OSRT: Omnidirectional Image Super-Resolution with Distortion-aware Transformer

Fanghua Yu[1*] Xintao Wang[2*] Mingdeng Cao[2,3] Gen Li[4] Ying Shan[2] Chao Dong[1,5†]
[1]ShenZhen Key Lab of Computer Vision and Pattern Recognition
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]ARC, Tencent PCG [3]The University of Tokyo
[4]Platform Technologies, Tencent Online Video [5]Shanghai Artificial Intelligence Laboratory
fanghuayu96@gmail.com, xintaowang@tencent.com, cmd@g.ecc.u-tokyo.ac.jp
{genli, yingsshan}@tencent.com, chao.dong@siat.ac.cn

Abstract

Omnidirectional images (ODIs) have obtained lots of research interest for immersive experiences. Although ODIs require extremely high resolution to capture details of the entire scene, the resolutions of most ODIs are insufficient. Previous methods attempt to solve this issue by image super-resolution (SR) on equirectangular projection (ERP) images. However, they omit geometric properties of ERP in the degradation process, and their models can hardly generalize to real ERP images. In this paper, we propose Fisheye downsampling, which mimics the real-world imaging process and synthesizes more realistic low-resolution samples. Then we design a distortion-aware Transformer (OSRT) to modulate ERP distortions continuously and self-adaptively. Without a cumbersome process, OSRT outperforms previous methods by about 0.2dB on PSNR. Moreover, we propose a convenient data augmentation strategy, which synthesizes pseudo ERP images from plain images. This simple strategy can alleviate the over-fitting problem of large networks and significantly boost the performance of ODISR. Extensive experiments have demonstrated the state-of-the-art performance of our OSRT.

1. Introduction

In pursuit of the realistic visual experience, omnidirectional images (ODIs), also known as 360° images or panoramic images, have obtained lots of research interest in the computer vision community. In reality, we usually view ODIs with a narrow field-of-view (FOV), *e.g.*, viewing in a headset. To capture details of the entire scene, ODIs require extremely high resolution, *e.g.*, 4K × 8K [1]. However, due to the high industrial cost of camera sensors with high precision, the resolutions of most ODIs are insufficient.

Recently, some attempts have been made to solve this problem by image super-resolution (SR) [12, 15, 28, 39, 40].

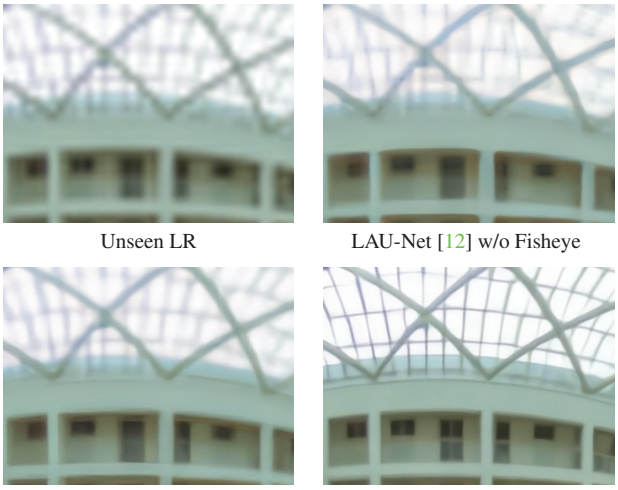

Figure 1. Visual comparisons of ×8 SR results on LR images[1] with unknown degradations. Fisheye denotes that the downsampling process in training stages is under Fisheye images.

As most of the ODIs are stored and transmitted in the equirectangular projection (ERP) type, the SR process is usually performed on the ERP images. To generate high-/low-resolution training pairs, existing ODISR methods [12, 15, 28, 39, 40] directly apply uniform bicubic downsampling on the original ERP images (called ERP downsampling), which is identical to general image SR settings [24, 43]. While omitting geometric properties of ERP in the degradation process, their models can hardly generalize to real ERP images. We can observe missing structures and blur textures in Fig. 1. Therefore, we need a more appropriate degradation model before studying SR algorithms. 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 into other storage types. Thus, to be conformed with real-world imaging processes, we propose to apply uniform bicubic

*Equal contribution
†Corresponding author (e-mail: chao.dong@siat.ac.cn)

[1]Photoed by Peter Leth on Flickr, with CC license.

downsampling on Fisheye images, which are the original format of ODIs. The new downsampling process (called Fisheye downsampling) applies uniform bicubic downsampling on Fisheye images before converting them to ERP images. Our Fisheye downsampling is more conducive to exploring the geometric property of ODIs.

The key issue of ODISR algorithm design is to utilize the geometric properties of ERP images, which is also the focus of previous methods. For example, Nishiyama *et al.* [28] add a distortion-related condition as an additional input. LAU-Net [12] splits the whole ERP image into patches by latitude band and learns upscaling processes separately. However, the separated learning process will lead to information disconnection between adjacent patches. SphereSR [40] learns different upscaling functions on various projection types, but will inevitably introduce multiple-time computation costs. To push the performance upper bound, we propose the first Transformer for Omnidirectional image Super-Resolution (OSRT), and incorporate geometric properties in a distortion-aware manner. Specifically, to modulate distorted feature maps, we implement feature-level warping, in which offsets are learned from latitude conditions. In OSRT, we introduce two dedicated blocks to adapt latitude-related distortion: distortion-aware attention block (DAAB), and distortion-aware convolution block (DACB). DAAB and DACB are designed to perform distortion modulation in arbitrary Transformers and ConvNets. These two blocks can directly replace the multi-head self-attention block and convolution layer, respectively. The benefit of DAAB and DACB can be further improved when being inserted into the same backbone network. OSRT outperforms previous methods by about 0.2dB on PSNR (Tab. 2).

However, the increase of network capacity will also enlarge the overfitting problem of ODISR, which is rarely mentioned before. The largest ODIs dataset [12] contains only 1K images, which cannot provide enough diversity for training Transformers. Given that acquiring ODIs requires expensive equipment and tedious work, we propose to generate distorted ERP samples from plain images for data augmentation. In practice, we regard a plain image as a sampled perspective, and project it back to the ERP format. Then we can introduce 146K additional training patches, 6 times of the previous dataset. This simple strategy can significantly boost the performance of ODISR (Tab. 4) and alleviate the over-fitting problem of large networks (Fig. 9). A similar data augmentation method is also applied in Nishiyama *et al.* [28], but shows marginal improvement on small models under ERP downsampling settings.

Our contributions are threefold. **1**) For *problem formulation*: To generate more realistic ERP low-resolution images, we propose Fisheye downsampling, which mimics the real-world imaging process. **2**) For *method*: Combined with the geometric properties of ERP, we design a distortion-aware Transformer, which modulates distortions continuously and self-adaptively without cumbersome process. **3**) For *data*: To reduce overfitting, we propose a convenient data augmentation strategy, which synthesizes pseudo ERP images from plain images. Extensive experiments have demonstrated the state-of-the-art performance of our OSRT[2].

2. Related Work

Single Image Super-Resolution (SISR). Deep learning for single image SR (SISR) is first introduced in [13]. Further works boost SR performance by CNNs [11, 14, 22, 24, 26, 29, 43], Vision Transformers (ViTs) [7, 8, 21, 23] and generative adversarial networks (GANs) [20, 35, 36, 42]. For instance, EDSR [24] removes Batch Normalization layers and applies a more complicated residual block. RCAN [43] introduces channel-wise attention mechanisms to a deeper network. SwinIR [23] proposes an image restoration Transformer based on [25]. To improve perceptual quality, adversarial training are performed as a tuning process to generate more realistic results [35, 36]. Moreover, various flexible degradation models are proposed in [35, 41] to synthesize more practical degradations.

Omnidirectional Image Super-Resolution (ODISR). Initially, ODISR models focus on the spherical assembling of LR ODIs under various projection types [2–4, 17, 27]. Recent ODISR models are performed on plane images and are fine-tuned from existing SISR models with L1 loss [15] or GAN loss [30, 44]. The improvements are limited, for they only concern the distribution gap between ODIs and plain images. Since LAU-Net [12] found pixel density in ERP ODIs is non-uniform, many studies attempt to design specific backbone networks to overcome this issue. LAU-Net [12] manually splits the whole ERP image into latitude-related patches and learns ERP distortion over different latitude ranges separately. While LAU-Net learns latitude-related ERP distortion somewhat, its non-overlapped patches lead to disconnection in whole ERP images. Nishiyama *et al.* [28] treats area stretching ratio as additional input. However, these conditions are tough to be utilized with an unmodified SISR backbone network. SphereSR [40] learns upsampling processes on various projection types to mitigate the influence of non-uniformity in specific projection types. Although SphereSR improves information consistency between various ODI projection types by LIIF [9], they apply multiple networks to learn the upscaling process of each projection type. Given that all other projection types in SphereSR are converted from ERP, patterns under various types are reusable when distortions are properly rectified. Moreover, the complex and unstructured image data in polyhedron projection hinders further research of ODISR.

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

Deformable Mechanism. Dai *et al.* [30] first propose deformable convolutions to obtain information out of its regular neighborhood. Xia *et al.* [38] further verified that Vision Transformers also benefit from applying deformable mechanisms on self-attention blocks. In Video SR tasks, the deformable mechanism can be adapted to align features between adjacent frames [5, 6, 34].

3. Method

In this section, we first analyze the cause of ERP and Fisheye distortions, as well as the relationship between these two distortions (Sec. 3.1). Then, we discuss the designs of Fisheye downsampling (Sec. 3.2), distortion-aware Transformer (OSRT) (Sec. 3.3), and the convenient data augmentation strategy (Sec. 3.4).

3.1. Revisiting Distortions in ODIs

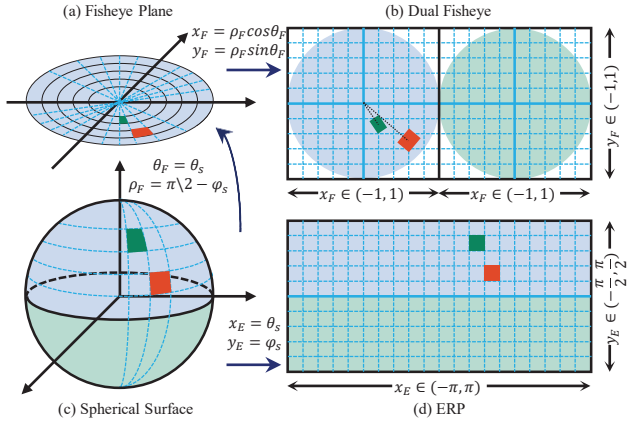

Figure 2. Geometric explanation of transforming between ERP, Fisheye, and the ideal spherical surface. To simplify, we discuss the horizontal spliced Fisheye with an aperture degree of π.

As ODIs under each projection type are constrained by different transforming equations, the distortion caused by each type is inconsistent, indicating that applying matrix operations under one projection type can introduce unexpected changes when being converted to other types. Specifically, applying uniformed bicubic downsampling on ERP images will affect the distribution of pixel density on Fisheye images, which are the original-formatted image type of imaging process in real-world scenarios. To analyze the specific effect of ERP downsampling on the Fisheye image, we revisit the cause of distortions in ERP and Fisheye.

As we assume that viewing directions are uniformly distributed, the data points in an ideal ODI should be uniformly distributed on a spherical surface. In practice, there is a trade-off between the uniformity of the spherical surface and the structural degree. ERP is the most convenient projection type for storage or transmission, but it is also the projection type that suffers the heaviest distortion. To better explain the causation of distortions, we follow the definition of stretching ratio ($\mathbf{K}$) in [31], which represents distortion degree at different locations from the target projection type to the ideal spherical surface. $\mathbf{K}$ is determined by area variation from one projection type to another. When the target type is uniforming spherical surface, $\mathbf{K}$ is defined as:

$$\mathbf{K}(x,y) = \frac{\delta S(\theta,\varphi)}{\delta P(x,y)} = \frac{\cos(\varphi)|d\theta d\varphi|}{|dxdy|} = \frac{\cos(\varphi)}{|J(\theta,\varphi)|}, \quad (1)$$

where $\delta S(\cdot,\cdot)$ and $\delta P(\cdot,\cdot)$ represent the area on the spherical surface and the projection plane, respectively. $|didj|$ represents plane microunit. $|J(\theta,\varphi)|$ is the Jacobian determinant from spherical coordinate to projection coordinate.

ERP distortion. The coordinate in ERP is defined as $x = \theta$ and $y = \varphi$. ERP stretching ratio can be derived as:

$$\mathbf{K}_{\text{ERP}}(x,y) = \cos(\varphi) = \cos(y), \quad (2)$$

where $x \in (-\pi, \pi)$, $y \in (-\frac{\pi}{2}, \frac{\pi}{2})$.

From Eq. (2), we conclude that ERP distortion is only determined by its latitude degree. $\mathbf{K}_{\text{ERP}}$ is reduced to zero when the absolute value of latitude degree increases to $\pi/2$, which represents that pixel density on the polar areas of ERP images is closer to zero. As shown in Fig. 2 (c), with the increasing of the absolutely value of latitude degree ($|\varphi_s|$), the corresponding area on the spherical surface of an ERP microunit is gradually decreased to zero. In conclusion, ERP distortion is caused by variable stretching ratios $\mathbf{K}_{\text{ERP}}$, and is the heaviest in the polar areas.

Fisheye distortion. The coordinate in Fisheye can be derived from $\theta = \arctan(\frac{y}{x})$ and $\varphi = (1 - \sqrt{x^2+y^2}) \times \frac{\pi}{2}$. The stretching ratios of Fisheye can be derived as[3]:

$$\mathbf{K}_{\text{Fisheye}}(x,y) = \frac{\frac{2}{\pi}\sin(\frac{\pi}{2}\sqrt{x^2+y^2})}{\sqrt{x^2+y^2}}, \quad (3)$$

where $\sqrt{x^2+y^2} \in (0,1)$.

$\mathbf{K}_{\text{Fisheye}}$ is determined by distance from the fisheye center. As $(\mathbf{K}_{\text{Fisheye}})^{-1}$ is bounded, fisheye projection is closer to uniform distribution than ERP. Moreover, it introduces much slighter distortion at the polar.

Relationship between ERP and Fisheye distortions. To simplify, here we only discuss a typical Fisheye with an aperture degree of π and a horizontal slicing plane[4]. In this case, the ERP coordinates and Fisheye's polar coordinates correspond linearly. We can quantize the relationship by:

$$\mathbf{K}_{\text{ERP|Fisheye}}(\theta,\varphi) = \frac{\mathbf{K}_{\text{ERP}}(x_E,y_E)}{\mathbf{K}_{\text{Fisheye}}(x_F,y_F)} = \frac{\pi}{2} - |\varphi|, \quad (4)$$

where θ, φ are spherical coordinates on the sphere, x_E, y_E (x_F, y_F) denotes the plain coordinate under ERP (Fisheye).

[3]Detailed derivative processes can be found in the supplementary file.

[4]The influence of Fisheye formats with arbitrary splicing plane is discussed in the supplementary file.

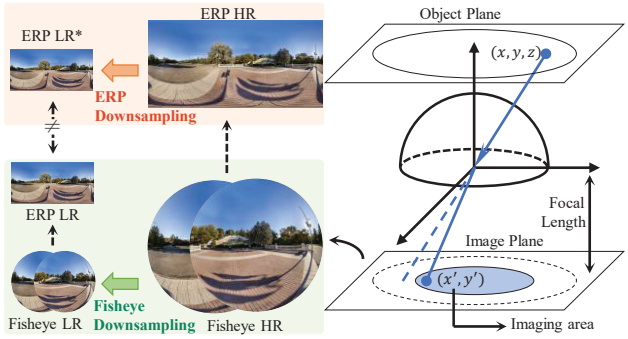

Figure 3. Downsampling process of ODIs (left) and imaging process in real world (right). * denotes that LR images synthesized from different downsampling processes are inconsistent.

From Eqs. (2) to (4), we conclude that when uniformed downsampling is performed on ERP, the kernel size of equivalent Fisheye downsampling is non-uniformed. Especially when fisheye projection is spliced horizontally, the kernel size is proportional with $\pi/2 - |\varphi|$.

3.2. Learning with More Realistic Degradation

As depicted in Fig. 3, the original-formatted projection type in ODI acquiring process is fisheye projection. Given that real-world low-resolution issues are caused by insufficient precision and density of sensors, we consider that the degeneration process should be directly applied to original-formatted images before the type conversion.

Ideally, as camera sensors are arranged in uniform arrays, pixel density on original-formatted images is consistent everywhere. Thus, for a realistic ODI, the pixel density on Fisheye should be a constant. As discussed in Sec. 3.1, applying uniformed downsampling on ERP means applying downsampling of variable kernel size on Fisheye. The variable kernel size leads to variable Fisheye pixel density, which results in unrealistic LR images. In conclusion, the ERP downsampling in previous methods influences the intrinsic distribution of pixel density in original-formatted images, which leads to unrealistic ODIs. When the downsampling process happens on Fisheye, the Fisheye pixel density is unchanged, which fits the real-world imaging process and synthesizes more realistic LR pairs.

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. 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 still learn the upscaling process under ERP. Thus we reconvert LR images to the ERP format. The overall process of Fisheye downsampling are described in Fig. 3.

3.3. OSRT: Modulate Distortion in ODIs

Overall. As discussed in Sec. 3.1, ERP images suffer a distortion caused by a non-consistency area stretching ratio from an ideal spherical surface. Referred from Eq. (2), for an LR input $X_i \in \mathbb{R}^{C\times M\times N}$, the distortion map $C_d \in \mathbb{R}^{1\times M\times N}$ is derived by:

$$C_d = \cos\left(\frac{m + 0.5 - M/2}{M}\pi\right), \tag{5}$$

where m is the current height of LR input.

Previous methods tend to treat C_d as an additional input of X_i [28], or re-weighting parameters by C_d [18]. Although these solutions can benefit from building awareness of distortion, continuous and amorphous distortions cannot be adequately fitted by scattering and structured convolution operations. While previous methods cannot fully explore the advantage of C_d, we intend to design a novel block for learning distorted patterns continuously. In VSR tasks, the deformable mechanism is proposed to align features between adjacent frames [32, 34]. Unlike standard DCN [10], which calculates offsets from the input feature map, offsets are calculated from bi-directional optical flow in VSR pipelines. Inspired by feature-level flow warping in VSR, we find that the deformable mechanism is a feasible solution for continuous mappings. Consequently, we modulate ERP distortion by feature-level warping operations. As shown in Fig. 4, C_d is only utilized to calculate the deformable offsets θ. To keep compatibility with arbitrary ConvNets and Transformers, we propose two blocks to modulate ERP distortion, which can directly replace the multi-head self-attention blocks in Transformers and the standard convolution layers in ConvNets, respectively.

Distortion-aware attention block (DAAB). As depicted in Fig. 4 (a), a distortion condition guided deformable self-attention is proposed to learn correlations between the distorted input X_i and its corresponding modulated feature map $\tilde{X}_i$. DAAB is formulated as:

$$\theta_i = H_{\text{offset}_i}(C_d, C_w), \tilde{X}_i = \phi(X_i; p_0 + \theta_i), \tag{6}$$

$$X_o = H_{\text{SA}}(X_i W_{q_i}, \tilde{X}_i W_{k_i}, \tilde{X}_i W_{v_i}), \tag{7}$$

where $H_{\text{offset}_i}(\cdot)$ denotes the i-th convolution block to calculate offset maps $\theta_i \in \mathbb{R}^{2\times H\times W}$, and H_{SA} denotes standard self-attention formula. $H_{\text{offset}}(\cdot)$ consists of 1×1 convolution block with two hidden layers. The input of $H_{\text{offset}}(\cdot)$ is concatenated by the latitude-related distortion condition $C_d \in \mathbb{R}^{1\times H\times W}$ and the window condition $C_w \in \mathbb{R}^{2\times H\times W}$. C_w is a linear position encoding within a self-attention kernel. $\phi(\cdot,\cdot)$ denotes a bilinear interpolation, and $W_{q_i}, W_{k_i}, W_{v_i}$ denote i-th weight matrix of query, key, and

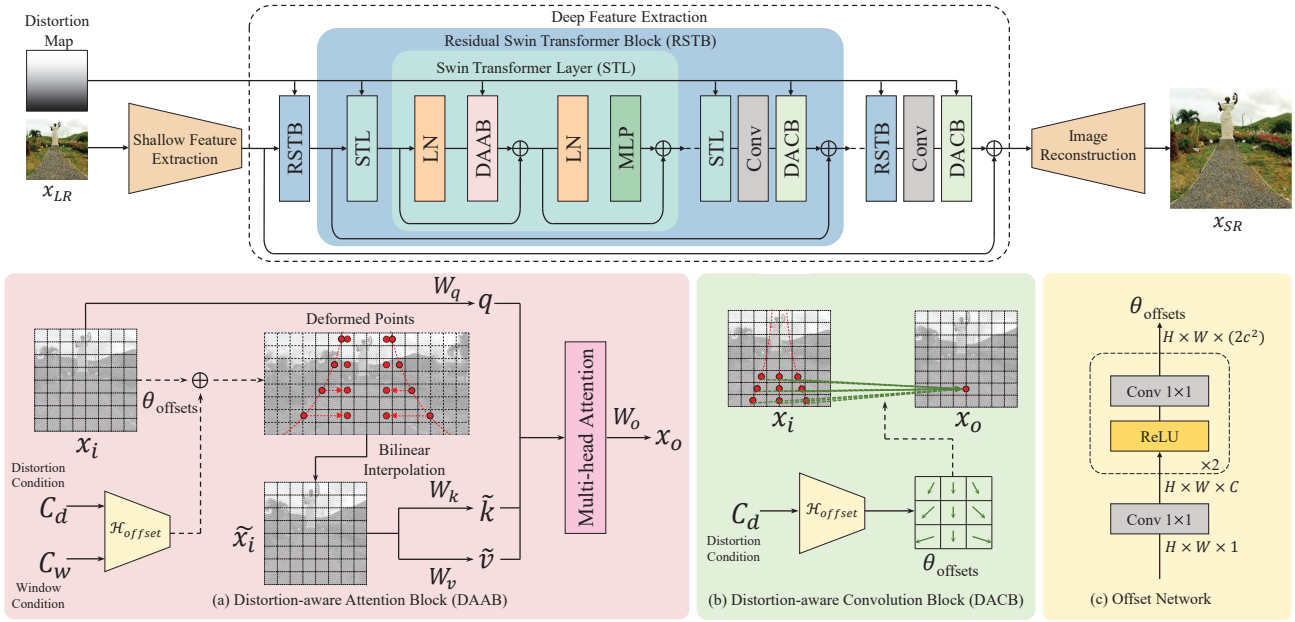

Figure 4. Overall illustration of OSRT. From SwinIR [23], we replace the standard multi-head self-attention block with DAAB and insert DACB behind the end of the RSTB. Channel dimensions of θ_{offsets} in DAAB and DACB are 2 and 18, respectively.

value, respectively. For multi-head self-attention blocks, $H_{\text{offset}_i}(\cdot)$ is identical in calculations of parallel heads.

Distortion-aware convolution block (DACB). As shown in Fig. 4 (b), we apply a standard deformable convolution layer with a substituted input for offset calculation. Modulated output X_o is extracted as:

$$\theta_i = H_{\text{offset}_i}(C_d), X_o = H_{\text{DCN}_i}(X_i, \theta_i), \tag{8}$$

where $H_{\text{DCN}}(X, \theta)$ denotes standard deformable convolution layer in [46]. The architecture of $H_{\text{offset}_i}(\cdot)$ is identical to that in DAAB. As the kernel size of DCN is 3×3 in DACB, the output channel dimension of offsets maps is 18.

OSRT. In practice, we propose an Omnidirectional image Super-Resolution Transformer, named OSRT. SwinIR [23] is selected as the basic architecture for its strong reconstruction ability in the SISR task. To learn distortion rectified representations, we stack a DACB after the last convolution layer of each residual swin Transformer block and replace all self-attention blocks as DAAB. The feature dimension of OSRT is reduced from 180 to 156 to maintain identical parameters with SwinIR.

3.4. Boosting ODISR Performance by Plain Images

As the capacity of OSRT is relatively large, it suffers overfitting for large upscaling factors (Fig. 9). Given that acquiring ODIs are expensive, we propose to generate pseudo ERP images from 2D plain images to tackle this issue. After being sampled by sliding windows, the patch of plain images is treated as a plain perspective. By converting from Perspective to ERP, plain images are distorted in the same way as ERP. Considering that distortion of a Perspective is enlarged by its FOV degree, a relatively small FOV degree of 90° is applied. For a given pseudo Perspective, θ_p is fixed at 0 and φ_p is derived by:

$$\Phi_p = \varphi_h + z_0, \tag{9}$$

where φ_h is determined by patch locations and z_0 is orderly sampled from $\{-15°, 0°, 15°\}$.

To maximize the approximate data distribution of ODIs, we horizontally split a plain image into three sub-images and define φ_h as $-30°, 0°, 30°$ respectively. Pseudo ERP images are cropped to remove the black border. As shown in Fig. 5, we get a new ERP dataset (called DF2K-ERP) by implementing the augmentation pipeline on widely-used plain image dataset DF2K [24, 33]. The DF2K-ERP dataset consists of 146K high-quality ERP image patches with a patch size larger than 256.

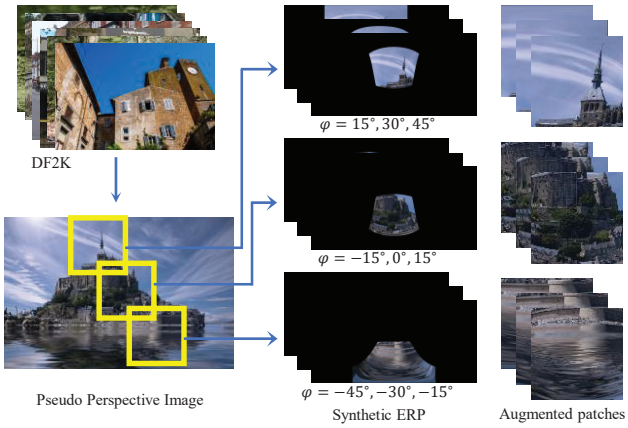

Figure 5. Synthetic process of DF2K-ERP.

Method	Scale	ODI-SR PSNR	ODI-SR SSIM	ODI-SR WS-PSNR	ODI-SR WS-SSIM	SUN 360 Panorama PSNR	SUN 360 Panorama SSIM	SUN 360 Panorama WS-PSNR	SUN 360 Panorama WS-SSIM
Bicubic	×2	28.21	0.8215	27.61	0.8156	28.14	0.8118	28.01	0.8321
RCAN [43]		30.08	0.8723	29.49	0.8714	30.56	0.8712	31.18	0.8969
SRResNet [36]		30.16	0.8717	29.59	0.8697	30.65	0.8714	31.20	0.8953
EDSR [24]		30.32	0.8770	29.68	0.8727	30.89	0.8784	31.42	0.8995
SwinIR [23]		30.52	0.8819	29.87	0.8772	31.21	0.8852	31.78	0.9051
SwinIR† [23]		30.64	0.8821	30.00	0.8777	31.33	0.8855	31.98	0.9059
OSRT†		**30.77**	**0.8846**	**30.11**	**0.8795**	**31.52**	**0.8888**	**32.14**	**0.9081**
Bicubic	×4	25.59	0.7118	24.95	0.6923	25.29	0.6993	24.90	0.7083
RCAN [43]		26.85	0.7621	26.15	0.7485	27.10	0.7660	26.99	0.7856
SRResNet [36]		26.91	0.7597	26.24	0.7457	27.10	0.7618	26.99	0.7812
EDSR [24]		26.97	0.7589	26.30	0.7458	27.19	0.7633	27.10	0.7827
SwinIR [23]		27.12	0.7663	26.44	0.7523	27.39	0.7707	27.30	0.7901
SwinIR† [23]		27.31	0.7735	26.61	0.7589	27.71	0.7804	27.64	0.7996
OSRT†		**27.41**	**0.7762**	**26.70**	**0.7609**	**27.84**	**0.7835**	**27.77**	**0.8020**

Table 1. SR results under Fisheye downsampling. † denotes applying DF2K-ERP as augmented dataset. Best results are shown in **Bold**.

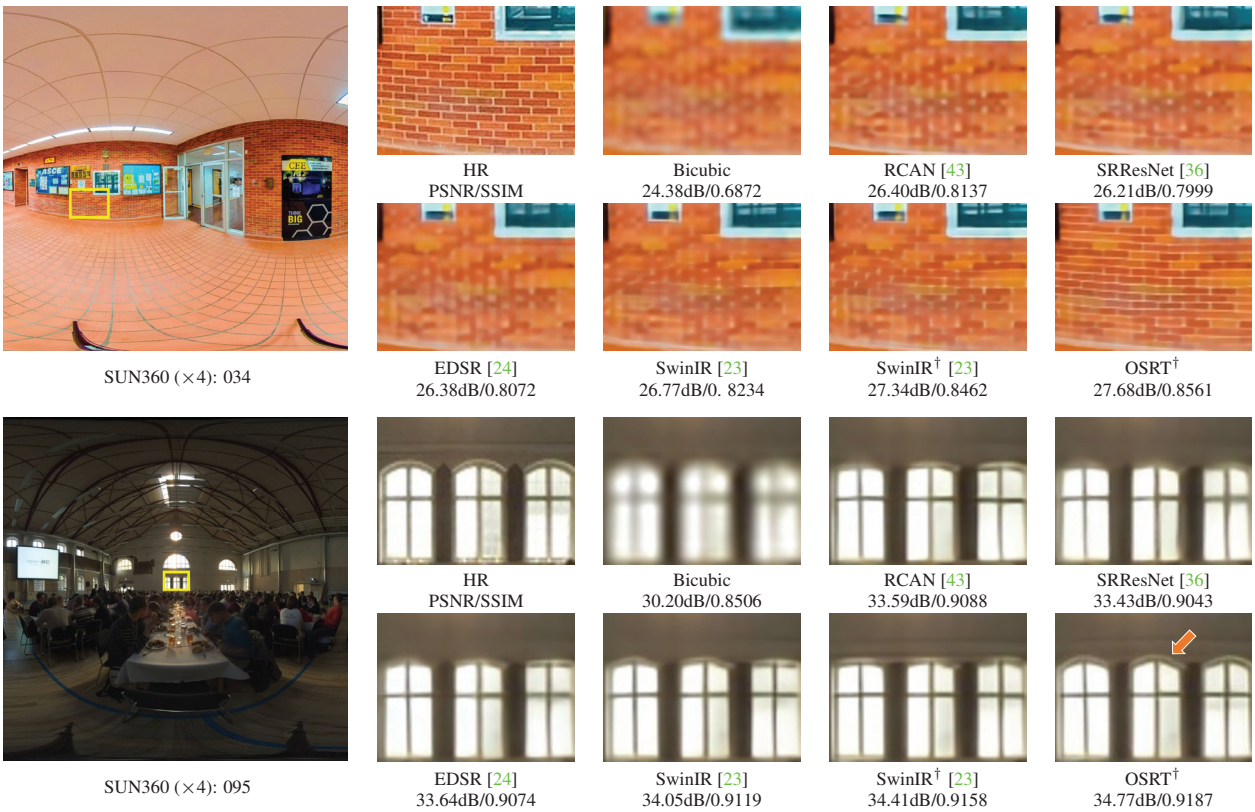

Figure 6. Visual comparisons of ×4 SR results under Fisheye downsampling.

4. Experiments

4.1. Experimental Setup

ODI-SR dataset [12] and SUN360 Panorama dataset [39] are used in our experiment. In the training phase, we follow the data split setting in [12] and train on the ODI-SR training set. The resolution of the ERP HR is 1024×2048, and the upscaling factors are $\times 2$ and $\times 4$. Fisheye downsampling is applied as our pre-defined downsampling kernel. Loss is calculated by L1 distance and optimized by Adam [19], with an initial learning rate of 2×10^{-4}, a total batch size of 32, and an input patch size of 64. We train OSRT for $500k$ iterations and halve the learning rate at $250k$, $400k$, $450k$ and $475k$. In evaluation, we test on the ODI-SR testing set and SUN360 dataset. PSNR [16], SSIM [37], and their distortion re-weighted versions (WS-PSNR [31], WS-SSIM [45]) are used as evaluation metrics.

4.2. Evaluation under Fisheye Downsampling

When the downsampling process is performed on Fisheye images, we train SRResNet [36], EDSR [24], RCAN [43], and SwinIR [23] for comparison.

Quantitative results. As shown in Tab. 1, with the help of additional DF2K-ERP training patches, OSRT out-

Scale	×8				×16			
Method	ODI-SR		SUN 360 Panorama		ODI-SR		SUN 360 Panorama	
	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM
Bicubic	19.64	0.5908	19.72	0.5403	17.12	0.4332	17.56	0.4638
SRCNN [13]	20.08	0.6112	19.46	0.5701	18.08	0.4501	17.95	0.4684
EDSR [24]	23.97	0.6417	22.46	0.6341	21.12	0.5698	21.06	0.5645
RCAN [43]	24.26	0.6628	23.88	0.6542	21.94	0.5824	21.74	0.5742
360-SS [30]	21.65	0.6417	21.48	0.6352	19.65	0.5431	19.62	0.5308
LAU-Net [12]	24.36	0.6801	24.02	0.6708	22.07	0.5901	21.82	0.5824
SphereSR [40]	24.37	0.6777	24.17	0.6820	22.51	**0.6370**	21.95	0.6342
OSRT	**24.53**	**0.6780**	**24.38**	**0.7072**	**22.69**	0.6261	**22.13**	**0.6388**

Table 2. SR results under ERP downsampling.

performs previous methods by 0.3dB on PSNR. Although directly applying SwinIR on the ODISR task has already reached SOTA performance, OSRT surpasses SwinIR over 0.1dB on two datasets for both ×2 and ×4 SR tasks, which demonstrates the effectiveness of its distortion modulation ability. The performance of RCAN degrades under Fisheye downsampling, which is caused by the incompatibility between channel attention and Fisheye downsampling[5].

Qualitative comparison. Fig. 6 shows the visualization results of ×4 ODISR task. While other methods struggle to understand the geometric transformation process in distorted images, OSRT can reconstruct sharp and accurate boundaries with the advantages of distortion modulation. It is observed that OSRT is skilled at reconstructing rigid texture. Moreover, benefiting from the distortion modulation ability, OSRT can preserve the original structure as most when being projected to other projection types (Fig. 7).

4.3. Evaluation under ERP Downsampling

To compare with previous ODISR methods [12, 30, 40], we train OSRT under the previous ERP setting. Regardless of over-fitting issues, we only train on the dataset provided by [12] for fairness. As shown in Tab. 2, OSRT still outperforms LAU-Net [12] and SphereSR [40] under large upscaling factor and ERP downsampling. Without a complicated training pipeline and discrete inference process, OSRT yields the best PSNR values and surpasses all previous methods on most SSIM-related metrics (three of four).

4.4. Ablation Study and Discussion

In this section, we prove the effectiveness of Fisheye downsampling, OSRT components, and augmented DF2K-ERP. We then explain the distortion modulation ability of OSRT by visualizing offsets in deformable blocks.

Fisheye downsampling. As shown in Fig. 1, the SR model trained under ERP downsampling is more likely to generate blur details and missing structures in real-world scenarios. These artifacts cannot be removed by a superior backbone network, but can be eliminated by a more realistic imaging process. More importantly, ERP downsampling directly covers the geometric property of ERP images

[5]The cause is discussed in the supplementary file.

feature dim	DACB	DAAB	ODI-SR		SUN360		Params. (M)
			PSNR	SSIM	PSNR	SSIM	
60	×	×	30.27	0.8739	30.78	0.8742	0.91
60	✓	×	30.41	0.8775	31.00	0.8793	1.16
60	×	w/o C_w	30.31	0.8746	30.83	0.8755	1.00
60	×	w/ C_w	30.32	0.8746	30.84	0.8753	1.01
60	✓	w/ C_w	**30.44**	**0.8780**	**31.04**	**0.8800**	1.26
72	×	×	30.32	0.8748	30.85	0.8755	1.29

Table 3. Ablation study on OSRT components. All models are trained on ×2 SR task under Fisheye downsampling.

and makes the ODISR task identical to the standard plain image super-resolution task. The evidence is that a standard SISR model (SwinIR) trained on a plain image dataset (DF2K) can outperform previous SOTA in the ODISR task, which yields WS-PSNR results of 24.63dB/24.49dB (22.68dB/22.13dB) on ×8 (×16) ODI-SR/SUN360 testing set, respectively. In conclusion, when the intrinsic property of ODIs is broken by ERP downsampling, the ODISR task degenerates into a plain image super-resolution task with a particular data distribution.

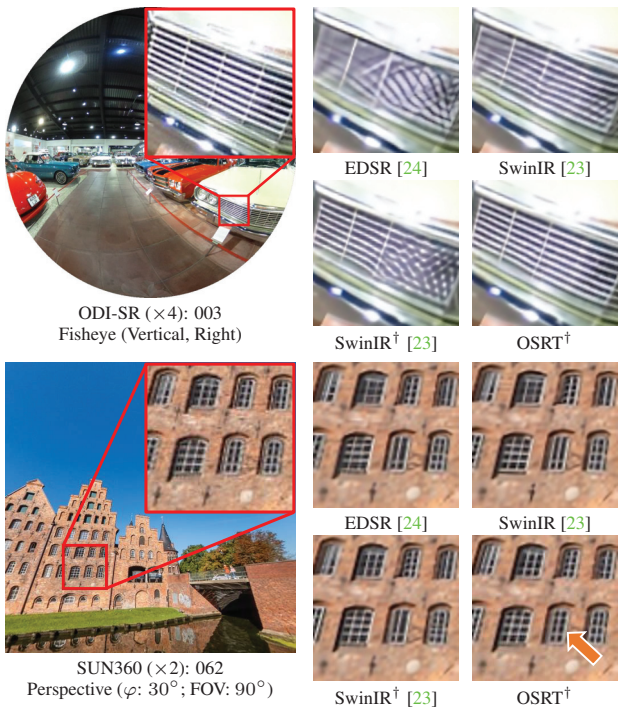

Figure 7. Visual comparisons for SR of Fisheye and Perspective images. † denotes applying DF2K-ERP as augmented dataset.

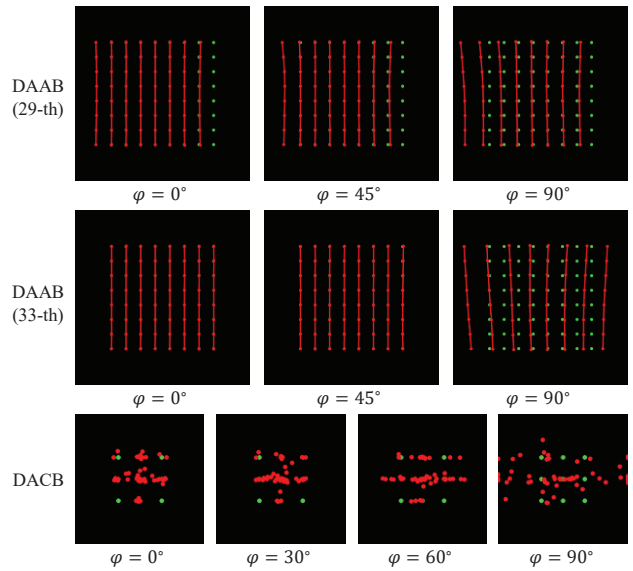

Figure 8. Visualizations of offset maps in OSRT. Reference and deformed points are depicted in green and red, respectively. The deformable kernel is sparse in the polar area.

OSRT components. To study the effectiveness of each component in OSRT, we propose a light version of OSRT (OSRT-light) for ablation study, which corresponds with the official SwinIR-light [23]. As proofed in Tab. 3, all components in OSRT are beneficial for modulating ERP distortion. The advantages of DACB and DAAB can be stacked when being applied in the same network. Compared with simply expanding the feature dimension of SwinIR to match the network complexity, the overall improvements of OSRT is more significant (+0.05dB vs. +0.2dB).

Offsets in OSRT. Offsets map in a well-trained OSRT are visualized in Fig. 8. Deformable kernels in both DAAB and DACB tend to gather at the equator and scatter at the polar, which conforms to the geometric distribution of pixel density in ERP images.

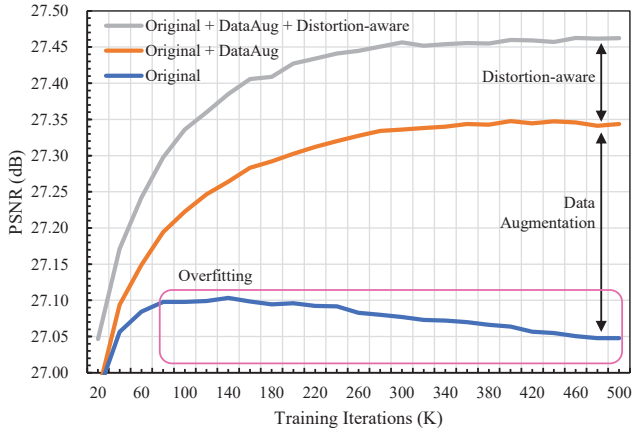

Figure 9. Training process of Transformers on ×4 ODISR task. The overfitting issue is tackled by our augmentation scheme.

Backbone network	Datasets	Training scheme	Scale	SUN360 PSNR	SUN360 SSIM
SwinIR	ODI-SR	N/A	×2	31.21	0.8852
SwinIR	DF2K/ODI-SR	one-stage	×2	31.26	0.8841
SwinIR	DF2K-ERP/ODI-SR	one-stage	×2	31.33	0.8855
SwinIR	DF2K-ERP/ODI-SR	two-stage	×2	31.17	0.8818
OSRT	DF2K-ERP/ODI-SR	one-stage	×2	**31.52**	**0.8888**
SwinIR	ODI-SR	N/A	×4	27.39	0.7707
SwinIR	DF2K/ODI-SR	one-stage	×4	27.59	0.7768
SwinIR	DF2K-ERP/ODI-SR	one-stage	×4	27.71	0.7804
SwinIR	DF2K-ERP/ODI-SR	two-stage	×4	27.74	0.7795
OSRT	DF2K-ERP/ODI-SR	one-stage	×4	**27.84**	**0.7835**

Table 4. Ablation study on data augmentation. The results of ODI-SR (In the supplementary file) are in the same trend as SUN360.

Pseudo ERP patches. In Sec. 3.4, we propose a distorted dataset DF2K-ERP to tackle over-fitting issues. We train a standard SwinIR on diverse datasets and training schemes to study the influence of data augmentation separately. As shown in Tab. 4, while training on ODI-SR and DF2K, distortion operations in DF2K lead to better performance. Compared with fine-tuning on DF2K-ERP pre-trained models (two-stage), training on two datasets jointly (one-stage) shows better results. We infer that there is a domain gap between ODI-SR and DF2K-ERP, which is caused by omitted Perspective distortion. Moreover, the advantage of distortion modulation mechanisms in OSRT is enlarged when additional training patches are applied. Fig. 9 proves that our data augmentation scheme overcomes the over-fitting issue and improves the reconstruction ability.

5. Conclusion

In this paper, we find that the previous downsampling process in the ODISR task harms the intrinsic distribution of pixel density in ODIs, which leads to poor generalization ability in real-world scenarios. To tackle this issue, we propose Fisheye downsampling, which mimics the real-world imaging process to preserve the realistic density distribution. After refining the downsampling process, we design a distortion-aware Transformer (OSRT) to modulate distortions continuously and self-adaptively. OSRT learns offsets from the distortion-related condition and rectifies distortion by feature-level warping. Moreover, to alleviate the over-fitting problem of large networks, we propose to synthesize additional ERP training data from the plain images. Extensive experiments have demonstrated the state-of-the-art performance of our OSRT.

Limitation. The process of sampling ERP images into viewing types also requires careful design.

Acknowledgement.

This work was supported in part by the National Key R&D Program of China (NO.2022ZD0160505), in part by the National Natural Science Foundation of China under Grant (62276251), the Joint Lab of CAS-HK, and in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (No.2020356).

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] 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] 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. 3

[6] 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. 3

[7] 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

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

[9] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8628–8638, 2021. 2

[10] 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. 4

[11] 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

[12] 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, 6, 7

[13] 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, 7

[14] 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

[15] 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. 1, 2

[16] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 6

[17] 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

[18] Renata Khasanova and Pascal Frossard. Geometry aware convolutional filters for omnidirectional images representation. In *International Conference on Machine Learning*, pages 3351–3359. PMLR, 2019. 4

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[20] 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

[21] 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

[22] 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

[23] 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. 2, 5, 6, 7, 8

[24] 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, 2, 5, 6, 7

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2

[26] 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

[27] 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

[28] 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. 1, 2, 4

[29] 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

[30] 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, 3, 7

[31] 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, 6

[32] 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. 4

[33] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 5

[34] 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. 3, 4

[35] 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

[36] 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. 2, 6

[37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[38] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4794–4803, 2022. 3

[39] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2695–2702. IEEE, 2012. 1, 6

[40] 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, 7

[41] 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

[42] 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

[43] 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, 2, 6, 7

[44] 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

[45] Yufeng Zhou, Mei Yu, Hualin Ma, Hua Shao, and Gangyi Jiang. Weighted-to-spherically-uniform ssim objective quality evaluation for panoramic video. In *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pages 54–57. IEEE, 2018. 6

[46] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 5