Omnidirectional Image Super-resolution via Bi-projection Fusion

Jiangang Wang¹, Yuning Cui², Yawen Li³, Wenqi Ren¹*, Xiaochun Cao¹
¹Shenzhen Campus of Sun Yat-sen University
²Technical University of Munich
³Beijing University of Posts and Telecommunications

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 equiangular 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 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 equiangular 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,
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 SR-CNN (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 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 upsampling network towards the non-uniformaly 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 \( \phi \) and \( \theta \), respectively, we have \((\phi, \theta) \in [-\pi, \pi] \times [-\frac{\pi}{2}, \frac{\pi}{2}]\) (Wang et al. 2023). The angular position \((\phi, \theta)\) can be converted to a coordinate \(Q_s = (q^x_s, q^y_s, q^z_s)\) on a standard sphere by:

\[
\begin{align*}
q^x_s &= \sin(\phi) \cos(\theta), \\
q^y_s &= \sin(\theta), \\
q^z_s &= \cos(\phi) \cos(\theta).
\end{align*}
\]

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

\[
Q_s = s \cdot R_i \cdot P_c,
\]

where \(P_c \in [0, r]\), \(P_c^z = \frac{r}{2}\), and the factor \(s = \frac{r}{|p_s|}\).

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 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 \(I_{ERP}\), we firstly transform it into the CMP form \(I_{CMP}\), and then use \(3 \times 3\) convolutions...
to separately extract shallow features for two projections as:

\[ I^{ir}_{CMP} = E2C(I^{ir}_{ERP}), \]  
\[ F^0_{ERP} = W^{3\times3}_1(I^{ir}_{ERP}), \]  
\[ F^0_{CMP} = W^{3\times3}_2(I^{ir}_{CMP}), \]

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

\[ F^i_{ERP} = HSAB_i(F^{i-1}_{ERP}), \]  
\[ F^i_{CMP} = PSAB_i(F^{i-1}_{CMP}), \]

where \( i \in [1, K] \) is the index of resulting features, and \( HSAB(\cdot) \) and \( PSAB(\cdot) \) 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:

\[ F^i_{FUS} = W^{fus}_{1\times1}(cat(F^i_{ERP}, F^i_{CMP})), \]
\[ F^i_{ERP} = W^{crt}_{1\times1}(cat(F^i_{ERP}, F^i_{FUS})), \]
\[ F^i_{CMP} = E2C(W^{cmp}_{1\times1}(cat(C2E(F^i_{CMP}), F^i_{FUS}))), \]

where \( cat \) is the concatenate operation, and \( W^{uti}_{1\times1} \) denotes a \( 1 \times 1 \) 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 \( F_j \) as:

\[ F_j = BAFM(cat(F^1_{FUS}, ..., F^{i-1}_{FUS}, \]
\[ F^i_{ERP}, C2E(F^i_{CMP}))), \]

\[ I^{ir}_{ERP} = F_{up}(F_j + F^0_{ERP} + C2E(F^0_{CMP})). \]

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 windows and apply the shift window self-attention...
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:

$$F_{CMP} = E2C(R(C2E(F_{CMP}))),$$

where $R$ 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:

$$Q = 3DConv_{Q}(F_{input}),$$
$$V = 3DConv_{V}(F_{input}),$$

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:

$$F_{m} = 3DConv(\frac{\text{Softmax}(Q \cdot Q^{\top})}{s} \cdot V),$$

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:

$$F_{out} = 3DConv(F_{input} + F_{m}) \in \mathbb{R}^{1 \times C \times H \times W}.$$
### Table 1: Quantitative comparisons (WS-PSNR/WS-SSIM) with SISR and ODISR algorithms on benchmark datasets. The best results are highlighted in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>ODI-SR</th>
<th>SUN360</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>×4</td>
<td>×8</td>
</tr>
<tr>
<td></td>
<td>×4</td>
<td>×8</td>
</tr>
<tr>
<td>SphereSR</td>
<td>26.55</td>
<td>0.6565</td>
</tr>
<tr>
<td>RCAN</td>
<td>21.19</td>
<td>0.6334</td>
</tr>
<tr>
<td>LapSRN</td>
<td>20.72</td>
<td>0.6214</td>
</tr>
<tr>
<td>MemNet</td>
<td>21.73</td>
<td>0.6284</td>
</tr>
<tr>
<td>MSRN</td>
<td>23.34</td>
<td>0.6496</td>
</tr>
<tr>
<td>EDSR</td>
<td>23.97</td>
<td>0.6483</td>
</tr>
<tr>
<td>D-DBPN</td>
<td>24.15</td>
<td>0.6573</td>
</tr>
<tr>
<td>RCAN</td>
<td>24.26</td>
<td>0.6554</td>
</tr>
<tr>
<td>DRN</td>
<td>24.32</td>
<td>0.6571</td>
</tr>
</tbody>
</table>

**Qualitative results.** In Figure 8 we show visual results for images obtained from the SUN360 dataset with a scale factor of ×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.

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

<table>
<thead>
<tr>
<th>Rotation ratio</th>
<th>w/o</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS-SSIM</td>
<td>0.6741</td>
<td>0.6758</td>
<td>0.6782</td>
<td>0.6774</td>
<td>0.6776</td>
</tr>
</tbody>
</table>

Table 5: Ablation studies for the Block Attention Fusion Module

<table>
<thead>
<tr>
<th>Method</th>
<th>mean</th>
<th>1 x 1 Conv</th>
<th>BAFM</th>
</tr>
</thead>
<tbody>
<tr>
<td>WS-PSNR</td>
<td>24.53</td>
<td>24.55</td>
<td>24.61</td>
</tr>
<tr>
<td>WS-SSIM</td>
<td>0.6735</td>
<td>0.6757</td>
<td>0.6782</td>
</tr>
</tbody>
</table>

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 x 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


