NTIRE 2023 Challenge on 360° Omnidirectional Image and Video Super-Resolution: Datasets, Methods and Results

Mingdeng Cao Chong Mou Fanghua Yu Xintao Wang Yinqiang Zheng
Jian Zhang Chao Dong Gen Li Ying Shan Radu Timofte Xiaopeng Sun
Weiqi Li Zhenyu Zhang Xuhan Sheng Bin Chen Haoyu Ma Ming Cheng
Shijie Zhao Wanwan Cui Tianyu Xu Chunyang Li Long Bao Heng Sun
Huaibo Huang Xiaoqiang Zhou Yuang Ai Ran He Renlong Wu Yi Yang
Zhilu Zhang Shuohao Zhang Junyi Li Yunjin Chen Dongwei Ren
Wangmeng Zuo Renlong Wu Yi Yang Zhilu Zhang Shuohao Zhang Junyi Li
Yunjin Chen Dongwei Ren Wangmeng Zuo Zhenyu Zhang Qian Wang
Weiqi Li Xuhan Sheng Bin Chen Hao-Hsiang Yang Yi-Chung Chen
Zhi-Kai Huang Wei-Ting Chen Yuan-Chun Chiang Hua-En Chang I-Hsiang Chen
Chia-Hsuan Hsieh Sy-Yen Kuo Zebin Zhang Jiaqi Zhang Yuhui Wang
Shuhao Cui Junshi Huang Li Zhu Shuman Tian Wei Yu Bingchun Luo

Abstract

This report introduces two high-quality datasets Flickr360 and ODV360 for omnidirectional image and video super-resolution, respectively, and reports the NTIRE 2023 challenge on 360° omnidirectional image and video super-resolution. Unlike ordinary 2D images/videos with a narrow field of view, omnidirectional images/videos can represent the whole scene from all directions in one shot. There exists a large gap between omnidirectional image/video and ordinary 2D image/video in both the degradation and restoration processes. The challenge is held to facilitate the development of omnidirectional image/video super-resolution by considering their special characteristics. In this challenge, two tracks are provided: one is the omnidirectional image super-resolution and the other is the omnidirectional video super-resolution. The task of the challenge is to super-resolve an input omnidirectional image/video with a magnification factor of ×4. Realistic omnidirectional downsampling is applied to construct the datasets. Some general degradation(e.g., video compression) is also considered for the video track. The challenge has 100 and 56 registered participants for those two tracks. In the final testing stage, 7 and 3 participating teams submitted their results, source codes, and fact sheets. Almost all teams achieved better performance than baseline models by integrating omnidirectional characteristics, reaching compelling performance on our newly collected Flickr360 and ODV360 datasets.

1. Introduction

The 360° or omnidirectional images/videos can provide users with an immersive and interactive experience, and have received much research attention with the popularity of AR/VR applications. Unlike planar 2D images/videos with a narrow field of view (FoV), 360° images/videos can represent the whole scene in all directions. However, 360° images/videos suffer from the lower angular resolution problem since they are captured by the fisheye lens with the same sensor size for capturing planar images. Although the 360° images/videos are high-resolution, their details are usually missing. In many application scenarios, increasing the resolution of 360° images/videos is highly demanded to achieve higher perceptual quality and boost the performance of downstream tasks.

Recently, considerable success has been achieved in image and video super-resolution (SR) tasks with the development of deep learning-based methods [19, 36, 38, 68]. Although 360° images/videos are often transformed into 2D planar representations by preserving omnidirectional information in practice, like equirectangular projection (ERP) and cube map projection (CMP), existing super-resolution methods still cannot be directly applied to 360° images/videos due to the distortions introduced by the projections. Thus some methods [18, 61, 62] specified for 360°

Ntire 2023 webpage: https://cvlai.net/ntire/2023/
360SR challenge homepage: https://github.com/360SR/360SR-Challenge

image super-resolution are developed by considering the omnidirectional characteristics. As for videos, the temporal relationships in a 360° video should be further considered since they differ from those in a planar perspective 2D video. Therefore, effectively super-resolving 360° image/video by considering these characteristics remains challenging.

The NTIRE 2023 challenge on 360° omnidirectional super-resolution steps forward in establishing high-quality benchmarks for 360° image and video SR, further highlighting the challenges and research problems. The challenge can also provide an opportunity for corresponding researchers to work together to show their insights and novel algorithms, significantly promoting the development of 360° image and video SR tasks. Two datasets termed Flickr360 and ODV360 are proposed for omnidirectional image and video super-resolution tasks, respectively. Realistic omnidirectional image downsampling methods are applied to generate LR image/video pairs to increase the generalization ability to the real world.

The challenge has 100 and 56 registered participants for the image and video tracks. In the final testing stage, 7 and 3 participating teams submitted their results, source codes, and fact sheets. They develop new methods to integrate omnidirectional characteristics, and introduce new technologies in network architecture design, data augmentation methods, and *etc*. We present detailed challenge results in Sec. 5.

Our challenge is one of the NTIRE 2023 Workshop [1] series of challenges on: night photography rendering [47], HR depth from images of specular and transparent surfaces [63], image denoising [34], video colorization [27], shadow removal [53], quality assessment of video enhancement [41], stereo super-resolution [54], light field image super-resolution [58], image super-resolution (×4) [71], 360° omnidirectional image and video super-resolution [5], lens-to-lens bokeh effect transformation [13], real-time 4K super-resolution [14], HR nonhomogenous dehazing [1], efficient super-resolution [33].

2. Related Work

2.1. Omnidirectional Image Super-Resolution

Deep learning for single image SR (SISR) is first introduced in [20]. Further works boost SR performance by CNNs [16, 21, 35, 39, 42, 45, 69], Vision Transformers (ViTs) [9, 11, 32, 37] and generative adversarial networks (GANs) [30, 56, 57, 66]. To improve perceptual quality, adversarial training is performed as a tuning process to generate more realistic results [56,57]. Moreover, various flexible degradation models are proposed in [56, 65] to synthesize more practical degradations.

Initially, ODISR models focus on the spherical assembling of LR ODIs under various projection types [2–4, 28, 43]. Recent ODISR models are performed on plane images and are fine-tuned from existing SISR models with L1 loss [22] or GAN loss [46, 70]. Since LAU-Net [18] found pixel density in ERP ODIs is non-uniform, many studies attempt to design specific backbone networks to overcome this issue. Nishiyama *et al*. [44] treats area stretching ratio as additional input. SphereSR [61] learns upsampling processes on various projection types to mitigate the influence of non-uniformity in specific projection types. Moreover, OSRT [62] modulates ERP distortions continuously and self-adaptively by learning deformable offsets from the ERP distortion maps.

2.2. Omnidirectional Video Super-Resolution

Video super-resolution (VSR) is a challenging task, which aims to gather complementary information across misaligned video frames for restoration. One prevalent approach is the sliding-window framework [24, 51, 55, 60], where each frame in the video is restored using the frames within a short temporal window. In contrast to the sliding-window framework, the recurrent framework [6, 8, 23, 26] attempts to exploit the long-term dependencies by propagating the latent features, which allows a more compact model compared to those in the sliding-window framework.

Unlike ordinary 2D videos, omnidirectional videos can provide users with a whole scene in all directions. Therefore, there is a large gap between omnidirectional video and ordinary 2D video in the degradation and restoration processes. Some prior works [17, 40] are proposed to solve this challenging task. However, how to make full use of the features of the omnidirectional format is still an open challenge.

3. Dataset Construction

3.1. Flickr360 Dataset for Omnidirectional Image Super-Resolution

To promote the development of this field, we construct a new 360° image dataset, which contains about 3150 ERP images with an original resolution larger than 5k. Specifically, 3100 images are collected from Flickr [2], and the other 50 images are captured by Insta 360° cameras. The images from Flickr are under either Creative Commons BY 2.0, Creative Commons BY-NC 2.0, Public Domain Mark 1.0, Public Domain CC0 1.0, or U.S. Government Works license. These licenses allow free use, redistribution, and adaptation for non-commercial purposes. The image contents vary both indoors and outdoors, containing a lot of natural scenery, human architecture, and street scenes. We first resize the original images into 2k resolution (2048 x

[1] https://cvlai.net/ntire/2023/

[2] https://www.flickr.com/

1024), serving as HR images. These HR images are further downsampled into LR images. These images are randomly partitioned into Training, Validation, and Testing sets, as shown in Tab. 1. The dataset is publicly available on the 360SR challenge homepage.

Table 1. The detailed data partition of the Flickr360 dataset.

	Training	Validation	Testing
Source	Flickr 360	Flickr 360	Flickr 360+capturing
Number	3000	50	50+50
Storage	8.1G (HR) 553M (LR)	137M (HR) 9.3M (LR)	271M (HR) 20M (LR)

3.2. ODV360 Dataset for Omnidirectional Video Super-Resolution

To rectify the lack of high-quality video datasets in the community of omnidirectional video super-resolution, we create a new high-resolution (4K-8K) 360° video dataset, including two parts:

- 90 videos collected from YouTube and public 360° video dataset. These videos are carefully selected and have high quality to be used for restoration. All videos have the license of Creative Commons Attribution license (reuse allowed), and our dataset is used only for academic and research proposes.
- 160 videos collected by ourselves with Insta360 cameras. The cameras we use include Insta 360 X2 and Insta 360 ONE RS. They can capture high-resolution (5.7K) omnidirectional videos.

These collected omnidirectional videos cover a large range of diversity, and the video contents vary indoors and outdoors. To facilitate the use of these videos for research, we downsample the original videos into 2K resolution (2160x1080) by OpenCV. The number of frames per video is fixed at about 100. We randomly divide these videos into training, validation, and testing sets, as shown in Tab. 2. The dataset is publicly available on the 360SR challenge homepage.

Table 2. The detailed data partition of ODV360 dataset.

	Training	Validation	Testing	All
Numbers	210	20	20	250
Storage	59G (GT) 4.9G (LR)	5.3G (GT) 446M (LR)	5.7G (GT) 485M (LR)	75.8G

3.3. Realistic Omnidirectional DownSampling

In practice, ODIs are acquired by the fisheye lens and stored in ERP. Given that the low-resolution issue in real-world scenarios is caused by insufficient sensor precision and density, the downsampling process should be applied to original-formatted images before converting them into other storage types. Thus, to be conformed with real-world imaging processes, we propose to apply uniform bicubic downsampling on Fisheye images, which are the original format of ODIs. As shown in Fig. 1, the new downsampling process (called Fisheye downsampling [62]) applies uniform bicubic downsampling on Fisheye images before converting them to ERP images. This downsampling kernel is more conducive to exploring the geometric property of ODIs.

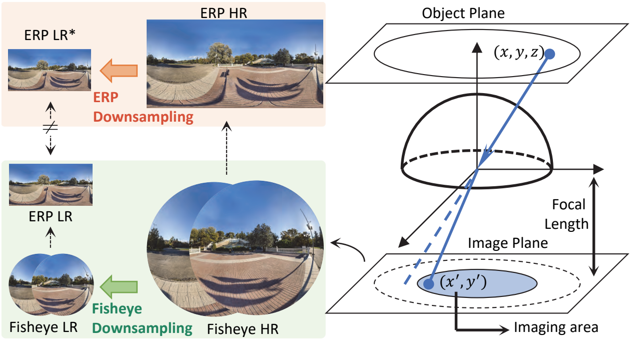

Figure 1. Downsampling process of ODIs (left) and imaging process in real-world (right). '*' denotes that LR images synthesized from different downsampling processes are inconsistent. This figure comes from [62].

Process of Fisheye downsampling. To generate more realistic LR ODIs, we mimic the real-world imaging process and apply bicubic downsampling on Fisheye images. One single Fisheye image can only store information about a hemisphere. Hence, ERP images are converted to dual Fisheye images. To keep the mean pixel density comparable, the resolutions of Fisheye images are slightly larger than that of ERP images. Before downsampling, Fisheye images are padded by a FOV larger than 180° to avoid edge disconnections. This padding operation will not influence the geometric transforming relation between ERP and Fisheye. As Fisheye data is unstructured and Fisheye distortion is more complicated than ERP distortion, we reconvert LR images to the ERP format for learning. All details of Fisheye downsampling can be found on GitHub [3].

We apply the realistic degradation in both the image and video tracks. For video, we further consider quality compression. More details can be found in the following section.

4. The NTIRE Challenge

This challenge has two tracks: omnidirectional image super-resolution (Track 1) and omnidirectional video super-resolution (Track 2).

[3] https://github.com/Fanghua-Yu/OSRT

4.1. Track 1: Omnidirectional Image Super-Resolution

Track 1 aims to super-resolve the LR omnidirectional images with a magnification factor of ×4.

Settings. Unlike previous settings that directly apply bicubic downsampling to generate low-resolution (LR) ERP images, we adopt a more realistic way to generate LR ERP images by considering the real acquisition process of 360° images. Since raw 360° images are captured by the fisheye lens and are then saved as fisheye formats, thus performing degradations on fisheye images is more realistic and reasonable.

Metrics. We evaluate the super-resolved 360° images by comparing them to the ground truth HR ERP images. To measure the fidelity, we adopt the widely used Weighted-to-Spherically-uniform Peak Signal to Noise Ratio (WS-PSNR) as the quantitative evaluation metric. We report the performance of these two baseline models on the validation/testing server.

Baseline Model. For the image track, we utilize the image super-resolution methods EDSR [38] and SwinIR [36] as the baseline. Meanwhile, we report the results of directly applying bicubic upsampling to super-resolve omnidirectional images.

4.2. Track 2: Omnidirectional Video Super-Resolution

Track 2 aims to super-resolve LR omnidirectional videos with a magnification factor of ×4.

Settings. In the process of generating low-resolution videos, we consider two main factors, *i.e.*, downsampling and video compression. For downsampling, we apply the same pipeline as track 1. For video compression, we use the H.264 codec rules in FFMPEG to generate the compressed video frames.

Metrics. We evaluate the super-resolved 360° video frames by comparing them to the ground truth HR ERP frames. To measure the fidelity, we adopt the widely used Weighted-to-Spherically-uniform Peak Signal to Noise Ratio (WS-PSNR) as the quantitative evaluation metric. The performances of the methods on the validation/testing server are reported.

Baseline Model. We utilize BasicVSR [6] as the baseline model for the video track.

4.3. Challenge phases

(1) Development and validation phase: The participants had access to all the training and validation pairs (LR/HR) of the Flickr360 dataset and ODV360 dataset. The details and guidelines are given on GitHub, allowing the participants to benchmark the performance of their models on their system. The participants could upload the HR validation results on the evaluation server to calculate the WSPSNR of the super-resolved image/video produced by their models to get immediate feedback.
(2) Testing phase: In the final test phase, the participants were granted access to the LR test set of Flickr360 and ODV360. The HR ground-truth images/videos are unreleased to participants. The participants then submitted their super-resolved results to the Codalab testing server and e-mailed the code and factsheet to the organizers. The organizers verified and ran the provided code to obtain the final results. Finally, the participants received the final results at the end of the challenge.

5. Challenge Results

There are 100 and 56 participants registered for the two challenge tracks, and 7 and 3 teams entered in the final testing phase and submitted their results, source codes, and fact sheets in the image track (track 1) and video track (track 2), respectively. Tab. 3 shows the main results on the validation and testing sets of these teams. The methods adopted by different teams are described in the following section, and the detailed information of these teams is listed in Appendix.

Track 1: Omnidirectional Image Super-Resolution From the upper part of Tab. 3, we see that most valid teams surpass the baseline models (*i.e.*, EDSR-M [38], SwinIR [36]) with a large margin. Most teams (except NTU607-360) adopt HAT [11] as the baseline model, and improve the performance by considering the characteristics of omnidirectional images and more efficient training strategies. Meanwhile, we observe that all teams' results are consistent in the validation and testing sets in terms of WS-PSNR metrics.

Track 2: Omnidirectional Video Super-Resolution
The results in Tab. 3 present that all valid teams outperform the baseline model, *i.e.*, BasicVSR [6]. At the same time, we obverse that all these methods are built based on BasicVSR++ [8]. The first team (MVideo) demonstrates the positive effect of the multi-stage training strategy. The second team (HIT-HL) presents a novel perspective by utilizing an adapter model to introduce the omnidirectional characteristics in the super-Resolution process. The third team (PKU_VILLA) proposes a multi-stage model to improve the model performance.

5.1. Conclusion

Omnidirectional Characteristics. The results show that all the teams in the image and video tracks considered the omnidirectional characteristics show performance improvement, which demonstrates the effectiveness of integrating

Table 3. Quantitative results of the NTIRE 2023 Challenge on 360° Omnidirectional Image and Video Super-Resolution.

Rank	Team Name	Author/Method	Flickr360/ODV360	
			WS-PSNR (Val)	WS-PSNR (Test)
Track 1: 360° Omnidirectional Image Super-Resolution (X4)				
1	BSR	Accusefive	30.43	28.64
2	Graphene	Bob072	30.20	28.49
3	HIT-IIL	Spider-Man	30.04	28.28
4	NTU607-360	HaoqiangYang	30.03	28.13
5	bee992	bee992	29.87	28.11
6	flowers	flowers	30.00	28.10
7	HIT-CVLab	luobingchun	29.60	27.65
	Baselines	Bicubic	27.45	25.74
		EDSR-M	29.18	27.30
		SwinIR	29.75	27.86
Track 2: 360° Omnidirectional Video Super-Resolution (X4)				
1	MVideo	cui666	25.89	26.35
2	HIT-IIL	Spider-Man	25.41	25.82
3	PKU_VILLA	eStarPro	25.04	25.47
	Baseline	BasicVSR	24.57	24.72

this information. Meanwhile, this also illustrates the differences between omnidirectional images/videos and ordinary images/videos, and considering the omnidirectional features is highly recommended for omnidirectional image processing.

Similarity to Plain Image/Video SR. From the results of the baseline models (*e.g.*, EDSR, SwinIR, BasicVSR++) and the methods improved by the participants, we see that the high-performance models for plain images also achieve high performance on omnidirectional images. Specifically, the HAT [12] achieves state-of-the-art performance in image super-resolution, and it also achieves highly competitive results on the Flickr360 dataset. Some teams further improve the performance by further considering the omnidirectional features.

5.2. Future Work

Omnidirectional image super-resolution aims to provide ordinary users with high-quality and immersive experiences in AR/VR applications. This challenge tries to promote the development of the field of omnidirectional image/video super-resolution, and the results show some promising results. However, some opening problems still required to be tackled and some promising directions can be considered to further improve the user experience.

Omnidirectional characteristics. Most existing methods still simply utilize the distortion map of ERP to improve the performance, while other characters like omnidirectional are still not considered for super-resolution, resulting the discontinuity between the left side and right side. Meanwhile, how to fuse the distortions to improve the performance effectively and efficiently still requires additional explorations.

High-resolution image processing. We have not restricted the runtime of the methods in this challenge, however, the high-quality VR/AR applications usually require much higher omnidirectional images (*i.e.*, larger than 8K), yet existing models still cannot super-resolve these images with such a high resolution. Meanwhile, processing these high-resolution images in head-mounted devices (HMD) is much more challenging with limited computational resources. Therefore, real-time omnidirectional processing is a promising direction for both the research community and industry.

Realistic degradation. In this challenge, we adopt a more realistic omnidirectional image downsampling method proposed in [62] to obtain better results in terms of real-captured images. However, this kind of simple degradation is far from the requirements, since there are many other degradations (*e.g.*, blur, noise) during capturing omnidirectional images. We should further consider the blind setting in plain image SR to obtain better performance in processing real-captured omnidirectional images. Meanwhile, the generative priors (*e.g.*, GAN, Diffusion models) can be further integrated into models to achieve better visual quality from the method perspective.

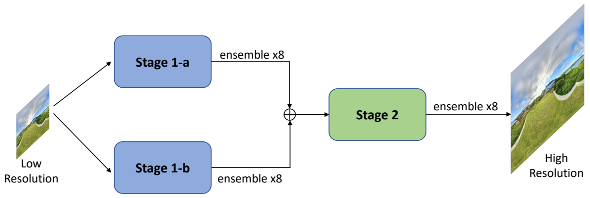

Figure 2. Overall pipeline of the method proposed by BSR team.

Omnidirectional image quality assessment. The WS-PSNR [50] is adopted to evaluate the performance of the super-resolved results. However, in real applications, the ERP images are usually projected into other types, like the perspective image for application, thus we may evaluate the quality under multiple projection types to better assess the quality of super-resolved omnidirectional images.

Future work on omnidirectional video super-resolution. In addition to the future work considered in the image domain, there are more factors that need to be considered in future work. First, there is a large gap in the temporal correlation between omnidirectional video and ordinary 2D video. How to take full advantage of the omnidirectional temporal correlation is still an open challenge. Second, compression coding is involved in omnidirectional video transmission, which is also different from the ordinary 2D video.

6. Challenge Methods

6.1. Track 1: Omnidirectional Image Super-Resolution

6.1.1 BSR Team

The BSR team [48] improved the results of the data processing and the SR model. The participants collected 360 image data from YouTube 360 videos and designed a degradation learning network inspired by AnimeSR [59]. For the degradation learning, a network was trained that downsamples by a factor of 4 using the given HR and LR samples to degrade the 360-degree data collected from YouTube.

Regarding the SR model, the authors adopted a two-stage model shown in Fig. 2, where the first stage is a super-resolution network, and the second stage is an equal-resolution enhancement network.

The SR First Stage Network (Fig.3) was designed based on HAT-L [12]. Specifically, inspired by DACB in [62] and BUSIFusion [31], a novel Omnidirectional Position-aware Deformable Block (OPDB) was proposed, which combines dimensional information and absolute position encoding information for 360-degree images. Additionally, a frequency fusion module was integrated at the end of the HAT blocks, and ultimately incorporated Fourier upsampling [31] to assist pixel shuffle. The SR Second Stage network shares the same structure as the first stage, with a pixel unshuffle downsampling operation, and is fine-tuned based on the first-stage weights.

6.1.2 Graphene Team

An Image Super-Resolution Transformer with Cross-Scale Attention (CSA) and Wavelet Hallucination (WT) was designed to solve the image super-resolution task. The framework is shown in Fig. 5.

In cross-scale attention, a 3 × 3 depth-wise convolution was added in the query transformation, and multi-scale depth-wise convolutions were added in the key and value transformation. Such a depth-wise convolution can extract local features and incorporate positional information into the visual tokens.

For the wavelet hallucination, input features were divided into different groups, each group was viewed as a frequency sub-band, feature enhancement was performed separately and then the feature was hallucinated with a wavelet reconstruction. After the wavelet hallucination, a downsampled convolution with stride 2 was applied to reduce the feature map size. By hallucinating features in different frequency sub-bands, a higher-resolution hidden feature was predicted. Such a hidden high-resolution feature helps to extract finer details.

Besides, the participants also presented horizontal attention. Considering the characteristic of the 360SR task, self-attention was introduced to perform in a horizontal way. Visual tokens in each X-axis were aggregated with the self-attention mechanism.

As for the training, a two-stage training strategy was used for efficiency. In the first stage, training images were cropped into patches with 64 × 64 randomly. The cropped image patch was used to train the network. In the second stage, the network was finetuned by changing the image patch size from 64 × 64 to 512 × 2048. The models were trained with L_1 loss.

6.1.3 HIT-IIL Team

For the super-resolution (SR) of equirectangular projection (ERP) images, it is not satisfactory when directly applying convolutional neural networks (CNN) with translational equivariance requirements. Instead, ERP-Adapter was proposed that injects a distortion-aware adapter into existing SR networks designed for perspective images, transferring the pre-trained perspective image SR network to fit ERP images with minimal additional effort.

Pre-trained HAT [12] was adopted as the perspective image SR network, which combines a self-attention and channel attention scheme for reconstruction. To transfer HAT into ERP domain, ERP-Adapter utilizes deformable convolution [15] for modulating ERP features. However, the di-

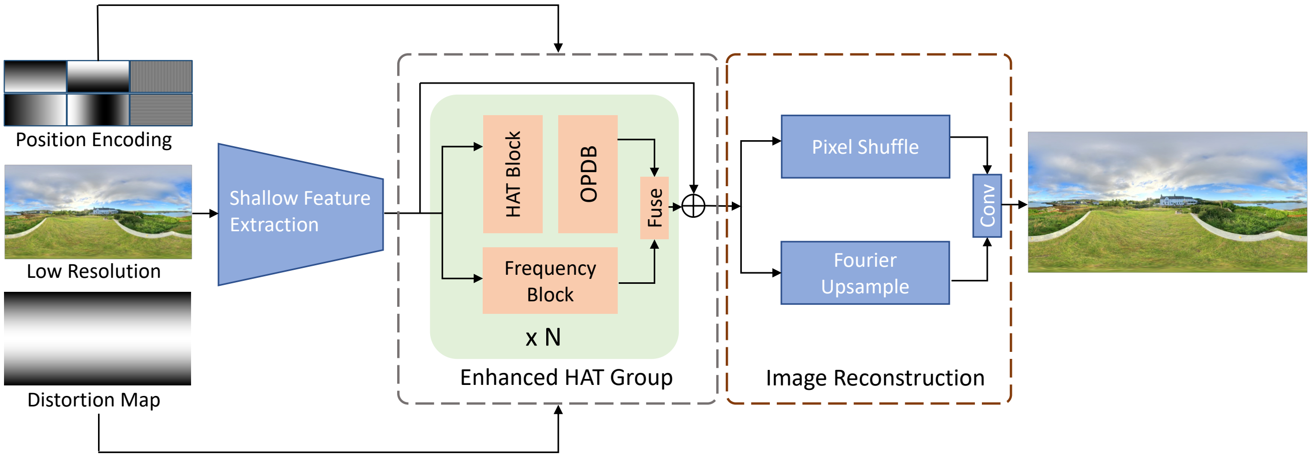

Figure 3. Network architecture proposed by BSR team for omnidirectional image super-resolution.

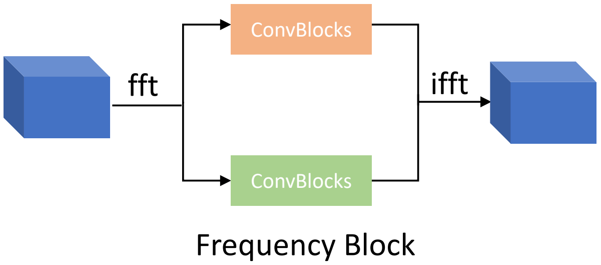

Figure 4. Illustration of our frequency block proposed by BSR team.

rect estimation of the offsets may bring instability to the network training. Inspired by SelfDZSR [72], an adaptive spatial transformer network (AdaSTN) was adopted to obtain offsets indirectly by estimating the pixel-level affine transformation matrix and translation vector. For every pixel, the predicted offset can be written as,

$$P = AG + b, \tag{1}$$

where $A \in \mathbb{R}^{2\times 2}$ is the estimated affine transformation matrix and $b \in \mathbb{R}^{2\times 1}$ is the translation vector. G is a positional grid, which can be expressed as,

$$\begin{bmatrix} -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \end{bmatrix}. \tag{2}$$

The sphere-to-plane projective distortion can be expressed as the distortion map [49] to some extent, which stores the per-pixel scaling factor from the 2D plane to the sphere and can be given by

$$C_{u,v} = cos((v - \frac{H}{2} + \frac{1}{2})\frac{\pi}{H}), \tag{3}$$

where H is the height of the low-resolution ERP image. The distortion map and the current features are fed into AdaSTN for generating offsets of deformable convolution. Finally, deformed features are added to the current features.

In the training stage, pre-trained HAT is fine-tuned on the given training dataset. Next, the ERP-Adapter branch is added to the HAT block (see Fig. 6), HAT is fixed and only the adapter is trained for additional iterations. Finally, a large patch size is found to be beneficial for performance improvement. Considering GPU memory, previous layers are fixed and tail layers are fine-tuned with the whole ERP image as input. In the testing stage, a self-ensemble strategy [52] is used for better performance.

6.1.4 NTU607-360 Team

The method is adapted from [67]. This model effectively extracts local structural information by shift convolution (shift-conv) while maintaining an identical level of complexity as a 1x1 convolution. The model also contains the group-wise multi-scale self-attention (GMSA) module, which calculates self-attention on non-overlapped groups of features by various window sizes to achieve long-range image dependency. A highly efficient long-range attention block (ELAB) is then built by simply cascading two shift-conv with a GMSA module, which is further accelerated by using a shared attention mechanism. Different from the image super-resolution model, the distortion map is also added as the input to capture the spatial information. The distortion map is written as

$$C_d = cos(\frac{m + 0.5 - M/2}{M}\pi) \tag{4}$$

where M and m are the height of the image and the current height of the image. The LR image and distortion map are concatenated as the input to train the overall network.

The proposed network contains a 3x3 convolution to extract features from the distortion map and the image. The

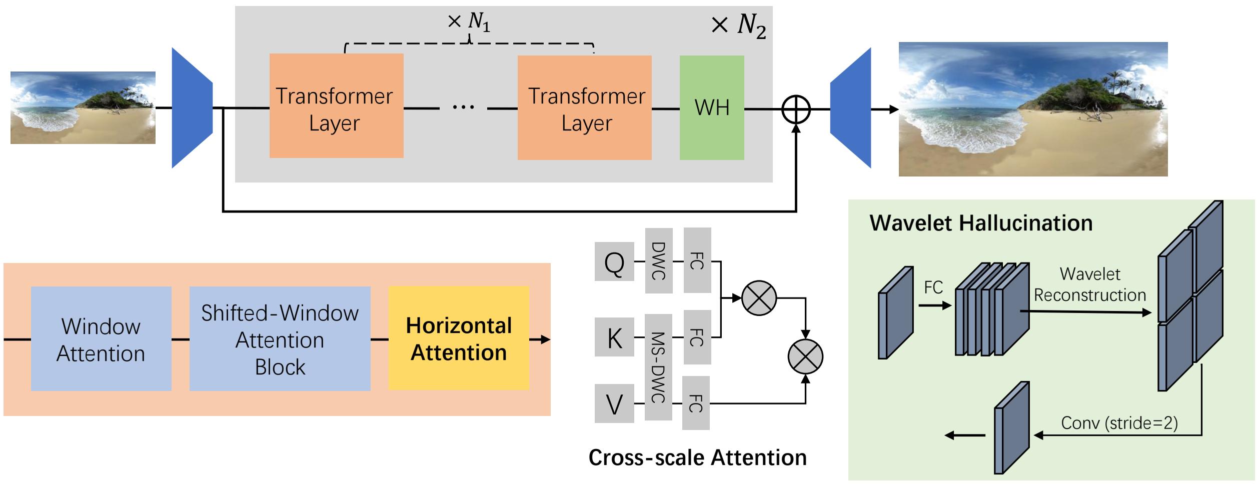

Figure 5. The pipeline of the method proposed by the Graphene team. A Transformer-based architecture is adopted. They propose horizontal attention, cross-scale attention, and wavelet hallucination to enhance feature extraction.

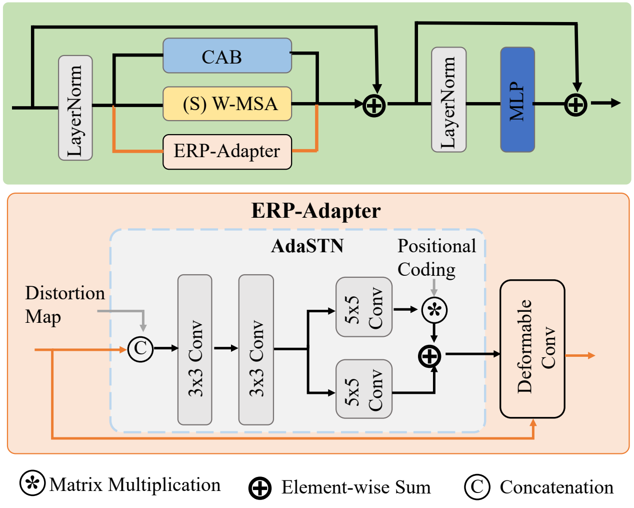

Figure 6. The architecture of ERP-Adapter. We inject an adapter into the basic block from HAT [12]. The adapter takes the distortion map and current features as input, and adopts AdaSTN [72] to estimate offsets of current features.

middle 36 ELAB [67] blocks refine features with efficient complexity. The final blocks are a 3x3 convolution and the pixel-shuffle operation to restore and magnify the image 4 times.

6.1.5 bee992 Team

The network used by the author is HAT (Hybrid Attention Transformer) [11]. The HAT method combines channel attention and self-attention together to show powerful performance in real images super-resolution. In the task of omnidirectional image super-resolution, the author simply tried to use the HAT method to solve this problem.

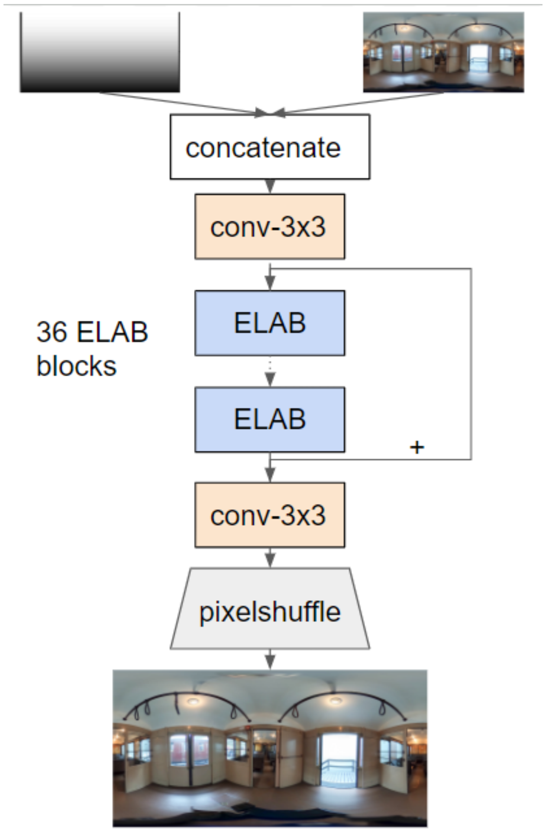

Figure 7. Systematic of the proposed model by NTU607-360 team.

6.1.6 flowers Team

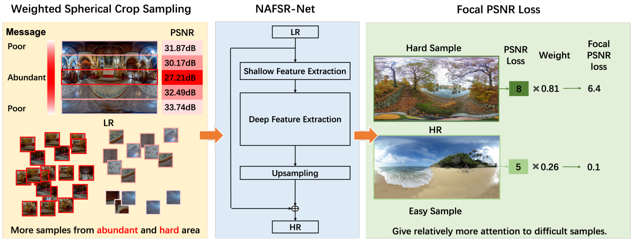

Figure 8. The method proposed by flowers team.

The authors' network is based on NAFNet [10] and consists of three modules: the shallow feature extraction module, the deep feature extraction network, and the upsampling module. To gather more detailed information, the global residuals is directly incorporated into the feature extraction process to avoid losing any valuable details. Additionally, To enhance the model's generalization and prevent overfitting, the droppath technique [25] has been used.

Weighted Spherical Crop Sampling (WS-Crop) method is proposed to solve the inconsistencies in density within panoramic regions problem. This method enhances the sampling rate in high-density areas, allowing the training process to emphasize challenging areas and extract more relevant information. WS-Crop samples with an additional sample drop probability P

$$\mathcal{P}(h_c) = 1 - \alpha \cos \frac{(h_c + 0.5 - H/2)\pi}{H}. \tag{5}$$

In the above, h_c is the height of the candidate sample's center point. α is set to 0.8 as a modulation coefficient, which is used to prevent the central area from always being preserved, and also to adjust the degree of difference in region drop probability. When sampling randomly, the weighted spherical coefficient calculated from the height of the sample center point determines the probability of retaining a sample. The closer to the middle area, the higher its probability of being retained.

Furthermore, their experimentation indicates a substantial disparity in texture complexity among images in the Flicker 360 dataset, with apparent distinctions between simple and challenging samples. The **Focal PSNR Loss** is proposed to enrich the quality of intricate texture generation. This function allocates higher weights to challenging samples with significant PSNR loss, and lower weights to simpler samples with lower PSNR loss.

$$\begin{aligned}
MSE &= \frac{1}{HW} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} [pred(i,j) - gt(i,j)]^2, \\
\mathcal{L}_{PSNR} &= 10 \cdot \log_{10} \left(\frac{MAX^2}{MSE} \right), \\
\mathcal{L}_{FocalPSNR} &= \sigma(\gamma \mathcal{L}_{PSNR} + \beta) \mathcal{L}_{PSNR}.
\end{aligned} \tag{6}$$

In the above H and W represent the height and width of the high-resolution image, and MAX represents the maximum pixel value of the image, usually 255. σ is the sigmoid function. Hyperparameters γ and β can either be manually adjusted to fit different tasks, or dynamically learned by the networks themselves. By observing the distribution of experimental data, hyperparameters are set as $\gamma = 0.66$, $\beta = 6.2$.

Due to the extensive training time required for the super-resolution network, a three-stage fine-tuning technique has been implemented, which gradually increases the image sizes, fine-tunes the learning rates, and utilizes WS-Crop sampling and Focal PSNR Loss at different stages. This approach enables the network to integrate complex image details effectively.

To ensure the effective performance of the models across a range of low-resolution image scenarios, an ensemble learning approach was employed. Several versions of each model with diverse initialization, hyperparameters, training data augmentation techniques, and sizes were combined to produce the final result.

6.1.7 HIT-CVLab Team

The adopted method is based on HAT [12]. The feature dimension was set to 192, the block numbers were set to [6,6,6,6,6,6], the number of multi-heads was set to [8,8,8,8,8,8], and the window size was set to 16. The method combines channel attention and shifted window attention schemes to exploit their complementary advantages. In addition, overlapping cross-attention modules were introduced to better modulate cross-window information adaptively and enhance the interaction between features.

6.2. Track 2: Omnidirectional Video Super-Resolution

6.2.1 MVideo Team

Similar to other common Video Super-Resolution (VSR) pipelines such as BasicVSR [7], the authors' models E2VSR also contain four basic functionalities: propagation, alignment, aggregation, and upsampling. The overall architecture is shown in Fig. 9.

Following BasicVSR++ [8], the authors utilized second-order grid propagation and flow-guided deformable align-

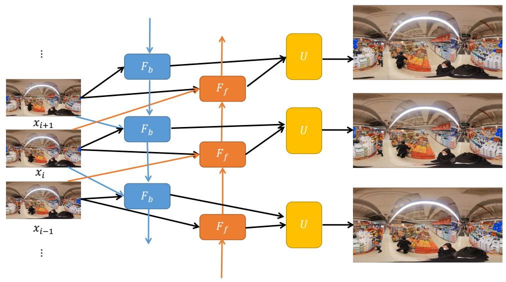

Figure 9. Architecture of E2VSR for 360 VSR Challenge

ment that allows information to be propagated and aggregated more effectively. The second-order grid propagation ameliorates information flow in the network and improves the robustness of the network against occluded and fine regions. The flow-guided deformable alignment reduces the burden of offset learning by using the optical flow field as base offsets refined by flow field residue.

The models adopt a multi-stage training strategy, and the training process includes four stages:

(1) Train the first stage model using 10 LR frames as inputs with batch size 8 and patch size 256 for 60k iterations. The model is optimized using Charbonnier loss with Adam optimizer and Cosine Annealing scheme, with an initial learning rate of 1e-4.

(2) Finetuning the model using the pre-trained weights from the stage (1) model. The batch size is 8 and the patch size is 256(HR), 10 LR frames are used as inputs, total iteration number is 60k. The model is optimized using MSE loss with Adam optimizer and Cosine Annealing scheme. And the initial learning rate is 1e-5.

(3) Continue to finetune the model from Stage (2) using MSE loss using 10 LR frames as inputs, batch size 8 and patch size 512(HR), and optimizer type stay the same. The initial learning rate of 1e-6.

(4) Finally, finetuning the model from Stage (3) using 30 LR frames as inputs to contain more information, with patch size 256 and batch size 8. This stage takes 60k iterations with a learning rate is 1e-6.

The training configurations, testing strategies, and hardware and software settings of E2VSR are described as follows. The Adam optimizer and Cosine Annealing scheme are adopted for training. The initial learning rate of the main network and the flow network are set to 1e-4 and 2.5e-5, respectively. The total number of iterations of each stage is 60K, and the weights of the flow network are fixed during the first 5,000 iterations. The batch size is 8 and the patch size is 256×256(HR) or 512×512(HR) randomly cropped from input frames. The network is trained on train datasets of ODV360, and has no additional data. In order to further improve the performance of the model, the number of residual blocks for each branch and feature channel is set to 15 and 256, respectively.

In the training process, Exponential Moving Average (EMA) is used to improve the robustness of the model in test datasets. In the testing process, Test-Time Augmentation(TTA) is used to improve the correction of predicted results and reduce the generalization error. Training is all done on the Nvidia A100 GPU server, with GPU memory of 40GB each.

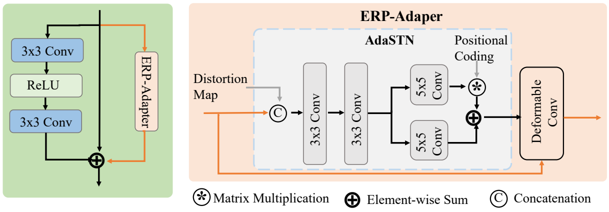

Figure 10. The architecture of ERP-Adapter. The authors inject an adapter into the basic block from BasicVSR++ [8]. The adapter takes the distortion map and current features as input, and adopts AdaSTN [72] to estimate offsets of current features.

6.2.2 HIT-IIL Team

Compared to perspective videos, Equirectangular projection (ERP) videos stretch the areas near the poles horizontally. Thus, for ERP video super-resolution (VSR), it is not satisfactory when directly applying convolutional neural networks (CNN) with translational equivariance requirements. Instead, we propose ERP-Adapter that injects a distortion-aware adapter into existing VSR networks designed for perspective videos, transferring the pre-trained perspective VSR network to fit ERP videos with minimal additional effort.

Detailly, BasicVSR++ [8] is adopted as the perspective VSR network, which adopts second-order grid propagation for better temporal modeling and flow-guided deformable alignment for stable offsets estimation. For performance improvement, the authors replace the reconstruction residual blocks with transformer blocks [64].

To transfer BasicVSR++ into the ERP domain, ERP-Adapter utilizes deformable convolution [15] for modulating ERP features. However, the direct estimation of the offsets may bring instability to the network training. Inspired by SelfDZSR [72], an adaptive spatial transformer network (AdaSTN) is designed to obtain offset indirectly by estimating the pixel-level affine transformation matrix and translation vector, as shown in Fig10. For every pixel, the pre-

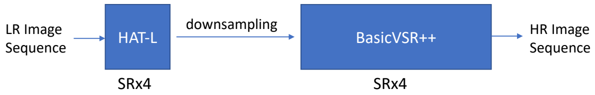

Figure 11. Pipeline of the PKU_VILLA Team.

dicted offset can be written as,

$$P = AG + b \tag{7}$$

where $A \in \mathbb{R}^{2\times2}$ is the estimated affine transformation matrix and $b \in \mathbb{R}^{2\times1}$ is the translation vector. G is a positional grid, which can be expressed as,

$$\begin{bmatrix} -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \end{bmatrix} \tag{8}$$

The sphere-to-plane projective distortion can be expressed as distortion map [49] to, which stores the per-pixel scaling factor from the 2-D plane to the sphere and can be given by

$$C_{u,v} = cos((v - \frac{H}{2} + \frac{1}{2})\frac{\pi}{H}) \tag{9}$$

where H is the height of the low-resolution ERP frame. The distortion map and the current features are fed into AdaSTN for generating offsets of deformable convolution. Finally, deformed features are added to the current features.

In the training stage, BasicVSR++ is first trained with Charbonnier loss [29]. Next, the BasicVSR++ is fixed and only the adapter is trained for additional iterations. In the testing stage, a self-ensemble strategy [52] is further adopted for better performance.

6.2.3 PKU_VILLA Team

A spatial-temporal two-stage model is developed, wherein the first stage is a 4x image super-resolution network, and the second stage is a 4x video super-resolution network. The overall pipeline has been shown in Fig.11.

In the first stage, the HAT-L [11] model architecture and its pre-trained weights are utilized for fine-tuning. Sequentially, video frames were input into the network for compressed image restoration and 4x image super-resolution tasks. The output images from the first stage underwent a PixelUnShuffle operation to ensure resolution consistency with the input.

During the second stage, BasicVSR++ [8] model architecture and pre-trained weights are employed for fine-tuning. Before feeding into the second-stage network, a 1x1 convolution is used to compress the channel number. To fully exploit the temporal information in the video, the authors input the first stage's output into the second stage video super-resolution network to obtain the predicted high-resolution video frames.

Acknowledgments We sincerely appreciate the generous donations contributed by Intellindust Information Technology Co., Ltd. We thank the NTIRE 2023 workshop and challenges sponsors: Sony Interactive Entertainment, Meta Reality Labs, ModelScope, ETH Zürich (Computer Vision Lab) and University of Würzburg (Computer Vision Lab).

A. Organizers of NTIRE 2023 Challenge

Title:
NTIRE 2023 Challenge on 360° Omnidirectional Image and Video Super-Resolution: Datasets, Methods and Results
Members:
Mingdeng Cao[1,2] *(mingdengcao@gmail.com)*, *Chong Mou*[2,3]*(eechongm@gmail.com)*, Fanghua Yu[4], Xintao Wang[2], Yinqiang Zheng[1], Jian Zhang[3], Chao Dong[4], Gen Li[5], Ying Shan[2], Radu Timofte[6]
$*$ Mingdeng Cao is mainly responsible for the image track, while Chong Mou is mainly responsible for the video track.
Affiliations:
[1] The University of Tokyo
[2] ARC Lab, Tencent PCG
[3] Peking University
[4] Shenzhen Institute of Advanced Technology, CAS
[5] Platform Technologies, Tencent Online Video
[6] Computer Vision Lab, IFI & CAIDAS, University of Würzburg
[7] Computer Vision Lab, ETHZürich

B. Track 1: Teams and Affiliations

BSR team

Title:
OPDN: Omnidirectional Position-awared Deformable Network
Members:
Xiaopeng Sun[1] *(sunxiaopeng.01@bytedance.com)*, Weiqi Li2, Zhenyu Zhang[2], Xuhan Sheng[2], Bin Chen[2], Haoyu Ma[1], Ming Cheng[1], Shijie Zhao[1]
Affiliations:
[1] ByteDance
[2] Peking University Shenzhen Graduate School

Graphene team

Title:
360° Omnidirectional Image Super-Resolution Transformer with Cross-Scale Attention and Wavelet Hallucination
Members:
Huaibo Huang[1,2]*, mail: huaibo.huang@cripac.ia.ac.cn*, Xiaoqiang Zhou[1,3], Yuang Ai[1,4], Ran He[1,2,5]
Affiliations:

[1] MAIS&CRIPAC, Institute of Automation, Chinese Academy of Sciences
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] University of Science and Technology of China
[4] Beijing Institute of Technology
[5] School of Information Science and Technology ShanghaiTech University

HIT-IIL team

Title:
ERP-Adapter: Distortion-Aware Adapter for Omnidirectional Image Super-Resolution
Members:
Renlong Wu, mail: hirenlongwu@gmail.com, Yi Yang, Zhilu Zhang, Shuohao Zhang, Junyi Li, Yunjin Chen, Dongwei Ren, Wangmeng Zuo
Affiliations:
Harbin Institute of Technology

NTU607-360 team

Title:
Efficient Long-Range Attention Network for 360° Omnidirectional Super-resolution
Members:
Hao-Hsiang Yang[2] (islike8399@gmail.com), Yi-Chung Chen[3], Zhi-Kai Huang[2], Wei-Ting Chen[1], Yuan-Chun Chiang[2], Hua-En Chang[2], I-Hsiang Chen[2], Chia-Hsuan Hsieh[4], Sy-Yen Kuo[2]
Affiliations:
[1]Graduate Institute of Electronics Engineering, National Taiwan University
[2]Department of Electrical Engineering, National Taiwan University
[3]Graduate Institute of Communication Engineering, National Taiwan University
[4]ServiceNow

bee992 team

Title:
Using Pretrained Hybrid Attention Transformer for 360° Omnidirectional Image Super-Resolution
Members:
Zebin Zhang, mail: bin06212213@gmail.com
Affiliations:
ShanghaiTech University

flowers team

Title:
Members:
Jiaqi Zhang, mail: zhangjiaqi23@meituan.com, Yuhui Wang, Shuhao Cui, Junshi Huang, Li Zhu, Shuman Tian
Affiliations:
Meituan

HIT-CVLab team

Title:
HAT-360
Members:
Wei Yu, mail: 20b903014@stu.hit.edu.cn, Bingchun Luo, mail: 2201110120@stu.hit.edu.cn
Affiliations:
Harbin Institute of Technology

C. Track 2: Teams and Affiliations

MVideo Team

Title:
Effective and Efficient Network for 360° Omnidirectional Video Super Resolution
Members:
Wanwan Cui, mail (cuiwanwan@xiaomi.com), Tianyu Xu, Chunyang Li, Long Bao, Heng Sun
Affiliations:
Xiaomi Inc

HIT-IIL Team

Title:
ERP-Adapter: Distortion-Aware Adapter for Omnidirectional Video Super-Resolution
Members:
Renlong Wu (hirenlongwu@gmail.com), Yi Yang, Zhilu Zhang, Shuohao Zhang, Junyi Li, Yunjin Chen, Dongwei Ren, Wangmeng Zuo
Affiliations:
Harbin Institute of Technology

PKU_VILLA Team

Title:
HAT BVSRPP
Members:
Zhenyu Zhang (zhenyuzhang@pku.edu.cn), Qian Wang, Weiqi Li, Xuhan Sheng, Bin Chen
Affiliations:
Peking University Shenzhen Graduate School

References

[1] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2023 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[2] Zafer Arican and Pascal Frossard. L1 regularized super-resolution from unregistered omnidirectional images. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 829–832. IEEE, 2009. 2

[3] Zafer Arican and Pascal Frossard. Joint registration and super-resolution with omnidirectional images. *IEEE Transactions on Image Processing*, 20(11):3151–3162, 2011. 2

[4] Luigi Bagnato, Yannick Boursier, Pascal Frossard, and Pierre Vandergheynst. Plenoptic based super-resolution for omnidirectional image sequences. In *2010 IEEE International Conference on Image Processing*, pages 2829–2832. IEEE, 2010. 2

[5] 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. 2

[6] Kelvin C.K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 2, 4

[7] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021. 9

[8] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5972–5981, 2022. 2, 4, 9, 10, 11

[9] 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. 2

[10] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Proceedings of the European Conference on Computer Vision*, pages 17–33. Springer, 2022. 9

[11] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*, 2022. 2, 4, 8, 11

[12] X Chen, X Wang, J Zhou, and C Dong. Activating more pixels in image super-resolution transformer. arxiv 2022. *arXiv preprint arXiv:2205.04437*, 2022. 5, 6, 8, 9

[13] Marcos V Conde, Manuel Kolmet, Tim Seizinger, Thomas E. Bishop, Radu Timofte, et al. Lens-to-lens bokeh effect transformation. NTIRE 2023 challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[14] Marcos V Conde, Eduard Zamfir, Radu Timofte, et al. Efficient deep models for real-time 4k image super-resolution. NTIRE 2023 benchmark and report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[15] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 6, 10

[16] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 2

[17] Mallesham Dasari, Arani Bhattacharya, Santiago Vargas, Pranjal Sahu, Aruna Balasubramanian, and Samir R Das. Streaming 360-degree videos using super-resolution. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 1977–1986. IEEE, 2020. 2

[18] 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. 1, 2

[19] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 184–199. Springer, 2014. 1

[20] 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. 2

[21] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *European conference on computer vision*, pages 391–407. Springer, 2016. 2

[22] Vida Fakour-Sevom, Esin Guldogan, and Joni-Kristian Kämäräinen. 360 panorama super-resolution using deep convolutional networks. In *Int. Conf. on Computer Vision Theory and Applications (VISAPP)*, volume 1, 2018. 2

[23] Dario Fuoli, Shuhang Gu, and Radu Timofte. Efficient video super-resolution through recurrent latent space propagation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3476–3485. IEEE, 2019. 2

[24] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3897–3906, 2019. 2

[25] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Pro-*

ceedings of the European Conference on Computer Vision, pages 646–661. Springer, 2016. 9

[26] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. *Advances in neural information processing systems*, 28, 2015. 2

[27] Xiaoyang Kang, Xianhui Lin, Kai Zhang, Zheng Hui, Wangmeng Xiang, Jun-Yan He, Xiaoming Li, Peiran Ren, Xuansong Xie, Radu Timofte, et al. NTIRE 2023 video colorization challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[28] Hiroshi Kawasaki, Katsushi Ikeuchi, and Masao Sakauchi. Super-resolution omnidirectional camera images using spatio-temporal analysis. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science)*, 89(6):47–59, 2006. 2

[29] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2599–2613, 2018. 11

[30] 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. 2

[31] 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. 6

[32] 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. 2

[33] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[34] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on image denoising: Methods and results. In *CVPRW*, 2023. 2

[35] Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Jinjin Gu, Yu Qiao, and Chao Dong. Blueprint separable residual network for efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 833–843, 2022. 2

[36] 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, 4

[37] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021. 2

[38] 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, 4

[39] 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. 2

[40] Hongying Liu, Zhubo Ruan, Chaowei Fang, Peng Zhao, Fanhua Shang, Yuanyuan Liu, and Lijun Wang. A single frame and multi-frame joint network for 360-degree panorama video super-resolution. *arXiv preprint arXiv:2008.10320*, 2020. 2

[41] Xiaohong Liu, Xiongkuo Min, Wei Sun, Yulun Zhang, Kai Zhang, Radu Timofte, Guangtao Zhai, Yixuan Gao, Yuqin Cao, Tengchuan Kou, Yunlong Dong, Ziheng Jia, et al. NTIRE 2023 quality assessment of video enhancement challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[42] 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. 2

[43] Hajime Nagahara, Yasushi Yagi, and Masahiko Yachida. Super-resolution from an omnidirectional image sequence. In *2000 26th Annual Conference of the IEEE Industrial Electronics Society. IECON 2000. 2000 IEEE International Conference on Industrial Electronics, Control and Instrumentation. 21st Century Technologies*, volume 4, pages 2559–2564. IEEE, 2000. 2

[44] Akito Nishiyama, Satoshi Ikehata, and Kiyoharu Aizawa. 360 single image super resolution via distortion-aware network and distorted perspective images. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1829–1833. IEEE, 2021. 2

[45] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European conference on computer vision*, pages 191–207. Springer, 2020. 2

[46] Cagri Ozcinar, Aakanksha Rana, and Aljosa Smolic. Super-resolution of omnidirectional images using adversarial learning. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2019. 2

[47] Alina Shutova, Egor Ershov, Georgy Perevozchikov, Ivan A Ermakov, Nikola Banic, Radu Timofte, Richard Collins, Maria Efimova, Arseniy Terekhin, et al. NTIRE 2023 challenge on night photography rendering. In *CVPRW*, 2023. 2

[48] Xiaopeng Sun, Weiqi Li, Zhenyu Zhang, Qiufang Ma, Xuhan Sheng, Ming Cheng, Haoyu Ma, Shijie Zhao, Jian Zhang, Junlin Li, and Li Zhang. OPDN: Omnidirectional position-aware deformable network for omnidirectional image super-resolution. In *CVPRW*, 2023. 6

[49] Y Sun, A Lu, and L Yu. Ahg8: Ws-psnr for 360 video objective quality evaluation. In *Joint Video Exploration Team*

of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0040, 4th Meeting, 2016. 7, 11

[50] 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. 6

[51] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3360–3369, 2020. 2

[52] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1865–1873, 2016. 7, 11

[53] Florin-Alexandru Vasluianu, Tim Seizinger, Radu Timofte, et al. NTIRE 2023 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[54] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, et al. NTIRE 2023 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[55] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[56] 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. 2

[57] 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 workshops*, pages 0–0, 2018. 2

[58] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2023 challenge on light field image super-resolution: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[59] Yanze Wu, Xintao Wang, Gen Li, and Ying Shan. Animesr: Learning real-world super-resolution models for animation videos. *arXiv preprint arXiv:2206.07038*, 2022. 6

[60] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127:1106–1125, 2019. 2

[61] 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. 1, 2

[62] Fanghua Yu, Xintao Wang, Mingdeng Cao, Gen Li, Ying Shan, and Chao Dong. Osrt: Omnidirectional image super-resolution with distortion-aware transformer. *arXiv preprint arXiv:2302.03453*, 2023. 1, 2, 3, 5, 6

[63] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, et al. NTIRE 2023 challenge on hr depth from images of specular and transparent surfaces. In *CVPRW*, 2023. 2

[64] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 10

[65] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 2

[66] 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. 2

[67] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 649–667. Springer, 2022. 7, 8

[68] 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*, pages 286–301, 2018. 1

[69] 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*, pages 286–301, 2018. 2

[70] 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. 2

[71] Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, Radu Timofte, et al. NTIRE 2023 challenge on image super-resolution (x4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2

[72] Zhilu Zhang, Ruohao Wang, Hongzhi Zhang, Yunjin Chen, and Wangmeng Zuo. Self-supervised learning for real-world super-resolution from dual zoomed observations. In *European Conference Computer Vision*, pages 610–627. Springer, 2022. 7, 8, 10