Omnidirectional Image Super-resolution via Bi-projection Fusion

Jiangang Wang[1], Yuning Cui[2], Yawen Li[3], Wenqi Ren[1*], Xiaochun Cao[1]

[1]Shenzhen Campus of Sun Yat-sen University
[2]Technical University of Munich
[3]Beijing University of Posts and Telecommunications
wangjg33@mail2.sysu.edu.cn, yuning.cui@in.tum.de
warmly0716@bupt.edu.cn, {renwq3, caoxiaochun}@mail.sysu.edu.cn

Abstract

With the rapid development of virtual reality, omnidirectional images (ODIs) have attracted much attention from both the industrial community and academia. However, due to storage and transmission limitations, the resolution of current ODIs is often insufficient to provide an immersive virtual reality experience. Previous approaches address this issue using conventional 2D super-resolution techniques on equirectangular projection without exploiting the unique geometric properties of ODIs. In particular, the equirectangular projection (ERP) provides a complete field-of-view but introduces significant distortion, while the cubemap projection (CMP) can reduce distortion yet has a limited field-of-view. In this paper, we present a novel Bi-Projection Omnidirectional Image Super-Resolution (BPOSR) network to take advantage of the geometric properties of the above two projections. Then, we design two tailored attention methods for these projections: Horizontal Striped Transformer Block (HSTB) for ERP and Perspective Shift Transformer Block (PSTB) for CMP. Furthermore, we propose a fusion module to make these projections complement each other. Extensive experiments demonstrate that BPOSR achieves state-of-the-art performance on omnidirectional image super-resolution. The code is available at https://github.com/W-JG/BPOSR.

Introduction

In recent years, omnidirectional images (ODIs), also known as 360° images or panoramic images, have gained significant attention due to their unique immersive experience. When viewed through headsets, ODIs provide a limited field-of-view through a small viewport (Elbamby et al. 2018). To accurately capture real-world details within this restricted viewport, ODIs require high resolutions ranging from 8K to 16K (Ai et al. 2022). Nonetheless, most existing ODIs have inadequate resolution due to limitations in acquisition, storage, and transmission.

As a typical low-level vision problem, super-resolution aims to generate high-resolution images with essential edge structures and texture details from low-resolution counterparts (Glasner, Bagon, and Irani 2009). Although conventional 2D image super-resolution methods have made remarkable advancements (Dong et al. 2014; Kim, Lee, and Lee 2016; Zhang et al. 2018; Chen et al. 2021; Liang et al. 2021; Zhang et al. 2022), directly applying these 2D methods to ODIs super-resolution is infeasible and suboptimal. This is due to the distortions and discontinuities that arise from projecting a spherical panoramic image onto a 2D plane (Deng et al. 2021). The different properties between image domains increase the complexity of ODIs reconstruction. Therefore, developing novel super-resolution methods that consider the unique geometric properties of ODIs is beneficial for high-quality omnidirectional image super-resolution.

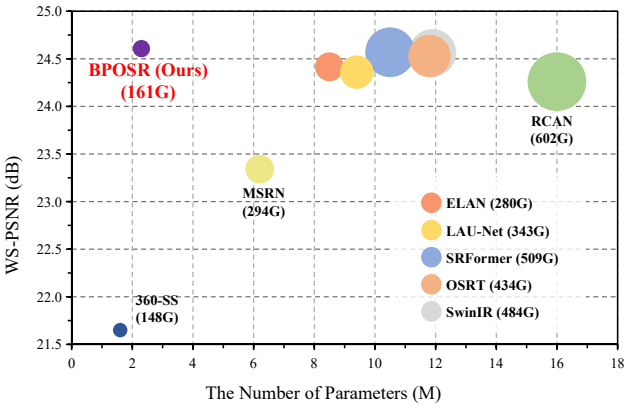


Figure 1: WS-PSNR *vs.* the number of parameters. The comparison is conducted on the ODI-SR test set with the $\times 8$ upscaling factor. BPOSR achieves a better trade-off than other algorithms.

Several studies have attempted to address the task of omnidirectional image super-resolution (ODISR), including LAU-Net (Deng et al. 2021), 360-SS (Ozcinar, Rana, and Smolic 2019), SphereSR (Yoon et al. 2022) and OSRT (Yu et al. 2023). However, these studies mainly focus on solving this task within the ERP domain without considering the various projection formats used in ODIs. The two most commonly used ODIs projection formats are equirectangular projection (ERP) and cubemap projection (CMP). Specifically, the ERP provides a wide global view but introduces significant distortion, while the CMP has less distortion but only provides a limited central view with discontinuous boundaries (Ai et al. 2022). Inspired by this fact,

*Corresponding author.

we aim to fully exploit the geometric properties and complementary information of these two projections to enhance the performance of ODISR. To achieve this, we develop the Bi-Projection Omnidirectional Image Super-Resolution (BPOSR) network, which enables the simultaneous information flow of ERP and CMP branches, and allows for the interaction and fusion of diverse projection features.

Furthermore, we conduct a comprehensive investigation into the geometric properties of ERP and CMP to better take advantage of different projections. As illustrated in Figure 2 (a), we observe a unique property of ERP, namely horizontal similarity, where objects at the same height in the real world exhibit similar appearances and features, creating horizontal similarity regions in the ERP. Moreover, as shown in Figure 2 (b), we also discover a characteristic of CMP, dubbed perspective variability, where we obtain diverse information under different perspectives when projecting and mapping the rotated spherical panoramic image. Based on these observations, we introduce the Horizontal Striped Transformer Block (HSTB) for ERP and the Perspective Shift Transformer Block (PSTB) for CMP to sufficiently leverage the intrinsic properties of different projections. Finally, we develop a block attention fusion module to facilitate information interactions between features from diverse projections and depths by assigning varying attention weights to them. As a result, the representation learning capability of the network is enhanced. Equipped with the above designs, the proposed BPOSR achieves state-of-the-art performance with fewer parameters, as shown in Figure 1.

The main contributions of our work are summarized as follows:

- We propose a Bi-Projection Omnidirectional Image Super-Resolution (BPOSR) network that takes advantage of both two omnidirectional projections, *i.e.,* ERP and CMP, to facilitate the interaction of information from both projections.
- By analyzing the image geometric properties of ERP and CMP, we introduce the Horizontal Striped Transformer Block (HSTB) and the Perspective Shift Transformer Block (PSTB) to utilize the inherent properties of both projections.
- We introduce a Block Attention Fusion Module (BAFM) to facilitate the fusion between features from different projections and depths. Extensive experiments demonstrate that the proposed network achieves state-of-the-art performance for omnidirectional image super-resolution.

Related Work

Single Image Super-Resolution

With the rapid development of deep learning, convolutional neural networks (CNNs) have dominated Single Image Super-Resolution (SISR) for many years. Since SRCNN (Dong et al. 2014) first introduced CNN to SR, a large number of CNN-based SR models have emerged. For instance, VDSR (Kim, Lee, and Lee 2016) adopts a deeper CNN-based architecture with residual learning to improve

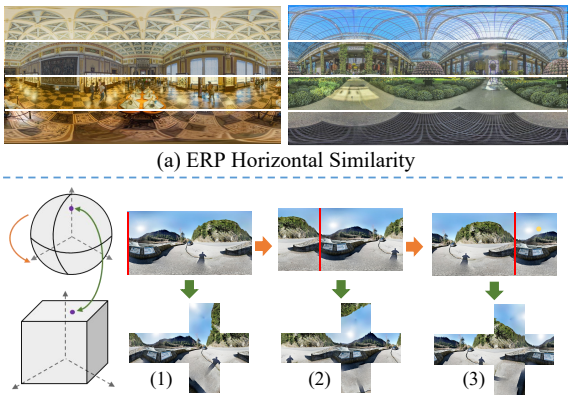


Figure 2: (a) ERP Horizontal Similarity. Upon dividing the ERP into regions along the horizontal direction, multi-scale similarities are observed within each region. (b) CMP Perspective Variability. Orange arrows represent spherical rotation, and green arrows represent projections onto CMP. By spherically rotating and projecting onto the CMP, the six surfaces of the CMP capture different information.

SR performance. RCAN (Zhang et al. 2018) utilizes a channel attention mechanism to adaptively modulate channels. ShuffleMixer (Sun, Pan, and Tang 2022) explores the large convolution and channel split-shuffle operation for SR. Recently, inspired by the success of ViT (Dosovitskiy et al. 2021) in high-level vision tasks, IPT (Chen et al. 2021) introduces Transformer into SISR, but it requires a large number of parameters. SwinIR (Liang et al. 2021) applies the Swin Transformer (Liu et al. 2021) framework to SR and achieves extremely powerful performance. ELAN (Zhang et al. 2022) simplifies the architecture of SwinIR (Liang et al. 2021) and performs self-attention among different window sizes to collect correlations between distant pixels. Despite the promising performance on 2D SR, these algorithms are inapplicable to ODISR.

Omnidirectional Image Super-Resolution

Several studies have explored the potential of deep learning for ODISR by fine-tuning traditional 2D planar image SR models. 360-SS (Ozcinar, Rana, and Smolic 2019) introduces a spherical loss function in the traditional 2D SR model, which is weighted according to the spherical geometric position of each pixel. Nishiyama *et al.* (Nishiyama, Ikehata, and Aizawa 2021) utilize 2D SR models to address ODISR by adding distortion maps as input to handle different distortions. LAU-Net (Deng et al. 2021) presents a latitude adaptive upscaling network towards the nonuniformaly distributed pixel density of ERP ODI. SphereSR (Yoon et al. 2022) utilizes icosahedral spherical data to extract features and uses a spherical local implicit image function to generate HR. Furthermore, OSRT (Yu et al. 2023) introduces deformable convolutions to learn the distortion of ERP. However, the above mentioned approaches mainly address ODISR using ERP, which introduces significant distortion. In this paper, we perform high-quality reconstruc-

tion by taking advantage of both ERP and CMP.

ODIs Analysis

For transmission convenience, the spherical panoramic projection is often transformed onto a 2D plane. In this part, we introduce the two widely used projections, *i.e.,* equirectangular projection (ERP) and cubemap projection (CMP), as well as our observations, based on which we establish our network.

Equirectangular Projection

ERP uniformly samples the sphere with longitude and latitude. Assuming the longitude and latitude are ϕ and θ, respectively, we have $(\phi, \theta) \in [-\pi, \pi] \times \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ (Wang et al. 2023). The angular position (ϕ, θ) can be converted to a coordinate $Q_s = (q_s^x, q_s^y, q_s^z)$ on a standard sphere by:

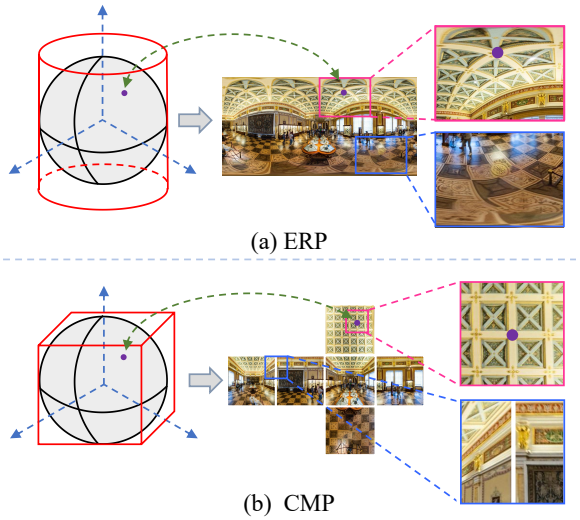
$$\begin{aligned} q_s^x &= \sin(\phi)\cos(\theta), \\ q_s^y &= \sin(\theta), \\ q_s^z &= \cos(\phi)\cos(\theta). \end{aligned} \tag{1}$$

As shown in Figure 3 (a), ERP projects a sphere onto a single surface, thus obtaining a wide field of view. However, due to the uniform spacing and parallel characteristics of latitude lines across the projection, the ERP introduces significant distortions, particularly near the poles. As the latitude lines converge towards the poles, the distortion becomes more pronounced, resulting in elongated shapes and stretching of the image.

ERP Horizontal Similarity. Through our observations, we investigate the inherent property of horizontal similarity within ERP. In the real world, objects at the same height exhibit similar appearance and characteristics due to their relative positions. ERP can capture comprehensive positional information by providing a full 360° view of the real world environment. Consequently, the relative positional relationship of objects in the real world is stored in ERP. As shown in Figure 2 (a), multi-scale similarities are prevalent in the horizontal regions of the ERP image. Therefore, the conventional global-scale isotropic attention mechanism becomes redundant for processing ERP image features. Instead, we propose a more suitable approach for ERP, which involves utilizing the horizontal window to model intra-image dependencies. Furthermore, by combining local perception and contextual information within these horizontal windows, we can introduce a limited spatial range to reduce the complexity of attention. It turns out that this approach is highly beneficial for ERP to enhance the capture of localized structures and complex image features.

Cubemap Projection

CMP projects a sphere onto the six surfaces of a cube. The resulting six surfaces are specific perspective images, corresponding to viewing directions: front, back, left, right, up, and down. The size of each surface is $r \times r$ and the focal length is $\frac{r}{2}$, where r is the radius of the source sphere. The front surface keeps the same coordinate system as the sphere, while the others are obtained by rotating the sphere



(a) ERP

(b) CMP

Figure 3: (a) ERP projects a sphere onto a single surface, resulting in a wide field of view but with distortion at high and low latitudes. (b) CMP projects a sphere onto a cube with six surfaces, which reduces distortion but results in discontinuities between the individual surfaces.

90° or 180° around a specific axis (Wang et al. 2023). Specifically, R_i denotes the rotation matrix that transforms from the coordinate system of the i-th surface to the spherical coordinate system. Then we can project the pixel $P_c = (p_c^x, p_c^y, p_c^z)$ as

$$Q_s = s \cdot R_i \cdot P_c, \tag{2}$$

where $p_c^x, p_c^y \in [0, r]$, $p_c^z = \frac{r}{2}$, and the factor $s = \frac{r}{|p_c|}$.

As shown in Figure 3 (b), compared with ERP, CMP exhibits a substantial reduction in image distortion. However, it introduces the discontinuity issue by disrupting the continuity of objects at the boundaries between different faces.

CMP Perspective Variability. The CMP projects the sphere onto six planes, each of which can obtain information about the sphere from different perspectives. As shown in Figure 2 (b), when the sphere is rotated and projected onto the CMP, the viewing angles of the six planes undergo changes. Based on this observation, we propose the perspective variability of CMP. The addition of new perspectives results in an augmented availability of information. By shifting perspectives on CMP, we effectively enhance the the feature representation of CMP and address the inherent limitations of image discontinuities in CMP.

Methodology

Overall Architecture

The overall architecture of the proposed network is illustrated in Figure 4, which mainly consists of three branches: ERP Branch, CMP Branch, and Fusion Branch.

Given a low-resolution input I_{ERP}^{lr}, we firstly transform it into the CMP form I_{CMP}^{lr}, and then use 3×3 convolutions

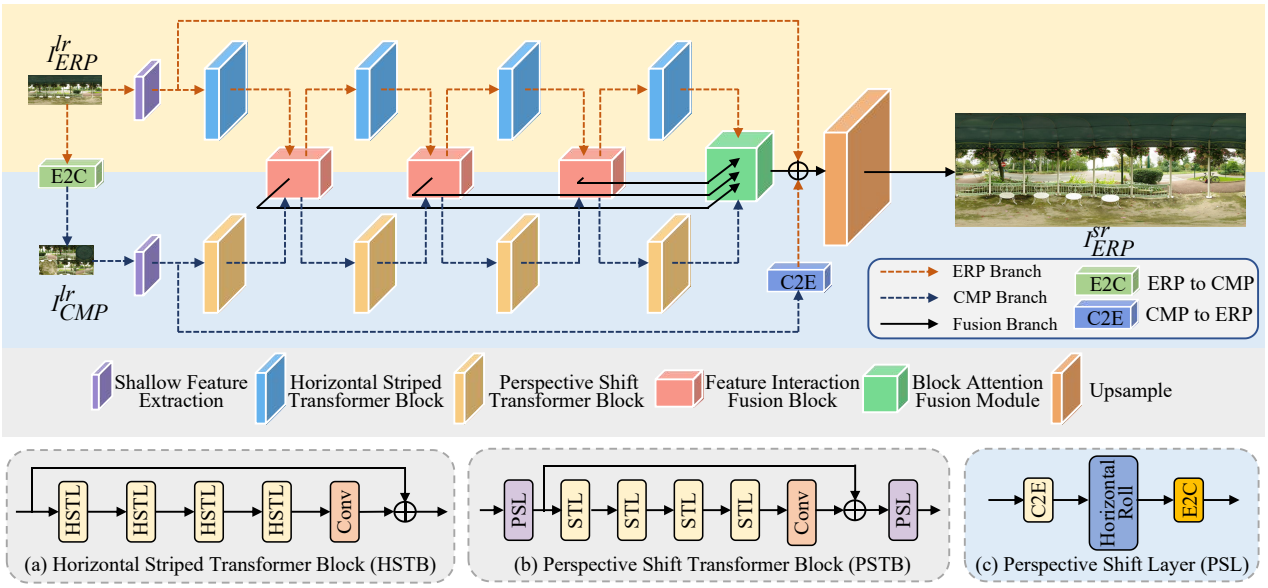


Figure 4: The overall diagram illustrates the architecture of BPOSR, which comprises three branches: ERP Branch, CMP Branch, and Fusion branch. In ERP Branch, HSTB employs horizontal striped self-attention to model the features of ERP. In CMP Branch, PSTB utilizes the PSL to obtain additional perspectives, enabling enhanced CMP features. In Fusion Branch, BAFM fuses features from diverse projections and depths.

to separately extract shallow features for two projections as:

$$I^{lr}_{CMP} = \mathrm{E2C}(I^{lr}_{ERP}), \tag{3}$$

$$F^{0}_{ERP} = W^{1}_{3\times3}(I^{lr}_{ERP}), \tag{4}$$

$$F^{0}_{CMP} = W^{2}_{3\times3}(I^{lr}_{CMP}), \tag{5}$$

where $\mathrm{E2C}(\cdot)$ represents the projection from ERP to CMP, and $W_{3\times3}$ denotes a 3×3 convolution. Next, we extract the deep features of ERP and CMP branches as:

$$F^{i}_{ERP} = \mathrm{HSAB_i}(F^{i-1}_{ERP}), \tag{6}$$

$$F^{i}_{CMP} = \mathrm{PSAB_i}(F^{i-1}_{CMP}), \tag{7}$$

where $i \in [1, K]$ is the index of resulting features, and $\mathrm{HSAB}(\cdot)$ and $\mathrm{PSAB}(\cdot)$ are the Horizontal Striped Transformer Block and Perspective Shift Transformer Block, respectively. To promote information interactions and feature fusion between two projections, we propose a feature interaction fusion block, which firstly generates the fused features using F^{i}_{ERP} and F^{i}_{CMP}, and then imposes resulting features on source features. This process can be formally expressed as:

$$F^{i}_{FUS} = W^{fus}_{1\times1}(cat(F^{i}_{ERP}, F^{i}_{CMP})), \tag{8}$$

$$F^{i}_{ERP} = W^{erp}_{1\times1}(cat(F^{i}_{ERP}, F^{i}_{FUS})), \tag{9}$$

$$F^{i}_{CMP} = \mathrm{E2C}(W^{cmp}_{1\times1}(cat(\mathrm{C2E}(F^{i}_{CMP}), F^{i}_{FUS}))), \tag{10}$$

where cat is the concatenate operation, and $W_{1\times1}$ denotes a 1×1 convolution.

Finally, in order to integrate the features from different branches and different depths, we develop a block attention fusion module (BAFM) to yield the final features F_f as:

$$F_f = \mathrm{BAFM}(cat(F^{1}_{FUS}, ..., F^{K-1}_{FUS}, F^{K}_{ERP}, \mathrm{C2E}(F^{K}_{CMP}))), \tag{11}$$

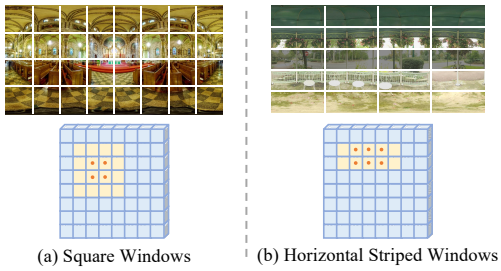


Figure 5: Different self-attention windows: (a) Square Windows (b) Horizontal Striped Windows. As can be seen, Horizontal Striped Windows are more effective in capturing the similarity within ERP compared to Square Windows.

where cat is the concatenate operation. Finally, the high-resolution is reconstructed via the upsampling module with a single 3×3 convolution and pixel shuffle operation (Shi et al. 2016) F_{up} as:

$$I^{sr}_{ERP} = F_{up}(F_f + F^{0}_{ERP} + \mathrm{C2E}(F^{0}_{CMP})). \tag{12}$$

We then delineate the core components of our network, *i.e.,* HSTB, PSTB, and BAFM.

Horizontal Striped Transformer Block (HSTB)

HSTB is designed by exploiting the horizontal similarity of ERP, which consists of numerous Horizontal Swin Transformer Layer (HSTL) and a convolutional layer, as shown in Figure 4 (a). In contrast to vanilla SwinIR square windows (Liang et al. 2021), we divide the input features into horizontal windowsand apply the shift window self-attention

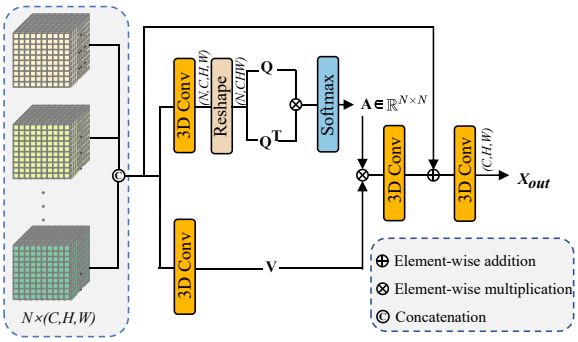


Figure 6: Block Attention Fusion Module. BAFM receives input from different projections and depths, employing a 3D self-attention mechanism to fuse all the features.

mechanism to these features. As shown in Figure 5, HSTL utilizes a self-attention mechanism within horizontal striped windows to establish long-term dependencies. By confining attention computation to horizontal windows, we enable the establishment of dependencies over a wider and more effective range, facilitating a comprehensive exploration of the contextual information within ERP.

Perspective Shift Transformer Block (PSTB)

PSTB is designed based on the perspective variability of the CMP. As shown in Figure 4 (b), PSTB consists of multiple Swin Transformer Layer (STL) (Liang et al. 2021) with shifted window self-attention and a convolutional layer. We introduce perspective shifts by deploying the Perspective Shift Layer (PSL) after the input and before the output. PSL first uses C2E to convert CMP features F_{CMP} to ERP, and then horizontally rolls the features in the ERP domain. The finally output of PSL is obtained by converting the features into CMP via E2C, which can be formally expressed as:

$$F_{CMP} = \mathrm{E2C}(\mathcal{R}(\mathrm{C2E}(F_{CMP})), \tag{13}$$

where $\mathcal{R}$ is the horizontally roll operation.

The modeling capacity of shift window self-attention modules is constrained by the absence of connections between different views. This limitation hinders their ability to fully exploit the characteristics of CMP. PSTB integrates the incorporation of interconnections among diverse perspectives, facilitating a broader and more effective range of modeling.

Block Attention Fusion Module (BAFM)

Although dense connections (Huang et al. 2017) and skip connections (He et al. 2016) facilitate the transfer of shallow information to deep layers, they do not effectively leverage the interdependencies among different blocks (Niu et al. 2020). As shown in Eq. 11, the input features to BAFM are derived from different depths and projections. To enhance the fusion effect, we develop BAFM, as illustrated in Figure 6. The core component of BAFM is a 3D self-attention mechanism, which selectively enhances feature blocks with significant contributions while suppresses redundant feature blocks. By doing this, the overall representation ability of the network is enhanced. More concretely, given any input $F_{input} \in \mathbb{R}^{N\times C\times H\times W}$, the query matrix Q and value matrix V are obtained by:

$$Q = \mathrm{3DConv}_Q(F_{input}), \tag{14}$$
$$V = \mathrm{3DConv}_V(F_{input}), \tag{15}$$

where 3DConv denotes a 3D convolution of size $1 \times 1 \times 1$. Next, the attention map is produced by the matrix multiplication between Q and $Q^\top$, followed by the Softmax function for normalization. Then, the modulated features via self-attention are yielded by:

$$F_m = \mathrm{3DConv}(\frac{\mathrm{Softmax}(Q \cdot Q^\top)}{s} \cdot V), \tag{16}$$

where s is the scaling factor. Finally, the output of BAFM is generated by compressing $F_m \in \mathbb{R}^{N\times C\times H\times W}$ via a 3D convolution layer as:

$$F_{out} = \mathrm{3DConv}(F_{input} + F_m) \in \mathbb{R}^{1\times C\times H\times W}. \tag{17}$$

Experiments

Dataset and Implementation Details

We verify the effectiveness of our method using the widely used datasets: ODI-SR (Deng et al. 2021) and SUN360 (Xiao et al. 2012), which contain various types of panoramic scenes. The model is trained using 1200 training images of ODI-SR and evaluated on the test sets of ODI-SR and SUN360, both containing 100 images.We adopted the L_1 loss function and use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The model is trained for 500k iterations with the initial learning rate as 2×10^{-4}, which is halved at 250k, 400k, 450k, and 475k iterations. In our model, K is set to 4, and the number of STL and HSTL is both set to 6. The attention window sizes of HSTB and PSTB are set as 4×16 and 8×8, respectively. The model feature dimension is set to 60, and the rotation magnification in PSTB is set to 3 times. For evaluation, we additionally employ Weighted-to-Spherical Uniform PSNR (WS-PSNR) and Weighted Spherical Uniform SSIM (WS-SSIM) (Sun, Lu, and Yu 2017) as metrics which are specially designed for ODIs quality measurement.

Comparisons with State-of-the-Art

To demonstrate the superiority of our proposed PBOSR, we compare it with 9 representative SISR methods, including SRCNN (Dong et al. 2014), VDSR (Kim, Lee, and Lee 2016), LapSRN (Ahn, Kang, and Sohn 2018a), MemNet (Tai et al. 2017), MSRN (Li et al. 2018), EDSR (Lim et al. 2017), D-DBPN (Haris, Shakhnarovich, and Ukita 2018), RCAN (Zhang et al. 2018), and DRN (Guo et al. 2020), and 4 state-of-the-art ODISR algorithms: 360-SS (Ozcinar, Rana, and Smolic 2019), LAU-Net (Deng et al. 2021), SphereSR (Yoon et al. 2022), and OSRT (Yu et al. 2023). More results can be found in the supplementary material.

Quantitative results. Table 1 presents the comparison results with state-of-the-art algorithms under $\times 4$, $\times 8$, and $\times 16$

	Dataset	ODI-SR						SUN360					
	Scale	×4		×8		×16		×4		×8		×16	
	Method	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM
SISR	Bicubic	24.62	0.6555	19.64	0.5908	17.12	0.4332	24.61	0.6459	19.72	0.5403	17.56	0.4638
	SRCNN	25.02	0.6904	20.08	0.6112	18.08	0.4501	26.30	0.7012	19.46	0.5701	17.95	0.4684
	VDSR	25.92	0.7009	21.19	0.6334	19.22	0.5903	26.36	0.7057	21.60	0.6091	18.91	0.5935
	LapSRN	25.87	0.6945	20.72	0.6214	18.45	0.5161	26.31	0.7000	20.05	0.5998	18.46	0.5068
	MemNet	25.39	0.6967	21.73	0.6284	20.03	0.6015	25.69	0.6999	21.08	0.6015	19.88	0.5759
	MSRN	25.51	0.7003	23.34	0.6496	21.73	0.6115	25.91	0.7051	23.19	0.6477	21.18	0.5996
	EDSR	25.69	0.6954	23.97	0.6483	22.24	0.6090	26.18	0.7012	23.79	0.6472	21.83	0.5974
	D-DBPN	25.50	0.6932	24.15	0.6573	22.43	0.6059	25.92	0.6987	23.70	0.6421	21.98	0.5958
	RCAN	26.23	0.6995	24.26	0.6554	22.49	0.6176	26.61	0.7065	23.88	0.6542	21.86	0.5938
	DRN	26.24	0.6996	24.32	0.6571	22.52	0.6212	26.65	0.7079	24.25	0.6602	22.11	0.6092
ODISR	360-SS	25.98	0.6973	21.65	0.6417	19.65	0.5431	26.38	0.7015	21.48	0.6352	19.62	0.5308
	LAU-Net	26.34	0.7052	24.36	0.6602	22.52	0.6284	26.48	0.7062	24.24	0.6708	22.05	0.6058
	SphereSR	—	—	24.37	0.6777	22.51	0.6370	—	—	24.17	0.6820	21.95	0.6342
	OSRT	26.89	0.7581	24.53	0.6780	22.69	0.6261	27.47	0.7985	24.38	0.7072	22.13	0.6388
	BPOSR	**26.95**	**0.7598**	**24.61**	**0.6782**	**22.72**	**0.6285**	**27.59**	**0.7997**	**24.47**	**0.7084**	**22.16**	**0.6433**

Table 1: Quantitative comparisons (WS-PSNR/WS-SSIM) with SISR and ODISR algorithms on benchmark datasets. The best results are highlighted in bold.

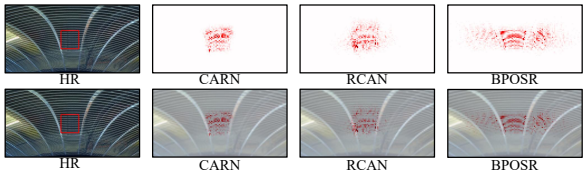


Figure 7: LAM results for different networks. The LAM attribution reflects the importance of each pixel in the input LR image when reconstructing the patch marked with a box.

Method	WS-PSNR	WS-SSIM
BPOSR	24.61	0.6782
Variant-CMP	24.30	0.6620
Variant-ERP	24.47	0.6716

Table 2: Ablation studies for Bi-Projection

upscaling factors on ODI-SR and SUN360. As seen, our model outperforms other competitors on both two datasets. Specifically, our model outperforms all SISR networks. On the ODI-SR dataset, our method achieves performance gains of 0.71 dB, 0.29 dB, and 0.2 dB WS-PSNR over the best SISR method DRN under $\times 4$, 8×8, and $\times 16$ factors, respectively. Furthermore, our model also achieves the best results compared to all ODISR models designed specifically for ODIs. On the SUN360 dataset, our method achieves a performance gain of 0.12 dB WS-PSNR over the recent algorithm OSRT (Yu et al. 2023) under $\times 4$ factor. These results provide strong evidence of the remarkable capability of our network in effectively leveraging the distinctive features inherent in panoramic images.

Qualitative results. In Figure 8 we show visual results for images obtained from the SUN360 dataset with a scale factor of $\times 8$. Both the full image and a cropped area are shown for comparisons. As shown, RCAN (Zhang et al. 2018) and LAU-Net (Deng et al. 2021) suffer from unpleasant blurring artifacts. OSRT (Yu et al. 2023) alleviates it to some extent, but still leaves out some details and structures. In contrast, our proposed BPOSR can effectively suppress artifacts and leverage scene details and internal natural image statistics to restore high-frequency content.

Window size	8×8	4×16	2×32	8×16
WS-PSNR	24.48	24.61	24.52	24.51
WS-SSIM	0.6747	0.6782	0.6745	0.6749

Table 3: Ablation studies for the Horizontal Striped Transformer Block

Ablation Study

To better understand BPOSR, we evaluate each key component under a completely fair setting. We use the same architecture and hyper-parameter for the following experiments and only vary one component for each ablation. The evaluation of these ablation experiments is conducted on the ODI-SR dataset, employing $\times 8$ upscaling factor.

Bi-Projection *vs*. Single Projection. To validate the effectiveness of the bi-projection mechanism used in our model, we introduce two alternative variants of our BPOSR: Variant-CMP and Variant-ERP, which leverage ERP or CMP in both two branches, respectively. In the experiments, we keep other configurations identical for a fair comparison. The results are presented in Table 2. We can see that the bi-projection strategy is superior to the other two versions, suggesting the effectiveness of our design that uses two different projections for high-fidelity reconstruction.

Effectiveness of the Horizontal Striped Transformer Block. We further verify the efficacy of our horizontal

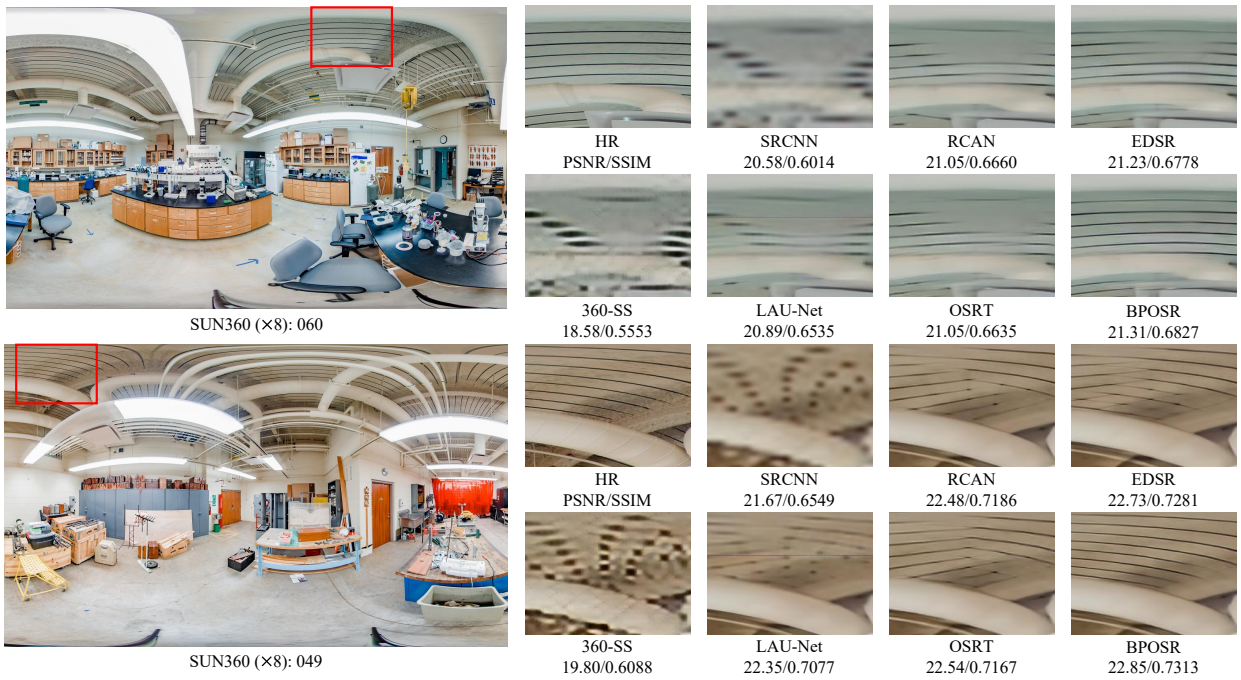


Figure 8: Visual comparisons with both SISR and ODISR methods on benchmark datasets. Our results are more visually faithful than other state-of-the-art algorithms.

Rotation ratio	w/o	2	3	4	5
WS-PSNR	24.49	24.52	24.61	24.62	24.62
WS-SSIM	0.6741	0.6758	0.6782	0.6774	0.6776

Table 4: Ablation studies for Perspective Shift Transformer Block. The rotation ratio r means that the angle of the spherical rotation is $\frac{360^\circ}{r}$.

Method	mean	1×1 Conv	BAFM
WS-PSNR	24.53	24.55	24.61
WS-SSIM	0.6735	0.6757	0.6782

Table 5: Ablation studies for the Block Attention Fusion Module

striped attention for ERP by changing the used window sizes. Table 3 shows that the horizontal choices outperform the square version when using the same region size for attention. This suggests that the horizontal window attention is more suitable for modeling ERP than the square variant.

Through the utilization of LAM (Gu and Dong 2021), an attribution method for super-resolution, we could tell which input pixels contribute most to the selected region. As shown in Figure 7, compared with CARN (Ahn, Kang, and Sohn 2018b) and RCAN (Zhang et al. 2018), BPOSR exhibits more pronounced horizontal regions and wider range of results in LAM analysis. This result implies that HSTB effectively captures the features in the horizontal region of ERP.

Effectiveness of the Perspective Shift Transformer Block. To evaluate the effectiveness of Perspective Shift Attention on CMP, we conduct experiments by varying the rotation magnifications applied to PSL. The results presented in Table 4 reveal a decrease of 0.14 dB in WS-PSNR when the Perspective Shift is not applied to CMP. This observation underscores the significance of view conversion in enhancing CMP's performance. Furthermore, through additional experiments, we find that as the rotation ratio increases, the effect of the model tends to converge. The model achieves the best performance when the rotation ratio is set to 3.

Effectiveness of the Block Attention Fusion Module. To further investigate the influence of BAFM, which fuses features from different projections and depths, we conduct experiments using a 1×1 convolution and mean operations to substitute for BAFM. Table 5 shows that the removal of BAFM leads to a performance decrease of 0.10 dB in terms of WS-PSNR, suggesting the effectiveness of our design.

Conclusion

In this paper, we present a novel Bi-Projection Omnidirectional Image Super-Resolution (BPOSR) network for ODISR. BPOSR performs ODISR based on the complementary information extracted from the ERP and CMP branches. To leverage the distinct geometric properties of these projections, we propose the Horizontal Striped Transformer Block (HSTB) for ERP and the Perspective Shift Transformer Block(PSTB) for CMP. Furthermore, we introduce the Block Attention Fusion Module (BAFM) to enhance the overall feature extraction capability by assigning varying attention weights to features from different projections and depths. Extensive quantitative and qualitative evaluations on multiple ODIs datasets demonstrate the superiority of our method over other state-of-the-art competitors.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No. 2022ZD0119202), the National Natural Science Foundation of China (Grant No.62322216, 62172409), Shenzhen Science and Technology Program (Grant No. JCYJ20220818102012025, RCYX20221008092849068, KQTD20221101093559018), 2023 CCF-Tencent Rhino-Bird Young Faculty Open Research Fund and CCF-Zhejiang Lab Joint Innovation Fund.

References

Ahn, N.; Kang, B.; and Sohn, K.-A. 2018a. Fast, Accurate, and Lightweight Super-Resolution with Cascading Residual Network. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision – ECCV 2018*, 256–272. Cham: Springer International Publishing. ISBN 978-3-030-01249-6.

Ahn, N.; Kang, B.; and Sohn, K.-A. 2018b. Fast, Accurate, and Lightweight Super-Resolution with Cascading Residual Network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 256–272.

Ai, H.; Cao, Z.; Zhu, J.; Bai, H.; Chen, Y.; and Wang, L. 2022. Deep Learning for Omnidirectional Vision: A Survey and New Perspectives. arXiv:2205.10468.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12299–12310.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9189–9198.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a Deep Convolutional Network for Image Super-Resolution. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 184–199. Cham: Springer International Publishing. ISBN 978-3-319-10593-2.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

Glasner, D.; Bagon, S.; and Irani, M. 2009. Super-resolution from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, 349–356.

Gu, J.; and Dong, C. 2021. Interpreting Super-Resolution Networks With Local Attribution Maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9199–9208.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5406–5415.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2018. Deep Back-Projection Networks for Super-Resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1664–1673.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1646–1654.

Li, J.; Fang, F.; Mei, K.; and Zhang, G. 2018. Multi-scale Residual Network for Image Super-Resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 527–542. Cham.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. SwinIR: Image Restoration Using Swin Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1132–1140.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Nishiyama, A.; Ikehata, S.; and Aizawa, K. 2021. 360° Single Image Super Resolution via Distortion-Aware Network and Distorted Perspective Images. In *2021 IEEE International Conference on Image Processing (ICIP)*, 1829–1833.

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 Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 191–207. Cham: Springer International Publishing. ISBN 978-3-030-58610-2.

Ozcinar, C.; Rana, A.; and Smolic, A. 2019. Super-resolution of omnidirectional images using adversarial learning. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 1–6. IEEE.

Shi, W.; Caballero, J.; Huszar, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-Time

Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1874–1883.

Sun, L.; Pan, J.; and Tang, J. 2022. ShuffleMixer: An Efficient ConvNet for Image Super-Resolution. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 17314–17326. Curran Associates, Inc.

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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4549–4557.

Wang, F.-E.; Yeh, Y.-H.; Tsai, Y.-H.; Chiu, W.-C.; and Sun, M. 2023. BiFuse++: Self-Supervised and Efficient Bi-Projection Fusion for 360° Depth Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5448–5460.

Xiao, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2012. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2695–2702.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5677–5686.

Yu, F.; Wang, X.; Cao, M.; Li, G.; Shan, Y.; and Dong, C. 2023. OSRT: Omnidirectional Image Super-Resolution With Distortion-Aware Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13283–13292.

Zhang, X.; Zeng, H.; Guo, S.; and Zhang, L. 2022. Efficient Long-Range Attention Network for Image Super-Resolution. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 649–667. Cham: Springer Nature Switzerland. ISBN 978-3-031-19790-1.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 294–310.