Spherical Pseudo-Cylindrical Representation for Omnidirectional Image Super-resolution

Qing Cai[1], Mu Li[2], Dongwei Ren[3], Jun Lyu[4], Haiyong Zheng[1]*, Junyu Dong[1]*, Yee-Hong Yang[5]

[1] Faculty of Computer Science and Technology, Ocean University of China
[2] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
[3] School of Computer Science and Technology, Harbin Institute of Technology
[4] School of Nursing, The Hong Kong Polytechnic University
[5] Department of Computing Science, University of Alberta
cq@ouc.edu.cn, {limuhit,rendongweihit}@gmail.com, junlyu@polyu.edu.hk, {zhenghaiyong,dongjunyu}@ouc.edu.cn,
herberty@ualberta.ca

Abstract

Omnidirectional images have attracted significant attention in recent years due to the rapid development of virtual reality technologies. Equirectangular projection (ERP), a naive form to store and transfer omnidirectional images, however, is challenging for existing two-dimensional (2D) image super-resolution (SR) methods due to its inhomogeneous distributed sampling density and distortion across latitude. In this paper, we make one of the first attempts to design a spherical pseudo-cylindrical representation, which not only allows pixels at different latitudes to adaptively adopt the best distinct sampling density but also is model-agnostic to most off-the-shelf SR methods, enhancing their performances. Specifically, we start by upsampling each latitude of the input ERP image and design a computationally tractable optimization algorithm to adaptively obtain a (sub)-optimal sampling density for each latitude of the ERP image. Addressing the distortion of ERP, we introduce a new viewport-based training loss based on the original 3D sphere format of the omnidirectional image, which inherently lacks distortion. Finally, we present a simple yet effective recursive progressive omnidirectional SR network to showcase the feasibility of our idea. The experimental results on public datasets demonstrate the effectiveness of the proposed method as well as the consistently superior performance of our method over most state-of-the-art methods both quantitatively and qualitatively.

Introduction

Omnidirectional images also referred to as 360° images, provide 360° × 180° field-of-view (FoV), and enable an excellent immersive experience. Recent years have garnered significant attention in many real-world applications, including robotics (Su and Grauman 2021; Scaramuzza 2007), computer vision (Khasanova and Frossard 2017; Ozcinar, Rana, and Smolic 2019), virtual reality (VR) and augmented reality (AR) (Su and Grauman 2019; Deng et al. 2021) and gaming (Tateno, Navab, and Tombari 2018). In general, the original omnidirectional image format, *i.e.*, the 3D sphere, must be transformed into a 2D planar representation to facilitate storage and transmission. Equirectangular projection (ERP) is the most popular projection form, in which the latitude and the longitude of the original spherical image are mapped to the horizontal and vertical grid coordinates. As a result, the distributed sampling density in the ERP is inhomogeneous and distorted across latitude (as shown in Fig. 1(a)), making it unfriendly to subsequent visual communication applications. Additionally, when considering the trade-off between resolution and ease of storage and transmission, omnidirectional images usually have low resolutions (Elbamby et al. 2018; Deng et al. 2021).

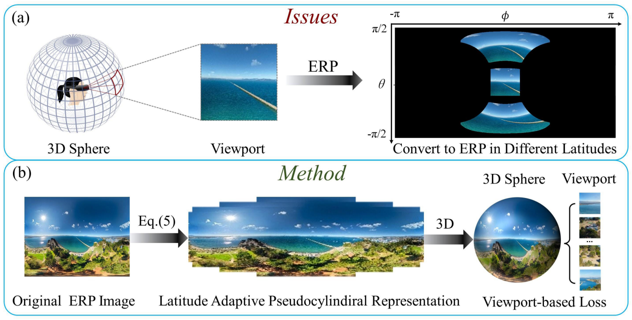

Figure 1: **(a)** Illustration of the inhomogeneous distributed sampling density and distortion issues of an ERP image. **(b)** Concept of the proposed spherical pseudo-cylindrical representation to address the above issues.

Image super-resolution (SR) is a technique that aims to recover high-resolution (HR) image from its corresponding degraded low-resolution (LR) version with algorithms alone, without the need for any hardware device. It plays an important and fundamental role in many computer vision tasks (Haris, Shakhnarovich, and Ukita 2018; Zhang et al. 2021; Xia et al. 2022; Cao et al. 2016; Cai et al. 2023; Lyu et al. 2023). Due to their superior feature representation capabilities, convolutional neural networks (CNNs) have achieved remarkable success in SR and many architectures have been presented so far, for example, residual learning (Kim, Lee, and Lee 2016; Nie et al. 2020), dense

*Corresponding authors

connections (Zhang et al. 2020; Song et al. 2020), UNet-like architectures with skip connections (Hu et al. 2019; Prajapati et al. 2021), dilated convolutions (Yang et al. 2017; Zhang et al. 2017), generative models (Ledig et al. 2017; Li et al. 2022) and other kinds of CNNs (Tai et al. 2017; Lu et al. 2021; Gao et al. 2022). Very recently, Transformer-based SR methods (Liang et al. 2021; Chu et al. 2021) have been proposed to fully utilize the advantages of Transformers in establishing long-range dependencies. However, directly applying these methods to omnidirectional images yields unsatisfactory performance, as they do not consider the inhomogeneous distributed sampling density and distortion across latitude in the ERP (Deng et al. 2021).

To address the above issues, two methods have been proposed. The first one utilizes priors of sphere-to-plane mapping, such as the LAU-Net (Deng et al. 2021), which not only mitigates the issues of ERP but also is model-agnostic to most existing methods. However, the ability of this method is limited by the intrinsic characteristics of ERP (Yoon et al. 2022). Additionally, it uses loss functions designed for 2D planar image SR, significantly impacting its performance for omnidirectional image SR by not considering the sampling issues of ERP. The other method designs spherical convolution for omnidirectional images, as see in SphereSR (Yoon et al. 2022), where a new kernel weight is proposed to adapt to the inhomogeneous distributed sampling density and distortion across latitude in ERP. While achieving impressive performance in omnidirectional image super-resolution, this method suffers from high computational costs due to the repeated switching between spherical and 2D planar coordinates. Additionally, it limits the amortization cost of convolution, as intermediate representations with different latitudes cannot be shared across 2D planar images since they are projected to different planes.

From the above discussions, one question arises immediately: Is there a simple yet effective method that can address the issues of ERP by fully utilizing the advantages of the above two methods and avoiding their weaknesses, and that can also be directly applied to most off-the-shelf SR network architectures? Motivated by the above questions, as shown in Fig. 1(b), this paper first introduces a novel latitude adaptive pseudo-cylindrical representation (LAPR) by designing a computationally tractable optimization algorithm to adaptively optimize the sampling density for each latitude of the ERP image. Then, a new viewport-based loss is proposed based on the original 3D sphere format of the omnidirectional image by transferring the final recovered HR ERP image back to the original 3D sphere format to avoid the distortion of ERP. Finally, we design an end-to-end recursive Transformer network based on CNN and Transformer to demonstrate the feasibility of the proposed idea. We conduct extensive comparisons with recently proposed state-of-the-art (SOTA) methods on benchmark omnidirectional datasets. The experimental results demonstrate that our method achieves SOTA performance.

Briefly, the contributions of this paper include:

- We propose a novel LAPR for the omnidirectional image by designing a computationally tractable optimization algorithm to adaptively obtain the optimal configuration of an ERP image. This approach not only addresses the sampling issues of ERP images but also can be easily applied to existing SR methods, directly improving their performance for omnidirectional images.
- We also proposes a new viewport-based training loss introduced into the field of omnidirectional image SR, which successfully avoids the distortion of ERP images, as it is defined on the original 3D sphere format of the omnidirectional image.
- We design a simple yet effective recursive omnidirectional backbone, which not only achieves SOTA performance but is also much more efficient in memory usage by recursively unfolding CNN and Transformer.

Related Work

2D-SR Methods: SRCNN (Dong et al. 2014), which is pioneering work in applying CNN to single image SR, uses only a three-layer CNN to represent the mapping between low-resolution (LR) and high-resolution (HR) images. Based on the SRCNN, many deeper and wider CNN-based SR methods have been proposed. For example, by introducing residual learning into a deeper network, Kim *et al.* propose the VDSR (Kim, Lee, and Lee 2016). Lim *et al.* propose the EDSR (Lim et al. 2017) by removing unnecessary modules in conventional residual networks. Guo *et al.* propose DRN (Guo et al. 2020) by learning an additional dual regression mapping to estimate the down-sampling kernel. Later, attention mechanism is introduced into SR to guide the CNN to selectively focus on some features where there is more information. For example, Niu *et al.* propose the HAN (Niu et al. 2020) by integrating a layer attention module and a channel-spatial attention module into the residual blocks. Mei *et al.* design a novel non-local sparse attention with dynamic sparse attention pattern and propose the NLSN (Mei, Fan, and Zhou 2021). Zhang *et al.* design a highly efficient long-range attention block by simply cascading two shift-conv with a group-wise multi-scale self-attention module and propose the ELAN (Zhang et al. 2022). Recently, inspired by the significant success of Transformer in natural language processing for its advantages in modeling long-range context (Vaswani et al. 2017), it is also introduced into SR (Chen et al. 2021; Liang et al. 2021; Chu et al. 2021), such as the SwinIR (Liang et al. 2021) and the Swin2SR (Chu et al. 2021). However, without taking into account the inhomogeneous distributed sampling density and distortion across latitude in the ERP, all of these existing 2D-SR methods yield unsatisfactory performance for omnidirectional image SR (Deng et al. 2021; Yoon et al. 2022).

360°-SR Methods: Since the sampling issue and distortion in omnidirectional images are caused by the transformation between the original spherical image and the 2D planar image, researchers try to address such issues from two aspects. On the one hand, some researchers focus on addressing them by fully utilizing priors of sphere-to-plane mapping. For example, Ozcinar *et al.* define a novel training loss by introducing the weighted-to-spherically uniform structural similarity to tackle the distortion issue of ERP images and propose the 360-SS (Ozcinar, Rana, and Smolic 2019). Deng *et al.* propose LAU-Net (Deng et al. 2021) by designing a latitude adaptive upscaling network. This network can dynamically upscale different latitude bands with varying upscaling factors, using a smaller upscaling factor for areas near the pole and a larger upscaling factor for areas around the equator of an ERP image. Although such method can mitigate the sampling issues of the ERP, it can not address them successfully because the method is based on the transformed format, *i.e.*, ERP, which is limited by the intrinsic characteristics of the ERP. Even worse, such method still uses the loss functions designed for 2D planar image SR, which seriously affects its performance for omnidirectional image SR because of not considering the sampling issues of the ERP. On the other hand, other researchers focus on the original spherical image and propose spherical CNN. For example, Coors *et al.* propose Spherenet (Coors, Condurache, and Geiger 2018) by designing a CNN filter based on its spatial location on a sphere to address the distortion issue of the ERP image. Yoon *et al.* apply convolution to the spherical structure constructed based on the subdivision of the icosahedron and propose SphereSR (Yoon et al. 2022). While this method has achieved superior results, it cannot be directly applied to existing 2D-SR architectures. This limitation arises because intermediate representations, extracted using spherical CNN with different latitudes, cannot be shared across 2D planar images. Additionally, it suffers from high computational costs due to the repeated switching between spherical and 2D planar coordinates.

Method

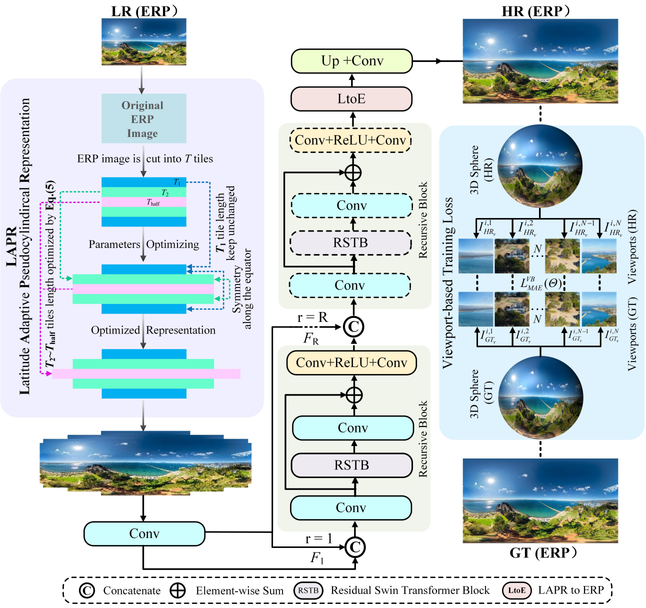

Figure 2: The overall framework of the proposed method mainly consists of three parts: latitude adaptive pseudo-cylindrical representation, viewport-based training loss, and recursive omnidirectional network.

To address the issues discussed above, this paper proposes a new method by designing a novel representation with optimized hyperparameter settings for the sphere-to-plane mapping and by defining a novel viewport-based training loss. The overall framework is shown in Fig. 2. Specifically, we first propose a latitude adaptive pseudo-cylindrical representation (LAPR) based on the sinusoidal projection since it satisfies all the three major projection requirements: equal-area, conformal and equidistant [1]. However, when we directly apply sinusoidal projection for image SR, it only achieves a relatively small improvement compared to ERP, as it downsamples each row of the input low-resolution (LR) ERP image and loses some important information for SR. Thus, we propose our LAPR by firstly upsampling each row of the input LR ERP image and then designing a computationally tractable optimization algorithm to adaptively obtain a (sub)-optimal configuration of the latitude representation for ERP. Then, we propose our novel viewport-based loss, relaying on the original 3D sphere format of the omnidirectional image. This effectively mitigates the distortion of ERP by defining loss on the 3D sphere rather than ERP. Finally, we design a simple yet effective recursive progressive backbone to demonstrate the feasibility of the proposed idea. Additionally, we discuss the significant differences between our method and the two 360°-SR methods.

LAPR

For an omnidirectional image $\mathbf{x} \in \mathbb{R}^{H \times W}$ represented in ERP with height H and width W, its plane-to-sphere coordinate conversion can be computed by:

$$\theta_i = \left(0.5 - \frac{i+0.5}{H}\right) \times \pi, (0 \le i < H), \tag{1}$$

$$\phi_j = \left(\frac{j+0.5}{W} - 0.5\right) \times 2\pi, (0 \le j < W), \tag{2}$$

where θ and ϕ, respectively, denote the latitude and the longitude. We also define our representation in a 2D image domain $\Omega = \{0, \ldots, H-1\} \times \{W_0^{new}, \ldots, W_{W-1}^{new}\}$, which is parameterized by $\{W_i^{new}\}_{i=0}^{W-1}$, where $W_i^{new} \in \{\mu_1 * W, \ldots, \mu_{W-1} * W\}$ denotes the width of the i-th row and μ_i the magnification of each row. To avoid information loss caused by performing down sampling, we define our representation by up sampling each row of the original LR image. Therefore, μ_i is defined as a positive integer greater than 1, and bicubic interpolation is adopted as the up sampling filter if necessary. Sampling image by interpolation may increase the incidences of other forms of distortion. Generally speaking, distortion (such as aliasing) caused by sampling can be mitigated by the subsequent convolution operation, which has been validated by the results without similar distortion. By varying W_i^{new}, our representation can achieve precise control over the sampling density of each row, and the beginning point of each row B_i is defined by:

$$B_i = \lfloor (\max(W_i^{new}) - W_i^{new})/2 \rfloor, \tag{3}$$

[1] Equal-area, conformal, and equidistant map projections preserve relative scales of things and stuff, local angles, and great-circle distances between points, respectively, on the sphere.

where $\lfloor \cdot \rfloor$ denotes the floor function. We call this data structure the parametric pseudo-cylindrical representation because it can generalize several pseudo-cylindrical map projections by specifying different magnifications μ_i for different rows. For example, the map projection would be the standard ERP when $\mu_i = 1$, the sinusoidal projection when $\mu_i = \cos(\theta_i)$ and Eq. 2 is replaced by:

$$\phi_j = \left(\frac{j - B_i + 0.5}{W_i^{new}} - 0.5 \right) \times 2\pi, \tag{4}$$

for $j = \{B_i, \ldots, B_i + W_i - 1\}$ as the longitude mapping.

By selecting different combinations of magnification μ_i for each row and the plane-to-sphere mapping, our LAPR not only includes a broad class of pseudo-cylindrical map projections as special cases but also opens the door to other novel representations that may be more suitable for omnidirectional image SR. Unless stated otherwise, in the remainder of the paper, we use Eq. 1 and Eq. 4 as the plane-to-sphere coordinate conversion.

Generally, different omnidirectional images possess different pseudo-cylindrical representations for optimal SR performance. To obtain the best optimization parameters, we propose a computationally tractable optimization algorithm. Specifically, we start by reducing the pseudo-cylindrical representation to a tiled representation (Yu, Lakshman, and Girod 2015) to further simplify and propose our LAPR. In this LAPR, the neighboring rows with the same magnification μ_i can be viewed as a *tile* and the tiled representation $\mathbf{z}$ for the original input $\mathbf{x}$ can be defined as $\{z_t\}_{t=0}^{T-1}$, where $z_t \in \mathbb{R}^{H_t \times W_t^{new}}$ denotes the t-th tile and $T = H/H_t$ denotes the total number of tiles. Then, we formulate the optimization problem of the pseudo-cylindrical representation parameters as:

$$\begin{aligned} \min_{\{W_t^{new}\}} \quad & \frac{1}{|\Psi|} \sum_{\mathbf{x} \in \Psi} F(\mathbf{x}, \mathtt{ISR}(\mathbf{x}), \{W_t^{new}\}) \\ \text{s.t.} \quad & W_t^{new} \in \{\overline{W}_0^{new}, \ldots, \overline{W}_{L-1}^{new}\},\ 0 \leq t < T, \\ & W_t^{new} = W_{T-1-t}^{new},\ 0 \leq t \leq T_{\text{half}}, \\ & W_t^{new} \leq W_{t'}^{new},\ \text{for } t \leq t' \text{ and } 0 \leq t, t' \leq T_{\text{half}}, \end{aligned} \tag{5}$$

where $\mathbf{x}$ denotes the given image; $F(\cdot)$ denotes a quantitative measure function for the SR performance; $\mathtt{ISR}(\cdot)$ denotes an existing pre-trained public 2D-SR method; $\overline{W}_t^{new} = (t + 1) \lfloor W/L \rfloor$, where L is the number levels of quantized width W of the ERP image and it is set such that $L \ll W$ to reduce the search space of the possible widths. As shown in Fig. 1 (b), we make the proposed LAPR symmetrical along the equator to double the search speed, and thus $T_{\text{half}} = \lfloor (T-1)/2 \rfloor - 1$. Finally, the tractable optimization algorithm is proposed (Please see the supplementary for details).

Fig. 3 shows a comparison experiment between the representative 2D-SR model EDSR (Lim et al. 2017) using and without using the proposed LAPR. From the results, it can be observed that EDSR using the proposed LAPR outperforms that without using it. This not only demonstrates the effectiveness of the proposed LAPR but also shows that the LAPR is model-agnostic and can be applied to most off-the-shelf SR methods.

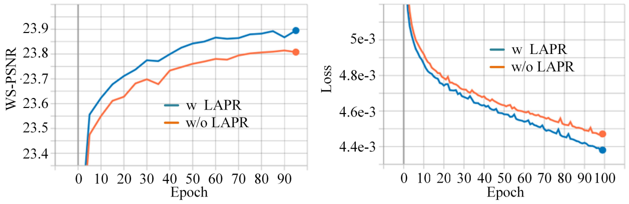

Figure 3: Comparison between existing methods with and without using the proposed LAPR.

Viewport-based Training Loss

As discussed in introduction section, almost all the existing training loss functions for omnidirectional image SR networks are designed for 2D planar image SR, which seriously limits their performance for omnidirectional image SR since they do not consider the distortion across the latitude of ERP images. When humans view an omnidirectional image using head-mounted displays, the ERP image is first transformed into a 3D sphere by using the plane-to-sphere coordinates defined in Eq. 1, Eq. 2. The visual content is then rendered as a viewport (as shown in Fig. 1 (a)), depending on the human's head position and the field-of-view (FoV) of the head-mounted display (Zhou et al. 2021).

Inspired by this observation, we define our training loss function based on the viewports of the omnidirectional image, reflecting how an omnidirectional image is viewed (Sui et al. 2021; Fang et al. 2022). Specifically, we first adopt rectilinear projections (Ye, Alshina, and Boyce 2017) to map the recovered HR image in ERP format back to the 3D sphere format and then sample 14 viewports uniformly distributed over the sphere for each omnidirectional image [2], which cover all spherical content. Each viewport is a $H_v \times W_v$ rectangle, where $H_v = \lceil \frac{H}{3} \rceil$ and $W_v = \lceil \frac{W}{4} \rceil$, with a FoV of $\frac{\pi}{3} \times \frac{\pi}{2}$. Given a training dataset $\{I_{HR_v}^{i,j}, I_{GT_v}^{i,j}\}_{i=1,j=1}^{N,14}$, which has N recovered images $I_{HR_v}^{i,j}$ (each of them has 14 viewports) and the corresponding ground truth images $I_{GT_v}^{i,j}$ (each of them has 14 viewports), our viewport-based loss function is defined as:

$$L_{MAE}^{VB}(\Theta) = \frac{1}{14} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{14} || I_{HR_v}^{i,j} - I_{GT_v}^{i,j} ||_1, \tag{6}$$

where Θ denotes the parameter set of the proposed network.

Recursive Omnidirectional Backbone

As shown in Fig. 2, our network is a progressive architecture designed by gradually unfolding recursive block (RB) constructed based on residual swin Transformer blocks (RSTB) (Liang et al. 2021), convolution layers, and ReLU layers. This progressive structure aims to recover the high-resolution (HR) omnidirectional image progressively from its low-resolution (LR) input. Specifically, we first represent the input LR omnidirectional image using the proposed

[2] The centers of the 14 viewports correspond to $(0, -\frac{\pi}{2})$, $(0, 0)$, $(0, \frac{\pi}{2})$, $(0, \pi)$, $(-\frac{\pi}{4}, -\frac{\pi}{2})$, $(-\frac{\pi}{4}, 0)$, $(-\frac{\pi}{4}, \frac{\pi}{2})$, $(-\frac{\pi}{4}, \pi)$, $(\frac{\pi}{4}, -\frac{\pi}{2})$, $(\frac{\pi}{4}, 0)$, $(\frac{\pi}{4}, \frac{\pi}{2})$, $(\frac{\pi}{4}, \pi)$, $(\frac{\pi}{2}, 0)$ and $(-\frac{\pi}{2}, 0)$, respectively.

LAPR to address the sampling issue of the ERP image. Then, it is input to the designed recursive network to extract deep features. Finally, the deep features are transformed back to the ERP form and then input into the upscale module and the reconstructed module to output the final HR omnidirectional image.

Following previous works (Zhang et al. 2018; Deng et al. 2021; Cai et al. 2022), we also use one convolution layer to extract the shallow feature (SF) F_0 from the represented LR omnidirectional image by our LAPR I_{LR}^{LAPR}:

$$F_0 = f_{SF}(I_{LR}^{LAPR}), \tag{7}$$

where f_{SF} is the convolution operation. Then, the extracted shallow feature is input to the proposed recursive network to further extract deep features (DF) F_{DF}:

$$F_{DF} = f_{RB}(f_{IN}(F_0, F_{s-1})), \tag{8}$$

where f_{IN} is the convolution operation; f_{RB} is the operation of the recursive block, which consists of *S* residual swin Transformer blocks (RSTB) (Liang et al. 2021) and four convolution layers and a ReLU layer. Motivated by the previous work (Ren et al. 2019), we implement the above progressive hybrid architecture by recursively unfolding it *R* times, as shown in Fig. 2. Considering that the parameters of our network mainly come from the recursive block, we adopt it in a recursive manner instead of directly stacking the recursive block to reduce the model size. Finally, the recovered HR omnidirectional image I_{HR} is obtained by mapping the deep features back to ERP F_{DF} and inputting them into an upscale module and a reconstruction module as follows:

$$I_{HR} = f_{Rec}(f_{UP}(F_{DF})), \tag{9}$$

where f_{UP} and f_{Rec} denote the operation of the upscale module and the reconstruction module, respectively.

Discussion

Below, we discuss the significant differences between our method and the two kinds of **360**°-SR methods discussed in the second paragraph of related work section.

Difference to the method based on priors of sphere-to-plane mapping: (i) This method only uses the 2D planar representation to mitigate the sampling issues of ERP, while we design our representation by integrating ERP format and sphere format into one model. Our approach fully utilizes the advantages of ERP format, making it model-agnostic to existing 2D-SR models, and also leverages the advantages of the sphere format to avoiding sampling issues in ERP. (ii) While this method still uses the loss functions designed for 2D planar image SR, it seriously affects its performance for omnidirectional image SR because it does not considering the sampling issues of ERP. In contrast, we design a novel loss for omnidirectional images based on the viewport of a 3D sphere.

Difference to the method based on designing spherical convolution: (i) This method needs to continuously switch between spherical and 2D planar coordinate to enable the proposed spherical convolution, resulting in high computational costs. In contrast, our method requires only one switch. (ii) This method presents a challenge to existing 2D-SR architectures as intermediate representations, obtained through spherical convolution with varying latitudes, cannot be effectively shared across 2D planar images. In contrast, our method is model-agnostic and compatible with most off-the-shelf SR models.

Experiments

Experiment Settings

Datasets: Following previous methods (Yoon et al. 2022), we also choose ODI-SR (Deng et al. 2021) as our training dataset, which contains 1200 training images, 100 validation images, and 100 testing images. We use the ODI-SR and SUN 360 Panorama (Xiao et al. 2012) as our test datasets.

Evaluation Metrics: To quantitatively compare the recovered HR results of the proposed model with that of the SOTA models, we use weighted-to-spherically-uniform PSNR (WS-PSNR) (Sun, Lu, and Yu 2017) and weighted-to-spherically-uniform SSIM (WS-SSIM) (Zhou et al. 2018). These are two widely used metrics for quantitatively evaluating the recovered omnidirectional image.

Implementation Details: Following previous works (Deng et al. 2021; Yoon et al. 2022), we train our model for the scales of $\times 8$ and $\times 16$, and all degraded datasets are obtained using bicubic interpolation. To avoid boundary artifacts between neighboring tiles, following previous work (Deng et al. 2021), an extra $\frac{H_t}{8}$ is added for neighboring tiles, where H_t denotes the height of each tile. The proposed model is trained by the ADAM optimizer (Kingma and Ba 2014) with a fixed initial learning rate of 10^{-4}. The whole process is implemented in the PyTorch platform with 4 RTX3090 GPUs, each with 24GB of memory (Please see the supplementary for more details).

Comparisons with State-of-the-art Methods

To validate the effectiveness and superior performance of the proposed method, we compare our method with 10 SOTA methods including 7 2D-SR methods: EDSR (Lim et al. 2017), HAN (Niu et al. 2020), DRN (Guo et al. 2020), NLSN (Mei, Fan, and Zhou 2021), SwinIR (Liang et al. 2021), ELAN (Zhang et al. 2022) and Swin2SR (Conde et al. 2023) (Note that, for a fair comparison, all the comparison methods are retrained on the ODI-SR dataset using their open-source codes with the same patch size as our method, dubbed as 2D-SR-Re), 3 360°-SR methods: 360-SS (Ozcinar, Rana, and Smolic 2019), LAU-Net (Deng et al. 2021) and SphereSR (Yoon et al. 2022).

Quantitative Comparison: Table 1 reports the quantitative comparisons between our method and 10 SOTA SISR methods on two benchmark datasets for scale factor $\times 8$ and $\times 16$. The best results are represented in bold and the second best in underlined. It can be found that, compared with these methods, our method achieves the best results on multiple benchmarks for all scaling factors and surpasses all of them in terms of WS-PSNR and WS-SSIM. In particular, our method improves the WS-PSNR value by 0.32 dB and 0.34 dB on the ODI-SR dataset for scale factor $\times 8$ and $\times 16$ compares with that of the second best method, respectively.

SR Methods		ODI-SR dataset				SUN 360 dataset			
		×8		×16		×8		×16	
		WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM
2D-SR-Re	EDSR	23.97	0.6483	22.24	0.6090	23.79	0.6472	21.83	0.5974
	DRN	24.32	0.6571	22.52	0.6212	24.25	0.6602	22.11	0.6092
	HAN	24.32	0.6620	22.53	0.6265	24.25	0.6681	22.12	0.6105
	NLSN	24.33	0.6684	22.53	0.6285	24.26	0.6709	22.14	0.6182
	SwinIR	24.34	0.6721	22.54	0.6288	24.27	0.6734	22.15	0.6273
	ELAN	24.35	0.6756	22.56	0.6390	24.28	0.6788	22.16	0.6355
	Swin2SR	<u>24.37</u>	0.6770	<u>22.58</u>	<u>0.6395</u>	<u>24.29</u>	<u>0.6822</u>	<u>22.18</u>	<u>0.6380</u>
360°-SR	360-SS	24.14	0.6539	22.35	0.6102	24.19	0.6536	22.10	0.5947
	LAU-Net	24.36	0.6602	22.52	0.6284	24.24	0.6708	22.05	0.6058
	SphereSR	<u>24.37</u>	<u>0.6777</u>	22.51	0.6370	24.17	0.6820	21.95	0.6342
	Ours	**24.72**	**0.6886**	**22.90**	**0.6480**	**24.53**	**0.6855**	**22.37**	**0.6475**

Table 1: Quantitative comparisons with state-of-the-art SR methods on two benchmark datasets for scale factor ×8 and ×16. The bold/underlined font represent the best/second best result.

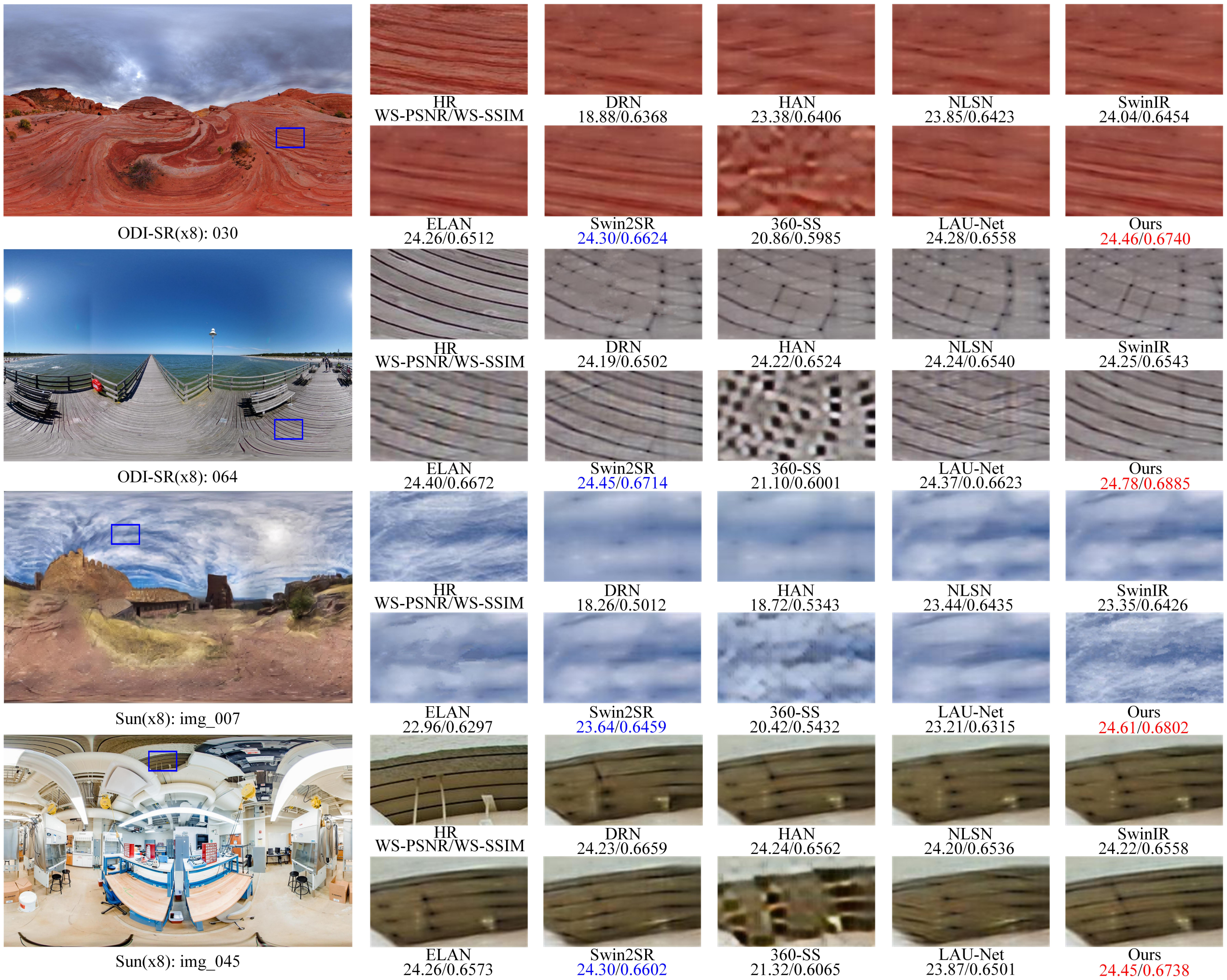

Figure 4: Visual comparisons with state-of-the-art SISR methods for 8× SR on the ODI-SR and the Sun 360 datasets. The colors red and blue represent the best and the second best methods. Best viewed on screen.

Qualitative Comparison: In Fig. 4, we also visually illustrate the zoomed-in comparison results with SOTAs on several images from the test datasets. From the results, we find that the proposed method can consistently obtain sharper results, recovering more high-frequency textures and details, while most competing models suffer from some unpleasant blurring artifacts. This successfully validates the effectiveness and efficiencies of the proposed method.

Further Comparison: Table 2 shows the FLOPs, the number of parameters, the running time and the WS-PSNR values comparisons between our method and SOTA methods on the ODI-SR dataset for scale factor of 16. It can be found that our method achieves the best performance with competitive efficiency and computation cost.

Model	FLOPs	Params	Time	WS-PSNR
SwinIR	900 G	11.5 M	0.982 s	22.54
ELAN	845 G	8.9 M	0.715 s	22.56
Swin2SR	952 G	12.3 M	1.041 s	22.58
360-SS	**15 G**	**1.6 M**	**0.025 s**	22.35
LAU-Net	685 G	9.4 M	0.443 s	22.52
SphereSR	587 G	8.7 M	0.401 s	22.51
Ours	372 G	7.8 M	0.312 s	**22.90**

Table 2: Computational complexity comparison on the ODI-SR dataset for scale factor $\times 16$.

Ablation Study

LAPR: As shown in Table 3, to evaluate the effectiveness of the proposed LAPR, we conduct a comparison experiment between our method using different representations: original ERP, sinusoidal projection and the proposed LAPR. It can be observed that the highest WS-PSRN and WS-SSIM values are obtained when the proposed LAPR is used.

Different Rep	ERP	Sinusoidal	Our LAPR
WS-PSNR	24.48	24.56	24.72
WS-SSIM	0.6688	0.6791	0.6886

Table 3: Different input representations comparison on the ODI-SR dataset for scale factor $\times 8$.

To validate the LAPR is model-agnostic to existing 2D-SR methods, as shown in Table 4, another comparison experiment between the Top 3 2D-SR models: SwinIR (Liang et al. 2021), ELAN (Zhang et al. 2022) and Swin2SR (Conde et al. 2023), using and without using the proposed LAPR is conducted. From the results, it can be observed that the values of all the three methods using the proposed LAPR outperform that of without using it. This not only validates the effectiveness of the proposed LAPR but also shows that the proposed LAPR is model-agnostic to most off-the-shelf SR methods and can improve their performance.

2D Models	SwinIR	ELAN	Swin2SR
w/o LAPR	24.34	24.35	24.37
w/ LAPR	24.48	24.50	24.52

Table 4: Comparison between exiting 2D-SR methods with and without using the proposed LAPR on the ODI-SR dataset for scale factor $\times 8$.

Viewport-Based Loss: To investigate the effectiveness of the proposed viewport-based loss, we conduct a comparison experiment between the proposed method using the widely used L_{MAE} loss designed for 2D planar image SR network and the proposed viewport-based loss L_{MAE}^{VB}. The corresponding WS-PSRN values are shown in Table 5. It can be found that our viewport-based loss achieves a better performance, which demonstrated its effectiveness.

Loss function	L_{MAE}	L_{MAE}^{VB}
WS-PSNR	22.80	22.90
WS-SSIM	0.6310	0.6480

Table 5: Influence of different training losses on the ODI-SR dataset for scale factor $\times 16$.

Recursive Network: To validate the superior performance of our method mainly come from the proposed LAPR and our viewport-based loss rather than the Transformer-based backbone, we train another version of our method by replacing the RSTB shown in Fig. 2 with 32-residual blocks similar to NLSA (Mei, Fan, and Zhou 2021), dubbed as Ours-C (Note that, for a fair comparison, all the parameter settings of residual blocks are the same as those in NLSA). As shown in Table 6, Ours-C can still achieve superior performance compared to previous SOTA methods, indirectly validating the effectiveness of our LAPR and viewport-based loss.

Model	Params (M)	WS-PSNR	WS-SSIM
SwinIR	11.5	24.34	0.6721
ELAN	8.9	24.35	0.6756
Swin2SR	12.3	24.37	0.6770
360-SS	**1.6**	24.14	0.6539
LAU-Net	9.4	24.36	0.6602
SphereSR	8.7	24.37	0.6777
Ours-C	7.5	**24.62**	**0.6779**

Table 6: Comparison on the ODI-SR dataset for scale $\times 8$.

Conclusion

In this paper, we present a novel method for accurate omnidirectional image super-resolution that effectively addresses sampling issues and distortion across the latitude of ERP images. Specifically, we introduce a latitude adaptive pseudo-cylindrical representation for omnidirectional images. This representations allows pixels at different latitudes to adaptively adopt the best distinct sampling density. This is achieved by employing the proposed computationally tractable optimization algorithm to search for the optimal width for each tile. Additionally, we propose a viewport-based loss, which reflects how humans view omnidirectional images, to mitigate the distortion of ERP. Finally, a recursive progressive backbone is designed to demonstrate the feasibility of our idea. Quantitative and qualitative evaluations on different benchmark datasets demonstrate the effectiveness of the proposed method, showcasing its the superior performance over most SOTA methods.

Acknowledgments

This work was supported in part by the National Science Foundation of China (No.62102338, 62102339, 62171421, 62371413); in part by the Natural Science Foundation of Shandong Province (No.ZR2020QF031); in part by the China Postdoctoral Science Foundation (No.2023M733342); in part by the Qingdao Postdoctoral Innovation Project under Grant (No.QDBSH20230101001); in part by TaiShan Scholars Youth Expert Program of Shandong Province (No.tsqn202306096); in part by the National Key R&D Program of China (No.2022ZD0117201) and in part by the University of Alberta, adn Natural Science and Engineering Research Council of Canada.

References

Cai, Q.; Li, J.; Li, H.; Yang, Y.-H.; Wu, F.; and Zhang, D. 2022. TDPN: Texture and Detail-Preserving Network for Single Image Super-Resolution. *IEEE TIP*, 31: 2375–2389.

Cai, Q.; Qian, Y.; Li, J.; Lyu, J.; Yang, Y.-H.; Wu, F.; and Zhang, D. 2023. HIPA: hierarchical patch transformer for single image super resolution. *IEEE TIP*, 32: 3226–3237.

Cao, L.; Ji, R.; Wang, C.; and Li, J. 2016. Towards domain adaptive vehicle detection in satellite image by supervised super-resolution transfer. In *AAAI*, volume 30, 1138–1144.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-trained image processing transformer. In *CVPR*, 12299–12310.

Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; and Shen, C. 2021. Twins: Revisiting the design of spatial attention in vision transformers. In *NIPS*, volume 34, 9355–9366.

Conde, M. V.; Choi, U.-J.; Burchi, M.; and Timofte, R. 2023. Swin2SR: Swinv2 transformer for compressed image super-resolution and restoration. In *ECCVW*, 669–687. Springer.

Coors, B.; Condurache, A. P.; and Geiger, A. 2018. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *ECCV*, 518–533.

Deng, X.; Wang, H.; Xu, M.; Guo, Y.; Song, Y.; and Yang, L. 2021. Lau-net: Latitude adaptive upscaling network for omnidirectional image super-resolution. In *CVPR*, 9189–9198.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *ECCV*, 184–199.

Elbamby, M. S.; Perfecto, C.; Bennis, M.; and Doppler, K. 2018. Toward low-latency and ultra-reliable virtual reality. *IEEE Network*, 32(2): 78–84.

Fang, Y.; Huang, L.; Yan, J.; Liu, X.; and Liu, Y. 2022. Perceptual quality assessment of omnidirectional images. In *AAAI*, volume 36, 580–588.

Gao, G.; Li, W.; Li, J.; Wu, F.; Lu, H.; and Yu, Y. 2022. Feature distillation interaction weighting network for lightweight image super-resolution. In *AAAI*, volume 36, 661–669.

Guo, Y.; Chen, J.; Wang, J.; Chen, Q.; Cao, J.; Deng, Z.; Xu, Y.; and Tan, M. 2020. Closed-loop matters: Dual regression networks for single image super-resolution. In *CVPR*, 5407–5416.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2018. Deep back-projection networks for super-resolution. In *CVPR*, 1664–1673.

Hu, X.; Naiel, M. A.; Wong, A.; Lamm, M.; and Fieguth, P. 2019. RUNet: A robust UNet architecture for image super-resolution. In *CVPRW*, 505–507.

Khasanova, R.; and Frossard, P. 2017. Graph-based classification of omnidirectional images. In *ICCVW*, 869–878.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate image super-resolution using very deep convolutional networks. In *ICCV*, 1646–1654.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 4681–4690.

Li, W.; Zhou, K.; Qi, L.; Lu, L.; and Lu, J. 2022. Best-buddy gans for highly detailed image super-resolution. In *AAAI*, volume 36, 1412–1420.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. SwinIR: Image restoration using swin transformer. In *ICCVW*, 1833–1844.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced deep residual networks for single image super-resolution. In *ICCVW*, 136–144.

Lu, L.; Li, W.; Tao, X.; Lu, J.; and Jia, J. 2021. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *CVPR*, 6368–6377.

Lyu, J.; Li, G.; Wang, C.; Cai, Q.; Dou, Q.; Zhang, D.; and Qin, J. 2023. Multicontrast MRI Super-Resolution via Transformer-Empowered Multiscale Contextual Matching and Aggregation. *IEEE TNNLS*.

Mei, Y.; Fan, Y.; and Zhou, Y. 2021. Image super-resolution with non-local sparse attention. In *CVPR*, 3517–3526.

Nie, S.; Ma, C.; Chen, D.; Yin, S.; Wang, H.; Jiao, L.; and Liu, F. 2020. A Dual Residual Network with Channel Attention for Image Restoration. In *ECCV*, 352–363.

Niu, B.; Wen, W.; Ren, W.; Zhang, X.; Yang, L.; Wang, S.; Zhang, K.; Cao, X.; and Shen, H. 2020. Single image super-resolution via a holistic attention network. In *ECCV*, 191–207.

Ozcinar, C.; Rana, A.; and Smolic, A. 2019. Super-resolution of omnidirectional images using adversarial learning. In *IEEE International Workshop on Multimedia Signal Processing*, 1–6.

Prajapati, K.; Chudasama, V.; Patel, H.; Sarvaiya, A.; Upla, K. P.; Raja, K.; Ramachandra, R.; and Busch, C. 2021. Channel Split Convolutional Neural Network (ChaSNet) for Thermal Image Super-Resolution. In *CVPR*, 4368–4377.

Ren, D.; Zuo, W.; Hu, Q.; Zhu, P.; and Meng, D. 2019. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, 3937–3946.

Scaramuzza, D. 2007. *Omnidirectional vision: from calibration to root motion estimation*. Ph.D. thesis, ETH Zurich.

Song, D.; Xu, C.; Jia, X.; Chen, Y.; Xu, C.; and Wang, Y. 2020. Efficient residual dense block search for image super-resolution. In *AAAI*, volume 34, 12007–12014.

Su, Y.-C.; and Grauman, K. 2019. Kernel transformer networks for compact spherical convolution. In *CVPR*, 9442–9451.

Su, Y.-C.; and Grauman, K. 2021. Learning Spherical Convolution for 360 Recognition. *IEEE TPAMI*.

Sui, X.; Ma, K.; Yao, Y.; and Fang, Y. 2021. Perceptual quality assessment of omnidirectional images as moving camera videos. *IEEE TVCG*, 28(8): 3022–3034.

Sun, Y.; Lu, A.; and Yu, L. 2017. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Processing Letters*, 24(9): 1408–1412.

Tai, Y.; Yang, J.; Liu, X.; and Xu, C. 2017. Memnet: A persistent memory network for image restoration. In *ICCV*, 4539–4547.

Tateno, K.; Navab, N.; and Tombari, F. 2018. Distortion-aware convolutional filters for dense prediction in panoramic images. In *ECCV*, 707–722.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

Xia, B.; Hang, Y.; Tian, Y.; Yang, W.; Liao, Q.; and Zhou, J. 2022. Efficient Non-Local Contrastive Attention for Image Super-Resolution. In *AAAI*.

Xiao, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2012. Recognizing scene viewpoint using panoramic place representation. In *CVPR*, 2695–2702.

Yang, W.; Tan, R. T.; Feng, J.; Liu, J.; Guo, Z.; and Yan, S. 2017. Deep joint rain detection and removal from a single image. In *CVPR*, 1357–1366.

Ye, Y.; Alshina, E.; and Boyce, J. 2017. JVET-G1003: Algorithm description of projection format conversion and video quality metrics in 360lib version 4. *Joint Video Exploration Team*.

Yoon, Y.; Chung, I.; Wang, L.; and Yoon, K.-J. 2022. SphereSR: 360deg Image Super-Resolution With Arbitrary Projection via Continuous Spherical Image Representation. In *CVPR*, 5677–5686.

Yu, M.; Lakshman, H.; and Girod, B. 2015. Content adaptive representations of omnidirectional videos for cinematic virtual reality. In *Proceedings of the International Workshop on Immersive Media Experiences*, 1–6.

Zhang, K.; Zuo, W.; Gu, S.; and Zhang, L. 2017. Learning deep CNN denoiser prior for image restoration. In *CVPR*, 3929–3938.

Zhang, X.; Zeng, H.; Guo, S.; and Zhang, L. 2022. Efficient Long-Range Attention Network for Image Super-resolution. *ECCV*.

Zhang, Y.; Li, K.; Li, K.; and Fu, Y. 2021. Mr image super-resolution with squeeze and excitation reasoning attention network. In *CVPR*, 13425–13434.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 286–301.

Zhang, Y.; Tian, Y.; Kong, Y.; Zhong, B.; and Fu, Y. 2020. Residual dense network for image restoration. *IEEE TPAMI*, 43(7): 2480–2495.

Zhou, Y.; Sun, Y.; Li, L.; Gu, K.; and Fang, Y. 2021. Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network. *IEEE TCSVT*, 32(4): 1767–1777.

Zhou, Y.; Yu, M.; Ma, H.; Shao, H.; and Jiang, G. 2018. Weighted-to-spherically-uniform SSIM objective quality evaluation for panoramic video. In *IEEE International Conference on Signal Processing*, 54–57.