imaging performance.

However, computational imaging for minimalist panoramic systems is scarcely explored. In a preliminary study, Jiang *et al.* [28] propose an Annular Computational Imaging (ACI) framework to break the optical limit of minimalist Panoramic Annular Lens (PAL), where the image processing is conducted on unfolded PAL images. To further develop a general framework for minimalist and high-quality panoramic imaging, this work fills the gap in high-quality PAL image datasets and designs the Panoramic Computational Imaging Engine (PCIE) directly on annular PAL images, in terms of both optical design and aberration-images recovery.

### C. Image Recovery of Aberration Images

The aberration-induced blur is always spatially-variant, *i.e.* Linear Shift Variant (LSV), due to the uneven thicknesses of the lenses. Several efforts have been made for the LSV system, spanning path-wise restoration [46], [47], experimental PSFs calibration and non-blind deconvolution [43], [48], and low-rank decomposition [49], [50], based on the degradation model of aberration-images [51], [52], [53].

Recent works tend to adopt the data-driven learning-based image restoration networks [54], [38]. These methods typically use a U-shaped network with an encoder-decoder structure [16], [55], [56] to achieve more efficient and robust recovery results, which can also be easily inserted into an end-to-end framework for joint optimization. To break the bottleneck of data-driven methods under scarce data, the PSF information is explored to design physical-informed networks, where model-based methods are characterized by Convolutional Neural Networks (CNNs) for learning ill-posed terms [14], [57].

Differently, we make a pioneering effort and investigate the potential of transformer-based SR models in aberration correction rather than conventional Deblur models. Then, the PSFs are transformed into an additional modality of the aberration image, based on which we design PSF-aware mechanisms for achieving better results. The proposed PSF-aware Aberration-image Recovery Transformer (PART) is a successful attempt to engage PSF information in the representation learning stage of SR models for recovering aberration images.

The overview of PCIE is shown in Fig. 3. It provides a powerful framework for minimalist and high-quality panoramic imaging, where optical design (detailed in Sec. III) and learning-based model (presented in Sec. IV) are intertwined for achieving impressive imaging results.

## III. MINIMALIST PANORAMIC IMAGING PROTOTYPE

In this section, we set up a universal prototype for minimalist panoramic imaging systems based on modern panoramic lens designs (Sec. III-A). To address the issues induced by the reduced lens numbers and limited image plane size, two settings of tasks and benchmarks are defined in Sec. III-B and Sec. III-C, respectively. In Sec. III-D, we describe the

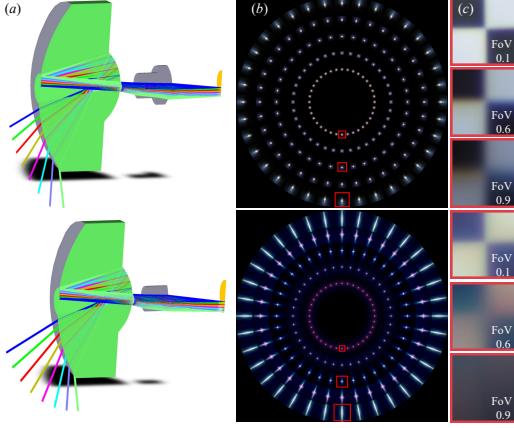


Fig. 4. Two prototype samples of MPIP. Up: MPIP-P1, Down: MPIP-P2. (a) Optical path diagram. (b) Visualized PSFs distributions. (c) The degraded checkerboard image patches of normalized FoVs 0.1, 0.6, and 0.9 are captured by two MPIP samples. The minimalist optical design brings spatially-variant aberration-induced blur, especially for MPIP-P2 equipped with fewer lenses.

constructed imaging simulation model to generate synthetic image pairs for training learning-based methods.

### A. Optical Design

To boost scene understanding with larger FoV, panoramic optical systems are emerging, including fisheye optical systems, refractive panoramic systems, panoramic annular optical systems, *etc.* [11]. In most modern designs of panoramic lenses [12], [21], [22], a panoramic head is applied for collecting incident rays of 360° FoV, while a set of relay lenses is designed to bend the rays and correct the aberrations. Based on the structure, we propose the Minimalist Panoramic Imaging Prototype (MPIP), including an essential panoramic head and a simple relay lens group, as shown in Fig. 1(a).

Specifically, we adopt a more compact and efficient solution, *i.e.* Panoramic Annular Lens (PAL) [23], [24], in MPIP samples, where a catadioptric PAL head is equipped for 360° annular imaging. For minimalist design and convenient manufacture, spherical lenses are applied in the relay lens group and the number is reduced to fewer than 3. To illustrate the imaging results of different relay lens groups, we define MPIP-P1 and MPIP-P2 in Fig. 4(a), whose relay lens group is composed of two lenses and a single lens, respectively.

The lack of enough lenses for aberration correction makes the imaging point spread from an ideal point, inducing spatially-variant PSFs with large kernel sizes, as shown in Fig. 4(b). The average geometric spot radius of MPIP-P1 is 13.78 $\mu\text{m}$ , whereas that of MPIP-P2 is 46.26 $\mu\text{m}$ . As a result, the captured images of MPIP suffer from spatially-variant aberration-induced blur, especially for MPIP-P2 with fewer lenses, as shown in Fig. 4(c).

### B. Definition of Tasks

In addition to the uncorrected optical aberrations, the limited image plane size due to the small aperture of MPIP presents the issue of image resolution. To fit the small image plane of the MPIP, an image sensor with a smaller pixel size can be applied for maintaining high resolution, but it makes the

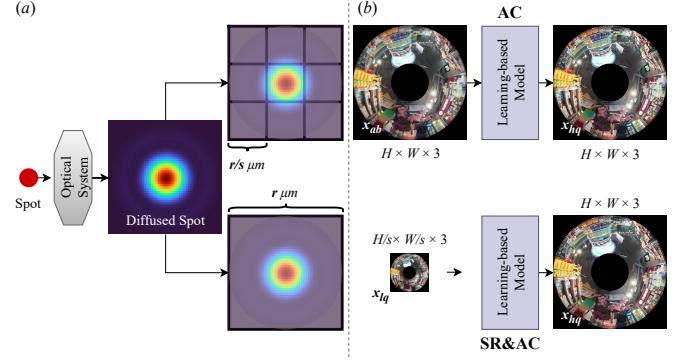


Fig. 5. Illustration of two pipelines for processing aberration-images of MPIP. (a) Comparison of image sensors with different pixel sizes. For a diffused spot through the optical system with a fixed size, more pixels of the sensor with smaller pixel sizes and higher resolutions are affected. (b) Raised two tasks based on sensors with different pixel sizes: Aberration Correction (AC) and Super-Resolution and Aberration Correction (SR&AC). In summary, we target the recovery of a high-quality image from an aberration image of MPIP.

system more sensitive to aberration-induced blur. As shown in Fig 5(a), the diffused optical spot of fixed physical size affects more pixels for the sensor with smaller pixel sizes and higher resolution. The opposite solution with large pixel sizes is less sensitive to the diffused spot, but the reduced image resolution also brings degradation to the images, which is especially harmful to panoramic images with large FoV [6].

To address this dilemma, we propose two pipelines for solving the contradictory problems, as shown in Fig. 5(b), where a learning-based model is applied to process different image recovery tasks. For image sensors with smaller pixel sizes, we define the Aberration Correction (AC) task, where the goal is to recover a clear image  $x_{hq} \in \mathbb{R}^{H \times W \times 3}$  from a high-resolution input aberration-image  $x_{ab} \in \mathbb{R}^{H \times W \times 3}$ . Whereas for image sensor with larger pixel size, the Super-Resolution and Aberration Correction (SR&AC) task is raised to recover a high-resolution aberration-free image  $x_{hq} \in \mathbb{R}^{H \times W \times 3}$  from a low-resolution input aberration-image  $x_{lq} \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times 3}$ , where  $s$  is the scale factor of SR.

### C. PALHQ: Established Dataset of High-Quality PAL Images

The lack of high-quality image datasets for PAL comes as a bottleneck to the above tasks. A piece of previous work for CI of PAL, ACI [28], unfolds the annular PAL images into perspective ones to utilize the publicly available datasets, *i.e.* DIV2K [58]. However, the asymmetrical interpolation during unfolding induces extra image degradation, which further complicates the image degradation factors for MPIP. In addition, the annular image is more appealing for the simulation of aberrations in the original image plane and necessary for some vision tasks like PAL-based SLAM [41], [42]. In the case of benchmarks for processing panoramic images, *e.g.*, the ODI-SR dataset [59] and the SUN360 panorama dataset [60], which are taken via fisheye cameras, the imaging process is also quite different from that of PAL [11]. These concerns raise an urgent request for high-quality panoramic annular image datasets.

To this intent, we propose PALHQ, a dataset of high-quality PAL images, to facilitate network training and evaluation of

PAL-based low-level vision tasks. A well-designed PAL of 11 lenses and a Sony α6600 camera are applied to capture high-resolution PAL images with negligible primary aberrations. PALHQ contains 550 clear PAL images with a resolution of  $3152 \times 3152$ , covering rich and varied scenes of indoor, natural, urban, campus, and scenic spots. We divide PALHQ into 500 images for the training set and 50 images for the validation set (refer to the appendix for sample images of PALHQ). In PCIE, we benchmark both AC and SR&AC on PALHQ, where the corresponding aberration images are generated by the imaging simulation model depicted below. Furthermore, PALHQ can be also transmitted to unfolded panoramas via equirectangular projection (ERP), which can support various panoramic image processing applications.

#### D. Imaging Simulation Model

To quantitatively benchmark the raised two tasks and enable supervised training of learning-based models, paired aberration images and clear images are required. Following previous super-resolution works [61], [62] and CI works [28], [56], we construct an imaging simulation model to generate synthetic aberration-images in batches.

The wave-based simulation pipeline with random perturbation in [28] is adopted to generate multiple aberration distributions directly on clear annular PAL images. Specifically, the clear raw image  $R$  is modulated by an optical system and then processed by ISP  $\Gamma(\cdot)$  to produce the final imaging result  $A$ :

$$A_\theta(x, y) = \Gamma\left[\left(\int r_\lambda R_\theta(x, y) \otimes K_\theta(x, y, \lambda) d\lambda\right) \downarrow + N\right], \quad (1)$$

where  $r_\lambda$  is the wave response of the sensor. The noise  $N$  and the sampling process  $(\cdot) \downarrow$  of the image sensor are also included in the model. We divide the image into patches for patch-wise convolution with PSFs  $K_\theta(x, y, \lambda)$  under different FoV  $\theta$ . Different from [28], the division of FoV is centrosymmetric for annular images as is shown in Fig. 4(b). Through scalar diffraction integral [63],  $K_\theta(x, y, \lambda)$  is calculated based on the wavefront  $\Phi_\theta(x', y', \lambda)$  on exit pupil plane, which is described by Zernike polynomials [64] mathematically:

$$\Phi_\theta(x', y', \lambda) = \sum_{n,m} C_n^m(\theta, \lambda) Z_n^m(x', y'), \quad (2)$$

where  $C(\theta, \lambda)$  denotes Zernike coefficients under FoV  $\theta$  and wavelength  $\lambda$  and  $Z$  refers to polynomials of the coordinate  $(x', y')$  on exit pupil. The combination of different  $m$  and  $n$  represents different orders. Finally, we apply the random disturbance strategy in [28] to fine-tune the ideal  $C(\theta, \lambda)$  from the Zemax software, generating synthetic aberration images with diverse aberration distributions.

## IV. LOW-QUALITY MPIP IMAGES RECOVERY

In this section, we describe the proposed learning-based model to recover low-quality MPIP images, as shown in Fig. 6. As a kind of representation of the image degradation process, the PSF information of applied MPIP is compressed into a PSF map, detailed in Sec. IV-A, serving as one additional modality for our model. With the PSF map, we design the PSF-aware

Feature Modulator (PFM) and the PSF-aware Mix-Attention Block (PMAB), elaborated in Sec. IV-B and Sec. IV-C, respectively. Then, the PSF-aware Aberration-image Recovery Transformer (PART) is established and introduced in Sec. IV-D as a transformer-based paradigm for the raised two tasks.

#### A. PSF Information of Optical System

For non-blind optimization-based recovery methods in aberration correction, *e.g.* Wiener filter [51], PSFs  $K$  of the system are exploited to predict the clear image  $x_{hg}$  from the aberration-image  $x_{ab}$  by deconvolution. However, this method often fails when the PSFs deviate from the design stage during manufacture and require time-consuming strategies [43], [48] for processing complex spatially-variant blur. Data-driven learning-based models [36], [27], which can be plugged directly into existing end-to-end frameworks of lens design [18], [19], [20], have demonstrated more powerful abilities in image recovery, but may hit a bottleneck when the training data is scarce.

This motivates us to break the bottleneck by utilizing PSFs of the optical system in a learning-based model. The PSFs of  $n$  sampled FoVs of the applied MIIP can be calculated based on the wavefront as depicted in Eq. (2):

$$K_i(x, y) = \mathcal{S}(\Phi_i(x', y')), i = 1, 2, 3, \dots, n. \quad (3)$$

The  $\mathcal{S}(\cdot)$  denotes scalar diffraction integral (refer to [63] for more details). Previous methods tend to use the kernel size of  $K_i$  to guide the network [28] or refine the ill-posed term in deconvolution through a learning-based model [57]. Although these attempts can improve the recovery benefiting from the applications of  $K_i$ , the spatial intensity distribution of PSFs is not fully exploited to guide the deep feature extraction of the general image recovery paradigm.

To this intent, we first map the spatial PSFs  $K_i \in \mathbb{R}^{k_i \times k_i \times 3}$  into the image feature shape  $H \times W \times C'$ , where  $k_i$  is the kernel size of PSF under the  $i_{th}$  FoV and  $C'$  denotes the channels of mapped PSFs, serving as an additional modality aligned with the aberration-image. As previously shown in [65], the spatial-to-channel arrangement helps transform spatially-variant kernels into a feature map. Similarly, to produce a PSF feature map, the spatial PSFs  $K_i$  are arranged into the channel dimension. Concretely, for a pixel at  $(x, y)$  of the image, we first calculate the vector  $\vec{p}$  from the image center  $(0, 0)$  to  $(x, y)$ , and define the vertical unit vector  $\vec{a}(1, 0)$ . The PSF of the corresponding FoV is located by  $|\vec{p}| / \max(|\vec{p}|)$ , and rotated by the angle  $\arccos\left(\frac{\vec{p} \cdot \vec{a}}{|\vec{p}| \cdot |\vec{a}|}\right)$ , producing the PSF  $K_{x,y}$  of the pixel. For memory-friendly computation, we pad all  $K_{x,y}$  into unified size of  $\max_i k_i$  and compress them into  $k' \times k' \times 1$  via adaptive average pooling:

$$\hat{K}_{x,y} = \text{AveragePool}(\text{padding}(K_{x,y})), \quad (4)$$

where the choice of compressed size  $k'$  is ablated in Sec. V-F. The  $\hat{K}_{x,y}$  is then reshaped into  $1 \times 1 \times (k'^2)$  and inserted into each pixel to produce the PSF feature map  $x_{int} \in \mathbb{R}^{H \times W \times (k'^2)}$ . In addition, considering the lost PSF size information during compressing, we also generate the size distribution map  $x_s \in \mathbb{R}^{H \times W \times 3}$  of RGB channels, where the

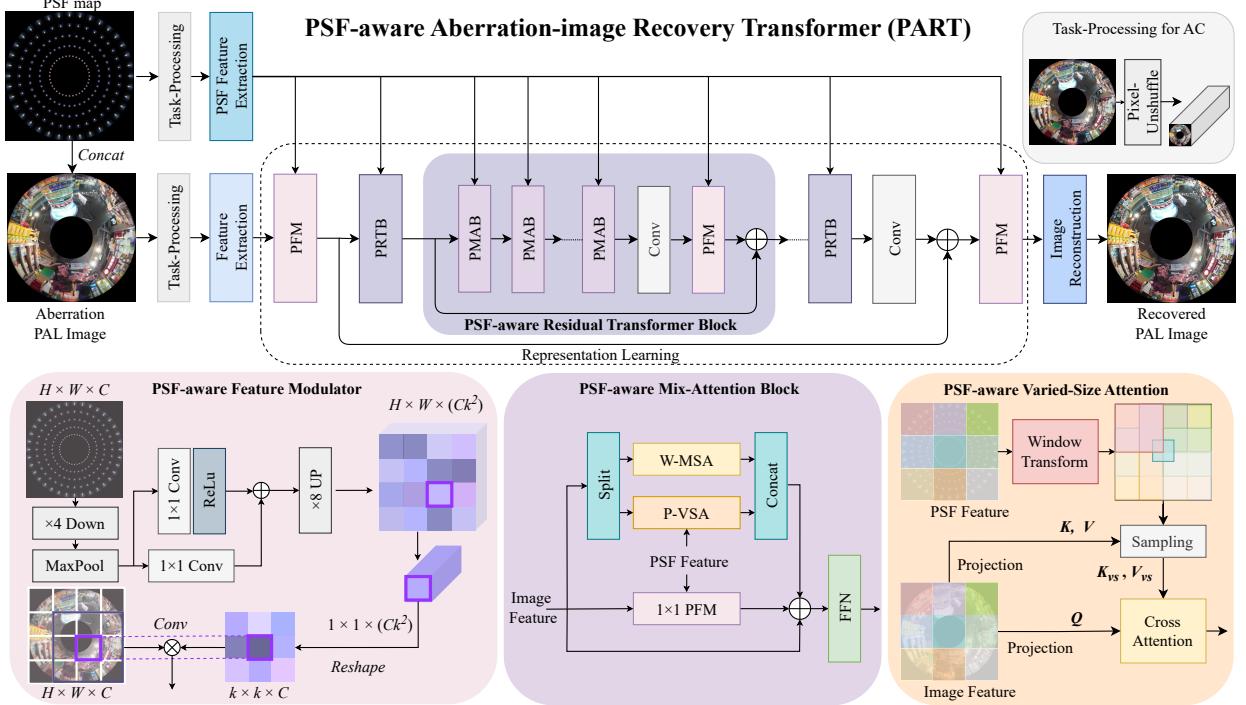


Fig. 6. PART: Proposed PSF-aware Aberration-image Recovery Transformer. PART is established on a classical super-resolution paradigm [27], [30], incorporating stages of feature extraction, representation learning, and image reconstruction, for dealing with both AC and SR&AC. Task-Processing leverages the pixel-unshuffle operation [62] for AC to reduce the spatial size of high-resolution images, whereas no operation is entailed for SR&AC. PSF-aware Residual Transformer Block (PRTB) is the basic block of representation learning, where we design PSF-aware Feature Modulator (PFM) and PSF-aware Mix-Attention Block (PMAB) to learn spatially-variant degradation features with the guidance of PSF features. The mixing of Window-based Multi-head Self-Attention (W-MSA), PSF-aware Varied-Size Attention (P-VSA), and PFM enable PMAB to capture both global and local dependencies adaptively. For an intuitive understanding of the PSF map, we visualize it via a form of PSFs distributions in Fig. 4.

value of each pixel represents the kernel size. Finally, the PSF map  $x_{psf}$  is produced via Eq. (5):

$$x_{psf} = \text{Concat}(x_{int}, x_s), \quad (5)$$

and the visualized PSF map is shown in Appendix. PSF map is an aligned modality of the aberration image characterizing the image degradation over FoVs, based on which we design a PSF-aware transformer, as described in the next subsection.

### B. PSF-aware Feature Modulator

CNN layers have shown impressive abilities of local feature extraction, but are restricted to the fixed spatially-invariant kernels. However, the mathematical imaging model in Eq. (1) reveals that the aberration-induced blur is just generated by convolution with spatially-variant PSFs kernels, whose inverse solution cannot be modeled by the traditional CNN layers.

To extract adaptive image features with the guidance of spatially-variant PSFs for dealing with aberration-induced blur, we propose the PSF-aware Feature Modulator (PFM), as shown in the lower left of Fig. 6. PFM builds on the idea of filter adaptive convolution [65], [66], where a kernel map of  $H \times W \times (Ck^2)$  is predicted from feature map of  $H \times W \times C$ . Unlike these pipelines, in PFM, the kernel map  $x_{kernel}$  is predicated on the features of PSF map  $x_{psf}$ , which has been compressed into a similar form as  $x_{kernel}$ . We first apply  $E_{psf}$  as a  $3 \times 3$  convolution layer to extract features of PSF map  $x_{psf}$ , as depicted in Eq. (6):

$$x'_{psf} = E_{psf}(x_{psf}), \quad (6)$$

where  $x'_{psf} \in \mathbb{R}^{H \times W \times C}$  is the extracted PSF feature map. Then, a lightweight kernel predictor composed of several convolution layers is proposed to output the kernel map  $x_{kernel} \in \mathbb{R}^{H \times W \times (Ck^2)}$  based on  $x'_{psf}$ , as in Eq. (7):

$$x_{kernel} = P(x'_{psf}). \quad (7)$$

To reduce the memory cost and inference latency of kernel prediction, the predictor  $P$  computes the kernel map on the downsampled features (by  $4 \times 4$  average pooling). Benefiting from that the PSF map shares a similar form with the kernel map, we further simplify the  $P$  where only one Max Pooling layer and one residual block of  $1 \times 1$  convolution layers are applied, to predict the kernel map  $x'_{kernel} \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times (Ck^2)}$  in a smaller resolution. The final kernel map  $x_{kernel}$  is then obtained by  $\times 8$  upsampling via bilinear interpolation. Finally, we reshape the  $x_{kernel}$  into a list of per-pixel kernels of  $k \times k \times C$  and apply them to the corresponding pixels of image feature  $x'_{img}$ . PFM can be interpreted as the inverse process of the aberration-induced blur, *i.e.* deconvolution, which promotes the recovery of the aberration image.

### C. PSF-aware Mix-Attention Block

We put forward the PSF-aware Mix-Attention Block (PMAB) as the basic unit of our PSF-aware transformer, to process aberration images assisted with the PSF map, as shown at the middle bottom of Fig. 6. The Window-based Multi-head Self-Attention (W-MSA) of the Swin-T block [27] is first adopted to be the baseline attention mechanism for mod-

eling spatially-variant convolution and long-range dependency, which is also important for stable training of the network.

To address the drawback of fixed window size in vanilla W-MSA, we further propose the PSF-aware Varied-Size Attention (P-VSA), shown on the lower right of Fig. 6. The vanilla varied-size attention [67] in high-level tasks predicts the sizes and locations of the windows from input features for computing self-attention on dynamic windows. To better adaptively modulate the windows according to the degradation information, we make use of the PSF map features  $x'_{psf}$  to generate PSF-aware varied-size windows. Concretely, the scale  $S$  and offset  $O$  of the varied-size windows are predicated on  $x'_{psf}$  by the Window Transform block, which is composed of a  $1 \times 1$  convolution layer. Then, we sample the projected key and value tokens  $K, V$  of image features  $x'_{img}$  based on the transformed window to obtain  $K_{vs}, V_{vs}$ . The cross-attention is computed between query  $Q$  of the default window and  $K_{vs}, V_{vs}$ . The operation of P-VSA can be expressed as:

$$Q, K, V = \text{Linear}(\text{WinPar}(x'_{img})), \quad (8)$$

$$S, O = \text{WinTrans}(x'_{psf}), \quad (9)$$

$$K_{vs}, V_{vs} = \text{Sample}(K, S, O), \text{Sample}(V, S, O) \quad (10)$$

$$\text{Attn}(Q, K_{vs}, V_{vs}) = \text{Softmax}\left(\frac{QK_{vs}^\top}{\sqrt{d}}\right)V_{vs}, \quad (11)$$

where WinPar denotes the window partition operation of Swin-T and  $d$  is the dimension of tokens.

Additionally, some works [30], [32] apply channel-attention-based convolution blocks in parallel with the self-attention to enhance the representation ability of the network. We insert the proposed PFM to PMAB in the same parallel way, where the filter adaptive convolution mechanism can better model the spatially-variant blur compared to channel-attention-based convolution.

Finally, PMAB is the mixing of W-MSA and P-VSA with a parallel  $1 \times 1$  PFM. For the self-attention module, the image feature map  $x'_{img}$  is equally split along the channel dimension and processed by parallel W-MSA and P-VSA, then concatenated along the channel dimension again. The modulated feature map by parallel PFM is multiplied by a constant  $\alpha$ , to be added to the result of self-attention and the original feature map as common practice for stable training [32]. The whole process of PMAB is computed as:

$$x'^{(1)}_{img}, x'^{(2)}_{img} = \text{Split}(x'_{img}), \quad (12)$$

$$x_{attn} = \text{Concat}(\text{M-WSA}(x'^{(1)}_{img}), \text{P-VSA}(x'^{(2)}_{img}, x'_{psf})), \quad (13)$$

$$x_{mix} = x_{attn} + \alpha \text{PFM}(x'_{img}, x'_{psf}) + x'_{img}, \quad (14)$$

$$y = x_{mix} + \text{FFN}(x_{mix}), \quad (15)$$

where FFN is a common Feed Forward Network composed of a LayerNorm and a Multi-Layer Perceptron (MLP) layer.

#### D. PSF-aware Aberration-image Recovery Transformer

Most previous networks [16], [55] for aberration corrections often utilize the architecture of image deblurring methods, *i.e.* U-Net. However, for MPIP images with a high resolution (*e.g.*, 3K), the U-Net methods incur unacceptable computational costs due to the large image sizes at shallow layers. Differently, we look into the tasks from the perspective of image super-resolution, which processes image features with low resolution and reconstructs the high-quality image via an upsampling module. The aberration-induced blur brings aliasing between pixels and losses of image details, which can also be interpreted as “low resolution”. Thereby, as shown in Fig. 6, the PSF-aware Aberration-image Recovery Transformer (PART) is set up based on the structure of SwinIR [27] and our proposed PSF-aware mechanisms.

A Task-Processing module is first applied to transform the input image and PSF map to a small spatial size, where pixel-unshuffle [62] is leveraged for AC and no operation is entailed for SR&AC. The PSF map is also concatenated with the aberration image as the input of the network. More precisely, PART contains three parts. (1) A feature extraction layer converts the input to image feature maps via a  $3 \times 3$  convolution. (2) The representation learning stage applies stacks of transformer-based blocks ending with a convolution layer to enrich the learned degradation information of aberration-induced blur progressively. We design the PSF-aware Residual Transformer Block (PRTB) with several PMAB layers and a convolution layer. The PFM is inserted into each PRTB to modulate the learned features and model the inverse process of the aberration-induced blur. We also implement PFM at the beginning and end of the representation learning stage, for adaptive feature extraction and feature fusion based on PSF information. (3) The image reconstruction module further fuses the extracted deep features and recovers a high-quality image with higher resolution. With PART, we can recover a high-quality aberration-free image  $x_{hq}$  from either a high-resolution aberration image  $x_{ab}$  or a low-resolution one  $x_{lq}$ , providing a general solution to AC and SR&AC:

$$x_{hq}^{AC} = \text{PART}(x_{ab}), x_{hq}^{SR\&AC} = \text{PART}(x_{lq}). \quad (16)$$

## V. EXPERIMENTS AND RESULTS

We conduct a comprehensive set of experiments to evaluate the proposed PCIE for minimalist and high-quality panoramic imaging. We first describe the implementation details of our work in Sec. V-A. The PCIE under different recovery models is then evaluated on both synthetic (Sec. V-B) and real (Sec. V-C) datasets. We further investigate the GAN-based training strategies for PCIE in Sec. V-D. At last, in Sec. V-E and Sec. V-F, ablation studies on training datasets and the architecture of PART are conducted.

#### A. Implementation Details

**Synthetic Datasets.** We apply the collected PALHQ dataset for training and evaluation. Based on PALHQ, the aberration images of two prototypes, *i.e.* PALHQ-SynMPIP-P1 and PALHQ-SynMPIP-P2, are generated by the simulation model

TABLE I

QUANTITATIVE EVALUATION OF PCIE WITH THE AC PIPELINE ON SYNTHETIC BENCHMARKS WITH MPIP-P1 AND MPIP-P2. WE HIGHLIGHT THE **BEST** AND **SECOND** RESULTS.

Method	PALHQ-SynMPIP-P1				PALHQ-SynMPIP-P2			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
SR	RRDB [29]	32.716	0.9265	0.0469	03.704	26.996	0.8503	0.0896
	RCAN [34]	32.496	0.9257	0.0459	04.327	26.456	0.8443	0.0956
	EDSR [33]	32.868	0.9282	0.0449	03.770	26.951	0.8500	0.0889
	SwinIR [27]	32.913	0.9291	0.0446	03.670	26.935	0.8509	0.0884
	EDT [31]	32.929	0.9288	0.0450	03.658	27.055	0.8518	0.0888
	HAT [32]	32.925	0.9288	0.0447	03.748	26.827	0.8500	0.0889
	GRL [30]	32.369	0.9256	0.0457	04.234	26.268	0.8424	0.0943
Deblur	HINet [37]	32.238	0.9234	0.0476	04.159	26.401	0.8428	0.0933
	NAFNet [38]	28.494	0.8853	0.0630	09.570	24.443	0.8181	0.1130
	Restormer [35]	27.424	0.8826	0.0595	10.795	24.374	0.8235	0.1012
	Uformer [36]	<b>32.999</b>	0.9290	<b>0.0442</b>	03.672	<b>27.133</b>	0.8525	0.0866
PSF-aware	PI <sup>2</sup> RNet [28]	32.682	0.9268	0.0448	03.638	26.656	0.8471	0.0874
	<b>RRDB+</b>	32.816	0.9271	0.0456	03.746	27.050	0.8505	0.0895
	<b>GRL+</b>	32.847	<b>0.9292</b>	0.0454	<b>03.627</b>	27.020	<b>0.8528</b>	<b>0.0864</b>
	<b>PART (Ours)</b>	<b>33.143</b>	<b>0.9304</b>	<b>0.0435</b>	<b>03.571</b>	<b>27.198</b>	<b>0.8540</b>	<b>0.0855</b>
								<b>16.281</b>
								<b>16.436</b>

of Eq. (1). Following [28], we set the random range of disturbance as 25% and generate 10 virtual MPIP samples for the training set (500 images) and 4 for the validation set (50 images) to simulate the synthetic-to-real gap. For image sensors, the MV-SUA1600C camera with a pixel size of  $1.34\mu\text{m}$  and the MV-SUA133GC camera with a pixel size of  $4\mu\text{m}$  are applied for the AC and SR&AC pipelines, respectively, where the ISP and wave response of them are simulated in the data generation. In addition, we use  $\times 3$  bicubic downsampling to produce low-resolution aberration images for SR&AC, considering the sensors' pixel sizes.

**Real-world Datasets.** As shown in Fig. 4, with only one more simple lens, the MPIP-P1 reveals much better image quality, which relieves the burden of the post-image processing pipelines. We manufacture MPIP-P1 and use it to record the RealMPIP3K-AC (58 images with a resolution of  $2912 \times 2912$ ) and RealMPIP1K-SR&AC (64 images with a resolution of  $992 \times 992$ ) with two cameras respectively, to provide real-world MPIP aberration-images for evaluating two pipelines of PCIE. We test models trained on PALHQ-SynMPIP-P1 (AC) and PALHQ-SynMPIP-P1 (SR&AC) with RealMPIP3K-AC and RealMPIP1K-SR&AC respectively.

**Evaluation Metrics.** For synthetic datasets with ground truth, PSNR and SSIM [68] are employed to evaluate the fidelity of the recovery results, whereas LPIPS [69] and FID [70] are employed to evaluate the perceptual quality.

For real datasets without reference clear image, we employ no-reference metrics, *i.e.* NIQE and BRISQUE, to evaluate the image quality of MPIP images in terms of natural images. The qualitative visual results are also provided for an intuitive evaluation. However, the NIQE [71] and BRISQUE [72] are built on the statistics of perspective natural images, which are challenging for assessing the MPIP images with the annular distribution of image content. Considering the specific tasks of correcting optical aberrations, we define the Optical-based Image Quality Evaluator (OIQE) for credible evaluation, based on the Modulation Transfer Function (MTF) of the imaging system calculated by a set of testing checkerboard images.

TABLE II  
QUANTITATIVE EVALUATION OF PCIE WITH SR&AC PIPELINE ON SYNTHETIC BENCHMARK.

Method	PALHQ-SynMPIP-P1			
	PSNR↑	SSIM↑	LPIPS↓	FID↓
RRDB [29]	28.856	0.8758	0.0733	09.957
RCAN [34]	28.238	0.8686	0.0787	12.689
EDSR [33]	28.817	0.8759	0.0715	10.670
SwinIR [27]	28.985	<b>0.8781</b>	<b>0.0714</b>	09.938
EDT [31]	29.008	0.8777	0.0726	10.750
HAT [32]	28.921	0.8771	0.0727	10.141
GRL [30]	28.695	0.8753	<b>0.0714</b>	09.829
<b>RRDB+</b>	<b>29.044</b>	0.8774	0.0724	10.068
<b>GRL+</b>	28.757	0.8768	0.0716	<b>09.709</b>
<b>PART (Ours)</b>	<b>29.310</b>	<b>0.8819</b>	<b>0.0681</b>	<b>09.648</b>

To be specific, we follow Spatial Frequency Response (SFR) [56] testing to calculate MTFs on image patches of “knife-edge” of different FoVs from different testing images.  $MTF_{50}$  and  $MTF_{area}$  are used to characterize the MTF curves, where the former is the frequency when the MTF drops 50% and the latter is the area under the MTF curve. We further define  $OIQE_{50}$  and  $OIQE_{area}$  as the ratio of the average  $MTF_{50}$  and  $MTF_{area}$  of the testing imaging pipeline to those of a well-designed panoramic imaging system. Accordingly, OIQE is defined as:

$$OIQE = \frac{OIQE_{50} + OIQE_{area}}{2}, \quad (17)$$

which measures the gap between the results of PICE and conventional panoramic lenses in terms of MTF. OIQE is only applied in the AC pipeline due to its specific design for evaluating the ability of the model to remove aberration-induced blur.

**Compared Methods.** For the AC pipeline, as shown in Table I, we compare PART with representative state-of-the-art SR models (RRDB [29], RCAN [34], EDSR [33], SwinIR [27], EDT [31], HAT [32], and GRL [30]), along with Deblur methods (HINet [37], NAFNet [38], Restormer [35], and UFormer [36]). Image restoration models with PSF-aware mechanisms, *i.e.* RRDB+, GRL+, and PI<sup>2</sup>RNet [28], are also included in the comparison. Here, “+” means that the methods are inserted with the designed PFM. For the SR&AC pipeline, SR methods RRDB+ and GRL+ are selected due to the specific task requirement for super-resolution, as shown in Table II.

All the models are retrained on PALHQ-SynMPIP-P1 and PALHQ-SynMPIP-P2 with their original optimizers, learning rates, and schedulers, where the number of training iterations and the batch size are set the same as PART for a fair comparison. Additionally, we apply task-processing for all the SR models the same as PART.

**Training Details.** The compressed kernel size  $k'$  of the PSF map is set to 5 in our experiments, where an ablation study is conducted in Sec. V-F. In addition, we set the kernel

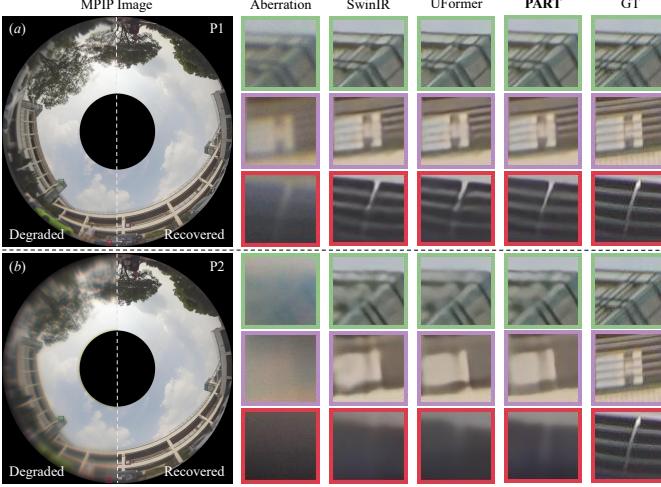


Fig. 7. Qualitative results of representative models on synthetic benchmarks for AC. We zoom in on image patches of different FoVs to show the details. (a) Results on PALHQ-SynMPIP-P1, where aberration-free images are produced by all models and PART provides more visually pleasant and clearer details. (b) Results on PALHQ-SynMPIP-P2. PCIE delivers clear aberration-free images, yet, with heavily corrupted detailed textures due to severe optical aberrations. Enabling much higher imaging quality with only one more spherical lens, MPIP-P1 is a superior choice for PCIE.

size  $k$  of PFM to 3 considering the computational efficiency. Following SwinIR [27], the PRTB number, PMAB number, channel number, attention head number, and window size are generally set to 6, 6, 180, 6, and 8, respectively.

PART is trained on L1Loss, while other loss functions are explored in Sec. V-D. We train the models with the Adam optimizer with an initial learning rate of  $2e-4$  and a batch size of 8 on a single A800 GPU. For data augmentation, random crop, flip, and rotation are applied, where the ground-truth crop size is  $256 \times 256$  for AC and  $196 \times 196$  for SR&AC to keep an image size of 64 in the representation learning stage. The number of training iterations is set to  $200k$  and the learning rate is halved at  $100k$ ,  $160k$ ,  $180k$ , and  $190k$ .

## B. Experiments on Synthetic Datasets

**AC Pipeline.** Table I shows numerical results of PCIE under different image recovery models on synthetic benchmarks of AC. We also present visual results of representative methods in Fig. 7. PCIE with most models achieves PSNR over  $30dB$  on PALHQ-SynMPIP-P1 and over  $25dB$  on PALHQ-SynMPIP-P2, producing impressive panoramic imaging results via a minimalist optical system. Compared to Deblur methods, SR methods overall deliver better results, illustrating the effectiveness of the SR framework in aberration correction. PSF-aware methods further outperform their baselines. Precisely, PI<sup>2</sup>RNet exceeds HINet, PART surpasses SwinIR, and RRDB+ and GRL+ outstrip their corresponding baselines by clear margins. We find that the models based on the window-attention mechanism (SwinIR, EDT, HAT, UFormer, GRL, and our proposed PART) realize more competitive results than CNN-based models, where the window-based self-attention can better model spatially-variant blur. Yet, the state-of-the-art SR model GRL performs poorly on the benchmarks, which is attributed to the stripe-based attention being difficult to adapt to MPIP images with annular distributions.

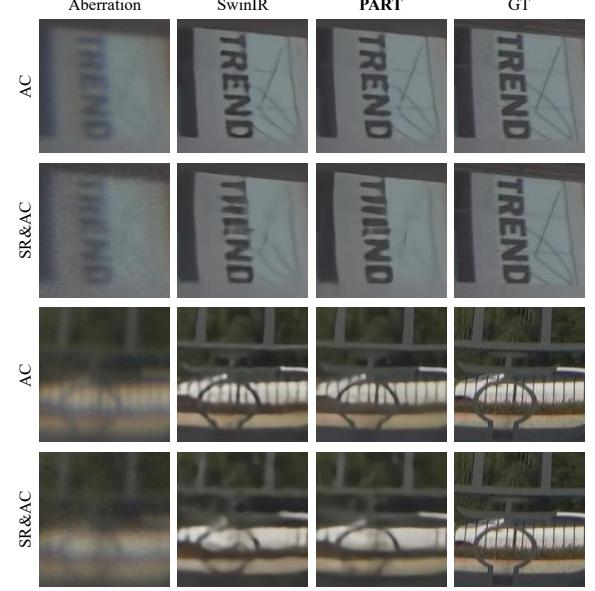


Fig. 8. Comparison between AC and SR&AC. The image patches are cropped from PALHQ-0532 (top) and PALHQ-0505 (bottom). We show the results of the proposed PART and its baseline SwinIR to illustrate the strengths of the AC pipeline, which produces richer and more realistic image details.

Overall, PART brings better results for PCIE, yielding state-of-the-art performance on two benchmarks, in terms of both fidelity-based metrics (PSNR and SSIM) and perceptual-based metrics (LPIPS and FID). As all the methods produce aberration-free visual results with some lost textures and artifacts, the recovered image of PART shows more visually pleasant details, as shown in Fig. 7(a), in all FoVs.

Further, applied with only one more spherical lens, the PCIE results of MPIP-P1 outperform those of MPIP-P2 by a large margin. For example, the PSNR drops by  $3.050dB \sim 6.101dB$  when MPIP-P2 is equipped. As shown in Fig. 7(b), PCIE with MPIP-P2 delivers moderate clear aberration-free images. Yet, suffering from severe aberrations, its detailed textures are heavily corrupted, especially for large FoVs. In this sense, MPIP-P1 is a superior choice for PCIE to achieve minimalist and high-quality panoramic imaging.

**SR&AC Pipeline.** The quantitative evaluation of PCIE with the SR&AC pipeline is shown in Table II. Consistent with the observations in AC, the methods with window-based attention and PSF-aware mechanisms lead to better performance. PART sets the state of the art in the SR&AC task, achieving improvements compared against the second best, e.g.  $0.266dB$  in PSNR,  $0.0038$  in SSIM, LPIPS from  $0.0714$  to  $0.0681$  (about 5%), and FID from  $9.709$  to  $9.648$  (about 6%).

Comparing SR&AC (Table II) with AC (Table I), we observe that the loss of spatial resolution in aberration-images causes significant deterioration to the imaging quality of PCIE, e.g., to an amount of  $-4.258dB \sim -3.674dB$  in PSNR. The visual quality comparison between the two pipelines is provided in Fig. 8, where the imaging results of AC reveal richer and more realistic details. In this case, AC is a more competitive pipeline for reconstructing high-resolution aberration-free images, where the real sampled pixels of the sensor offer more convincing imaging features than super-resolved ones despite more aberration-induced blur.

TABLE III

QUANTITATIVE EVALUATION OF PCIE ON REAL-WORLD BENCHMARKS REALMPIP. THE OIQE OF ORIGINAL ABERRATION IMAGES IS 55.22%. WE ALSO LIST THE RANKS ON EACH METRIC IN “()” AND THE AVERAGE RANK (A.R.) OF EACH METHOD FOR AN INTUITIVE EVALUATION.

Method	RealMPIP3K-AC			RealMPIP1K-SR&AC		A.R.
	OIQE↑	NIQE↓	BRISQUE↓	NIQE↓	BRISQUE↓	
RRDB [29]	66.94% (4)	04.930 (8)	45.692 (2)	04.848 (4)	50.729 (5)	4.6
SwinIR [27]	67.51% (3)	04.816 (6)	46.380 (5)	04.833 (2)	50.555 (4)	4.0
GRL [30]	64.63% (7)	04.665 (1)	46.587 (6)	04.824 (1)	51.057 (6)	4.2
Uformer [36]	68.55% (2)	04.914 (7)	45.427 (1)	n.a.	n.a.	3.3
PI <sup>2</sup> RNet [28]	64.86% (6)	04.710 (3)	47.148 (8)	n.a.	n.a.	5.7
<b>RRDB+</b>	58.28% (8)	04.783 (5)	46.675 (7)	04.857 (5)	50.383 (2)	5.4
<b>GRL+</b>	65.31% (5)	04.724 (4)	46.138 (4)	04.842 (3)	50.061 (1)	3.4
<b>PART (Ours)</b>	77.87% (1)	04.707 (2)	45.968 (3)	04.933 (6)	50.422 (3)	3.0

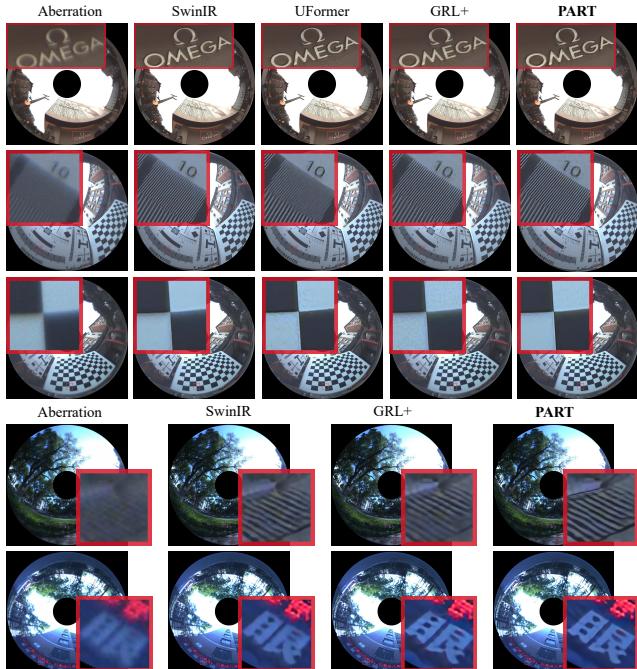


Fig. 9. Visual results of PCIE on real MPIP with state-of-the-art models and our proposed PART. Top three rows: results of the AC pipeline. Bottom two rows: results of the SR&AC pipeline. We choose the top-four performing methods to show the results, while the UFormer is not applicable for SR&AC.

### C. Experiments on Real-World Datasets

As shown in Table III, PCIE with representative models makes significant contributions to the removal of the aberration-induced blur of real-world MPIP images. To be specific, OIQE improves from 55.22% to 58.28%~77.87%. The results on NIQE and BRISQUE reveal a large variance, which is attributed to that these metrics are designed for perspective natural images rather than annular MPIP images. For a comprehensive and intuitive evaluation, we rank each method on each metric and provide the average rank (A.R.). In the real-world case, PART outperforms other models, achieving the best OIQE (77.87%), and the best A.R. (3.0). The visual results of PCIE on real-world scenes are provided in Fig. 9. PCIE enables most methods to deliver high-quality panoramic images with few aberrations and high resolution, where PART sets the state of the art in terms of higher contrast, sharper edges, and fewer artifacts. Additionally, consistent with experiments on synthetic data, the recovered images of

TABLE IV

QUANTITATIVE EVALUATION OF GAN-BASED TRAINING ON OUR BENCHMARKS OF AC AND SR&AC.

Task	Method	Training Strategy	PALHQ-SynMPIP-P1			
			PSNR↑	SSIM↑	LPIPS↓	FID↓
AC	RRDB [29]	PSNR-oriented	32.716	0.9265	0.0469	03.704
		+GAN [29]	28.929	0.8840	0.0392	04.919
		+LDL [73]	31.864	0.9166	0.0338	04.559
	SwinIR [27]	PSNR-oriented	<b>32.913</b>	<b>0.9291</b>	0.0446	03.670
		+GAN [29]	29.916	0.8920	0.0449	04.254
		+LDL [73]	31.770	0.9130	<b>0.0297</b>	<b>03.444</b>
SR&AC	PART	PSNR-oriented	<b>33.143</b>	<b>0.9304</b>	0.0435	03.571
		+GAN [29]	30.965	0.9045	0.0410	04.402
		+LDL [73]	31.854	0.9148	<b>0.0264</b>	<b>03.541</b>
	RRDB [29]	PSNR-oriented	28.856	0.8758	0.0733	13.989
		+GAN [29]	25.842	0.8276	0.0638	15.112
		+LDL [73]	28.112	0.8633	<b>0.0561</b>	13.351
	SwinIR [27]	PSNR-oriented	<b>28.985</b>	<b>0.8781</b>	0.0714	13.736
		+GAN [29]	26.596	0.8385	0.0634	14.476
		+LDL [73]	27.875	0.8575	0.0686	13.208
	PART	PSNR-oriented	<b>29.310</b>	<b>0.8819</b>	0.0681	13.826
		+GAN [29]	28.382	0.8688	0.0682	<b>12.706</b>
		+LDL [73]	28.608	0.8720	<b>0.0508</b>	<b>12.805</b>

the SR&AC pipeline reveal perceptually unpleasant artifacts.

### D. Investigation on GAN-based Training Strategies

To generate richer details for recovered images, we investigate GAN-based training strategies on classical models RRDB, SwinIR, and our PART. Following [73], the GAN-based loss functions in ESRGAN [29] and Local Discriminative Learning (LDL) [73] are adopted, where the former is a classical GAN-based framework and the latter is an improved strategy to remove artifacts. We take models trained with L1Loss, *i.e.* PSNR-oriented models, as pre-training generators, then apply GAN and LDL loss functions to enable these networks to generate more textures respectively. As shown in Table IV, on synthetic data, both GAN and LDL lead to a decrease in recovery accuracy (PSNR and SSIM), while bringing great gains under the perceptual quality metrics. LDL is a more competitive strategy that outperforms GAN with higher fidelity and fewer visual artifacts, especially with PART.

Regarding real-world data, we present the OIQE and qualitative results in Fig. 10. GAN-based training further contributes to the removal of the aberration-induced blur, achieving better OIQE with higher image contrast. Aside from this, GAN-based models deliver more realistic imaging results with richer textures, which also bring some perceptually unpleasant artifacts and fake details despite being trained with LDL. We point out that the GAN-based strategies offer the potential for learning a more realistic high-quality MPIP image. Still, the local statistics in LDL of perspective images may need to be adapted to annular images for better suppression of artifacts.

### E. Effectiveness of PALHQ

The collected PALHQ demonstrates an impressive ability to train the model for recovering both synthetic and real-world MPIP images in previous experiments. In this section, we explore whether PALHQ is necessary for PCIE. As an

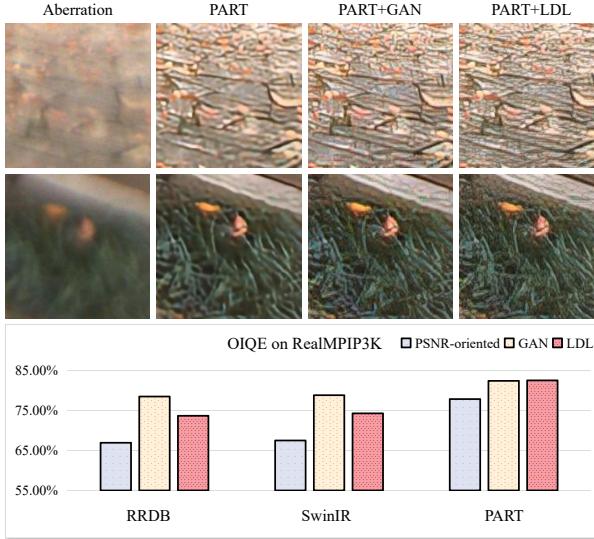


Fig. 10. Evaluation of GAN-based training strategies on real-world data. We take the AC pipeline as an example, where image patches from RealMPIP3K-0031 and RealMPIP3K-0057 are presented.

TABLE V

QUANTITATIVE COMPARISON BETWEEN THE EFFECTIVENESS OF PALHQ AND AVAILABLE HQ DATASET IN PCIE.

Task	Method	Training Dataset	PALHQ-SynMPIP-P1			
			PSNR↑	SSIM↑	LPIPS↓	FID↓
AC	Uformer [36]	PanoFlickr2K	32.340	0.8253	0.0484	04.558
		<b>PALHQ</b>	32.999	0.9290	0.0442	03.672
	RRDB [29]	PanoFlickr2K	32.128	0.9235	0.0508	04.717
		<b>PALHQ</b>	32.716	0.9265	0.0469	03.704
	SwinIR [27]	PanoFlickr2K	32.292	0.9255	0.0485	04.521
		<b>PALHQ</b>	32.913	0.9291	0.0446	03.670
SR&AC	PART	PanoFlickr2K	32.498	0.9259	0.0480	04.348
		<b>PALHQ</b>	33.143	0.9304	0.0435	03.571
	RRDB [29]	PanoFlickr2K	27.943	0.8688	0.0906	14.123
		<b>PALHQ</b>	28.856	0.8758	0.0733	13.989
	SwinIR [27]	PanoFlickr2K	28.129	0.8705	0.0880	13.980
		<b>PALHQ</b>	28.985	0.8781	0.0714	13.736
	PART	PanoFlickr2K	28.558	0.8748	0.0854	14.309
	PART	<b>PALHQ</b>	29.310	0.8819	0.0681	13.826

alternative to PALHQ, we simulate the aberrations of MPIP-P1 directly on the publicly available perspective image dataset, *i.e.* Flickr2K [74], creating PanoFlickr2K for training.

We compare representative models trained on PanoFlickr2K and PALHQ on both synthetic and real-world benchmarks in Table V and Fig. 11. It becomes clear that PALHQ contributes significantly to high-quality panoramic imaging, where the numerical results in all metrics are improved by a large margin and the visual results are more perceptually pleasant with sharper edges, fewer artifacts, and fewer noises.

#### F. Ablation Study

We conduct ablation studies to investigate how PSF-aware mechanisms contribute to high-quality MPIP image reconstruction. In all cases, the experiments are implemented with the AC pipeline on PALHQ-SynMPIP-P1, evaluated by PSNR and SSIM, and set up on the baseline model SwinIR.

**Physical Information.** As reported in Table VI, the different types of physical information are concatenated with the input

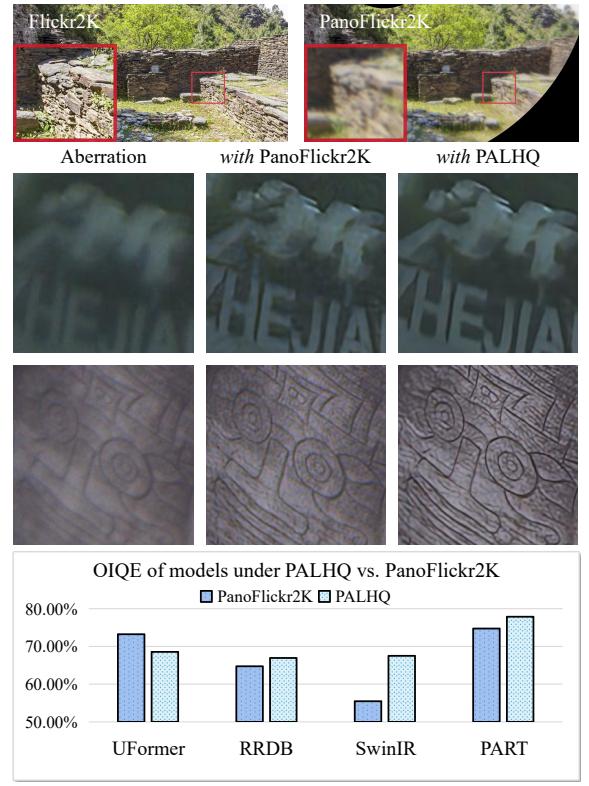


Fig. 11. PALHQ vs. PanoFlickr2K. The top row: illustration of PanoFlickr2K. We simulate the aberrations and FoV distributions of MPIP on perspective images of Flickr2K. Bottom two rows: image patches from RealMPIP3K-0020 and RealMPIP3K-0027. We take PART as an example to show the results trained on different datasets.

image respectively for an intuitive evaluation. The PSF map contains rich information characterizing aberration-induced blur, providing better results compared to the FoV map. Then, we set the optimal  $k'$  to 5. A larger  $k'$  tends to improve the model's scores, but the performance becomes saturated when  $k'$  is too large with redundant and sparse information.

**PSF-aware Mechanisms.** Table VII shows that all the designed PSF-aware mechanisms contribute to reaching better scores, *i.e.* 0.230dB and 0.0013 improvements in PSNR and SSIM. PFM attains the highest gains of 0.201dB in PSNR and 0.0012 in SSIM. In addition, the performances of RRDB+ and GRL+ in Table I and Table II verify the consistent effectiveness of the plug-and-play PFM in other models, bringing improvements of 0.100dB~0.478dB in PSNR and 0.0006~0.0036 in SSIM on SynMPIP-P1. Regarding the attention block, PMAB with  $1 \times 1$  PFM and P-VSA enable adaptive self-attention guided by PSF information, outperforming the vanilla window-based self-attention.

**Position of PFM.** We further investigate the optimal position to insert PFM. As shown in Table VIII, we apply PFM after the feature extraction, at the last of each PRTB, and before the image reconstruction, for ablations. The PFM on shallow features reveals more competitive results, while increasing the number of PFM during the representation learning stage also leads to significant improvements. Using PFM in all ablated positions helps to reach the best performance, which is also corroborated by the observation in omnidirectional image super-resolution [6].

TABLE VI  
ABLATIONS ON PHYSICAL INFORMATION.

Physical Information	$k'$	PSNR	SSIM
w/o FoV map	-	32.913	0.9291
	-	32.992	0.9293
PSF map	1	33.012	0.9300
	5	<b>33.021</b>	<b>0.9301</b>
	9	<b>33.021</b>	0.9299

### G. Summary

The extensive experiments illustrate the critical points in the proposed PCIE for achieving minimalist and high-quality panoramic imaging. We summarize the following primary findings of our experiments:

- The proposed PCIE presents impressive high-quality imaging results, where the MPIP-P1 and AC pipeline are superior choices for delivering aberration-free panoramic images with much more realistic details.
- In PCIE, we find that SR and transformer-based models reveal better results. Furthermore, PSF-aware mechanisms are effective for improving the performance of SR models, where the proposed PSF-aware transformer, *i.e.* PART, sets state of the art.
- Regarding the training strategies, GAN-based methods contribute to more realistic recovered images, but with some visually unpleasant artifacts and fake details. The generative model appears to be more competitive in PCIE if a good balance is struck when generating rich details and suppressing artifacts.
- Comparing with the adaptation of perspective images, the collected high-quality panoramic annular images dataset, *i.e.* PALHQ, brings considerable improvements. PALHQ serves as the cornerstone of our PCIE for training a robust model to process MPIP images.

We hope that the PCIE can bring inspiration from optical design, network architecture, sensor choice, data preparation, and training strategies, for minimalist and high-quality panoramic imaging in mobile and wearable applications.

## VI. CONCLUSION

In this paper, we design PCIE to present a general solution to minimalist and high-quality panoramic imaging. Based on the idea of PAL, the MPIP is proposed for 360° panoramic imaging with less than three lenses. Then, learning-based models, which are trained on synthetic aberration images from simulation, are applied to solve the aberration-induced blur and low resolution of MPIP images. A new dataset PALHQ is collected to fill the gap of high-quality PAL images for low-level vision. We explore utilizing PSF information of the optical system to improve the performance of models and design a PSF-aware transformer PART with PSF-aware mechanisms. The plug-and-play mechanism PFM can enhance modern SR models for removing aberration-induced blur, while PART with PMAB delivers state-of-the-art performance on both synthetic and real-world benchmarks. Extensive experiments are conducted to investigate how to improve PCIE, providing

TABLE VII  
ABLATIONS ON PSF-AWARE MECHANISMS.

PSF-aware Mechanism	Params	PSNR	SSIM
w/o	-	11.97M	32.913
concat	-	12.02M	33.021
PFM	-	16.72M	33.114
PMAB	1×1 PFM	14.32M	33.069
	P-VSA	12.18M	32.999
	both	14.53M	33.082
	all	19.27M	<b>33.143</b>
			<b>0.9304</b>

TABLE VIII  
ABLATIONS ON THE POSITION OF PFM.

Position	Params	PSNR	SSIM
w/o	11.97M	32.913	0.9291
first conv	12.61M	33.032	0.9297
PRTB	15.54M	33.071	0.9299
last conv	12.61M	32.971	0.9298
all	16.72M	<b>33.114</b>	<b>0.9303</b>

heuristic findings for constructing a computational-imaging-based minimalist panoramic system with impressive imaging quality, in terms of optical design, network architecture, sensor selection, training strategies, and data preparation.

## REFERENCES

- [1] K. Yang, X. Hu, and R. Stiefelhagen, “Is context-aware CNN ready for the surroundings? Panoramic semantic segmentation in the wild,” *TIP*, vol. 30, pp. 1866–1881, 2021.
- [2] K. Tateno, N. Navab, and F. Tombari, “Distortion-aware convolutional filters for dense prediction in panoramic images,” in *ECCV*, vol. 11220, 2018, pp. 732–750.
- [3] Z. Shen, C. Lin, K. Liao, L. Nie, Z. Zheng, and Y. Zhao, “PanoFormer: Panorama transformer for indoor 360° depth estimation,” in *ECCV*, vol. 13661, 2022, pp. 195–211.
- [4] K. Liao, X. Xu, C. Lin, W. Ren, Y. Wei, and Y. Zhao, “Cylinder-Painting: Seamless 360° panoramic image outpainting and beyond with cylinder-style convolutions,” *arXiv preprint arXiv:2204.08563*, 2022.
- [5] Y. Yoon, I. Chung, L. Wang, and K.-J. Yoon, “SphereSR: 360° image super-resolution with arbitrary projection via continuous spherical image representation,” in *CVPR*, 2022, pp. 5667–5676.
- [6] F. Yu, X. Wang, M. Cao, G. Li, Y. Shan, and C. Dong, “OSRT: Omnidirectional image super-resolution with distortion-aware transformer,” in *CVPR*, 2023, pp. 13 283–13 292.
- [7] X. Sun *et al.*, “OPDN: Omnidirectional position-aware deformable network for omnidirectional image super-resolution,” in *CVPRW*, 2023, pp. 1293–1301.
- [8] S. Yang, C. Lin, K. Liao, and Y. Zhao, “FishFormer: Annulus slicing-based transformer for fisheye rectification with efficacy domain exploration,” *arXiv preprint arXiv:2207.01925*, 2022.
- [9] S. Yang, C. Lin, K. Liao, C. Zhang, and Y. Zhao, “Progressively complementary network for fisheye image rectification using appearance flow,” in *CVPR*, 2021, pp. 6348–6357.
- [10] S. Yang, C. Lin, K. Liao, and Y. Zhao, “Dual diffusion architecture for fisheye image rectification: Synthetic-to-real generalization,” *arXiv preprint arXiv:2301.11785*, 2023.
- [11] S. Gao, K. Yang, H. Shi, K. Wang, and J. Bai, “Review on panoramic imaging and its applications in scene understanding,” *TIM*, vol. 71, pp. 1–34, 2022.
- [12] D. Cheng, C. Gong, C. Xu, and Y. Wang, “Design of an ultrawide angle catadioptric lens with an annularly stitched aspherical surface,” *OE*, vol. 24, no. 3, pp. 2664–2677, 2016.
- [13] S. Gao, E. A. Tsyanok, and X. Xu, “Design of a compact dual-channel panoramic annular lens with a large aperture and high resolution,” *AO*, vol. 60, no. 11, pp. 3094–3102, 2021.
- [14] Q. Jiang *et al.*, “Computational optics meet domain adaptation: Transferring semantic segmentation beyond aberrations,” *arXiv preprint arXiv:2211.11257*, 2022.
- [15] Y. Peng, Q. Fu, F. Heide, and W. Heidrich, “The diffractive achromat full spectrum computational imaging with diffractive optics,” *TOG*, vol. 35, no. 4, pp. 1–11, 2016.
- [16] Y. Peng, Q. Sun, X. Dun, G. Wetzstein, W. Heidrich, and F. Heide, “Learned large field-of-view imaging with thin-plate optics,” *TOG*, vol. 38, no. 6, pp. 1–14, 2019.
- [17] X. Li, J. Suo, W. Zhang, X. Yuan, and Q. Dai, “Universal and flexible optical aberration correction using deep-prior based deconvolution,” in *ICCV*, 2021, pp. 2593–2601.
- [18] Q. Sun, C. Wang, Q. Fu, X. Dun, and W. Heidrich, “End-to-end complex lens design with differentiate ray tracing,” *TOG*, vol. 40, no. 4, pp. 1–13, 2021.

- [19] C. Wang, N. Chen, and W. Heidrich, “dO: A differentiable engine for deep lens design of computational imaging systems,” *TCI*, vol. 8, pp. 905–916, 2022.
- [20] X. Yang, Q. Fu, and W. Heidrich, “Curriculum learning for ab initio deep learned refractive optics,” *arXiv preprint arXiv:2302.01089*, 2023.
- [21] K. Zhang, X. Zhong, L. Zhang, and T. Zhang, “Design of a panoramic annular lens with ultrawide angle and small blind area,” *AO*, vol. 59, no. 19, pp. 5737–5744, 2020.
- [22] J. Wang, J. Bai, K. Wang, and S. Gao, “Design of stereo imaging system with a panoramic annular lens and a convex mirror,” *OE*, vol. 30, no. 11, pp. 19 017–19 029, 2022.
- [23] P. Greguss, “Panoramic imaging block for three-dimensional space,” Jan. 28 1986, US Patent 4,566,763.
- [24] I. Powell, “Panoramic lens,” *AO*, vol. 33, no. 31, pp. 7356–7361, 1994.
- [25] S. Thibault, J. Gauvin, M. Doucet, and M. Wang, “Enhanced optical design by distortion control,” in *SPIE*, vol. 5962, 2005, pp. 307–314.
- [26] D. Geng, H.-t. Yang, C. Mei, and Y.-h. Li, “Optical system design of space fisheye lens and performance analysis,” in *SPIE*, vol. 10462, 2017, pp. 1276–1282.
- [27] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using swin transformer,” in *ICCVW*, 2021, pp. 1833–1844.
- [28] Q. Jiang, H. Shi, L. Sun, S. Gao, K. Yang, and K. Wang, “Annular computational imaging: Capture clear panoramic images through simple lens,” *TCI*, vol. 8, pp. 1250–1264, 2022.
- [29] X. Wang *et al.*, “ESRGAN: Enhanced super-resolution generative adversarial networks,” in *ECCVW*, vol. 11133, 2018, pp. 63–79.
- [30] Y. Li *et al.*, “Efficient and explicit modelling of image hierarchies for image restoration,” in *CVPR*, 2023, pp. 18 278–18 289.
- [31] W. Li, X. Lu, J. Lu, X. Zhang, and J. Jia, “On efficient transformer and image pre-training for low-level vision,” *arXiv preprint arXiv:2112.10175*, 2021.
- [32] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, “Activating more pixels in image super-resolution transformer,” in *CVPR*, 2023, pp. 22 367–22 377.
- [33] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in *CVPRW*, 2017, pp. 1132–1140.
- [34] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in *ECCV*, vol. 11211, 2018, pp. 294–310.
- [35] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, “Restormer: Efficient transformer for high-resolution image restoration,” in *CVPR*, 2022, pp. 5718–5729.
- [36] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, “Uformer: A general U-shaped transformer for image restoration,” in *CVPR*, 2022, pp. 17 662–17 672.
- [37] L. Chen, X. Lu, J. Zhang, X. Chu, and C. Chen, “HINet: Half instance normalization network for image restoration,” in *CVPRW*, 2021, pp. 182–192.
- [38] L. Chen, X. Chu, X. Zhang, and J. Sun, “Simple baselines for image restoration,” in *ECCV*, vol. 13667, 2022, pp. 17–33.
- [39] H. Ai, Z. Cao, J. Zhu, H. Bai, Y. Chen, and L. Wang, “Deep learning for omnidirectional vision: A survey and new perspectives,” *arXiv preprint arXiv:2205.10468*, 2022.
- [40] J. Zhang, K. Yang, C. Ma, S. Reiß, K. Peng, and R. Stiefelhagen, “Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation,” in *CVPR*, 2022, pp. 16 896–16 906.
- [41] H. Chen, W. Hu, K. Yang, J. Bai, and K. Wang, “Panoramic annular SLAM with loop closure and global optimization,” *AO*, vol. 60, no. 21, pp. 6264–6274, 2021.
- [42] Z. Wang, K. Yang, H. Shi, P. Li, F. Gao, and K. Wang, “LF-VIO: A visual-inertial-odometry framework for large field-of-view cameras with negative plane,” in *IROS*, 2022, pp. 4423–4430.
- [43] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf, “Non-stationary correction of optical aberrations,” in *ICCV*, 2011, pp. 659–666.
- [44] Y. Liu, C. Zhang, T. Kou, Y. Li, and J. Shen, “End-to-end computational optics with a singlet lens for large depth-of-field imaging,” *OE*, vol. 29, no. 18, pp. 28 530–28 548, 2021.
- [45] G. Côté, F. Mannan, S. Thibault, J.-F. Lalonde, and F. Heide, “The differentiable lens: Compound lens search over glass surfaces and materials for object detection,” in *CVPR*, 2023, pp. 20 803–20 812.
- [46] H. Trussell and B. Hunt, “Image restoration of space variant blurs by sectioned methods,” in *ICCASP*, vol. 3, 1978, pp. 196–198.
- [47] J. Kim, A. Tsai, M. Cetin, and A. S. Willsky, “A curve evolution-based variational approach to simultaneous image restoration and segmentation,” in *ICIP*, vol. 1, 2002, pp. I–I.
- [48] E. Kee, S. Paris, S. Chen, and J. Wang, “Modeling and removing spatially-varying optical blur,” in *ICCP*, 2011, pp. 1–8.
- [49] Y. Xue, Q. Yang, G. Hu, K. Guo, and L. Tian, “Deep-learning-augmented computational miniature mesoscope,” *Optica*, vol. 9, no. 9, pp. 1009–1021, 2022.
- [50] L. Denis, E. Thiébaut, F. Soulez, J.-M. Becker, and R. Mourya, “Fast approximations of shift-variant blur,” *IJCV*, vol. 115, pp. 253–278, 2015.
- [51] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. MIT press Cambridge, MA, 1949, vol. 113, no. 21.
- [52] W. H. Richardson, “Bayesian-based iterative method of image restoration,” *JOSA*, vol. 62, no. 1, pp. 55–59, 1972.
- [53] L. B. Lucy, “An iterative technique for the rectification of observed distributions,” *The Astronomical Journal*, vol. 79, p. 745, 1974.
- [54] J. Sun, W. Cao, Z. Xu, and J. Ponce, “Learning a convolutional neural network for non-uniform motion blur removal,” in *CVPR*, 2015, pp. 769–777.
- [55] S. Chen, H. Feng, K. Gao, Z. Xu, and Y. Chen, “Extreme-quality computational imaging via degradation framework,” in *ICCV*, 2021, pp. 2612–2621.
- [56] S. Chen, T. Lin, H. Feng, Z. Xu, Q. Li, and Y. Chen, “Computational optics for mobile terminals in mass production,” *TPAMI*, vol. 45, no. 4, pp. 4245–4259, 2023.
- [57] T. Lin, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, “Non-blind optical degradation correction via frequency self-adaptive and finetune tactics,” *OE*, vol. 30, no. 13, pp. 23 485–23 498, 2022.
- [58] E. Agustsson and R. Timofte, “NTIRE 2017 challenge on single image super-resolution: Dataset and study,” in *CVPRW*, 2017, pp. 1122–1131.
- [59] X. Deng, H. Wang, M. Xu, Y. Guo, Y. Song, and L. Yang, “LAU-net: Latitude adaptive upscaling network for omnidirectional image super-resolution,” in *CVPR*, 2021, pp. 9189–9198.
- [60] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba, “Recognizing scene viewpoint using panoramic place representation,” in *CVPR*, 2012, pp. 2695–2702.
- [61] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, “Designing a practical degradation model for deep blind image super-resolution,” in *ICCV*, 2021, pp. 4771–4780.
- [62] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data,” in *ICCVW*, 2021, pp. 1905–1914.
- [63] E. Huggins, “Introduction to fourier optics,” *Physics Teacher*, vol. 45, no. 6, pp. 364–368, 2007.
- [64] V. N. Mahajan, “Zernike circle polynomials and optical aberrations of systems with circular pupils,” *AO*, vol. 33, no. 34, pp. 8121–8124, 1994.
- [65] D. Li *et al.*, “Involution: Inverting the inheritance of convolution for visual recognition,” in *CVPR*, 2021, pp. 12 321–12 330.
- [66] Y. Jiang, B. Wronski, B. Mildenhall, J. T. Barron, Z. Wang, and T. Xue, “Fast and high quality image denoising via malleable convolution,” in *ECCV*, vol. 13678, 2022, pp. 429–446.
- [67] Q. Zhang, Y. Xu, J. Zhang, and D. Tao, “VSA: Learning varied-size window attention in vision transformers,” in *ECCV*, 2022, pp. 466–483.
- [68] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *TIP*, vol. 13, no. 4, pp. 600–612, 2004.
- [69] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *CVPR*, 2018, pp. 586–595.
- [70] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in *NeurIPS*, vol. 30, 2017, pp. 6626–6637.
- [71] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “completely blind” image quality analyzer,” *SPL*, vol. 20, no. 3, pp. 209–212, 2012.
- [72] A. Mittal, A. Moorthy, and A. Bovik, “Referenceless image spatial quality evaluation engine,” in *ACSSC*, vol. 38, 2011, pp. 53–54.
- [73] J. Liang, H. Zeng, and L. Zhang, “Details or artifacts: A locally discriminative learning approach to realistic image super-resolution,” in *CVPR*, 2022, pp. 5647–5656.
- [74] R. Timofte *et al.*, “NTIRE 2017 challenge on single image super-resolution: Methods and results,” in *CVPRW*, 2017, pp. 1110–1121.

## APPENDIX A SAMPLE IMAGES OF PALHQ

We show the shooting device and sample images of PALHQ in Fig. A.1. The high-quality PAL images dataset covers a wide variety of scenes. The PAL can present 360° imaging of the surroundings, but with a blind area in the center of the image due to the reflective surface in the center FoV. As illustrated on the left of Fig. A.1, the PAL is usually placed toward the sky during the application, where the occlusion of the blind area causes little influence on the acquisition of panoramic information. PALHQ serves as the cornerstone of our PCIE for training a robust model to process MPIP images. Additionally, PALHQ can be transmitted to unfolded panoramas via equirectangular projection (ERP) for other various panoramic image processing applications. More types of minimalist PAL design, *e.g.*, using fewer lenses or applying meta surface, would also benefit from PALHQ for training learning-based recovery models to improve imaging quality.



Fig. A.1. The shooting device and sample images of PALHQ. With a well-designed PAL of 11 lenses and a Sony α6600 camera (on the left), we capture high-quality panoramic images covering various scenes including indoor, natural, urban, campus, and scenic spots (on the right).

## APPENDIX B VISUALIZATION OF THE PSF MAP

The visualization of the process of producing PSF maps is illustrated in Fig. B.1. As depicted in Sec. IV-A, we locate the corresponding FoV of the target pixel and obtain the R/G/B PSFs. Then, they are rotated according to their locations and compressed into a shape of  $k' \times k' \times 1$  via adaptive average pooling, where  $k'$  is set to 5 as an example. Finally, we reshape the compressed kernel into channel dimension and insert it into the pixel to produce the PSF map.

## APPENDIX C PIPELINE FOR OIQE CALCULATION

The detailed pipeline for calculating the defined OIQE is shown in Fig. C.1. We capture 8 testing images of the checkerboard under different ISP settings with our MPIP-P1 and AC pipeline. For evaluating different models in terms

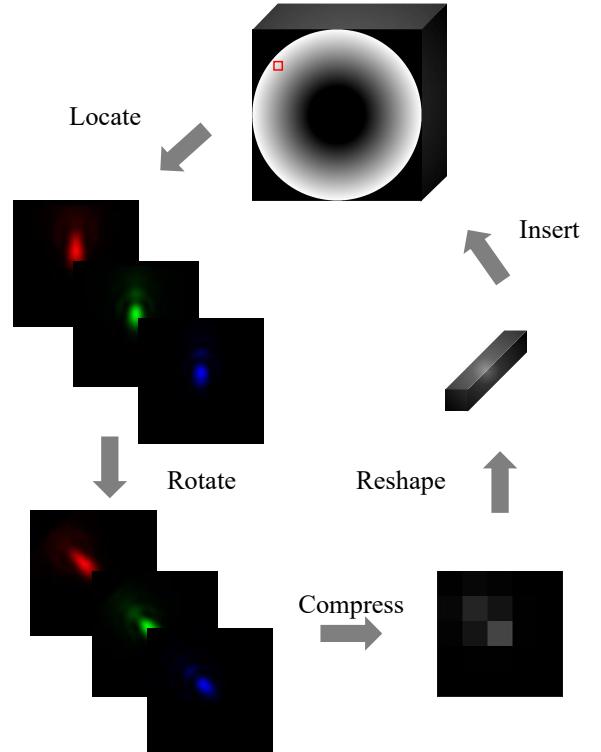


Fig. B.1. The visualization of compressing PSFs into PSF maps.

of aberration correction in real-world scenes, the processed testing images are fed into the OIQE pipeline, where sample knife-edges from different FoVs are cropped for SFR testing. In OIQE, the comprehensive results of different shooting settings and FoVs present a more credible evaluation of the ability to correct optical aberrations.

## APPENDIX D IMPLEMENTATION DETAILS OF SIMULATION MODEL

In this section, we introduce how to generate aberration images based on the simulation model and *Zemax* software in specific. To apply Eq. (1) for simulating aberrations, a set of PSFs under all FoVs of the target optical lens are required. We input the structure of MPIP into *Zemax*, then calculate the Zernike standard coefficients under different FoVs (128 FoVs from the minimum to the maximum FoV) and wavelengths (31 wavelengths from 400nm to 700nm), where the first 37 polynomials are kept as a common practice. In this way, the Zernike coefficients matrix with a shape of  $31 \times 128 \times 37$  is produced. Then, we plug the coefficients into Eq. (2) to describe the wavefronts under all FoVs and wavelengths. The random disturbance strategy is applied here to fine-tune the coefficients for multiple virtual aberration distributions. Before calculating PSFs, we also need to have access to the spot diagram and illumination distribution of the MPIP in *Zemax*, where the sizes of spots determine the kernel sizes (the ratio of the spot size to the pixel size) of PSFs, and the illumination provides the relative amplitude of PSFs.

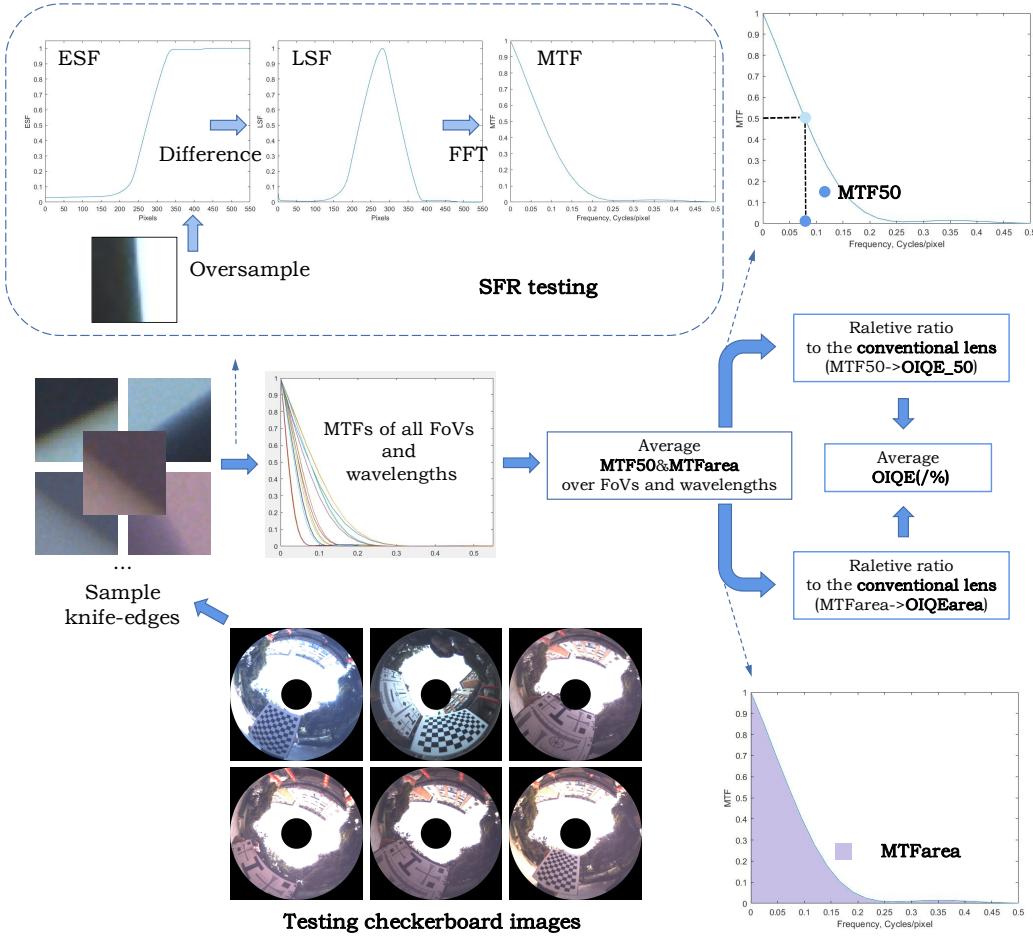


Fig. C.1. Pipeline for OIQE calculation. ESF: Edge Spread Function. LSF: Line Spread Function.

Finally, the wavefronts are transformed to PSFs  $K_\theta(x, y, \lambda)$  via Eq. (D.1) to Eq. (D.3):

$$\mathcal{P}_\theta(x, y, \lambda) = P(x, y)e^{i\Phi_\theta(x, y, \lambda)}, \quad (\text{D.1})$$

$$E_\theta(x, y, \lambda) = \frac{E_0}{\lambda d} \iint \mathcal{P}_\theta(x', y', \lambda) e^{-i\frac{2\pi}{\lambda d}(x'x + y'y)} dx' dy', \quad (\text{D.2})$$

$$K_\theta(x, y, \lambda) = |E_\theta(x, y, \lambda)|^2, \quad (\text{D.3})$$

where  $P(x, y)$  is the circ function and  $d$  is the distance from exit pupil to image plane. With multiple sets of  $K_\theta(x, y, \lambda)$  under different random disturbances, we generate aberration images of multiple virtual MPIPs samples via Eq. (1), where the high-quality MPIP images are transformed to raw images by invert-ISPs (Gamma Decompression, Invert Color Correction Matrix, and Invert White Balance), and the aberrated raw images are further processed by ISP (Mosaiced, Adding Noise, Demosaiced, White Balance, Color Correction Matrix, and Gamma Compression) to obtain the final results.

direct way, which improves the performance with extra parameters and computational overhead. More efficient PSF-aware architectures or training strategies are expected for better results. Then, we investigate state-of-the-art GAN-based training strategies, while there is open research space for further suppressing artifacts. Aside from this, the results of PCIE on real-world data are not as good as on synthetic data, where artifacts and fake details exist in some recovered images. The synthetic-to-real gap is huge and needs research on domain adaptation. The image number of PALHQ is also limited due to the difficulties of capturing high-quality PAL images under various scenes. We intend to design a hybrid training approach to take advantage of the large data size of the publicly available perspective datasets while improving the training with PALHQ. Finally, an end-to-end framework for joint optimization of MPIP design and recovery model will be focused to present a more general engine of minimalist and high-quality panoramic imaging.

## APPENDIX E

### LIMITATION ANALYSIS AND FUTURE PERSPECTIVES

There are still some limitations in PCIE, which call for further investigation into extremely high-quality imaging. Firstly, we only design the PSF-aware mechanisms in a simple and