OPDN: Omnidirectional Position-aware Deformable Network for Omnidirectional Image Super-Resolution

Xiaopeng Sun*[1] Weiqi Li*[1,2] Zhenyu Zhang[1,2] Qiufang Ma[1] Xuhan Sheng[2] Ming Cheng[1]
Haoyu Ma[1] Shijie Zhao†[1] Jian Zhang[2] Junlin Li[1] Li Zhang[1]
[1]ByteDance Inc., [2]Peking University Shenzhen Graduate School
sunxiaopeng.01@bytedance.com, liweiqi@stu.pku.edu.cn

Abstract

360° omnidirectional images have gained research attention due to their immersive and interactive experience, particularly in AR/VR applications. However, they suffer from lower angular resolution due to being captured by fisheye lenses with the same sensor size for capturing planar images. To solve the above issues, we propose a two-stage framework for 360° omnidirectional image super-resolution. The first stage employs two branches: model A, which incorporates omnidirectional position-aware deformable blocks (OPDB) and Fourier upsampling, and model B, which adds a spatial frequency fusion module (SFF) to model A. Model A aims to enhance the feature extraction ability of 360° image positional information, while Model B further focuses on the high-frequency information of 360° images. The second stage performs same-resolution enhancement based on the structure of model A with a pixel unshuffle operation. In addition, we collected data from YouTube to improve the fitting ability of the transformer, and created pseudo low-resolution images using a degradation network. Our proposed method achieves superior performance and wins the NTIRE 2023 challenge of 360° omnidirectional image super-resolution.

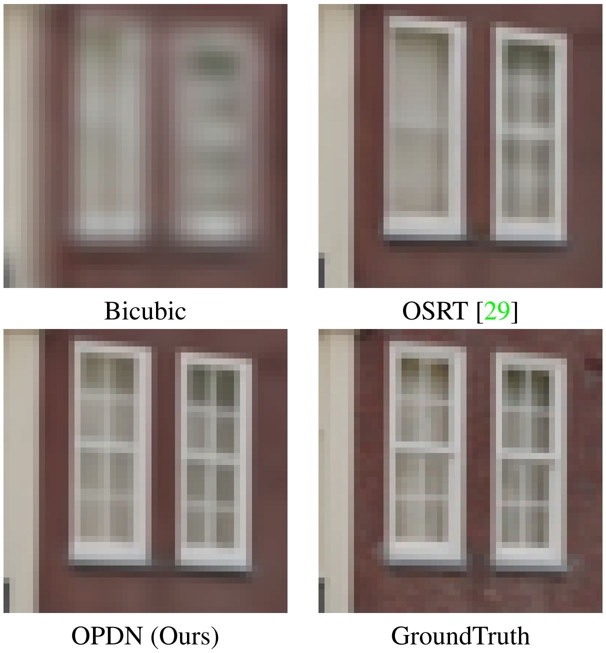


Figure 1. Visual comparisons of ×4 SR results on one image from Flickr360 validation set. We fine-tuned the OSRT [29] model on Flickr360 training set for fair comparison.

1. Introduction

With the rising popularity of AR/VR applications, 360° images, also known as omnidirectional or panoramic images, have garnered significant research interest in the computer vision community due to their immersive and interactive properties. Although the viewport range of 360° images is 360 × 180°, only a narrow field-of-view (FOV) is visible through head-mounted displays (HMDs). Consequently, extremely high resolutions, such as 4K × 8K [1], are required to ensure a small viewport with sufficient details. Thus, it is crucial to super-resolve low-resolution (LR) 360° images to provide high visual quality.

Recently, deep learning (DL) has significantly contributed to the success of single-image super-resolution (SISR). Following the initial introduction of DL in [7], subsequent studies have enhanced SR performance using convolutional neural networks (CNNs) [5, 13, 16–20, 31], generative adversarial networks (GANs) [10, 11, 23–25, 30], Vision Transformers (ViTs) [3, 4, 14, 15] and diffusion models [26]. Notably, SwinIR [15] achieved remarkable performance utilizing the Swin Transformer architecture, incorporating a shifted window mechanism for modeling long-range dependencies. HAT [4] further expanded

*Equal contribution. Weiqi Li is an intern in MMLab, ByteDance.
†Corresponding author. (e-mail: zhaoshijie.0526@bytedance.com)

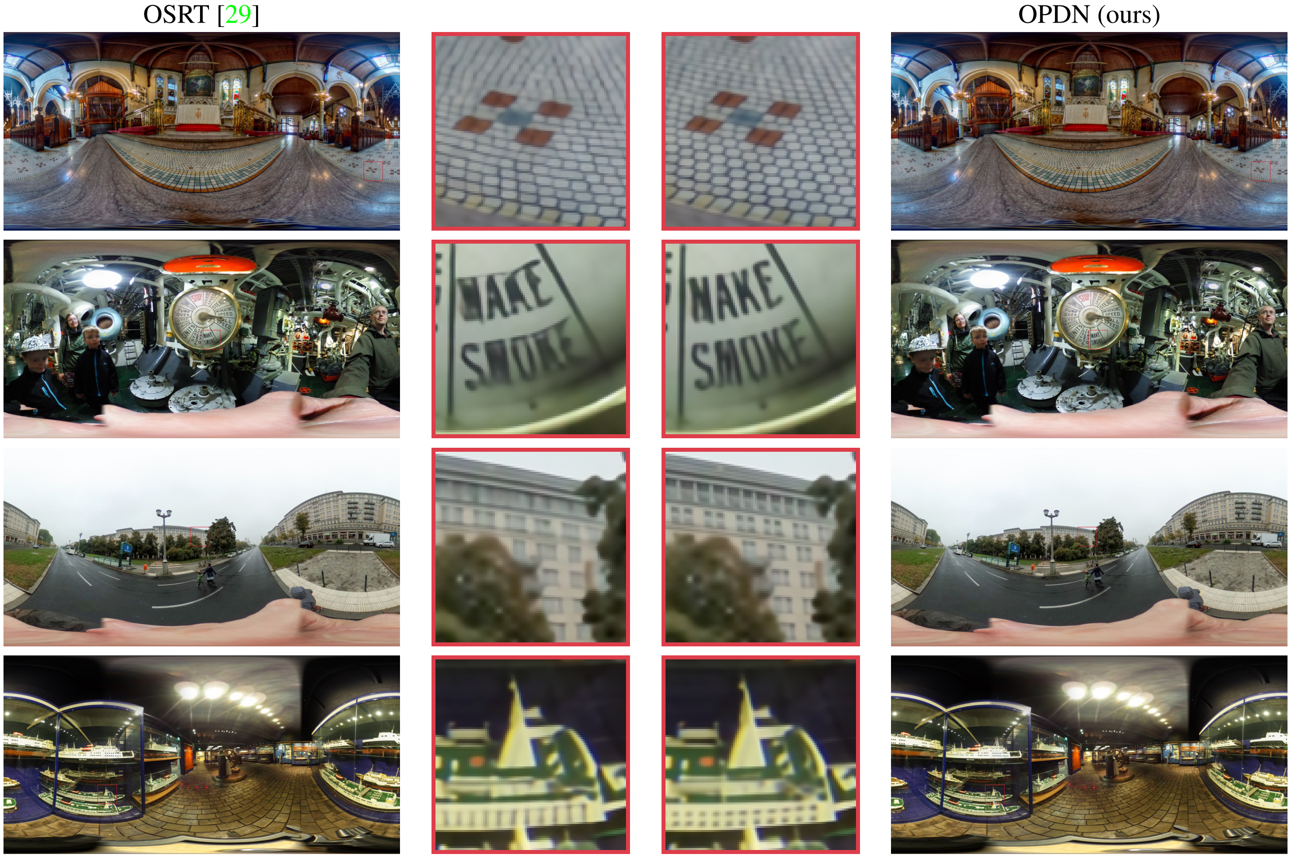


Figure 2. Visual comparisons of four images from Flickr360 validation set.

the model's scale, introducing a novel overlapping cross-attention mechanism to activate more pixels in the transformer, thereby achieving state-of-the-art (SOTA) performance in the SISR task. However, these methods cannot be directly applied to 360° images due to nonuniform pixel density and texture complexity across latitudes.

Consequently, several attempts have been made to address 360° SR problems. LAU-Net [6] divides the entire ERP image into patches based on latitude and upscales them separately. However, this approach obstructs the information connection between adjacent patches. SphereSR [28] introduces a Spherical Local Implicit Image Function (SLIIF) and a novel feature extraction module to leverage information from arbitrary projection types, but this results in extremely high computational complexity. Recently, OSRT [29] presented a distortion-aware transformer to address dimension-related distortions in ERP images, achieving state-of-the-art performance. Additionally, OSRT proposes fisheye downsampling and pseudo-ERP image generation methods to simulate real-world settings and mitigate network overfitting. However, position information remains unclear in OSRT, and only spatial information is considered throughout the pipeline, which constrains the final performance.

To address the aforementioned issues, we propose a two-stage framework that combines two super-resolution networks and a same-resolution enhancement network. Specifically, two models are employed in the first stage (referred to as model A and model B, respectively). Model A is designed based on the HAT architecture, incorporating proposed omnidirectional position-aware deformable blocks (OPDB) and Fourier upsampling, while model B adds a spatial frequency fusion (SFF) module to model A. In the second stage, we perform a same-resolution enhancement based on model A's structure, incorporating a pixel unshuffle operation at the beginning of the network. Moreover, various data augmentation and training strategies are implemented to improve the final restoration performance. In summary, our contributions include:

- We propose a two-stage framework that obtains high-resolution images in the first stage and performs same-resolution enhancement in the second stage, effectively preserving more image details and eliminating artifacts.
- We introduce a novel omnidirectional position-aware deformable block (OPDB), which efficiently lever-

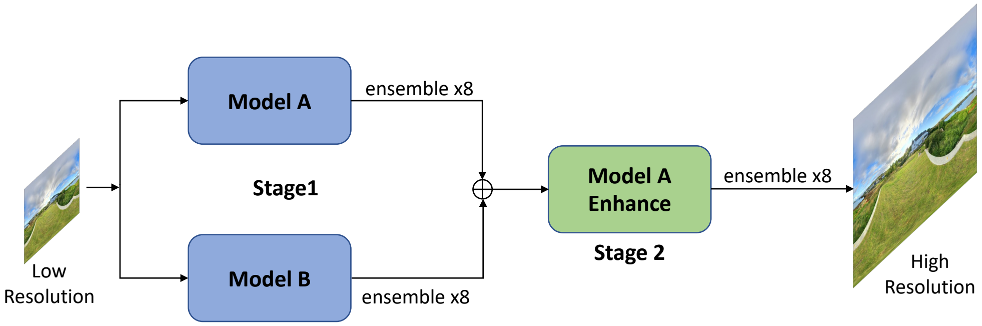


Figure 3. The overall pipeline of proposed two-stage framework, in which two ×4 SR models are employed in stage 1, while stage 2 performs a same-resolution enhancement.

ages position encoding information and ERP geometric properties.

- We implement a spatial frequency fusion module (SFF) and Fourier upsampling to explore information in the frequency domain.
- Our proposed method achieves superior performance and wins the NTIRE 2023 challenge of 360° omnidirectional super-resolution [2].

2. Related Work

2.1. Single Image Super-Resolution (SISR)

Since the introduction of deep learning to single-image super-resolution tasks by SRCNN [7], it has outperformed many traditional algorithms. Consequently, various CNN architectures have been extensively investigated by researchers to further improve the performance of image super-resolution algorithms. For example, EDSR [16] initially employed residual blocks without batch normalization as the fundamental building blocks, forming a deeper super-resolution network. RDN [32] combined residual blocks with dense connections, introducing residual dense blocks. RCAN [31] integrated channel attention into residual blocks, proposing residual attention modules and deepening the network.

Recently, Vision Transformers have overcome the inductive biases inherent in CNNs and effectively modeled long-range dependencies, achieving optimal performance in numerous high-level visual tasks. ViT-like structures have also been applied to low-level tasks. For instance, IPT [3] proposed a network structure akin to ViT, pre-trained on large-scale datasets and multiple different low-level tasks, yielding impressive performance. SwinIR [15] adopted the Swin Transformer's structure, using a shifted window mechanism to model long-range dependencies and achieving enhanced performance with fewer parameters. EDT [14] further improved single-image super-resolution performance by employing self-attention mechanisms and multi-related-task pre-training strategies.

HAT [4] combines self-attention, channel attention, and a novel overlapping cross-attention mechanism, introducing the residual hybrid attention groups (RHAG), which are composed of hybrid attention blocks (HAB) and overlapping cross-attention blocks (OCAB). This approach activates more pixels to facilitate reconstruction. In contrast to previous methods, HAT employs same-task pre-training on large-scale datasets, demonstrating the effectiveness of this strategy. Additionally, HAT expands the model's scale, establishing new state-of-the-art benchmarks for the single-image super-resolution task.

2.2. Omnidirectional Image Super-Resolution (ODISR)

Initial research on omnidirectional image super-resolution (ODISR) concentrated on stitching and optimizing multiple low-resolution omnidirectional images using various projection types, such as spherical and hyperbolic. To assess the efficacy of omnidirectional image super-resolution qualitatively, Sun et al. [21] proposed a weighted-to-spherically-uniform quality evaluation method (WS-PSNR) for spheres. Recently, GANs have been incorporated into ODISR, with models operating on planar images, fine-tuning existing SISR models using L1 loss [8] or GAN loss [33], and employing WS-SSIM to evaluate model performance. LAU-Net [6] identified pixel density non-uniformity in ERP omnidirectional images, prompting numerous studies to develop dedicated base networks addressing this issue. LAU-Net [6] partitions the entire ERP image into latitude-related patches manually, learning ERP distortions within distinct latitude ranges. Instead of processing the whole ERP image end-to-end, LAU-Net separately processes non-overlapping patches of varying latitudes, resulting in discontinuities throughout the ERP image. Nishiyama et al. incorporated the area stretching ratio as an additional condition input to the network; however, this necessitates modifications to existing SISR backbone networks. SphereSR introduces an algorithm

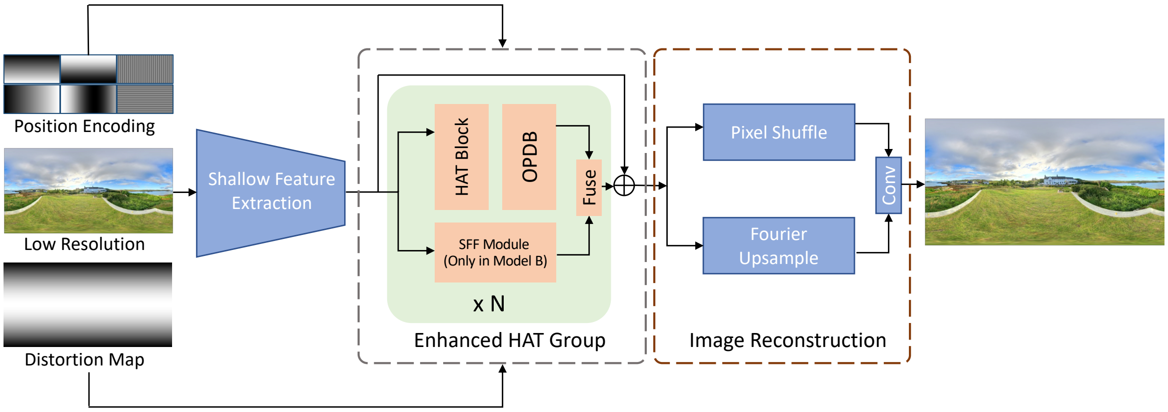


Figure 4. The network architecture of our proposed model A.

capable of handling omnidirectional image super-resolution for arbitrary projection types. Specifically, SphereSR [28] proposes a feature extraction module that extracts features on a spherical surface from various projection types (such as CP, ERP, and polyhedra). Based on the extracted spherical features, SphereSR employs a spherical local implicit image function (SLIIF) to predict RGB values corresponding to spherical coordinates, yielding high-resolution reconstruction results for arbitrary projection types. Nonetheless, ERP remains the most commonly utilized projection type for omnidirectional image editing, transmission, and display.

In realistic scenarios, omnidirectional images (ODIs) are typically captured using two or more fisheye lenses, leading to distortions during the fisheye projection process. Recognizing this, Yu et al. [29] proposed a degradation process—Fisheye downsampling—that emulates the imaging process in real-world settings. To more effectively address dimension-related distortions in equirectangular projection (ERP) images, Yu et al. introduced a Distortion-aware Transformer designed to adaptively perform super-resolution on omnidirectional images. Notably, panoramic image datasets tend to be smaller in size compared to conventional 2D images. To mitigate network overfitting, OSRT synthesizes pseudo-panoramic images from 2D images during training.

In this context, the OSRT approach aims to address the challenges inherent in omnidirectional image super-resolution, which differ from those in traditional 2D image super-resolution. By simulating real-world distortions during the degradation process and incorporating a Distortion-aware Transformer, this method is more adept at managing the distinctive features of omnidirectional images. Furthermore, employing synthesized pseudo-panoramic images alleviates the problem of limited panoramic image datasets, mitigating overfitting and enhancing the model's performance.

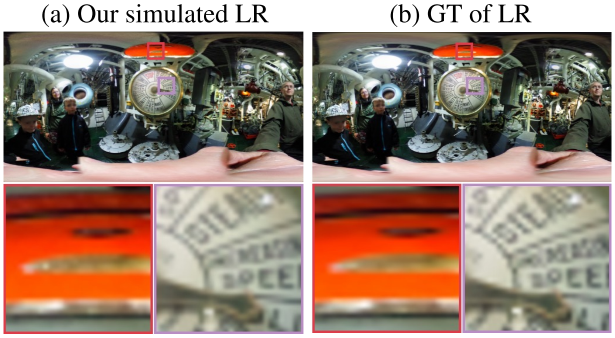


Figure 5. Visualization comparisons of (a) our simulated LR images with applying our degradation network to one HR image in Flickr360 and (b) the corresponding groudtruth LR image.

3. Method

3.1. Simulating Data Degradation

Transformer-based models necessitate substantial data input for training to enhance performance [3]. Therefore, we amassed a vast collection of panoramic video data from YouTube[1] to train our model. Due to the lack of HR-LR degradation settings provided by the competition organizers, we adopted the LBO module from AnimeSR [27] and implemented a degradation network identical to LBO to learn the mapping from HR to LR on the training set. The fully converged degradation model achieved a WS-PSNR [21] of 43 when evaluated on the Flickr360 validation set. As illustrated in Fig. 5, the generated pseudo-LR images subjectively exhibit increased clarity in polar regions compared to LR, yet diminished clarity in regions with low dimensions.

3.2. Proposed Two-Stage Framework

We propose a two-stage super-resolution (SR) model for 360-degree image SR, as illustrated in Fig. 3. The first stage

[1] www.youtube.com

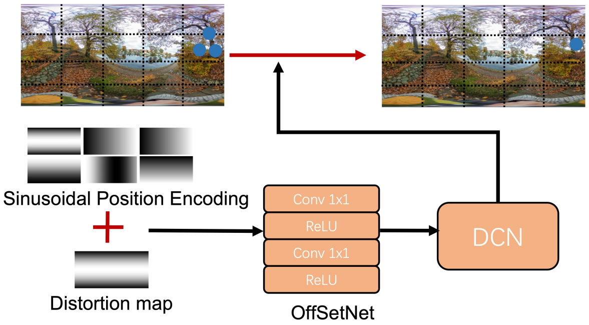


Figure 6. OPDB Module.

comprises an SRx4 model, while the second stage features a same-resolution enhancement model. In the first stage, we utilize two models (Model A and Model B) for ensemble purposes. Model A is built on the Hybrid Attention Transformer (HAT) architecture, integrating Orthogonal Projection Depthwise Block (OPDB) and Fourier upsampling, whereas Model B incorporates a Spatial Frequency Fusion (SFF) module into Model A. In fact, Model A alone is sufficient to ensure our victory in this competition. In the second stage, we employ a structure based on Model A and its weight parameters, introducing a pixel unshuffle at the beginning to downsample the input image by a factor of four, ultimately producing a result with the same resolution as the input.

3.3. First Stage with OPDB

Model A and Model B were combined through model ensemble to form the first stage. In Model A, we based our network structure on HAT and added OPDB after each RHAG, along with a frequency domain module for spatial frequency fusion in Fig. 4. Finally, Fourier upsampling was added to the upsampling module [35]. Specifically, inspired by DACB in [29] and BUSIFusion [12], a novel Omnidirectional Position-aware Deformable Block (OPDB) was proposed, which combines dimensional information and position encoding information for 360-degree images. As shown Fig. 6, OPDB uses Sinusoidal Position Encoding [22] and distortion map [29] to encode the position of ERP projections: absolute positions are represented using sine and cosine functions, and relative positions are obtained by multiplying the two. This design allows positional encoding to be linearly represented by position, reflecting its relative position relationship:

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/d_{model}}}), \tag{1}$$

$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{2i/d_{model}}}). \tag{2}$$

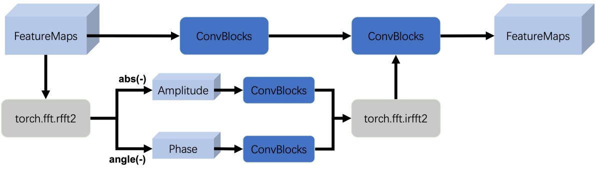


Figure 7. SFF Module.

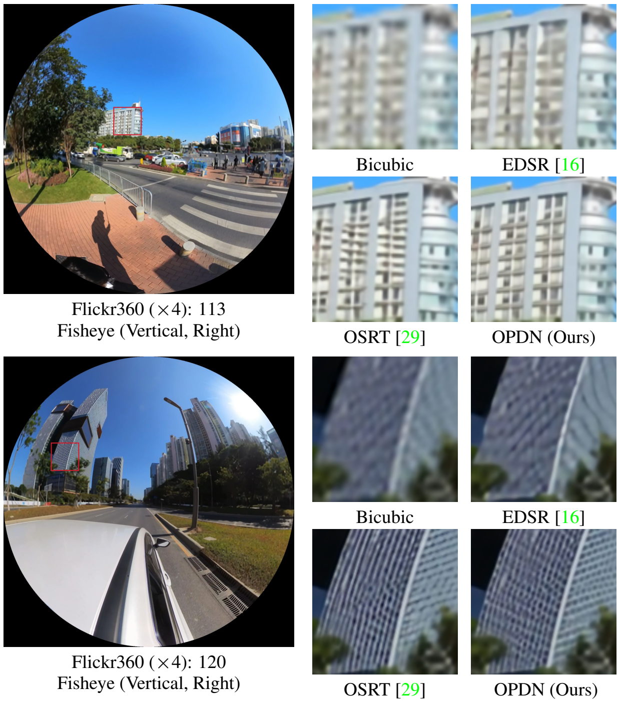
Here, pos denotes the position in the input sequence, which is the coordinate on the ERP image. We encode the latitude and longitude of the ERP image separately using the above encoding method. i represents the index of the sine function, and d_{model} signifies the dimension of the model. In this manner, each position is encoded into a d_{model}-dimensional vector, with each dimension corresponding to a sine function. When encoding the position of Equirectangular Projection (ERP) representations, the ERP representation can be regarded as a sequence, and Sinusoidal Position Encoding can be employed to encode its position. Consequently, each ERP representation is encoded into a d_{model}-dimensional vector, which encompasses its position information in the sequence. This encoding technique aids the model in better understanding the relationships among different ERP representations, thereby enhancing the model's performance. The positional encoding information is transmitted to the offsetNet, allowing the deformable convolution to utilize this offset information to adjust the convolution kernel based on the spherical position correlation. In this manner, the spherical coordinate information can be aggregated through deformable convolution, achieving a larger receptive field and superior reconstruction effect.

The aforementioned modules constitute Model A, while the subsequent Spatial Frequency Fusion (SSF) module is a novel addition to Model B. Inspired by the work of Zhou et al. [34], the SSF module, as depicted in Fig. 7, conducts a Fourier transform on the feature map, extracts features from the amplitude and phase components individually. After that, the processed amplitude and phase components are transformed back into the complex domain. Finally, an inverse Fast Fourier Transform (IFFT) is performed on the complex feature to merge the feature map with spatial domain information.

3.4. Second Stage and Overall Strategy

In the second stage model, we introduced a pixel unshuffle operation before the model, based on the weights of Model A, to ensure that the input and output resolutions remain consistent. The training data for the second stage is generated by performing inference on the training data using Model A, without employing self-ensemble. We discovered that utilizing self-ensemble during the second stage training resulted in performance degradation, as demonstrated in the table.

The overall training process comprises the following steps:


Flickr360 (×4): 113
Fisheye (Vertical, Right)

Bicubic

EDSR [16]

OSRT [29]

OPDN (Ours)

Flickr360 (×4): 120
Fisheye (Vertical, Right)

Bicubic

EDSR [16]

OSRT [29]

OPDN (Ours)

Figure 8. Visual comparisons for SR of Fisheye images.

Stage I: Initially, we fine-tune the pre-trained HAT-L model on the competition data, followed by fine-tuning it on the generated degraded 360-degree data, and finally fine-tuning it once more on the competition data, while increasing the window size in the Hybrid Attention Block (HAB) from 16 to 32. This process yields two branches: Model A, which does not include the Spatial Frequency Fusion (SSF) module, and Model B, which incorporates the SSF module. Both models are employed for ensemble purposes.

Stage II: We adhere to the approach delineated in Section 3.4.

Testing stage: Initially, we conduct inference on Models A and B from Stage I and apply a novel self-ensemble x8 technique. Taking into account the characteristics of Equirectangular Projection (ERP) images, in Fig 9, we perform horizontal and vertical flipping, roll 1/4 of the width (We have tested other parameters, and only 1/4 showed improvement.), and subsequently conduct horizontal and vertical flipping once more. Next, we average the results of Models A and B, and derive the final result from inference using the second stage model.

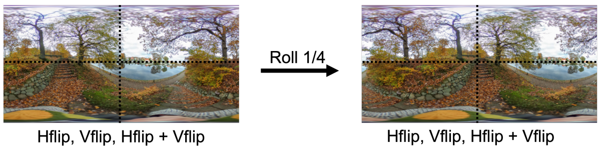


Figure 9. Self-ensemble strategy for ERP images.

Table 1. Final results of the NTIRE2023 Challenge on 360° Image SR track [2].

Methods	WS-PSNR (Val)	WS-PSNR (Test)
1st (Ours)	**30.43**	**28.64**
2nd	30.20	28.49
3rd	30.04	28.28
4th	30.03	28.13
5th	29.87	28.11
6th	30.00	28.10
7th	28.32	27.65

4. Experiments

4.1. Datasets

We utilized two datasets in our experiments. The first dataset is the Flickr360 dataset, which serves as the official data for the NTIRE 2023 competition and is used for model training and testing. In addition to Flickr360, we collected an extra 260 high-resolution panoramic videos from YouTube[2]. We extracted all the I-frames from these videos, converted frames in non-ERP projection formats (such as EAC, Cubemap, and Stereo EAC) to the ERP projection format, and subsequently downsampled these I-frames to a resolution of 2048×1024. We manually removed low-quality frames or those influenced by transition effects in the videos. Lastly, we employed the LBO degradation network to downsample the selected approximately 7,000 frames, obtaining their corresponding low-resolution images.

4.2. Implementation Detail

For Stage I, we initially fine-tune the official pre-trained HAT-L model for 450K iterations using the Charbonnier loss. The Adam optimizer is employed with an initial learning rate of 1×10^{-4}. We implement the MultiStepLR strategy for learning rate adjustment, progressively reducing it to 1×10^{-6}. The same training strategy is applied to all training instances in Stage I. Ultimately, we fine-tune our model using the L2 loss for 10K iterations.

For Stage II, we train our model using the L2 loss for 300K iterations, with an initial learning rate of 5×10^{-5}. All experiments are conducted using four NVIDIA A100 GPUs.

4.3. Quantitative Results

We compare our Omnidirectional Position-aware Deformable Network (OPDN) with previous Single Image Super-Resolution (SISR) methods [14, 15] and the Omnidirectional Deep Image Super-Resolution (ODISR)

[2] https://github.com/sxpro/OPDN/blob/main/youtube_video_data_list.txt

Table 2. Quantitative results of PSNR on Flickr360 dataset.

Methods	WS-PSNR (Val)	WS-PSNR (Test)
Bicubic	27.45	25.74
EDSR-M [16]	29.18	27.30
SwinIR [15]	29.75	27.86
OSRT [29]	30.05	-
OPDN (Ours)	**30.43**	**28.64**

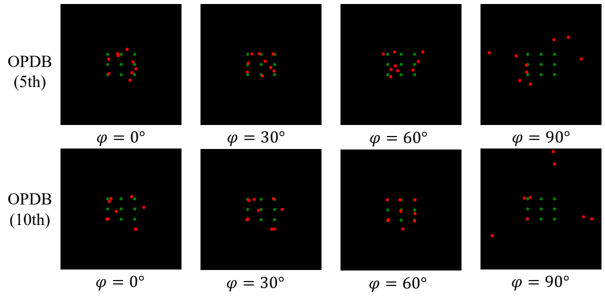


Figure 10. Visualizations of offset maps in OPDB. Reference and deformed points are depicted in green and red, respectively.

Table 3. Ablation results in Flickr360 validation.

Method	WS-PSNR
HAT-L imageNet pretrained	30.15
+ Fourier upsampling	30.16
+ YouTube data	30.17
+ windowsize 32	30.17
+ SSF	30.18
+ OPDB	30.28
OPDN	**30.37**

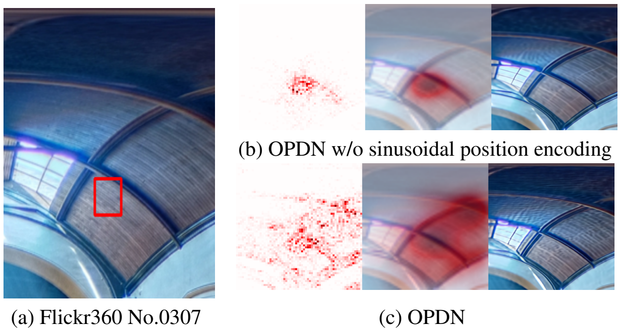
method [29]. For fair comparison, we fine-tune OSRT [29] on the Flickr360 training set. Since the test set is not yet publicly available, the WS-PSNR of OSRT on test sets is not obtained. As shown in Table 2, our OPDN surpasses the previous method by 0.38dB in terms of WS-PSNR.

4.4. Qualitative Results

We present our results on the official validation set of the NTIRE 2023 challenge (i.e., Flickr360) in Fig. 2. It is evident that our proposed Omnidirectional Position-aware Deformable Network (OPDN) restores more reliable texture details, which are not captured by OSRT [29]. Furthermore, our OPDN exhibits superior recovery of lines and stripes, with fewer visible artifacts and a reduced blurring effect. From Fig. 8, one can see that OPDN can preserve the original structure when being projected to other projection types.

4.5. Ablation Study

The ablation experiments primarily evaluate the effectiveness of Omnidirectional Position-aware Deformable Block (OPDB), Spatial Spectral Fusion (SSF), Fourier upsampling, and YouTube data. These experiments are conducted by fine-tuning the HAT-L ImageNet pre-trained weights using identical settings and assessing the PSNR on the validation set. As shown in Table 3, our methods consistently achieve substantial performance improvements. Figs. 10 and 11 illustrate that OPDB is capable of adapting to dimensional changes in 360-degree images and leveraging a wider range of information to reconstruct SR results.



(b) OPDN w/o sinusoidal position encoding

(a) Flickr360 No.0307

(c) OPDN

Figure 11. Results of LAM [9] visualization. From left to right, (b) and (c) show the LAM contribution, area of contribution and SR results. The LAM maps illustrate the significance of each pixel in the input LR image with respect to the SR of the patch indicated by a red box. The LAM results demonstrate that OPDB can effectively adapt to changes in dimensions of 360-degree images and utilize a broader range of information to reconstruct SR results.

Table 4. Ensemble results in Flickr360 dataset.

Method	WS-PSNR(Val)	WS-PSNR(Test)	Infer Time
HAT-L	30.15	-	4.4s/per
Model A	30.37	-	4.5s/per
Model B	30.37	-	5s/per
Model A + se x8	30.41	28.61	35s/per
Model B + se x8	30.41	28.61	37s/per
Stage1	30.42	28.62	66s/per
Stage2	**30.43**	**28.64**	73s/per

5. NTIRE 2023 Challenge

We participate in Track 1 of the NTIRE 2023 360° Omnidirectional Super-Resolution Challenge [2]. Quantitative results are presented in Table 1. During the competition, the new self-ensemble is utilized in all stages, along with the model ensemble. In fact, Model A alone would have sufficed to secure the championship; however, we pursued excellence and achieved an additional 0.03 dB improvement, as shown in Table 4. The testing condition is to infer LR images of size 512 × 256 on A100.

6. Conclusion

In this paper, we propose an Omnidirectional Position-aware Deformable Network for 360-degree image super-resolution. Specifically, we introduce a two-stage framework and OPDB, which incorporates a frequency block and Fourier upsampling to enhance the final performance of image improvement. Our method strikes a favorable balance between enhancement performance and model complexity, ultimately winning the championship in the 360° Omnidirectional Super-Resolution category of NTIRE 2023.

References

[1] Hao Ai, Zidong Cao, Jinjing Zhu, Haotian Bai, Yucheng Chen, and Ling Wang. Deep learning for omnidirectional vision: A survey and new perspectives. *arXiv preprint arXiv:2205.10468*, 2022. 1

[2] Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Ying Shan, Gen Li, Radu Timofte, et al. NTIRE 2023 challenge on 360° omnidirectional image and video super-resolution: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 3, 6, 7

[3] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 1, 3, 4

[4] X Chen, X Wang, J Zhou, and C Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 3

[5] Jian Zhang Chong Mou. Transcl: Transformer makes strong and flexible compressive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 1

[6] Xin Deng, Hao Wang, Mai Xu, Yichen Guo, Yuhang Song, and Li Yang. Lau-net: Latitude adaptive upscaling network for omnidirectional image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9189–9198, 2021. 2, 3

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 1, 3

[8] Vida Fakour-Sevom, Esin Guldogan, and Joni-Kristian Kämäräinen. 360 panorama super-resolution using deep convolutional networks. 2018. 3

[9] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021. 7

[10] Yujie Hu, Yinhuai Wang, and Jian Zhang. Dear-gan: Degradation-aware face restoration with gan prior. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2023. 1

[11] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

[12] Jiabao Li, Yuqi Li, Chong Wang, Xulun Ye, and Wolfgang Heidrich. Busifusion: Blind unsupervised single image fusion of hyperspectral and rgb images. *IEEE Transactions on Computational Imaging*, 9:94–105, 2023. 5

[13] Weiqi Li, Bin Chen, and Jian Zhang. D3c2-net: Dual-domain deep convolutional coding network for compressive sensing. *arXiv preprint arXiv:2207.13560*, 2022. 1

[14] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 1, 3, 6

[15] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1, 3, 6, 7

[16] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. 1, 3, 6, 7

[17] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3517–3526, 2021. 1

[18] Chong Mou, Yanze Wu, Xintao Wang, Chao Dong, Jian Zhang, and Ying Shan. Metric learning based interactive modulation for real-world super-resolution. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 723–740. Springer, 2022. 1

[19] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 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*, pages 1874–1883, 2016. 1

[20] Xiaopeng Sun, Wen Lu, Rui Wang, and Furui Bai. Distilling with residual network for single image super resolution. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1180–1185. IEEE, 2019. 1

[21] Yule Sun, Ang Lu, and Lu Yu. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE signal processing letters*, 24(9):1408–1412, 2017. 3, 4

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[23] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 1

[24] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 1

[25] Yinhuai Wang, Yujie Hu, Jiwen Yu, and Jian Zhang. Gan prior based null-space learning for consistent super-resolution. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 1

[26] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations (ICLR)*, 2023. 1

[27] Yanze Wu, Xintao Wang, Gen Li, and Ying Shan. Animesr: Learning real-world super-resolution models for animation videos. In *Advances in Neural Information Processing Systems*, 2022. 4

[28] Youngho Yoon, Inchul Chung, Lin Wang, and Kuk-Jin Yoon. 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*, pages 5677–5686, 2022. 2, 4

[29] Fanghua Yu, Xintao Wang, Mingdeng Cao, Gen Li, Ying Shan, and Chao Dong. Osrt: Omnidirectional image super-resolution with distortion-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 4, 5, 6, 7

[30] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 1

[31] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 1, 3

[32] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018. 3

[33] Yupeng Zhang, Hengzhi Zhang, Daojing Li, Liyan Liu, Hong Yi, Wei Wang, Hiroshi Suitoh, and Makoto Odamaki. Toward real-world panoramic image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 628–629, 2020. 3

[34] Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. Spatial-frequency domain information integration for pan-sharpening. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 274–291. Springer, 2022. 5

[35] Man Zhou, Hu Yu, Jie Huang, Feng Zhao, Jinwei Gu, Chen Change Loy, Deyu Meng, and Chongyi Li. Deep fourier up-sampling. *arXiv preprint arXiv:2210.05171*, 2022. 5