Dual Enhancement in ODI Super-Resolution: Adapting Convolution and Upsampling to Projection Distortion

Xiang Ji[1], **Changqiao Xu**[1*], **Lujie Zhong**[2], **Shujie Yang**[1], **Han Xiao**[1] and **Gabriel-Miro Muntean**[3]

[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications

[2]Information Engineering College, Capital Normal University

[3]School of Electronic Engineering, Dublin City University

{jxiang_2019, cqxu, sjyang, xiaohan}@bupt.edu.cn, zhonglj@cnu.edu.cn, gabriel.muntean@dcu.ie

Abstract

Omnidirectional images (ODIs) demand considerably higher resolution to ensure high quality across all viewports. Traditional convolutional neural networks (CNN)-based single image super-resolution (SISR) networks, however, are not effective for spherical ODIs. This is due to the uneven pixel density distribution and varying texture complexity in different regions that arise when projecting from a sphere to a plane. Additionally, the computational and memory costs associated with large-sized ODIs present a challenge for real-world application. To address these issues, we propose an efficient distortion-adaptive super-resolution network (ODA-SRN). Specifically, ODA-SRN employs a series of specially designed Distortion Attention Block Groups (DABG) as its backbone. Our Distortion Attention Blocks (DABs) utilize multi-segment parameterized convolution to generate dynamic filters, which compensate for distortion and texture fading during feature extraction. Moreover, we introduce an upsampling scheme that accounts for the dependence of pixel position and distortion degree to achieve pixel-level distortion offset. A comprehensive set of results demonstrates that our ODA-SRN significantly improves the super-resolution performance for ODIs, both quantitatively and qualitatively, when compared to other state-of-the-art methods.

1 Introduction

Omnidirectional images (ODIs), as a novel visual form of virtual reality (VR) [Zhang *et al.*, 2023], are anticipated to become integral components of future social environments, enriching and enhancing the ways in which we share experiences. Consequently, ODIs are being introduced rapidly into people's lives. As their name suggests, ODIs offer a $360^\circ \times 180^\circ$ field of view, and the expansive visual range naturally demands high resolution. The level of detail and sharpness in ODIs directly influences the realism of user experiences. However, due to storage and transmission constraints, omnidirectional images often lack the necessary resolution to provide users with a truly immersive experience. Super-resolution (SR) technology transforms low-resolution (LR) images into high-resolution (HR) ones. As a long-standing computer vision problem, single-image super-resolution (SISR) has attracted extensive attention from researchers [Anwar *et al.*, 2021]. In recent years, several SISR studies based on convolutional neural networks (CNN) have achieved remarkable results [Senmao *et al.*, 2023; Wang *et al.*, 2021]. These studies utilize CNNs to extract features from images and subsequently employ the extracted features for upsampling. CNNs use convolution kernels to sense images through sliding windows, allowing them to capture local patterns. However, these works primarily target two-dimensional (2D) planar images, neglecting the global distortion information present in omnidirectional images (ODIs). For ODIs, equirectangular projection (ERP) from sphere to plane is widely employed for convenient storage and transportation. Regrettably, the distortion caused by ERP results in unevenly distributed features within ODIs. Fig. 1 illustrates the correlation between distortion degree and pixel position in ODI projection. Notably, stretch distortion is particularly severe in non-horizontal regions, often leading to unsatisfac-

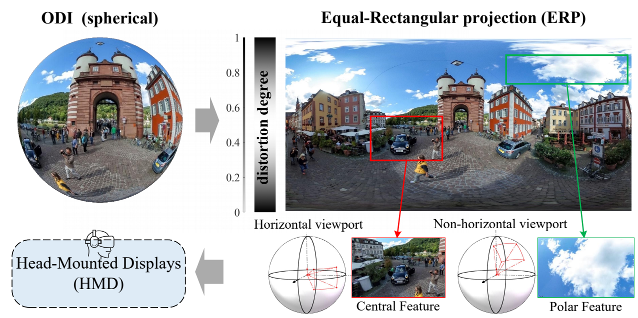

Figure 1: ERP projection from sphere to plane. The distortion degree is related to the location of the regions. Polar features of ERP format images have more stretch distortion than central features.

*Corresponding author.

tory feature extraction results for ODIs. This phenomenon raises the question: How can we achieve super-resolution in the presence of distortion and improve CNNs to effectively extract features from ODIs?

For ODI super-resolution (SR), traditional CNN-based feature extraction strategies have proven ill-suited for ODIs in practice. The issue arises from the fact that both existing CNN filters and image data are "flat", making it challenging for these filters to achieve the right balance with ODIs that exhibit varying pixel densities and texture complexities. Recently, [Deng *et al.*, 2021] proposed a multilevel CNN network that allocates additional computational resources to the central regions of ODIs. Their approach modified existing SR networks but did not fundamentally enhance the sensitivity of CNNs to ODI distortion. Conversely, most existing ODI SR research works [Ozcinar *et al.*, 2019; Su and Grauman, 2017; Sun *et al.*, 2021] tend to focus on the local resolution of ODIs. To mitigate the effects of distortion, these methods process tangent planes surrounding the sphere repeatedly. While this boosts accuracy, it also significantly heightens computational costs. These methods are heavily dependent on hardware performance, which poses a considerable burden for portable devices like head-mounted displays and cell phones.

In an effort to strike a balance between accuracy and cost, numerous CNN-based SR methods [Dai *et al.*, 2019; Liu *et al.*, 2020; Mei *et al.*, 2020; Zhang *et al.*, 2021] have introduced attention mechanisms. [Dai *et al.*, 2019] proposed a deep second-order attention network to capture spatial context information over long distances. [Liu *et al.*, 2020] incorporated spatial attention into the residual network to enhance model efficiency. Drawing on recent neuroscience research, [Zhang *et al.*, 2021] proposed CRAN, which utilizes semantic knowledge as a factor influencing network attention. However, these attention mechanisms appear to be ineffective in addressing the distortion caused by projection. This is primarily because these methods focus on the comprehensive utilization of global information without considering distortion shifts. These challenges limit the representational power of CNNs, and as a result, the development of an optimal single ODI SR network remains an ongoing endeavor.

Motivated by the observations and analyses discussed above, we propose a novel method named Omnidirectional Distortion Attention Super-Resolution Networks (ODA-SRN). Contrary to existing strategies, the core idea of our ODA-SRN is the dual enhancement of two fundamental components in SR, convolution and upsampling, rather than designing intricately structured integrated network models. Our innovation resides in considering the global and local differences of ERP during the convolution process, and introducing pixel-level offsets in the upsampling phase to accommodate distortion. This focus on enhancing fundamental components, rather than simply stacking larger models, enables our ODA-SRN to efficiently conserve computational and memory resources.

Our main contributions can be summarized as follows:

- We have introduced novel convolution and upsampling techniques for ODIs. These components have been integrated into our ODA-SRN, achieving superior SR results.
- We propose extracting the potential representation of distortion information to obtain a set of routing weights. By leveraging the relationship between routing weights and segment positions, we enhance the distortion sensitivity of CNN.
- We introduce a Jacobian matrix of spatially transformed elements and mathematically adjust the sampling window of the upsampling layer. The reusable Jacobian matrix improves the efficiency of our ODA-SRN.

2 Related Work

Deep CNN for Super Resolution. Over the past decade, SR has been one of the most extensively studied inverse problems [Chen *et al.*, 2022]. The combination of SR and CNN technology has led to significant progress in reconstruction error and practical applications [Chen *et al.*, 2021; Ma *et al.*, 2020; Senmao *et al.*, 2023]. Since [Dong *et al.*, 2014] first proposed SRCNN, numerous SR works based on deep CNN have been introduced. These works have made improvements in various aspects, such as increasing network depth to utilize more parameters, incorporating residual attention block mechanisms to enhance performance accuracy, and modifying the loss function [Li *et al.*, 2022]. All of these works have achieved remarkable results in SR for 2D images. However, a common characteristic of CNN-based SISR methods is that their "flat" convolutional layers are inefficient for ODIs in feature extraction. Specifically, due to the uneven pixel density and varying texture complexity of ODIs, it is challenging to optimize a convolutional kernel that can account for the effects of all degrees of distortion.

Super Resolution for VR. With the rise of virtual reality (VR) technology, omnidirectional super-resolution has emerged as a challenging task for the research community. In recent years, viewport-based methods have been proposed as a promising approach for enhancing large-sized ODIs [Li *et al.*, 2020]. [Su and Kristen, 2017] utilized different convolution kernels to extract features along various latitudes on ODIs; however, this approach required training a large number of convolution kernels. Inspired by the success of the Transformer architecture, [Shen *et al.*, 2022] proposed PanoFormer, a deep neural network capable of estimating depth in ODIs. SphereSR [Yoon *et al.*, 2022] incorporated RGB prediction into feature extraction and learned the relationships between different ODI formats to achieve more flexible SR. However, using Transformers for ERP-format ODIs poses challenges: their global attention may overlook local nuances critical for SR, and their computational demands can be burdensome given the lage-size nature of ODIs.

Techniques for Projection Distortion. Projection distortion in ODIs leads to distinct regions showing varied distortion degrees. In the realm of efficient adaptability, parameterized convolutions have emerged as a promising technique [Brandon Yang, 2020]. Unlike traditional convolutional layers that utilize static filters, parameterized convolutions change their filters based on the input. However, current

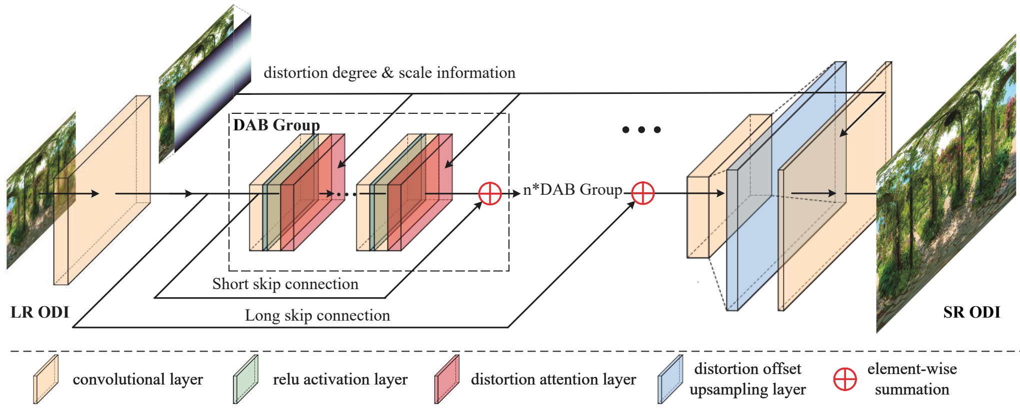

Figure 2: Overview of our proposed ODA-SRN architecture. Details of our distortion attention layer and distortion offset upsampling layer are illustrated in Fig.3 and Fig.4, respectively.

attention mechanisms are not guided by distortion information, which limits their ability to address the regional differences in ODIs caused by projection.

3 Proposed Method

Overview. We designed a concise network architecture, illustrated in Fig. 2. It demonstrates how a low-resolution ODI input ($\mathbf{I_{LR}} \in \mathcal{R}^{c\times h\times w}$) generates a high-resolution ODI ($\mathbf{I_{SR}} \in \mathcal{R}^{c\times sh\times sw}$), where s is a multiple of the super-resolution, c represents the number of channels, and h and w denote the ODI size.

3.1 Network Architecture

ODA-SRN mainly consists of two parts: the feature extraction part and the upsampling recovery part. First, shallow features $\mathbf{F_0}$ are extracted from the input LR ODI:

$$\mathbf{F_0} = f_{conv}\left(\mathbf{I}_{LR}\right), \tag{1}$$

where f_{conv} represents a 3×3 convolution operation, and $\mathbf{F_0}$ is used as input to DABGs (Distortion Attention Block Groups). Furthermore, each DABG contains several DAB modules. We adopted a residual-in-residual (RIR) structure as the backbone of our network. In this way, we aim to superimpose low-frequency and high-frequency feature information, producing an SR image with more texture details. The i-th DAB group is represented as follows:

$$\mathbf{F_i} = \mathbf{F_{i-1}} \oplus W_{SSC} * f_{DAB}^{m}\left(\dots f_{DAB}^{1}\left(\mathbf{F}_{i-1}, M_d\right)\dots, M_d\right), \tag{2}$$

The operation $\oplus$ can be expressed element-wise, where $*$ represents dot multiplication, and W_{SSC} is the weight associated with the DAB group output. f_{DAB}^{k} represents the k-th distortion attention block operation. Distortion information M_d contains all ODI distortion information $D_{x,y}$. The mathematical representation of the ERP distortion information is as follows:

$$D_{x,y} = \cos\left(\left(y - \frac{H}{2} + \frac{1}{2}\right)\frac{\pi}{H}\right), \tag{3}$$

H represents the height of the ODI, and we denote $1-D_{x,y}$ as the distortion degree of pixel (x, y). Note that the higher the $D_{x,y}$ value, the less the distortion. Eq. (3) is derived from the JVET standard D0040 [Sun *et al.*, 2016], which defines evaluation metrics for omnidirectional images. After the data stream passes through the DABGs, the deep information of the ODI is extracted and provided to the subsequent module for upsampling and reconstruction.

$$\mathbf{I}_{SR} = f_{REC}\left(f_{UP}\left(\mathbf{F}0 + W_{LSC} * \mathbf{F}_n\right)\right), \tag{4}$$

Eq. (4) shows the upsampling and reconstruction part. Similar to Eq. (2), W_{LSC} represents the weight associated with the deep feature $\mathbf{F}_n$. Eq. (4) indicates how, after fusion, the features are sent to the distortion offset upsampling layer f_{UP}. Finally, through the reconstruction process f_{REC}, the upsampled features are converted to SR ODI, which can be upscaled by $s\times$ (s is an integer, e.g., $2, 3, \dots$).

3.2 Enhanced Distortion Attention Block (DAB) Methodology

The main challenge of feature extraction is the distortion caused by ERP. Previous CNN-based feature extraction methods treat the features of all LR regions equally and cannot sense the distortions in a targeted way. As shown in Fig. 3, we designed a Distortion Attention Block (DAB) mixed with general convolution layers and a distortion attention layer to learn features in a "total-local" manner. The DA layer is responsible for piecewise learning according to the distortion degree. Meanwhile, the general convolutional layer appears alternately to supplement the global association information.

Distortion Attention layer. We attempt to incorporate the global distortion information into the convolution process. As shown in Fig. 3, the input features and distortion degree & scale information are fed into the Distortion Mask (DM) module. DM cuts the feature maps into k segments (Section 4 provides a more comprehensive analysis and guidance on the choice of this parameter). The calculation of the segmentation position p follows the equal distortion principle, which means that the difference between the maximum and

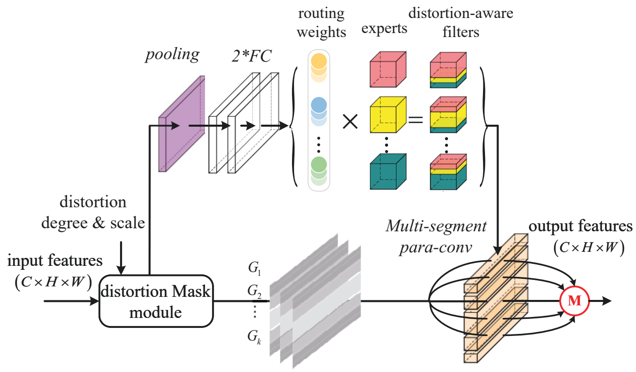

Figure 3: An illustration of our distortion attention layer.

the minimum distortion degree of each segment are equal. After segmenting transversely $k-1$ times, we represent the set of segments as $\{G_1, G_2, \ldots, G_k\}$. As depicted in Fig. 3, this set of segments is introduced to a network to generate a two-dimensional routing weights matrix. The learning process of the routing weights matrix $W \in \mathbb{R}^{k\times d}$ from our network is defined as:

$$W = \sigma(\mathrm{BN}(\mathrm{ReLU}(\mathrm{FC2}(\mathrm{BN}(\mathrm{ReLU}(\mathrm{FC1}(\mathrm{Pool}(G))))))) \tag{5}$$

where σ can be a Softmax function, BN signifies batch normalization, ReLU represents the Rectified Linear Unit activation function, and Pool indicates pooling. For each segment G_i, we extract a weight vector w_i from the routing matrix W. Each element w_{ij} in w_i quantifies the influence of the j-th expert (a conditionally applied convolutional kernel) on the i-th segment. To normalize weights, we employ a softmax operation over each expert pertaining to segment G_i:

$$\alpha_{ij} = \frac{\exp(w_{ij})}{\sum_{j=1}^{d}\exp(w_{ij})} \tag{6}$$

Given this weight normalization, we proceed to combine these weights with the routing matrix W through a Hadamard product. This results in a tailored distortion-aware filter P_i. Through the procedures delineated above, we've refined traditional convolution layers. A standout feature of the DA Layer is its adaptability: it can adjust its convolutional filters based on the nuances of the incoming feature maps. Importantly, it ensures efficiency, circumventing the computational overhead associated with training a larger number of filters thus marrying efficiency with model precision.

A comparative analysis was conducted against the feature extraction attention block incorporated in RCAN [Zhang *et al.*, 2018]. As shown in Fig. 5, with the deepening of the network, the traditional feature extraction block has poor performance for ODIs affected by distortion, and the features in the severely distorted regions appear to fade (Feature vanishing in the polar region.). We found that traditional feature extraction blocks cannot find a balance between regions affected by different distortion degrees. In contrast, our DABs can still efficiently extract features from different regions affected by distortion at the same time in deep networks.

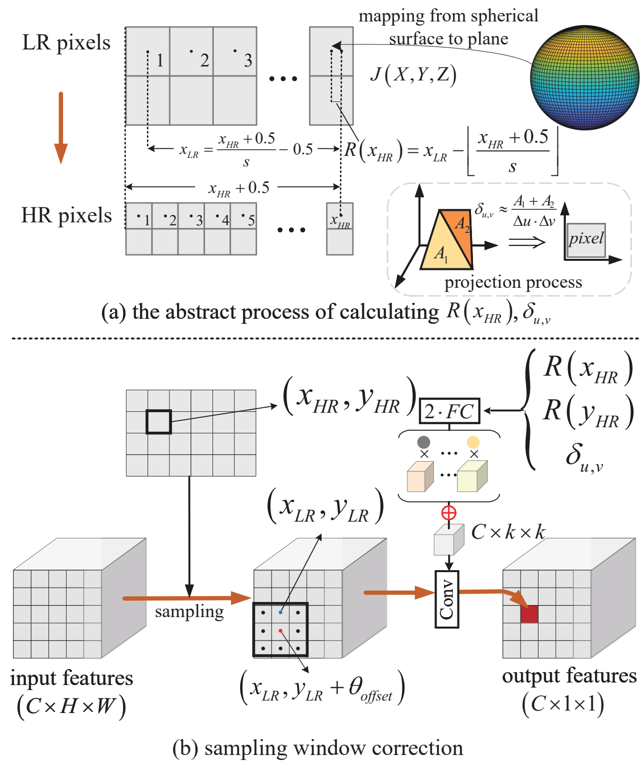

Figure 4: An illustration of our distortion offset upsampling layer.

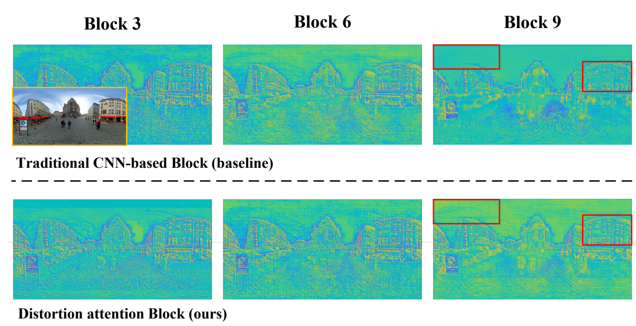

Figure 5: The feature maps output of block 3, 6 and 9. Comparison of Distortion Attention Block and baseline. (Zoom-in for best view)

3.3 Distortion Offset Upsampling (DOU) Layer

The upsampling layer is a pivotal component, leveraging deep information to reconstruct larger-sized feature maps. For ODIs, the increased sampling density at the sphere's poles engenders information redundancy in these polar regions. We conceptualize the SR process for ODIs as a three-tiered progression: from the sphere to a low-definition plane, and to a high-definition plane, as illustrated in Fig. 4(a). Our analysis seeks to determine the tile-to-pixel area ratio before and after projection, termed as the zoom factor δ. This metric is instrumental in gauging the influence of planar pixels on spherical visual perceptions. With ERP mapping a point (X, Y, Z) on

Scale	×4 (ODI-SR & SUN 360)				×8 (ODI-SR & SUN 360)			
	ERP Format		Viewport Format		ERP Format		Viewport Format	
Method	WS-PSNR	WS-SSIM	PSNR	SSIM	WS-PSNR	WS-SSIM	PSNR	SSIM
Bicubic	27.64	0.6065	23.30	0.5762	21.34	0.5241	19.09	0.5198
SRCNN[Dong *et al.*, 2014]	28.08	0.6148	23.39	0.5828	21.38	0.5317	19.21	0.5218
MemNet[Tai *et al.*, 2017]	28.31	0.6442	24.41	0.6128	22.55	0.5356	20.23	0.5294
RCAN[Zhang *et al.*, 2018]	28.78	0.6581	24.89	0.6145	22.86	0.5385	21.35	0.5332
EDSR[Lim *et al.*, 2017]	28.74	0.6571	24.71	0.6112	22.84	0.5378	21.29	0.5329
360-SS[Ozcinar *et al.*, 2019]	28.19	0.6486	23.69	0.5992	22.75	0.5313	21.01	0.5255
LAU-Net+[Deng *et al.*, 2023]	29.13	0.6751	25.32	0.6302	**23.23**	**0.5449**	**21.88**	**0.5441**
OSRT[Yu *et al.*, 2023]	**29.25**	**0.6760**	**25.34**	**0.6306**	22.92	0.5384	21.68	0.5393
ODA-SRN(ours)	**29.47**	**0.6786**	**25.59**	**0.6331**	**23.52**	**0.5480**	**22.11**	**0.5472**

Table 1: Comparisons Between the SR results. The 1st and the 2nd best performances are highlighted in **red** and **blue**

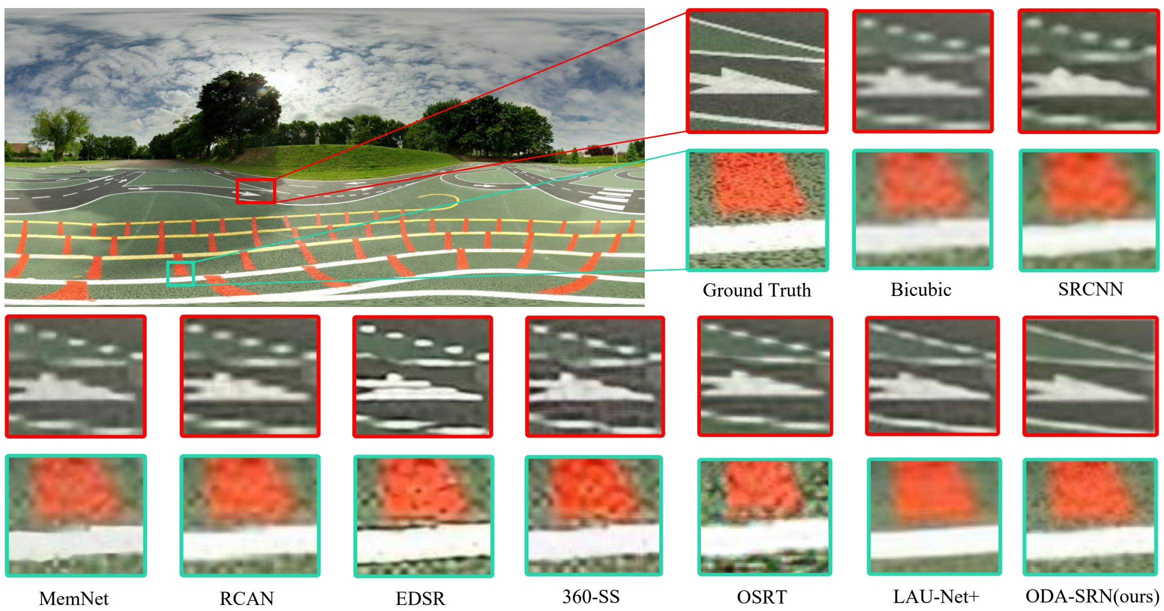

Figure 6: Visual comparison (ERP form) of $\times 4$ SR results with different methods in the central and polar regions of ODI.

the sphere to (x_{LR}, y_{LR}) on the plane, we delve into its reverse process as delineated below:

$$\begin{aligned} X &= \cos(\pi(0.5 - v)) \cdot \cos(2\pi(u - 0.5)), \\ Y &= \sin(\pi(0.5 - v)), \\ Z &= -\cos(\pi(0.5 - v)) \cdot \sin(2\pi(u - 0.5)), \end{aligned} \tag{7}$$

where (u, v) represents the normalized (x_{LR}, y_{LR}), the derivation of Eq. (5) is omitted for brevity. Based on linear algebra, we calculate the Jacobian $J(u, v)$ of the space projection using the following approach:

$$\vec{x}_{\text{sphere}} = \begin{bmatrix} \frac{\partial X}{\partial u} & \frac{\partial Y}{\partial u} & \frac{\partial Z}{\partial u} \\ \frac{\partial X}{\partial v} & \frac{\partial Y}{\partial v} & \frac{\partial Z}{\partial v} \end{bmatrix}^T \vec{x}_{LR} = J^T(u, v)\vec{x}_{LR}, \tag{8}$$

The Jacobian $J(u, v)$ can be used to obtain the mapping of any 2D vector in 3D space. We use (u, v) and $(\Delta u, \Delta v)$ to represent any pixel in the plane. Through Eq. (6), we can obtain the four vectors corresponding to the tile in 3D space. We approximate the tile area as the sum of two triangle areas, as shown in Fig. 4(a). We use the cross product of two vectors to compute the areas of A_1 and A_2. It should be noted that the zoom factor $\delta_{u,v}$ and the Jacobian are reusable, making this approach computationally efficient and cost-friendly. In the process of converting low-resolution pixels to high-resolution pixels, we calculate the relative distance between LR pixels and HR pixels. The value of the HR pixel (x_{HR}, y_{HR}) is unknown, but we can obtain the position of the LR pixel (x_{LR}, y_{LR}) corresponding to the HR pixel. Further, we can determine their correlation distances $R(x_{HR})$ and $R(y_{HR})$. As shown in Fig. 4(a), where $\lfloor\rfloor$ indicates integer rounding down. $R(y_{HR})$ are calculated in the same way. The correlation distance represents the bias relationship between HR pixels and LR pixels. Crucially, excessive reconstruction of

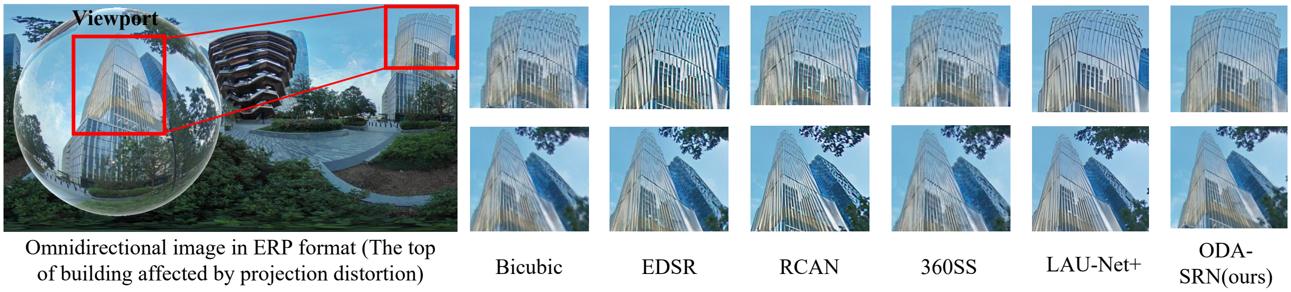

Figure 7: Comparative Visualization of ODI SR methods: ERP vs. Viewport Formats.

redundant information could affect the ODI SR process. To overcome this, we propose DOU for correcting the sampling window. Pixels near the polar regions tend to perceive information near the center of the ODIs, which helps better understand the surrounding valuable information. Specifically, we add an offset $\theta_{\text{offset}} = \gamma/\delta_{u,v}$ to the sampling window, where γ is a coefficient. The purpose of this design is to add more sampling bias in areas with smaller zoom factors. As shown in Fig. 4(b), a convolution layer is used to generate output features of $(C \times sH \times sW)$ size. $R\left(x_{HR}\right)$, $R\left(y_{HR}\right)$, and $\delta_{u,v}$ are fed into a bottleneck network structure consisting of two FC layers to obtain the weights, which are combined with the experts to generate an adaptive filter of size $(C \times k \times k)$. Finally, the output is a $(C \times 1 \times 1)$ size feature. This step is repeated $x_{HR} \times y_{HR}$ times during the upsampling of an ODI.

Loss function. For ODA-SRN, we design a loss function similar to the L_1 norm:

$$\mathcal{L}(\sigma) = \frac{1}{N}\sum_{i=1}^{N}\left|\delta\left(I_{SR}^{i} - I_{HR}^{i}\right)\right|_1, \tag{9}$$

where σ represents the weight parameters of our network. The training goal is to minimize $\mathcal{L}(\sigma)$. I_{SR} and I_{HR} stand for the generated high-resolution ODI and the original high-resolution ODI, respectively. δ is the zoom factor, applied to each pixel, which is used to guide the backpropagation under the influence of distortion.

4 Experiment

4.1 Datasets and Metrics

Dataset. ODI SR is currently underexplored, and few works focus on this new challenge. Consequently, there are only a few datasets available for evaluation. We selected 2600 high-quality ODIs from the 360 Panorama Database [Xiao *et al.*, 2012] and ODI-SR Database [Deng *et al.*, 2021]. All selected ODIs are in ERP format and have a resolution of 2048$\times$1024. To enhance the robustness of the dataset, the selected ODIs include indoor scenes, outdoor scenes, natural scenes, and cityscapes. After collecting the HR ODI dataset, we generated LR ODIs by processing HR ODIs with bicubic downsampling. The LR-HR pairs were randomly divided into 2000 for training, 300 for testing, and 300 for validation.

Method	FLOPs(G)	WS-PSNR	Params(M)	Time(s)
EDSR	2494.59	28.74	43.1	3.320
RCAN	616.59	28.78	160.0	0.982
SRCNN	52.48	28.08	0.8	0.249
MemNet	1445.43	28.31	31.1	1.208
360-SS	14.99	28.19	22.7	0.438
LAU-Net+	402.20	29.13	12.0	0.551
OSRT	3235.93	29.25	119.3	2.115
ODA-SRN	239.88	29.47	8.9	0.298

Table 2: Efficiency Comparison with Advanced Algorithms.

Metrics. To address both the plane and the sphere, we used four metrics to evaluate SR performance. In our experiments, on one hand, the widely used metrics PSNR and SSIM were employed to evaluate the performance of solutions in viewport format. On the other hand, the Weighted to Spherically Uniform PSNR (WS-PSNR) and Weighted to Spherically Uniform SSIM (WS-SSIM) [Sun *et al.*, 2016] were used as metrics specifically designed to evaluate the performance of methods in ERP format.

4.2 Implementation Details

For the model setup, the final version of ODA-SRN consists of 5 DABGs, with each DABG containing 2 DABs. The number of experts in the DA layer is set to 3, Coefficent γ is set to 0.1. Our model was trained using the ADAM optimizer [Kingma and Ba, 2015] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$. The learning rate was set to 10^{-4}, decaying by half every 50 epochs. The weight initialization of the model follows [He *et al.*, 2015]. We used PyTorch [Paszke *et al.*, 2017] to implement our model. During training, random crops were not used to preprocess ODIs in order to retain complete distortion information.

4.3 Comparison with State-of-the-Art Methods

In our experiments, we compared our ODA-SRN with other SOTA methods. To ensure consistent conditions across various models, each was deployed within its dedicated Docker container. This methodology was adopted to mitigate the inconsistencies stemming from divergent code implementations.

Feature/Result	Settings							
Loss function with δ	×	✓	×	✓	×	✓	×	✓
Distortion Attention Layer	×	×	✓	✓	×	×	✓	✓
Distortion-Aware Unsampling	×	×	×	×	✓	✓	✓	✓
WS-PSNR	27.83	27.95	28.09	28.63	28.01	28.35	28.94	29.47
WS-SSIM	0.6131	0.6161	0.6242	0.6474	0.6395	0.6596	0.6628	0.6786

Table 3: Ablation study. Result achieved by our ODA-SRN with different settings on validation set

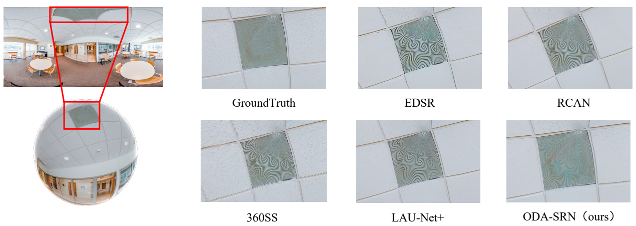

Figure 8: Comparing the performance of different methods in dealing with distortion caused by projection.

Segments Num.	3	4	5	6	7
WS-PSNR	28.38	29.14	29.47	29.30	29.23
WS-SSIM	0.6291	0.6593	0.6786	0.6752	0.6739
FLOPs (G)	201.4	220.7	239.9	259.1	278.7

Table 4: Influence of segment number on ODI SR ($\times 4$)

Quantitative Analysis. Table 1 shows the ODA-SRN consistently outperforms its contemporaries in a wide spectrum of evaluations. This overarching superiority of our proposed method becomes particularly pronounced when diving deeper into the results of the viewport format, a pivotal metric that simulates what a user would witness through their Head-Mounted Display (HMD). Such an unequivocal lead testifies to the meticulous engineering of the ODA-SRN, tailored to replicate and enhance the user's visual experience. An underpinning idea of our model is the segmental processing of the ODI. By partitioning the ODI regionally, each segment is processed through individualized Distortion-aware filter. This strategy addresses a perennial challenge: mediating the inherent tug-of-war between varying distortion regions during the feature extraction phase. By localizing the processing, the ODA-SRN can give specialized attention to each region's idiosyncratic distortions. The results, as illuminated in the Table 1, attest to the efficacy of this approach.

Qualitative Analysis. Fig. 6 visually presents the results of the $\times 4$ super-resolution performances by various algorithms. Distinct regions from the image, central and polar, are highlighted using red and green boxes for clearer examination. On the other hand, methods such as EDSR, 360-SS, and OSRT, despite their sharp outputs, occasionally exhibit discrepancies in their outline representations. On the left of Fig. 7, the transformation of an image before and after ERP projection is showcased. Through VR equipment, users experience a spherical field of view. However, when this image undergoes ERP projection onto a two-dimensional plane, perceptual distortions arise. For instance, the skyscrapers that span across both the central and polar regions of the image exhibit window outlines that, while straight in the original three-dimensional context, become curved upon projection.

The right portion of Fig. 7 offers a comparative analysis of the results generated by various algorithms. Traditional 2D-oriented methods, like EDSR and RCAN, tend to restore features by treating the ERP-projected image as a regular 2D grid. This method, however, fails to accommodate the curvatures introduced by the ERP projection. As a result, when their outputs are translated back to a spherical perspective, the previously straight window contours appear misshapen and curved. While methods explicitly designed for ODIs, such as 360SS and LAU-Net+, make commendable efforts to tackle the anomalies caused by ERP projection, they still face challenges, particularly in extreme polar regions. These areas, depicted by the side of the blue building and the apex of the white structure in Fig. 7, are marked by results that are overly smoothed and lacking in contour fidelity. This limitation stems from the inherent sparsity of feature information in these polar regions, making the precise restoration of details a formidable task. ODA-SRN is designed to address the scarcity of features in the polar regions, the distortion-offset upsampling layer of ODA-SRN adopts a unique approach. By leveraging a receptive field offset mechanism, it pulls richer feature details from more central areas. This allows ODA-SRN to supplement the sparse features in the extremities and bring a marked improvement in restoring details. The result is a more accurate representation of even subtle elements, such as the high-rise window outlines, when they are projected back to the spherical viewpoint.

Anti-distortion Analysis. ERP projection significantly stretches polar regions in images, where inaccuracies can magnify during spherical conversion. We illustrate this with a comparison against various SR algorithms. Fig.8 shows that vents in ERP images are notably stretched. Our method effectively reduces this distortion, with ODA-SRN outperforming other tested methods in anti-distortion capability.

Model Efficiency Analysis. We have included Table 2. This table provides a comprehensive performance comparison of our method with other algorithms, covering various aspects such as FLOPs, WS-PSNR, network parameters, and runtime. By presenting this multidimensional comparison, we aim to offer a thorough evaluation of the effectiveness and efficiency of our approach compared to existing methods. ODA-SRN strikes a balance between performance and efficiency, delivering the highest WS-PSNR using fewer parameters, reduced FLOPs, and enhanced speed.

4.4 Ablation Study

Ablation Study on Weighted Loss Function. This study examines the zoom factor δ in loss calculation, which reflects a pixel's projection scale from planar to spherical views. Table 3 demonstrates the superiority of the δ-weighted loss function over conventional $L1$ loss in SR.

Ablation Study on Distortion Attention Layer (DA). In the absence of the DA layer. Considering the unique patterns in ODIs in ERP format, Table 3 shows that integrating the DA layer significantly enhances WS-PSNR and WS-SSIM. Our DA layer, focusing on specific ODI features, improves structural similarity in SR outcomes. This is due to ODA-SRN's adaptive feature extraction for different ODI regions.

Ablation Study on Distortion Offset Upsampling (DOU). Data in Table 3 emphasizes the effectiveness of DOU in ODI super-resolution. Standard upsampling doesn't adequately address ODI specifics, leading to the creation of the DOU layer. This layer computes θ_{offset} to amend inherent distortions, allowing polar-proximate pixels to utilize central ODI information and enhance output quality. Compared to the pixel shuffle baseline, DOU consistently achieves superior WS-PSNR and WS-SSIM results.

Effect of Segments Number. We propose DAB for extracting features of ODI. Unlike typical CNNs, DAB adopts a segmented strategy to handle regions with varying degrees of distortion. We attempted to change the number of segments in DAB to observe the impact of different segment numbers on network performance. Table 4 shows the experimental results. With the increase in the number of segments, the performance of the network improves. However, the improvement in FLOPs is linear, which means that more segments do not always yield better results. We need to choose an appropriate number of segments to achieve a balance between performance and cost. When the number of segments increases to more than 5, the model becomes difficult to converge and performance decreases. Therefore, considering multiple factors, we finally chose 5 as the number of segments in ODA-SRN.

5 Conclusion

We addressed the SR challenges of ODIs. Our solution was not rooted in proposing a completely new network. Rather, our main contribution centers on the innovative redesign of two plug-and-play modules: the Distortion Attention Block and the Distortion Offset Upsampling Layer. These modules, crafted to counter the unique challenges posed by ODIs, not only enhance super-resolution performance but also offer the flexibility of being integrated into various existing network architectures. Comprehensive experiments highlight that our modular approach excels in metrics such as PSNR, SSIM, WS-PSNR, and WS-SSIM, and also stands out in terms of model efficiency.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under grant No. 62225105, 62394323 and 62301070. Han Xiao wishes to acknowledge support from the Postdoctoral Science Foundation of China under grant No. 2022M720518. G.-M. Muntean wishes to acknowledge the Science Foundation Ireland's support via grant nos. 21/FFP-P/10244 (FRADIS) and 12/RC/2289_P2 (Insight).

References

[Anwar *et al.*, 2021] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM Comput. Surv.*, 53(3):34, 2021.

[Brandon Yang, 2020] Quoc V. Le Jiquan Ngiam Brandon Yang, Gabriel Bender. Condconv: Conditionally parameterized convolutions for efficient inference. *arxiv*, 2020.

[Chen *et al.*, 2021] Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12021–12030, 2021.

[Chen *et al.*, 2022] Guang Chen, Haitao Wang, Kai Chen, Zhijun Li, Zida Song, Yinlong Liu, Wenkai Chen, and Alois Knoll. A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(2):936–953, 2022.

[Dai *et al.*, 2019] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019.

[Deng *et al.*, 2021] Xin Deng, Hao Wang, Mai Xu, Yichen Guo, Yuhang Song, and Li Yang. Lau-net: Latitude adaptive upscaling network for omnidirectional image super-resolution. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9194, 2021.

[Deng *et al.*, 2023] Xin Deng, Hao Wang, Mai Xu, Li Li, and Zulin Wang. Omnidirectional image super-resolution via latitude adaptive network. *IEEE Transactions on Multimedia*, pages 4108–4120, 2023.

[Dong *et al.*, 2014] Chao Dong, Change Loy Chen, He Kaiming, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. *ECCV. Springer International Publishing, 2014*, pages 184–199, 2014.

[He *et al.*, 2015] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *2015 International Conference on Learning Representations (ICLR)*, 2015.

[Li *et al.*, 2020] Shiyuan Li, Chunyu Lin, Kang Liao, Yao Zhao, and Xue Zhang. Panoramic image quality-enhancement by fusing neural textures of the adaptive initial viewport. *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 816–817, 2020.

[Li *et al.*, 2022] Wenbo Li, Kun Zhou, Lu Qi, Liying Lu, and Jiangbo Lu. Best-buddy gans for highly detailed image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1412–1420, 2022.

[Lim *et al.*, 2017] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017.

[Liu *et al.*, 2020] Jie Liu, Wenjie Zhang, Yuting Tang, Jie Tang, and Gangshan Wu. Residual feature aggregation network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2359–2368, 2020.

[Ma *et al.*, 2020] Dizhi Ma, Andrew Lumsdaine, and Wenhui Zhou. Flexible spatial and angular light field super resolution. *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2970–2974, 2020.

[Mei *et al.*, 2020] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5690–5699, 2020.

[Ozcinar *et al.*, 2019] Cagri Ozcinar, Aakanksha Rana, and Aljosa Smolic. Super-resolution of omnidirectional images using adversarial learning. *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2019.

[Paszke *et al.*, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *Workshop Autodiff Submission in Conference on Neural Information Processing Systems (NIPS-W)*, 2017.

[Senmao *et al.*, 2023] Tian Senmao, Lu Ming, Liu Jiaming, Guo Yandong, Chen Yurong, and Zhang Shunli. Cabm: Content-aware bit mapping for single image super-resolution network with large input. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[Shen *et al.*, 2022] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360 depth estimation. *ECCV*, pages 23–27, 2022.

[Su and Grauman, 2017] Yu-Chuan Su and Kristen Grauman. Making 360 video watchable in 2d: Learning videography for click free viewing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1368–1376. IEEE, 2017.

[Su and Kristen, 2017] Yuchuan Su and Grauman. Kristen. Learning spherical convolution for fast features from 360 imagery. *NIPS*, 2017.

[Sun *et al.*, 2016] Yule Sun, Ang Lu, and Lu Yu. Ahg8: Ws-psnr for 360 video objective quality evaluation. *Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC*, JTC1/SC29/WG11(JVET-D0040), 2016.

[Sun *et al.*, 2021] Yuqi Sun, Ri Cheng, Bo Yan, and Shili Zhou. Space-angle super-resolution for multi-view images. *Proceedings of the 29th ACM International Conference on Multimedia*, page 750–759, 2021.

[Tai *et al.*, 2017] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4549–4557, 2017.

[Wang *et al.*, 2021] Hu Wang, Peng Chen, Bohan Zhuang, and Chunhua Shen. Fully quantized image super-resolution networks. *Proceedings of the 29th ACM International Conference on Multimedia*, page 639–647, 2021.

[Xiao *et al.*, 2012] Jianxiong Xiao, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702, 2012.

[Yoon *et al.*, 2022] Youngho. Yoon, Inchul. Chung, Lin Wang, and Kuk-Jin Yoon. Spheresr: 360 image super-resolution with arbitrary projection via continuous spherical image representation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5667–5676, 2022.

[Yu *et al.*, 2023] Fanghua Yu, Xintao Wang, Mingdeng Cao, Gen Li, Ying Shan, and Chao Dong. Osrt: Omnidirectional image super-resolution with distortion-aware transformer. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[Zhang *et al.*, 2018] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. *ECCV 2018: Computer Vision*, pages 294–310, 2018.

[Zhang *et al.*, 2021] Yulun Zhang, Donglai Wei, Can Qin, Huan Wang, Hanspeter Pfister, and Yun Fu. Context reasoning attention network for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4278–4287, 2021.

[Zhang *et al.*, 2023] Yunjian Zhang, Yanwei Liu, Jinxia Liu, Antonios Argyriou, Liming Wang, Zhen Xu, and Xiangyang Ji. Social vr platform: Building 360-degree shared vr spaces. *IEEE Transactions on Image Processing*, 32:2552–2567, 2023.