Geometric Distortion Guided Transformer for Omnidirectional Image Super-Resolution

Cuixin Yang, Rongkang Dong, Jun Xiao, Cong Zhang,
Kin-Man Lam, *Senior Member, IEEE,* Fei Zhou, Guoping Qiu, *Senior Member, IEEE*

***Abstract*—As virtual and augmented reality applications gain popularity, omnidirectional image (ODI) super-resolution has become increasingly important. Unlike 2D plain images that are formed on a plane, ODIs are projected onto spherical surfaces. Applying established image super-resolution methods to ODIs, therefore, requires performing equirectangular projection (ERP) to map the ODIs onto a plane. ODI super-resolution needs to take into account geometric distortion resulting from ERP. However, without considering such geometric distortion of ERP images, previous methods only utilize a limited range of pixels and may easily miss self-similar textures for reconstruction. In this paper, we introduce a novel Geometric Distortion Guided Transformer for Omnidirectional image Super-Resolution (GDGT-OSR). Specifically, a distortion modulated rectangle-window self-attention mechanism, integrated with deformable self-attention, is proposed to better perceive the distortion and thus involve more self-similar textures. Distortion modulation is achieved through a newly devised distortion guidance generator that produces guidance for the rectangular windows by exploiting the variability of distortion across latitudes. Furthermore, we propose a dynamic feature aggregation scheme to adaptively fuse the features from different self-attention modules. We present extensive experimental results on public datasets and show that the new GDGT-OSR outperforms methods in existing literature.**

***Index Terms*—Omnidirectional image, Super-resolution, Distortion, Rectangle-window, Transformer.**

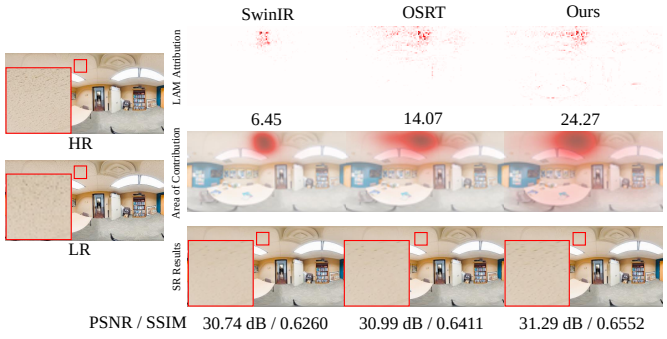


Fig. 1. Comparison of local attribution maps [6] and SR results among different methods. The local attribution maps represent the importance of each pixel in reconstructing the patch in the red box. The Diffusion Index (DI) is shown below the local attribution maps. A higher DI value indicates a wider range of the involved pixels. The second row shows the Area of Contribution, which implies the areas involved and their contributions. The local attribution maps, DI values, and Area of Contribution collectively demonstrate that our proposed method engages more pixels in the reconstruction. This contributes to restoring more realistic details, leading to improved SR performance.

I. Introduction

OMNIDIRECTIONAL imaging, also known as 360° imaging, is a fundamental technology for developing immersive virtual reality (VR) and augmented reality (AR) applications. In practice, omnidirectional images (ODIs) are viewed through head-mounted display devices, which means that the viewport will have a limited range. To visualize the details of a scene from a narrow field-of-view, the images need to be of very high resolutions. However, camera systems for capturing high-resolution ODIs are expensive, as are the costs of storing and transmitting high-resolution ODIs [1].

One way to tackle this issue is through image super-resolution (SR), which reconstructs a high-resolution (HR) image from a low-resolution (LR) input [2]–[4]. For convenience in storage and transmission, raw ODIs are generally projected into 2D planar representations. Equirectangular projection (ERP) is the most common method for representing ODIs [5].

Transformer has emerged as a powerful and versatile computational paradigm [7]. Local square-window self-attention is proposed to reduce the computational complexity of the global self-attention mechanism in vision transformer, leading to a limitation in the receptive field [8], [9]. However, for omnidirectional image super-resolution (ODISR), the ERP expands the ODIs and introduces distortion into ERP images. For example, a circle in the ODI is distorted into an oval in the ERP image. Small squared windows struggle to capture the whole oval, while large squared windows may encompass irrelevant patterns. Therefore, square-window self-attention is a suboptimal option for reconstructing ERP images. Rectangular windows [9] can calibrate and expand the receptive field by involving more self-similar textures along the direction of stretching distortion, which is more appropriate

This work was supported by the Hong Kong Research Grants Council (RGC) Research Impact Fund (RIF) under Grant R5001-18. *(Corresponding authors: Cuixin Yang and Kin-Man Lam.)*

Cuixin Yang, Rongkang Dong, Jun Xiao, Cong Zhang, and Kin-Man Lam are with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: cuixin.yang@connect.polyu.hk; rongkang97.dong@connect.polyu.hk; jun.xiao@connect.polyu.hk; cong-clarence.zhang@connect.polyu.hk; enkm-lam@polyu.edu.hk).

Fei Zhou is with the College of Electronic and Information Engineering, Shenzhen University, Shenzhen 518060, China, and also with the Guangdong-Hong Kong Joint Laboratory for Big Data Imaging and Communication, Shenzhen 518060, China (email: flying.zhou@163.com).

Guoping Qiu is with the School of Computer Science, University of Nottingham, NG8 1BB Nottingham, United Kingdom, and Ningbo 315100, China (email: guoping.qiu@nottingham.ac.uk).

for modeling features of ERP images than squared windows. However, existing ODISR methods [1], [10], [11] fail to consider extracting features of distorted ERP images from the perspective of the window's shape, leading to the limitation of involved pixels and self-similar textures. To solve this issue, we propose a Geometric Distortion Guided Transformer for Omnidirectional image Super-Resolution (GDGT-OSR), which aggregates features from windows of diverse shapes to calibrate and expand the attention area, involving more self-similar textures. Specifically, we propose a Distortion Modulated Rectangle-window Self-Attention (DMRSA) mechanism, which takes into account more self-similar regions of ERP images within a rectangular window. We integrate a deformable self-attention mechanism, which considers irregular neighborhoods and captures out-of-window similar patterns [12], with DMRSA. Self-similarity is essential because it significantly contributes to the reconstruction of HR images [13]–[16]. In DMRSA, Rectangle-window Self-Attention (Rwin-SA) is modulated by the distortion guidance that is generated through a newly devised Distortion Guidance Generator (DGG). DGG transforms the geometric distortion into distortion guidance, making the DMRSA adapt to the stretched ERP images. Furthermore, we dynamically aggregate the features from two self-attention modules by exploiting the information of windows with various shapes in Dynamic Feature Aggregation (DFA). To reveal the effects on the range of the involved area, we resort to Local Attribution Map (LAM) [6], which is an attribution method for analyzing and visualizing attribution in SR. As shown in Fig. 1, our proposed method can utilize a wider range of self-similar information and more pixels to recover the patch in the red box, achieving a higher Diffusion Index (DI) and better SR results.

In summary, our main contributions are as follows:

- We propose a novel framework, GDGT-OSR, designed for omnidirectional image super-resolution. By leveraging the distortion as guidance, GDGT-OSR effectively captures self-similar textures with windows of different shapes, leading to enhanced SR performance.
- To leverage the distortion information, we propose a Distortion Modulated Rectangle-window Self-Attention (DMRSA) mechanism paired with a Distortion Guidance Generator (DGG). The DGG transforms geometric distortion into distortion guidance, empowering DMRSA to adapt to the inherent distortion in ERP images through the newly designed distortion-guided rectangular windows.
- We devise a Dynamic Feature Aggregation (DFA) module to aggregate and complement features from different self-attention mechanisms based on the differences between them.
- Our GDGT-OSR framework achieves superior performance in ODISR, outperforming other state-of-the-art methods, including 2D plain SR and ODISR methods, in terms of quantitative and qualitative results.

The remaining parts of this paper are organized as follows. In Section II, we briefly review some related works. Section III introduces preliminaries about the related knowledge, followed by a detailed introduction of the proposed GDGT-OSR framework. Section IV presents the experimental settings, experiment results, and analysis of ablation studies. Finally, Section V-B provides a summary, including the limitations and a brief conclusion of this paper.

II. Related Works

In this section, we briefly review related works, including single image super-resolution (SISR), omnidirectional image super-resolution (ODISR), and Vision Transformer (ViT).

A. Single Image Super-Resolution

Long before the emergence of deep learning, researchers developed learning-based image resolution enhancement methods using image pyramids [17] and look-up tables [18]–[20]. After deep learning models demonstrated powerful capabilities in many related applications, researchers began to apply deep learning models, such as convolutional neural networks (CNNs), to image SR [2]–[4], [21]. As recent literature is dominated by deep-learning-based methods, we review three types of popular deep-learning-based SR methods, including GAN-based SR, Transformer-based SR, and Diffusion-based SR.

Generative Adversarial Networks (GANs) [22] have shown powerful generative modeling abilities in many computer vision tasks, including SR [23]–[26]. SRGAN [23] introduces an adversarial loss for GAN training to overcome the constraints associated with the PSNR-focused image SR. GLEAN [25], [26] uses pre-trained GANs, such as StyleGAN [27] and BigGAN [28], as a latent bank to exploit their diverse priors.

Since 2017, Transformers [29] have been widely used in computer vision. Researchers have applied Transformers to SR and demonstrated their efficacy in low-level computer vision tasks [8], [9], [30]–[32]. IPT [30] introduces a backbone model based on the standard Transformer to address diverse restoration challenges. Based on Swin Transformer [33], SwinIR [8] is proposed to solve different image restoration problems and has become a strong and popular backbone architecture for image restoration. However, using local squared windows, SwinIR suffers from a lack of direct interaction among windows, restricting the potential of establishing long-range dependencies. To solve this issue, CAT [9] proposes the Rwin-SA mechanism to broaden the attention area and increase the interaction across various windows by utilizing horizontal and vertical rectangle window attention.

Recently, diffusion models [34]–[38] have achieved unprecedented success in image/video synthesis and restoration. Due to the high computational cost resulting from pixel-space operation, LDM [39] suggests training diffusion models in the latent space to save computational resources. Since there is a strong connection between the LR inputs and the ground-truths when generating the outputs of SR, DiffIR [40] only employs the diffusion model to estimate a compact image restoration prior representation with much fewer iterations than traditional diffusion models. Another way to make the diffusion model efficient is to increase the speed of inference [41].

The above methods focus on SR for 2D plain images, which are inappropriate for ODIs. In this paper, we propose a novel geometric distortion guided framework for ODISR.

B. Omnidirectional Image Super-Resolution

Initially, researchers began tackling the issue of SR for ODIs through spherical assembling [42]–[44]. These methods attempted to super-resolve an HR image from a series of consecutive LR ODIs under diverse projections. Subsequently, researchers shifted their research focus and began to study ODISR on plain images, i.e., equirectangular panorama images. Meanwhile, deep learning was introduced into ODISR by training the deep-learning-based model SRCNN [21] and adapting it to equirectangular panorama images [45]. As GANs become popular in the computer vision community, they have been employed in ODISR [46], [47]. In the work of [46], the authors introduce a fast PatchGAN discriminator trained with a loss function designed for spherical images. In [47], an efficient multi-frequency GAN architecture is proposed to solve the SR of real-world panoramic images. However, these methods only address the distribution discrepancy between plain images and panoramic images and only fine-tune the plain image SR methods.

Deng et al. [1] pointed out that the pixel density is non-uniform and varies across latitudes in ERP projected ODIs. Therefore, they propose a progressive pyramid network, namely LAU-Net, to super-resolve the pixels at different latitude bands hierarchically, rather than making the model adapt to ODIs. However, training multiple levels of networks for different latitude bands is computationally expensive, and it can also lead to inconsistencies between latitude bands. The work [48] directly concatenates a distortion map and an LR image as the input, which is fed into a network designed for 2D plain images. However, the information of the distortion map is not exploited sufficiently, and it is difficult for a 2D-image SR network to learn to restore ODIs without modification. SphereSR [10] generates a continuous spherical image representation and employs the local implicit image function (LIIF) [49] to predict the RGB values under various projection types continuously. Although SphereSR can resolve ODIs with arbitrary projection types flexibly, it requires training multiple network branches for different projection types. TCCL-Net [50] proposes a Transformer and convolution collaborative learning network for ODISR, which extracts both long-range and short-range dependencies in an end-to-end manner. Furthermore, in the work [51], the same authors further explore the SR of panoramic images from the left-right view by combining binocular information and panoramic characteristics. When generating training pairs, all of these ODISR methods apply uniform bicubic downsampling on the ERP images, neglecting the geometric properties of ERP in the degradation process. OSRT [11] utilizes Fisheye downsampling that applies uniform bicubic downsampling on the original ODIs. Even though OSRT makes use of the deformable Transformer to leverage the distortion information, the attention area is still small and inefficient for reconstructing ODIs and it is difficult to adaptively calibrate the features across latitudes. To solve this issue, we propose to enlarge the attention area by involving different kinds of self-attention with diverse window shapes. Furthermore, we introduce a distortion guidance generator to modulate the features based on the variability of distortion across latitudes.

C. Vision Transformer

Transformer [29] was first introduced to process sequences in natural language processing. It has been developed for various computer vision tasks and has achieved extraordinary performances, including image recognition [7], [52], object detection [12], [53], and segmentation [54], [55]. Due to their great potential in dealing with high-level computer vision tasks, Transformers have been adopted in the field of image restoration, such as IPT [30] and SwinIR [8]. Uformer [56] applies self-attention in an 8×8 local windows and adopts the U-net architecture to capture both local and global dependencies. However, the local windows used in SwinIR and Uformer limit the range of long-range dependency representation and the receptive field of the Transformer model. CAT [9] adopts the Rwin-SA to enlarge the attention area. Furthermore, deformable attention is exploited in ViTs [11], [57] to capture the content in irregular neighborhoods.

The contents are non-uniformly stretched in ERP images. The squared windows, which are commonly utilized for regular 2D images, restrict the attention area from accommodating the distortion caused by the stretch in ERP images. To address this limitation, we propose a distortion modulated Rwin-SA that incorporates deformable self-attention. By considering the shape and direction of distortion, this approach calibrates the attention area, enabling it to capture more similar textures and patterns in ERP images.

III. Methodology

A. Preliminaries

In this section, we introduce the stretching ratio and distortion map for a better understanding of the projection relationship between ODIs and the projection plane, as well as the derivation of the distortion map.

*1) **Stretching Ratio**:* As shown in Fig. 2, each coordinate (θ, φ) on the ideal spherical surface corresponds to a point (x, y) on the projection plane. The relationship between (θ, φ) and (x, y) can be formulated as follows:

$$x = h(\theta, \varphi), y = t(\theta, \varphi), \tag{1}$$

where $h(\cdot)$ and $t(\cdot)$ are coordinate transformation functions from the spherical surface to the projection plane. Given a microunit $\delta S(\theta, \varphi)$ centered at (θ, φ) on the spherical surface and its corresponding microunit $\delta P(x, y)$ centered at (x, y) on the projection plane, the stretching ratio is defined as follows:

$$R(x, y) = \frac{\delta S(\theta, \varphi)}{\delta P(x, y)} = \frac{\cos(\varphi)|d\theta d\varphi|}{|dxdy|} = \frac{\cos(\varphi)}{|J(\theta, \varphi)|}, \tag{2}$$

where $J(\theta, \varphi)$ is a Jacobian determinant. Considering Eq. (1), the Jacobian determinant is defined as follows:

$$J(\theta, \varphi) = \frac{\partial(x, y)}{\partial(\theta, \varphi)} = \begin{vmatrix} \frac{\partial x}{\partial \theta} & \frac{\partial x}{\partial \varphi} \\ \frac{\partial y}{\partial \theta} & \frac{\partial y}{\partial \varphi} \end{vmatrix}. \tag{3}$$

The projection relationship for ERP, which is a commonly used projection method for ODIs, is defined as follows [58]:

$$x = h(\theta, \varphi) = \theta, y = t(\theta, \varphi) = \varphi. \tag{4}$$

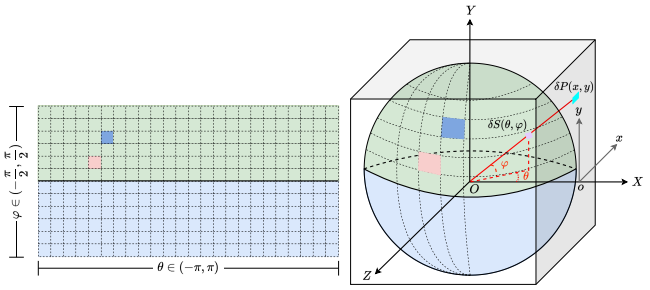


Fig. 2. Geometric explanation of the relationship between ERP (left) and the sphere, as well as the relationship between the sphere and the tangential cube (right).

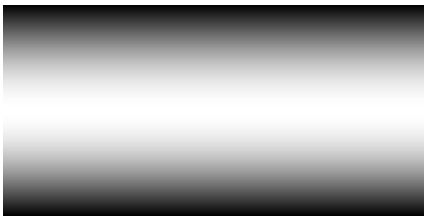


Fig. 3. Distortion map. A lighter area represents less distortion, while a darker area represents higher distortion.

Thus, derived from Eqs. (2)-(4), the stretching ratio of ERP can be formulated as follows:

$$R_{ERP}(x, y) = \cos(\varphi) = \cos(y), \tag{5}$$

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

*2) **Distortion Map**:* The projection from raw ODIs to ERP images leads to distortion in the latter. This distortion varies along the latitude and is symmetric in the two hemispheres. According to Eq. (5), given an LR image $I^{LR} \in \mathcal{R}^{H\times W\times C_{in}}$ (H, W, and C_{in} represent the height, width, and number of channels, respectively), the corresponding distortion map $D \in \mathcal{R}^{H\times W\times 1}$ is defined as follows [11], [58]:

$$D(h, 1:W) = \cos\frac{(h + 0.5 - H/2)\pi}{H}, \tag{6}$$

where $D(h, 1:W)$ represents the stretching ratio from an ideal spherical surface to the 2D ERP image at the current height of h. As shown in Fig. 3, the distortion around the equator area is the smallest, while the distortion intensifies as the latitude increases.

B. Architecture

Unlike regular 2D images, ERP images are non-uniformly stretched in the projection space. Local squared windows in traditional Transformers make it difficult for the attention area to adapt to the stretched distortion. An attention area that takes this distortion into account is more appropriate for distorted ERP images. To address this issue, we propose a geometric distortion guided framework called GDGT-OSR, which includes Distortion Modulated Rectangle-window Self-Attention (DMRSA) and Distortion-aware Deformable Self-Attention (DDSA). By considering the distortion characteristics of ERP images, GDGT-OSR calibrates and expands the attention area. This allows it to capture more similar textures and patterns, which are crucial for SR reconstruction.

*1) **Overview**:* An overview of the proposed method is shown in Fig. 4. It consists of three modules: shallow feature extraction, deep feature extraction, and image reconstruction. Our model takes an LR ODI $I^{LR} \in \mathcal{R}^{H\times W\times C_{in}}$ and the corresponding distortion map $D \in \mathcal{R}^{H\times W\times 1}$ as input and reconstructs an HR ODI. We utilize only one convolutional layer in the shallow feature extraction module to obtain the low-level feature $F^0 \in \mathcal{R}^{H\times W\times C}$ (C is the number of channels). This feature is then fed into the deep feature extraction module. This module comprises K Dual-attention Aggregation Blocks (DABs), each of which includes several Dual-attention Aggregation Layers (DALs), a convolutional layer, and a Distortion-Aware Convolution Block (DACB) [11].

The structure of the proposed DAB is depicted at the bottom of Fig. 4. The Distortion Modulated Rectangle-window Self-Attention (DMRSA) and Distortion-aware Deformable Self-Attention (DDSA) mechanisms generate features characterized by different local windows. Specifically, in DMRSA, we leverage the proposed Distortion Guidance Generator (DGG) to adaptively modulate the features according to the latitudes. Furthermore, for better feature fusion, features from DMRSA and DDSA are dynamically aggregated within a Dynamic Feature Aggregation (DFA) module according to their respective importance.

*2) **Distortion Guidance Generator (DGG)**:* The DGG is designed to extract a representative feature, i.e., the distortion guidance, that encapsulates the distortion. This distortion guidance is utilized to adaptively modulate the key and value features from the Rwin-SA to better adapt to the unevenly stretched content in ERP images. Thus, the output feature of the self-attention mechanism is calibrated by the distortion guidance. The process of DGG can be expressed as follows:

$$G = \mathrm{DGG}\left(D\right), \tag{7}$$

where D and G denote the distortion map and the distortion guidance, respectively.

From the geometric property of the distortion map, it can be noted that the distortion for locations at the same latitude is identical. Based on this fundamental property, we propose to model the distortion map in a latitude-wise manner. Specifically, the distortion map $D \in \mathcal{R}^{H\times W\times 1}$ is first convolved to extract feature maps $f \in \mathcal{R}^{H\times W\times C}$. Then, in the Latitude-wise branch, each feature map undergoes latitude-wise pooling (LWP). In LWP, as shown in Fig. 5, the pixel values of each feature map are averaged by row, transforming a feature map into a vector. The process is expressed as follows:

$$v^i = \mathrm{LWP}(f^i), \tag{8}$$

$$\text{where } v^i_j = \frac{1}{W}\sum_{w=1}^{W}\left(f^i_{j,w}\right), j \in [1, H]. \tag{9}$$

In Eq. (8) and (9), v^i and f^i denote the i-th vector and feature map, respectively, v^i_j represents the j-th element of the i-th vector, and $f^i_{j,w}$ represents the w-th element of the j-th row in the i-th feature map. After LWP, the column vectors are

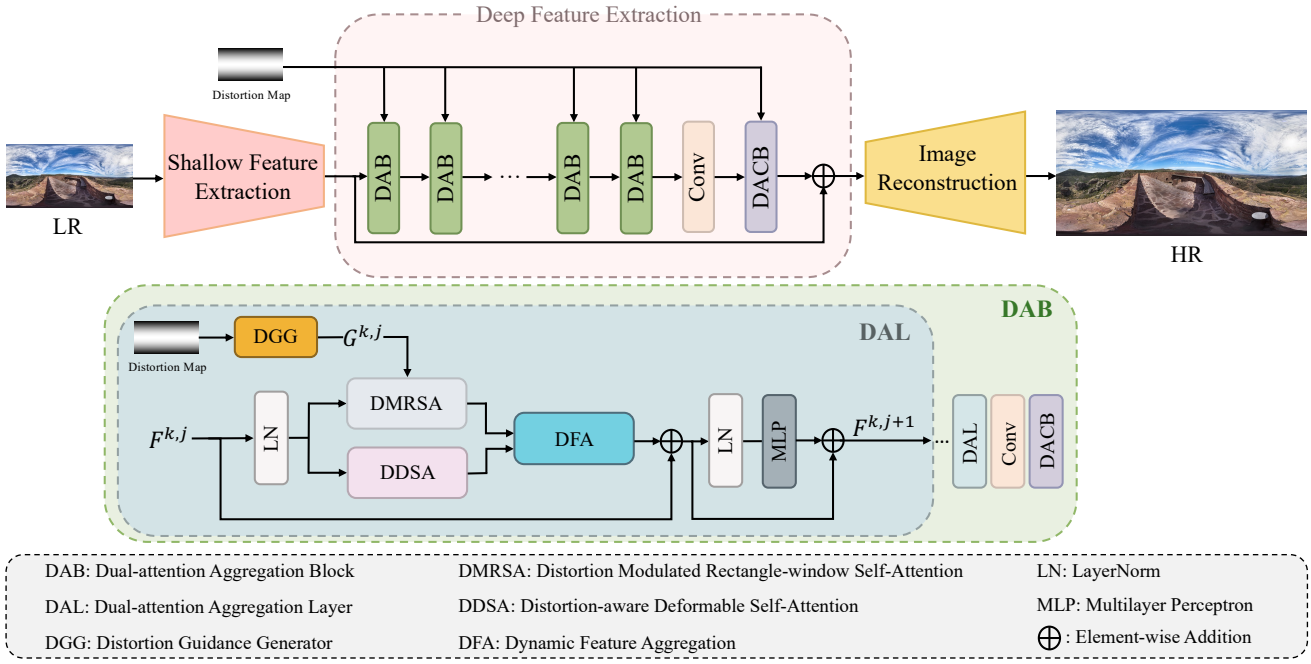


Fig. 4. Overview of the GDGT-OSR architecture (upper part) and the detailed structure of the DAB (bottom part).

then convolved by a 1×1 convolutional layer followed by an activation layer, as follows:

$$v' = \text{ReLU}(\text{Conv}(v)). \tag{10}$$

Latitude-wise expansion (LWE) is then applied to the learned vectors v'. In latitude-wise expansion, each element in the column vector is duplicated W times to restore the size $(H \times W)$ of the original feature map.

$$g^i = \text{LWE}((v')^i), \tag{11}$$

$$\text{where } g^i_{j,1:W} = (v')^i_j, j \in [1, H]. \tag{12}$$

In Eq. (11) and (12), g^i denotes the i-th feature map of the expanded features g generated by LWE, $g^i_{j,1:W}$ denotes all the elements of the j-th row in the i-th feature map of g.

The other branch, i.e., the attention branch, aims to calculate the attention to adaptively weight the expanded features from latitude-wise expansion based on the distortion feature maps f. Specifically, the attention branch consists of a Global Average Pooling (GAP) layer, a convolutional layer, followed by an activation layer. The expanded features from the latitude-wise branch are multiplied by the learned attention weights. In this way, the learned attention weights adaptively weigh the expanded features. This process can be formulated as follows:

$$\hat{g}^i = w^i \cdot g^i, \tag{13}$$

where w represents the learned attention weights, i represents the i-th element, and $\hat{g}$ denotes the distortion guidance.

Depthwise separable convolution [59] is then adopted to refine $\hat{g}$, which consists of a depthwise convolutional layer and a pointwise convolutional layer. The advantage of utilizing the depthwise separable convolution in the refinement is three-fold. Firstly, for a specific latitude, its distortion is similar to that of its neighboring latitudes. A depthwise convolutional layer can leverage the features of local neighboring latitudes to complement and enrich the features at the current latitude. Secondly, a pointwise convolutional layer fuses the features of different channels at the same location, which means that the features from different channels are integrated pixel by pixel and latitude by latitude. Thirdly, the computational cost is reduced by adopting depthwise separable convolution, compared to traditional convolution.

The advantages of the proposed DGG are two-fold. Firstly, DGG can encode the distortion map in latent space and utilize the distortion in an implicit way. Secondly, DGG can leverage the distortion prior to effectively guide the expansion of the attention area. We will demonstrate the effectiveness of DGG in Section IV-C.

*3) **Distortion Modulated Rectangle-Window Self-Attention (DMRSA)**:* As mentioned in Section I, the ERP images are non-uniformly stretched in the projection. Specifically, Fig. 3 shows the stretching ratio of a whole ERP image. From Eq. (2), we understand that the stretching ratio represents the ratio between the area of a microunit on the ideal spherical surface and that on the projection plane. Therefore, it can be inferred that at high latitudes, a small area on the ideal spherical surface is mapped to a large area on the projection plane, while at low latitudes, the distortion is smaller but still exists. Considering this geometric distortion, we propose to adopt the rectangle window self-attention scheme, which can also expand the attention area by providing more informative textures and details [9]. Moreover, based on the intrinsic geometric distortion, we leverage the proposed DGG generator to provide modulations for the key and value features according to the distortion of different latitudes.

Rectangle Window Self-Attention (Rwin-SA). The rectangle window self-attention scheme is illustrated in Fig. 6(a). If $rh < rw$, the rectangle windows are horizontal windows,

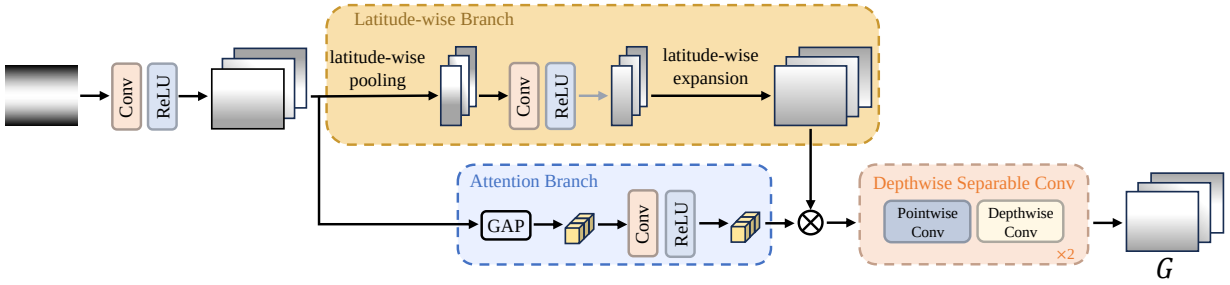
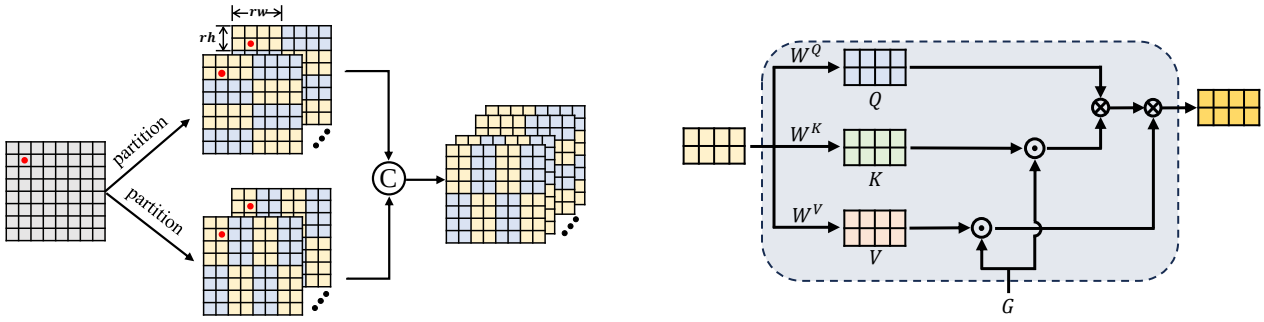


Fig. 5. Illustration of the Distortion Guidance Generator (DGG).



(a) Rectangle Window Self-Attention (b) Distortion Modulated Rectangle Window Self-Attention

Fig. 6. (a) Illustration of the rectangle window self-attention scheme. (b) The overall structure of the distortion modulated rectangle-window self-attention. The key K and value V features are modulated by G learned from the distortion map by the DGG module. W^Q, W^K, W^V are transformation matrices for Q, K, V.

denoted as H-Rwins. Conversely, if $rh > rw$, the rectangle windows are vertical windows, denoted as V-Rwins. For a given input $X \in \mathcal{R}^{H\times W\times C}$, where C is the number of channels, it is divided into a number of $rh \times rw$ non-overlapping H-Rwins or V-Rwins in each attention head. Specifically, given N attention heads, where N is even, these attention heads are equally divided into two parts. The outputs of H-Rwins and V-Rwins are concatenated along the channel dimension. The process for the i-th rectangle window can be expressed as follows:

$$Y^i = \text{Concat}\left(Y_1^i, Y_2^i, ..., Y_N^i\right), \tag{14}$$

where $Y_1^i, ..., Y_{\frac{N}{2}}^i$ are the outputs of the $\frac{N}{2}$ attention heads using H-Rwin, while $Y_{\frac{N}{2}+1}^i, ..., Y_N^i$ are the outputs of the remaining attention heads using V-Rwin.

Distortion Guided Rwin-SA. In ODIs, the pixel density is non-uniform across latitudes due to latitude-wise distortions. As depicted in Fig. 3, the pixel density exhibits the highest level of compactness, with the distortion being mildest around the equator area. Conversely, in the high-latitude areas, particularly in the polar area, the pixel density appears considerably sparser, accompanied by a noticeable distortion. The distorted pixels with different distortions across latitudes should contribute to the reconstruction of the current patch differently. We argue that the neighboring latitude-wise pixels with similar distortions are more related to the current patch because of the projection, which should be paid more attention to and contribute more. Therefore, it is necessary to leverage the geometric distortion information of ERP images for reconstructing a high-quality image. As shown in Fig. 6(b), the output of the DGG generator G is utilized to modulate the key and value features of the Rwin-SA based on the distortion map. The element-wise dot product is conducted between the output modulation features and the key features, as well as the value features, which can be expressed as follows:

$$\tilde{K}_n^i = K_n^i \odot G, \tilde{V}_n^i = V_n^i \odot G, \tag{15}$$

where $\tilde{K}_n^i$ and $\tilde{V}_n^i$ denote the modulated key and value features, respectively, G is the distortion guidance, and $\odot$ represents the element-wise multiplication. Note that D, i.e., the distortion map, is shared across different heads. With query Q_n^i, the modulated key $\tilde{K}_n^i$ and value $\tilde{V}_n^i$, DMRSA is formulated as follows:

$$\tilde{Y}_n^i = \text{SoftMax}\left(\frac{Q_n^i(\tilde{K}_n^i)^T}{\sqrt{d}} + B\right)\tilde{V}_n^i, \tag{16}$$

$$\tilde{Y}^i = \text{Concat}\left(\tilde{Y}_1^i, \tilde{Y}_2^i, ..., \tilde{Y}_N^i\right), \tag{17}$$

$$\text{DMRSA}(X) = \left(\tilde{Y}^1, \tilde{Y}^2, ..., \tilde{Y}^{\frac{H\times W}{rh\times rw}}\right)W^f, \tag{18}$$

where B represents dynamic relative position encoding [60], d is the channel dimension of each head, $\tilde{Y}^i$ denotes the modulated output of the i-th window, and W^f is the projection matrix for feature aggregation. In DMRSA, the key and value features are adaptively modulated by geometric distortion across latitudes.

DMRSA can calibrate and expand the attention area along the direction of stretching distortion, involving most related and self-similar textures and patterns. For those similar and out-of-window textures, we adopt the Distortion-aware Deformable Self-Attention (DDSA) mechanism [11] as shown in Fig. 7, which considers irregular neighborhoods and provides

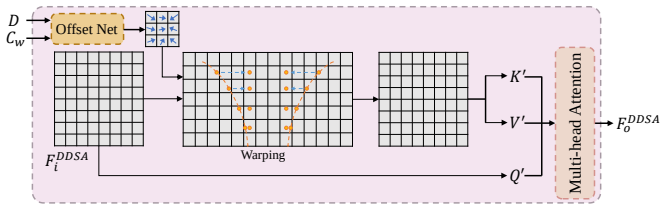


Fig. 7. Illustration of Distortion-aware Deformable Self-Attention (DDSA).

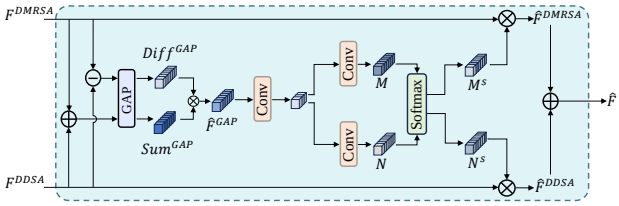


Fig. 8. Illustration of Dynamic Feature Aggregation (DFA).

a distortion-dependent attention area for flexibly modeling features, to complement DMRSA in our DAL.

*4) **Dynamic Feature Aggregation (DFA)**:* Inspired by [61], we propose to dynamically aggregate the features from DMRSA and DDSA in a DFA module to better fuse the features from the dual-attention mechanism. Fig. 8 illustrates the DFA scheme, where features from DMRSA (F^{DMRSA}) and DDSA (F^{DDSA}) are firstly summed element-wisely. The difference between F^{DMRSA} and F^{DDSA} reveals the distinction between DMRSA and DDSA. This distinction implies the features neglected by either DMRSA or DDSA, such as the out-of-window features neglected by DMRSA, to which we should pay more attention. Therefore, we propose to compute the difference between F^{DMRSA} and F^{DDSA}. The difference, denoted as $Diff$, is then utilized to weigh the summed features. Specifically, the Global Average Pooling (GAP) is performed on the $Diff$ and summed features. Thus, two features, i.e., $Diff^{GAP}$ and SUM^{GAP}, which condense the information along the channel dimension, are obtained. The weighted feature $\hat{F}^{GAP}$ is calculated as follows:

$$\hat{F}^{\mathrm{GAP}} = Diff^{\mathrm{GAP}} \odot SUM^{\mathrm{GAP}}. \tag{19}$$

After that, $\hat{F}^{GAP}$ is further condensed by reducing the dimension through a convolutional layer, and then the compact feature is convolved to generate attention vectors, i.e., M and N, for F^{DMRSA} and F^{DDSA}, respectively. Specifically, softmax is conducted on M and N as follows:

$$M_i^s = \frac{e^{M_i}}{e^{M_i} + e^{N_i}}, N_i^s = \frac{e^{N_i}}{e^{M_i} + e^{N_i}}, \tag{20}$$

where M_i and N_i denote the i-th element of M and N, respectively. M_i^s and N_i^s denote the i-th element of M and N after softmax, respectively. The original input features are weighed by the attention vectors:

$$\hat{F}^{DMRSA} = M^s \cdot F^{DMRSA}, \hat{F}^{DDSA} = N^s \cdot F^{DDSA}. \tag{21}$$

Finally, we obtain the output of DFA by summing them, which can be expressed as follows:

$$\hat{F} = \hat{F}^{DMRSA} + \hat{F}^{DDSA}. \tag{22}$$

DFA takes into account the distinction between the features from different self-attention mechanisms, which should be addressed in feature fusion. By complementing these features, DFA learns to adjust them adaptively.

C. Loss Function

Due to the non-uniform pixel densities, it is unreasonable to calculate the pixel-wise errors in ERP images using the standard $l1$ or $l2$ loss. The pixel-wise errors at different locations in the ERP image should have different significance to reflect such uneven distribution. Therefore, we adopt the Weighted-to-Spherically $l1$ loss (WS-$l1$ loss [62]) as the loss function in the training process. The WS-$l1$ loss can calibrate the pixel-wise errors based on the distortion map as follows:

$$\text{WS-}l1 = \sum_{h=1}^{H}\sum_{w=1}^{W} |I^{gt}(h,w) - I^{o}(h,w)| D(h,w), \tag{23}$$

where I^{gt} and I^{o} denote the ground-truth image and the output image, respectively. $D(h,w)$ represents the value of the distortion map at the coordinate (h,w). From Eq. (6), we can know that the weights are larger at lower latitudes, while the weights are smaller at higher latitudes. From the perspective of non-uniform pixel density, this design is reasonable: the pixel density is more compact at lower latitudes while it is sparser at higher latitudes, which means more pixels are distributed around lower latitudes, so these pixels should contribute more to the loss function, and vice versa.

IV. Experiments

A. Experimental Settings

Data and Evaluation. In our experiments, we adopted the ODI-SR dataset [1], which consists of 800 HR images, as our training dataset. The resolution of the HR ERP images is 1024×2048. However, when training on such a small dataset, the model will easily suffer from overfitting. Therefore, in order to avoid the overfitting problem, we follow OSRT [11] and include the DF2K-ERP dataset as a part of our training dataset. The DF2K-ERP dataset generates synthetic ERP images from 2D plain images in the DF2K dataset. The DF2K-ERP dataset consists of 146,000 HR ERP image patches with a patch size larger than 256×256. The evaluation datasets include the ODI-SR testing dataset [1] and the SUN360 Panorama dataset [63], both of which contain 100 images. The resolution of the HR ERP images from these two different testing datasets is 1024×2048. In our experiments, we implement scaling factors of 2× and 4× by downsampling the HR ERP images with the corresponding factors. Furthermore, following OSRT [11], fisheye downsampling is applied to the HR ERP images to generate LR-HR image pairs. Besides the commonly used PSNR and SSIM, WS-PSNR and WS-SSIM re-weighted with distortion are also adopted as our evaluation metrics.

Implementation Details. We conducted our experiments with PyTorch [67]. During the training stage, the batch size is 16, and the patch size is 64×64. If there is no extra

TABLE I
QUANTITATIVE COMPARISON OF DIFFERENT METHODS IN TWO PUBLIC DATASETS, I.E., ODI-SR [1] AND SUN 360 PANORAMA [63]. THE SCALING FACTORS ARE 2 AND 4. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Scale	ODI-SR [1]				SUN 360 Panorama [63]			
		PSNR	SSIM	WS-PSNR	WS-SSIM	PSNR	SSIM	WS-PSNR	WS-SSIM
Bicubic	×2	28.21	0.8215	27.61	0.8156	28.14	0.8118	28.01	0.8321
RCAN [64]		30.26	0.8777	29.61	0.8739	30.84	0.8793	31.39	0.9008
SRResNet [24]		30.12	0.8703	29.56	0.8685	30.57	0.8692	31.13	0.8937
EDSR [65]		30.18	0.8740	29.57	0.8708	30.70	0.8743	31.24	0.8970
SwinIR [8]		30.64	0.8821	30.00	0.8777	31.33	0.8855	31.98	0.9059
HAT [31], [66]		30.67	0.8821	30.05	0.8780	31.37	0.8858	32.06	0.9065
RGT [32]		30.46	0.8781	29.86	0.8753	31.10	0.8793	31.79	0.9023
OSRT [11]		30.77	0.8846	30.11	0.8795	31.52	0.8888	32.14	0.9081
GDGT-OSR (ours)		**30.87**	**0.8863**	**30.21**	**0.8811**	**31.67**	**0.8910**	**32.33**	**0.9099**
Bicubic	×4	25.59	0.7118	24.95	0.6923	25.29	0.6993	24.90	0.7083
RCAN [64]		26.90	0.7618	26.21	0.7486	27.10	0.7649	27.01	0.7851
SRResNet [24]		26.91	0.7592	26.24	0.7447	27.10	0.7613	26.99	0.7802
EDSR [65]		26.87	0.7612	26.18	0.7467	27.11	0.7643	26.97	0.7830
SwinIR [8]		27.31	0.7735	26.61	0.7589	27.71	0.7804	27.64	0.7996
HAT [31], [66]		27.29	0.7717	26.61	0.7578	27.67	0.7783	27.60	0.7982
RGT [32]		27.31	0.7711	26.63	0.7571	27.69	0.7782	27.65	0.7982
OSRT [11]		27.41	0.7762	26.70	0.7609	27.84	0.7835	27.77	0.8020
GDGT-OSR (ours)		**27.44**	**0.7776**	**26.72**	**0.7623**	**27.92**	**0.7850**	**27.81**	**0.8036**

TABLE II
QUANTITATIVE COMPARISON OF DIFFERENT METHODS UNDER LARGE SCALING FACTORS, I.E., SPECIFICALLY ×8 AND ×16. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Scale	ODI-SR [1]				SUN 360 Panorama [63]			
		PSNR	SSIM	WS-PSNR	WS-SSIM	PSNR	SSIM	WS-PSNR	WS-SSIM
Bicubic	×8	24.24	0.6525	23.51	0.6204	23.73	0.6369	23.22	0.6324
SwinIR [8]		25.09	0.6873	24.38	0.6619	24.97	0.6881	24.57	0.6953
HAT [31], [66]		25.15	0.6909	24.42	0.6653	25.12	0.6936	24.69	0.7008
RGT [32]		24.86	0.6889	24.08	0.6632	24.68	0.6893	24.23	0.6963
OSRT [11]		25.28	0.6929	24.53	0.6664	25.32	0.6964	24.86	0.7027
GDGT-OSR (ours)		**25.32**	**0.6949**	**24.60**	**0.6687**	**25.42**	**0.6994**	**25.00**	**0.7068**
Bicubic	×16	22.66	0.6171	21.90	0.5785	22.04	0.6025	21.42	0.5899
SwinIR [8]		23.33	0.6388	22.58	0.6046	22.86	0.6292	22.28	0.6230
HAT [31], [66]		23.32	0.6401	22.55	0.6057	22.87	0.6306	22.27	0.6246
RGT [32]		23.17	0.6391	22.40	0.6049	22.70	0.6297	22.10	0.6236
OSRT [11]		23.49	0.6421	22.71	0.6074	23.11	0.6339	22.48	0.6278
GDGT-OSR (ours)		**23.54**	**0.6425**	**22.78**	**0.6087**	**23.21**	**0.6356**	**22.60**	**0.6303**

demonstration, the default rh and rw are 8 and 64 for H-Rwins, while for V-Rwins, the default rh and rw are 64 and 8. The sizes of rh and rw will be explored in the later Section IV-C. The initial learning rate is set as 2×10^{-4}, and reduced by half at 250K, 400K, 450K, and 475K iterations. The total number of training iterations is 500K. The whole model consists of 6 DABs, and each DAB contains 6 DALs. The dimension of the embedding feature is 156. Adam [68] is adopted as the optimizer in the training process, with $\beta_1 = 0.9$ and $\beta_2 = 0.99$.

B. Comparisons with State-of-the-Art Methods

Quantitative Results. Table I shows the quantitative comparisons among different methods under $\times 2$ and $\times 4$ scaling factors. OSRT is an ODI-SR method, and we report the results as in [11]. The remaining compared methods, e.g., RCAN [64], SRResNet [24], EDSR [65], SwinIR [8], HAT [31] and RGT [32] are SR methods for 2D plain images. Thus, we retrained them on the ODI-SR dataset plus the auxiliary DF2K-ERP dataset. RCAN, SRResNet, EDSR, and SwinIR have inferior performance to OSRT and our GDGT-OSR, which demonstrates that the key to ODISR is the adaptiveness of the distortion information. With the dual-attention aggregation and distortion-guided components, our GDGT-OSR has strong distortion-awareness and distortion transformation abilities and achieves state-of-the-art (SOTA) performance on the two public datasets for both $\times 2$ and $\times 4$ tasks. This illustrates that our proposed method can better exploit and transform the distortion map to enhance the SR performance than OSRT.

To further evaluate the effectiveness of GDGT-OSR on larger scaling factors, we conducted experiments with scaling factors of $\times 8$ and $\times 16$. The results, presented in Table II, compare GDGT-OSR with other SOTA SR methods, including SwinIR [8], HAT [31], RGT [32], and OSRT [11]. As shown in Table II, despite the increased difficulty associated with larger scaling factors, our proposed GDGT-OSR consistently outperforms other SOTA 2D plain image SR and ODISR methods. This demonstrates the robustness and generalization capability of GDGT-OSR across different scaling factors, highlighting its effectiveness.

Qualitative Results. Fig. 9 presents a visual comparison of different SR methods on the ODI-SR dataset with a scaling factor of 4. The red-boxed patches are zoomed in for better

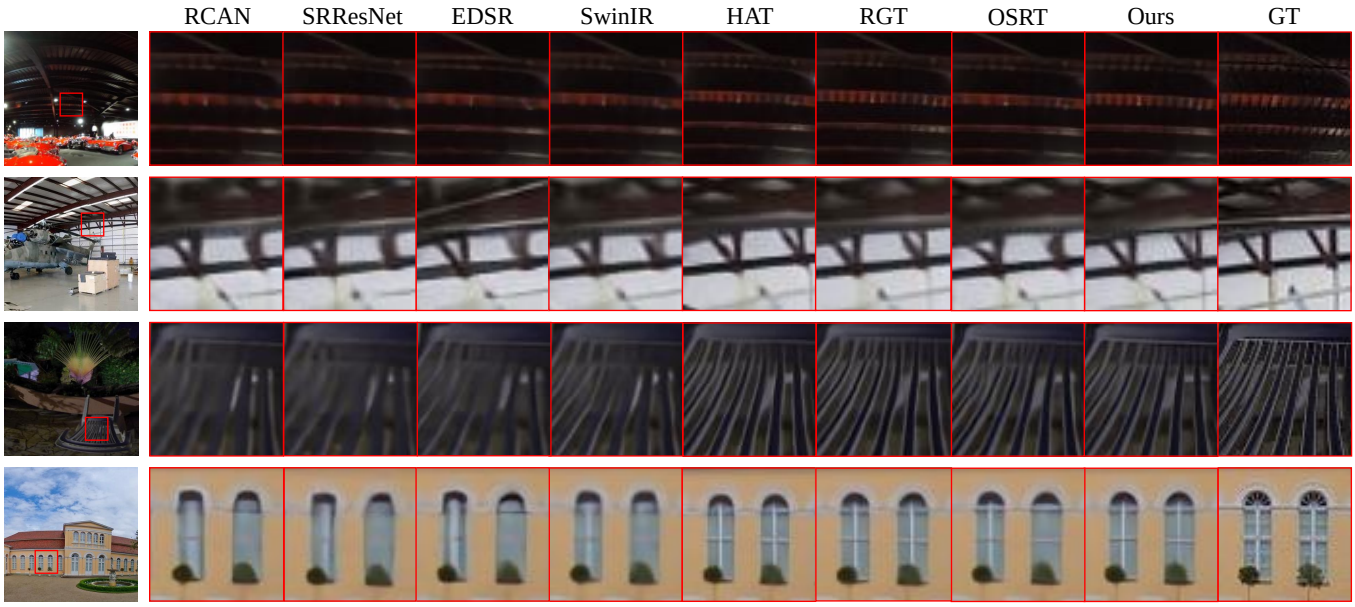


Fig. 9. Visual comparison of different methods on the ODI-SR dataset [1] with a scaling factor of 4. The red boxes highlight zoomed-in patches for better visualization.

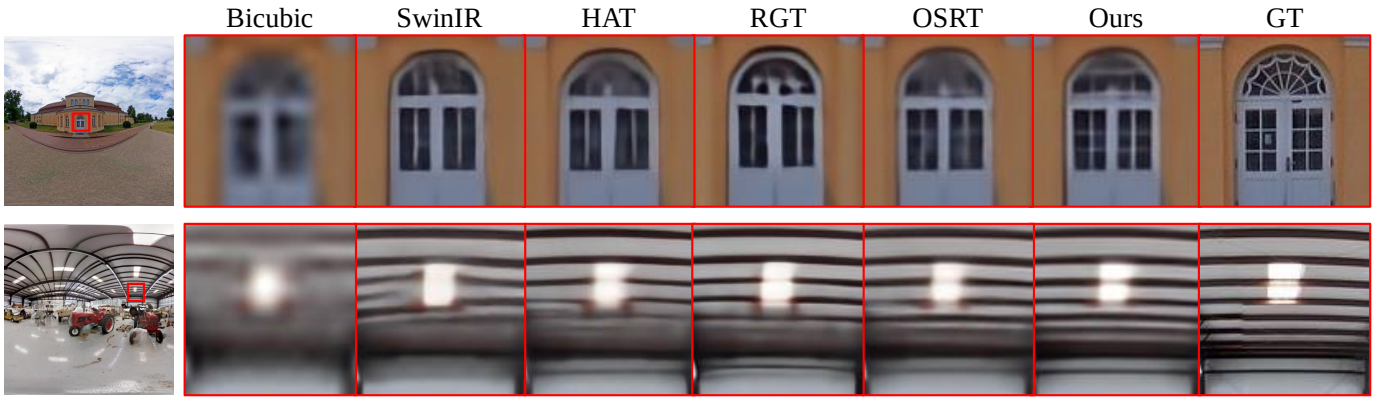


Fig. 10. Visual comparison of different methods on the ODI-SR dataset [1] with a scaling factor of 8. The red boxes highlight zoomed-in patches for better visualization.

visualization. Although the 2D plain image SR methods are retrained with the ODI datasets, their performances are still worse than those of OSRT and our GDGT-OSR. For example, as shown in the third sample in Fig. 9, the ODISR methods, i.e., OSRT and our GDGT-OSR, can restore more visually pleasing results with less blurriness and distortion of the chair compared to the preceding four methods. This demonstrates that the inclusion and exploration of the distortion map contribute to the superiority and effectiveness of the last two ODISR methods. Although OSRT represents SOTA performance in ODISR methods, our proposed method surpasses it by restoring more details and richer textures. For example, in the last sample in Fig. 9, our method can restore comparatively complete window frames, while some parts of the window frames are missing in the visual result of OSRT.

Fig. 10 shows the visual comparison of different methods under the $\times 8$ scaling factor. It is more difficult to super-resolve the images under a larger scaling factor because more details are lost in the LR images. As shown in Fig. 10, under the $\times 8$ scaling factor, most of the SOTA SR methods, including SwinIR [8], HAT [31], RGT [32] and OSRT [11], struggle to recover the details and textures, leading to unsatisfactory visual results. With the proposed dual-attention and distortion-guided mechanisms, GDGT-OSR effectively manages distortion transformation and can recover more details for the ERP images even under the challenging large scaling factor. To summarize, the above visual results demonstrate the superiority and effectiveness of the proposed GDGT-OSR under various scaling factors.

C. Ablation Study

Impacts of Self-Attentions. We investigate the influence of DMRSA and DDSA. As shown in Table III, compared to the variant without both DMRSA and DDSA, the performance

TABLE III
ABLATION STUDY OF DMRSA AND DDSA. THE SCALING FACTOR IS 4.

DMRSA	DDSA	ODI-SR [1]				SUN 360 Panorama [63]			
		PSNR	SSIM	WS-PSNR	WS-SSIM	PSNR	SSIM	WS-PSNR	WS-SSIM
✗	✗	27.10	0.7666	26.39	0.7503	27.46	0.7714	27.28	0.7883
✗	✓	27.38	0.7753	26.66	0.7599	27.77	0.7819	27.70	0.8005
✓	✗	27.39	0.7759	26.66	0.7605	27.83	0.7830	27.72	0.8015
✓	✓	**27.44**	**0.7776**	**26.72**	**0.7623**	**27.92**	**0.7850**	**27.81**	**0.8036**

TABLE IV
ABLATION STUDY OF DGG AND DIFF. THE SCALING FACTORS ARE 2 AND 4.

Scale	DGG	Diff	ODI-SR [1]				SUN 360 Panorama [63]			
			PSNR	SSIM	WS-PSNR	WS-SSIM	PSNR	SSIM	WS-PSNR	WS-SSIM
×2	✗	✗	30.83	**0.8863**	30.17	0.8810	31.60	0.8908	32.25	0.9097
	✗	✓	30.84	0.8860	30.19	0.8809	31.63	0.8904	32.29	0.9095
	✓	✗	**30.87**	**0.8863**	30.20	0.8809	**31.67**	0.8909	32.32	0.9098
	✓	✓	**30.87**	**0.8863**	**30.21**	**0.8811**	**31.67**	**0.8910**	**32.33**	**0.9099**
×4	✗	✗	27.43	0.7775	26.71	0.7621	27.87	0.7847	27.78	0.8032
	✗	✓	**27.44**	0.7775	**26.72**	0.7621	27.88	0.7848	27.79	0.8033
	✓	✗	27.43	**0.7779**	26.71	**0.7626**	27.89	**0.7853**	27.80	**0.8038**
	✓	✓	**27.44**	0.7776	**26.72**	0.7623	**27.92**	0.7850	**27.81**	0.8036

TABLE V
RESULT COMPARISON AMONG VARIANTS OF DGG. THE SCALING FACTOR IS 4.

Variants	ODI-SR [1]				SUN 360 Panorama [63]			
	PSNR	SSIM	WS-PSNR	WS-SSIM	PSNR	SSIM	WS-PSNR	WS-SSIM
Direct	27.42	0.7769	26.71	0.7618	27.86	0.7841	27.79	0.8031
One conv	27.43	0.7765	26.71	0.7614	27.88	0.7841	27.80	0.8029
w/o atten	**27.44**	0.7771	**26.72**	0.7619	27.88	0.7843	27.79	0.8029
GDGT-OSR	**27.44**	**0.7776**	**26.72**	**0.7623**	**27.92**	**0.7850**	**27.81**	**0.8036**

of the variant with DMRSA improves significantly. The variant with only DMRSA performs better than that with only DDSA. This illustrates that the proposed DMRSA mechanism dominantly contributes to the ODISR. When both DMRSA and DDSA are incorporated into the network architecture, the performance is further improved, which demonstrates the effectiveness of the proposed GDGT-OSR framework. Moreover, we investigate the impacts of the offsets in DDSA. Fig. 11 shows the visual comparison of offset maps in DDSAs of both OSRT [11] and GDGT-OSR. We can see that the offsets from our GDGT-OSR are larger than those from OSRT around the polar area where the distortion is the most severe, while the offsets from both OSRT and GDGT-OSR around the equator area are small due to negligible distortion. This observation demonstrates that, when combined with DMRSA, DDSA can adapt better to the distortion and further broaden its attention areas to involve a larger range of pixels. The above results and analysis show the effectiveness of collaboration between DMRSA and DDSA.

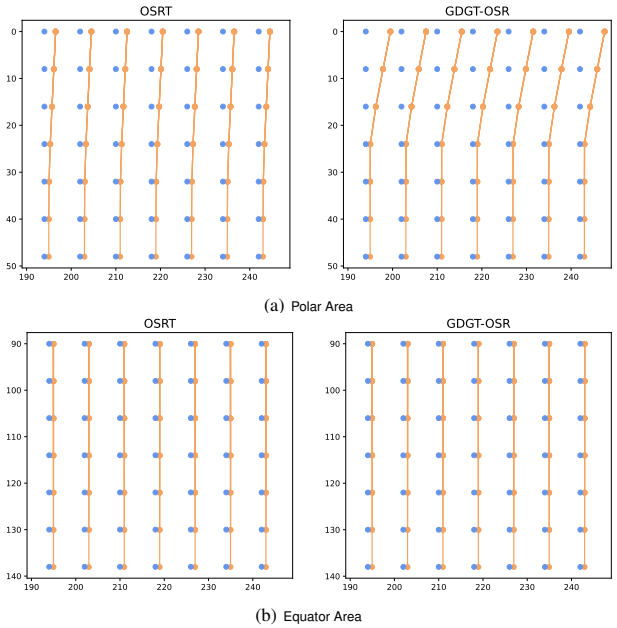


Fig. 11. Visualization of offset maps in OSRT [11] and GDGT-OSR. Reference and deformed points are colored in blue and orange. The horizontal and vertical axes denote the x and y coordinates in the image, respectively. (a) Visual comparison of offset maps around the polar area. (b) Visual comparison of offset maps around the equator area.

Impacts of Distortion Guidance Generator. DGG is designed to thoroughly exploit the modulation information of the distortion map for Rwin-SA. As shown in Table IV, we present the results of our model without and with DGG in the second and fourth rows under both $\times 2$ and $\times 4$ scaling factors. With DGG, the model performance surpasses its counterpart (without DGG) across all PSNR-related and SSIM-related metrics, demonstrating the effectiveness of DGG. Fig. 12 shows the qualitative comparison of our model without and with DGG. As shown in the red cropped patches, while the visual results of the model without DGG differ from the ground truth and appear distorted, the model with DGG can restore more visually pleasing results. It demonstrates that our

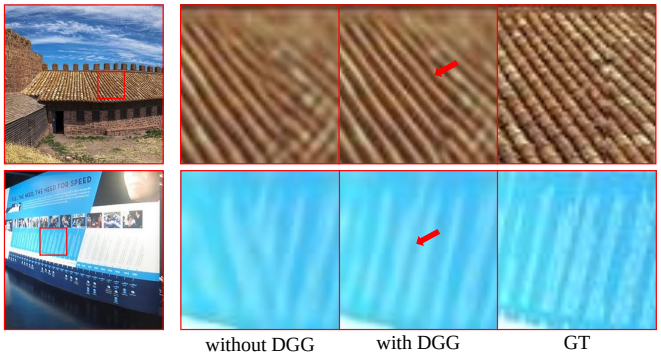


Fig. 12. Visual comparison of our model without and with DGG. The scaling factor is 4.

model benefits from DGG in enhancing SR performances.

To further investigate the effects of DGG, we conducted some ablation experiments on the architecture of DGG, as shown in Table V, i.e., different ways of exploiting the distortion map. 'Direct' represents that the distortion map is replicated along the channel dimension, and then directly utilized to modulate the feature maps. 'One conv' means that we use one convolutional layer to replace the proposed DGG. 'w/o atten' denotes the DGG without the attention branch. It can be seen that the performance deteriorates most in the 'direct' variant. Although we use one convolutional layer to encode the distortion map, the performance is almost the same as that of the 'direct' variant. Thus, we can conclude that exploiting the distortion map in simple ways brings little improvement. Moreover, the performance of the 'w/o atten' variant is inferior to that of GDGT-OSR, which involves a complete DGG. DGG not only encodes the distortion map into feature maps that match the dimension of the key/value features in Rwin-SA, but also effectively exploits the information of the distortion map.

Impacts of Diff. In DFA, we calculate the difference between F^{DMRSA} and F^{DDSA}, which is used to adaptively weigh the combined features. As shown in Table IV, the performance slightly degrades without Diff. Fig. 13 shows the visual comparison between our model without and with the Diff design in DFA. We can see that some details are missing without Diff, while more details are restored with Diff. It is worth noting that computing the Diff does not require additional parameters, but it can bring improvements in restoring details. In Table IV, we can observe that the SR performance is moderately impacted when both DGG and Diff are absent simultaneously, especially under the $\times 2$ scaling factor. This demonstrates the necessity and importance of both utilizing the distortion map and leveraging the distinction between features from DMRSA and DDSA.

Impacts of Dynamic Feature Aggregation. We propose to aggregate features from DMRSA and DDSA in a dynamic way, based on the learned weights. To verify the efficacy of DFA, we compare the SR performance of different aggregation methods on the SUN 360 panorama dataset, as illustrated in Table VI. 'Addition' denotes that the two features are summed directly to obtain the final feature. 'Concatenation' denotes

TABLE VI
RESULT COMPARISON AMONG DIFFERENT WAYS OF AGGREGATION ON SUN 360 PANORAMA [63]. THE SCALING FACTOR IS 4.

Ways of Aggregation	PSNR	SSIM	WS-PSNR	WS-SSIM
Addition	27.88	**0.7850**	27.79	**0.8037**
Concatenation	27.89	0.7849	27.80	**0.8037**
DFA	**27.92**	**0.7850**	**27.81**	0.8036

Fig. 13. Visual comparison of our model without and with the difference calculation in DFA. The scaling factor is 4.

that the two features are concatenated along the channel dimension, and then convolved by a convolutional layer to reduce the number of channels to match that of the input features. From Table VI, we can see that DFA achieves the best or comparative performance across all evaluation metrics compared with other aggregation methods.

Impacts of Window Size. We investigate the effects of different sizes of rectangle windows, as shown in Table VII. A window size of (32/8) means that the height and width are 32 and 8 for vertical windows, respectively, and 8 and 32 for horizontal windows. As demonstrated in Table VII, the network with a window size of (64/8) outperforms those with window sizes of (32/8) and (64/4). This improvement may be because the rectangle window with size (64/8) can effectively capture features of distorted content in a more appropriate range of area.

Impacts of Window Shape. Besides the window size, the effects of the window shapes of the rectangular windows in DMRSA are also worth exploring. To this end, we conducted experiments using various configurations of rectangular window orientations, i.e., vertical only, horizontal only, and a combination of both vertical and horizontal windows with 50% each. The results, as presented in Table VIII, indicate that the configuration with only horizontal rectangular windows outperforms the one with only vertical rectangular windows across both testing datasets. This advantage is likely attributable to the substantial horizontal stretch of projected ERP images, enabling horizontal rectangular windows to capture more self-similar features for enhanced reconstruction. Nonetheless, as demonstrated in Table VIII, the variant incorporating both vertical and horizontal rectangular windows achieves superior performance. Despite the primary horizontal distortion of ERP images, certain self-similar textures and patterns may still exist vertically, which can be effectively captured using vertical windows. Consequently, our DMRSA mechanism employs a combination of both vertical and horizontal rectangular windows.

Impacts of Training Loss. Given the non-uniform pixel distribution, we adopt the WS-$l1$ loss as our training loss function to calculate pixel-wise errors in ERP images. From

TABLE VII
ABLATION STUDY OF THE WINDOW SIZE OF DMRSA. THE SCALING FACTOR IS 4.

Window size	ODI-SR [1]				SUN 360 Panorama [63]			
	PSNR	SSIM	WS-PSNR	WS-SSIM	PSNR	SSIM	WS-PSNR	WS-SSIM
32/8	27.39	0.7768	26.67	0.7616	27.82	0.7837	27.73	0.8024
64/8	**27.44**	**0.7776**	**26.72**	**0.7623**	**27.92**	**0.7850**	**27.81**	**0.8036**
64/4	27.39	0.7746	26.67	0.7594	27.81	0.7817	27.71	0.8001

TABLE VIII
ABLATION STUDY OF THE WINDOW SHAPE OF DMRSA. THE SCALING FACTOR IS 4.

Window shape	ODI-SR [1]				SUN 360 Panorama [63]			
	PSNR	SSIM	WS-PSNR	WS-SSIM	PSNR	SSIM	WS-PSNR	WS-SSIM
Vertical	27.40	0.7773	26.68	0.7619	27.83	0.7842	27.74	0.8027
Horizontal	27.42	0.7771	26.70	0.7616	27.86	0.7842	27.76	0.8026
Vertical & Horizontal	**27.44**	**0.7776**	**26.72**	**0.7623**	**27.92**	**0.7850**	**27.81**	**0.8036**

TABLE IX
RESULT COMPARISON AMONG DIFFERENT TRAINING LOSSES. THE SCALING FACTOR IS 4.

loss	ODI-SR [1]				SUN 360 Panorama [63]			
	PSNR	SSIM	WS-PSNR	WS-SSIM	PSNR	SSIM	WS-PSNR	WS-SSIM
$l1$	27.43	**0.7779**	26.70	**0.7625**	27.89	**0.7854**	27.78	**0.8039**
$l1$-(WS-$l1$)	27.41	0.7763	26.70	0.7611	27.82	0.7834	27.75	0.8019
WS-$l1$	**27.44**	0.7776	**26.72**	0.7623	**27.92**	0.7850	**27.81**	0.8036

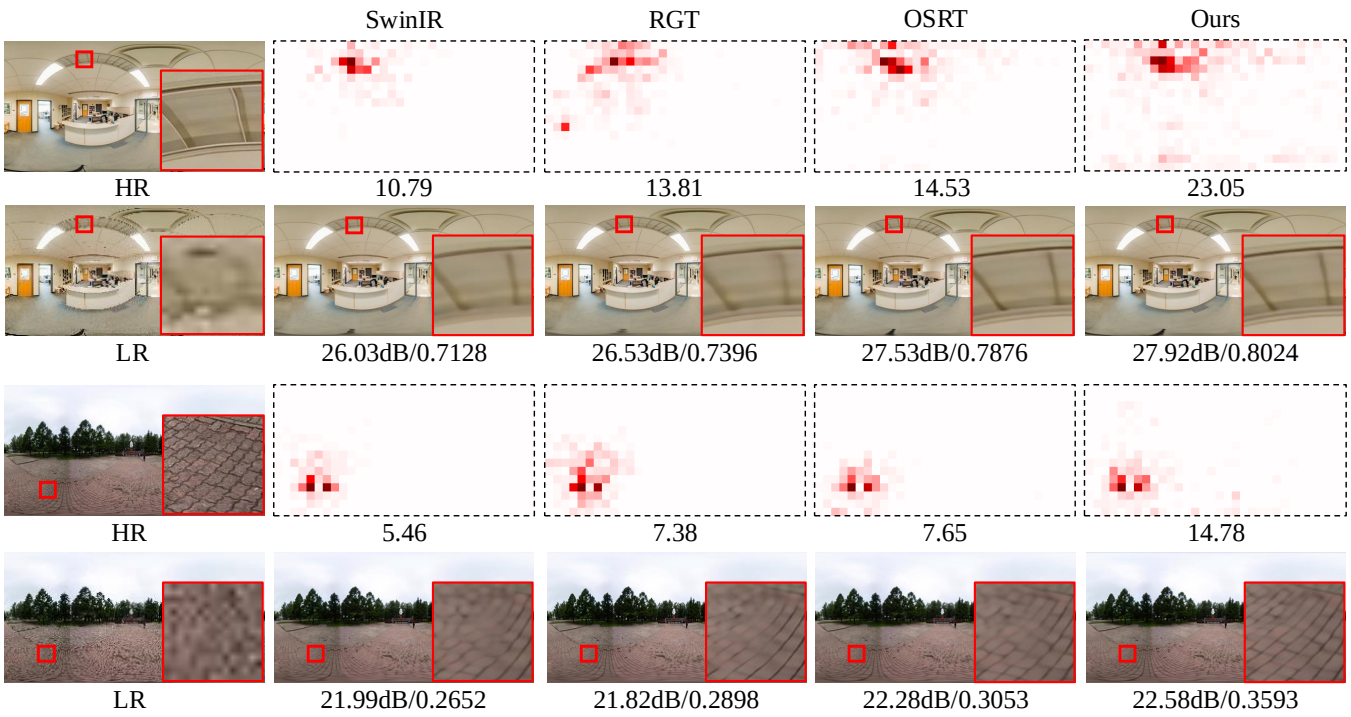


Fig. 14. Visual comparison of local attribution maps and SR results among different methods. The Diffusion Index (DI) is given below the local attribution maps. A higher DI indicates a wider range of the involved pixels, and vice versa.

Table IX, we can see that WS-$l1$ outperforms $l1$ in the evaluation metrics of PSNR and WS-PSNR, while $l1$ performs better than WS-$l1$ in the evaluation metrics of SSIM and WS-SSIM. Referred to Eq. (23), it is probable that in the WS-$l1$ loss, luminance and contrast are affected by the multiplication of the distortion map in the training process, which leads to the degradation of SSIM and WS-SSIM. However, the performance of PSNR and WS-PSNR demonstrates that WS-$l1$ loss can improve and relieve pixel-wise errors.

Eq. (23) indicates that the WS-$l1$ loss is obtained by weighing the $l1$ loss with the distortion map D. To further study the effectiveness and rationality of WS-$l1$ loss as the

TABLE X
MODEL COMPLEXITY COMPARISON. THE SCALING FACTOR IS 8.

Model	#Params(M)	#Multi-Adds(G)	PSNR	WS-PSNR
SwinIR	12.05	60.59	25.09	24.38
HAT	20.92	96.05	25.15	24.42
RGT	13.51	62.17	24.86	24.08
OSRT	12.08	54.50	25.28	24.53
GDGT-OSR	15.99	69.74	25.32	24.60

TABLE XI
COMPARISON OF AVERAGE DI VALUES AT FOUR DIFFERENT PIXELS IN THE SUN 360 PANORAMA DATASET [63].

	(50, 90)	(100, 125)	(150, 200)	(200, 110)
EDSR [65]	1.22	1.23	1.16	1.03
RCAN [64]	12.29	8.77	8.09	8.25
SwinIR [8]	5.63	5.09	4.88	3.97
OSRT [11]	8.30	6.88	6.84	8.61
GDGT-OSR	**16.18**	**11.15**	**10.37**	**19.10**

training loss, we train the model with a loss that weighs the $l1$ loss with (1-D). This loss assigns larger weights to the high-latitude areas and smaller weights to the low-latitude areas, which is opposite to the WS-$l1$ loss. It can easily be written as $l1 - (\text{WS-}l1)$. As shown in Table IX, the model trained with $l1 - (\text{WS-}l1)$ yields inferior results compared to those trained with $l1$ and WS-$l1$. These findings suggest that prioritizing lower-latitude regions, where pixels are more densely distributed, is crucial during the training process of ODISR.

D. Model Complexity

We compare the model complexity of our methods with other SOTA methods, as summarized in Table X. We evaluate the model performance on the ODI-SR dataset under the $\times 8$ scaling factor. The number of Multi-Adds is calculated for an input size of 64×64. As shown in Table X, the number of parameters and Multi-Adds of HAT [31] are much larger than those of other methods. However, despite its substantial computational complexity, the performance of HAT [31] remains inferior to both OSRT [11] and our GDGT-OSR. On the contrary, compared to other methods, GDGT-OSR achieves significant performance improvements with a comparable or only marginally increased number of parameters. This highlights the superiority and effectiveness of our GDGT-OSR.

E. Exploration on Involved Area.

To investigate the effects of the range of involved pixels, we randomly selected four different pixels from each image in the SUN 360 Panorama testing dataset and calculated their average DI values, as shown in Table XI. Higher DI values indicate a wider range of pixels involved. Note that the coordinates of the upper left corner of an image are (0,0). As shown in Table XI, the proposed method achieves the highest DI values at the four selected pixels from different latitudes compared to other methods, which demonstrates that our method has the widest range of involved area. Fig. 1 visually shows that our method's area of contribution for reconstructing the patch in the red box is the widest with more self-similar textures, resulting in superior SR performance. More visual results of local attribution maps are illustrated in Fig. 14, demonstrating that the proposed GDGT-OSR method effectively leverages a broader range of pixels. This capability enables it to produce more visually pleasing and higher-quality SR results. The above results indicate that the proposed mechanisms can enlarge the attention area and improve SR performance.

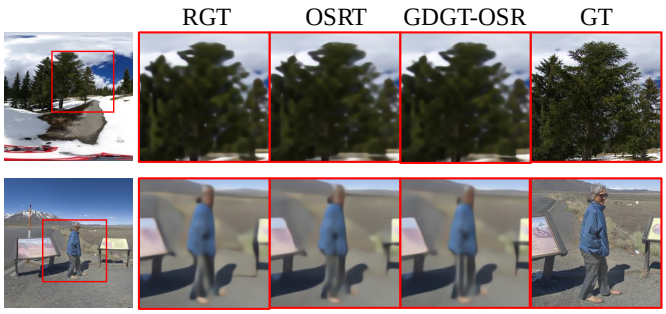


Fig. 15. Failed cases under the $\times 16$ scaling factor.

V. SUMMARY

A. Limitations

Inference Speed. Typically, ODI images possess high resolutions, such as 2K, 4K, or even 8K. However, the inference speed of the proposed method is limited when processing these high-resolution ODI images due to the computationally intensive self-attention mechanisms employed. Furthermore, during inference, if the model size is excessively large, GPU memory may become insufficient to handle such high-resolution inputs. Although GDGT-OSR can produce high-quality SR results, its practical application is constrained in scenarios requiring real-time super-resolution of high-resolution ODI images with limited computational resources.

Large Scaling Factors. When super-resolving objects with numerous high-frequency details, e.g., trees and humans, under the large scaling factor of $\times 16$, the qualitative results produced by GDGT-OSR are not visually satisfactory, as illustrated in Fig. 15. The primary reason is the significant loss of fine-grained information in the LR images when using a large scaling factor. With limited available information in the LR images, it becomes challenging for GDGT-OSR, as well as other SOTA methods, such as RGT [32] and OSRT [11], to reconstruct visually pleasing results for these high-frequency components in such demanding conditions.

B. Conclusion

In this paper, we propose a Geometric Distortion Guided Transformer for Omnidirectional image Super-Resolution, named GDGT-OSR. Considering the geometric distortion in ERP images, we propose a Distortion Modulated Rectangle-window Self-Attention (DMRSA) mechanism, integrated with Distortion-aware Deformable Self-Attention (DDSA), to adapt to the unevenly distorted content. In this way, GDGT-OSR captures features from the attention areas with various shapes,

aiming to calibrate the attention regions and facilitate their expansion, capturing more self-similar and related textures. To exploit the distortion map, we propose a Distortion Guidance Generator (DGG) to transform geometric distortion into distortion guidance, which is leveraged to modulate the key and value features in Rwin-SA. Furthermore, we adaptively aggregate two features generated by DMRSA and DDSA through a Dynamic Feature Aggregation (DFA) module. Experiment results demonstrate that GDGT-OSR can restore more details and richer textures over other methods, achieving SOTA performance on omnidirectional image super-resolution.

REFERENCES

[1] X. Deng, H. Wang, M. Xu, Y. Guo, Y. Song, and L. 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*, 2021, pp. 9189–9198.

[2] Z.-S. Liu, W.-C. Siu, and Y.-L. Chan, "Photo-realistic image super-resolution via variational autoencoders," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1351–1365, 2020.

[3] Z. Wu, W. Liu, J. Li, C. Xu, and D. Huang, "Sfhn: spatial-frequency domain hybrid network for image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 11, pp. 6459–6473, 2023.

[4] Y. Huang, J. Li, Y. Hu, H. Huang, and X. Gao, "Deep convolution modulation for image super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 5, pp. 3647–3662, 2024.

[5] R. G. d. A. Azevedo, N. Birkbeck, F. De Simone, I. Janatra, B. Adsumilli, and P. Frossard, "Visual distortions in 360° videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2524–2537, 2019.

[6] J. Gu and C. Dong, "Interpreting super-resolution networks with local attribution maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9199–9208.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*, 2020.

[8] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.

[9] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yuan *et al.*, "Cross aggregation transformer for image restoration," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 25 478–25 490.

[10] Y. Yoon, I. Chung, L. Wang, and K.-J. 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*, 2022, pp. 5677–5686.

[11] F. Yu, X. Wang, M. Cao, G. Li, Y. Shan, and C. Dong, "Osrt: Omnidirectional image super-resolution with distortion-aware transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 283–13 292.

[12] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *International Conference on Learning Representations*, 2021.

[13] F. Zhou, S.-T. Xia, and Q. Liao, "Nonlocal pixel selection for multisurface fitting-based super-resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 12, pp. 2013–2017, 2014.

[14] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5197–5206.

[15] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2009, pp. 349–356.

[16] T. Michaeli and M. Irani, "Nonparametric blind super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2013, pp. 945–952.

[17] G. Qiu, "A progressively predictive image pyramid for efficient lossless coding," *IEEE Transactions on Image Processing*, vol. 8, no. 1, pp. 109–115, 1999.

[18] ——, "Interresolution look-up table for improved spatial magnification of image," *Journal of Visual Communication and Image Representation*, vol. 11, no. 4, pp. 360–373, 2000.

[19] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

[20] X. Li, K. M. Lam, G. Qiu, L. Shen, and S. Wang, "Example-based image super-resolution with class-specific predictors," *Journal of Visual Communication and Image Representation*, vol. 20, no. 5, pp. 312–322, 2009.

[21] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[23] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4681–4690.

[24] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision Workshops*, 2018, pp. 0–0.

[25] K. C. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, "Glean: Generative latent bank for large-factor image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 245–14 254.

[26] K. C. Chan, X. Xu, X. Wang, J. Gu, and C. C. Loy, "Glean: Generative latent bank for image super-resolution and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3154–3168, 2022.

[27] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[28] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in *Proceedings of the International Conference on Learning Representations*, 2018.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 30, 2017.

[30] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[31] X. Chen, X. Wang, W. Zhang, X. Kong, Y. Qiao, J. Zhou, and C. Dong, "Hat: Hybrid attention transformer for image restoration," *arXiv preprint arXiv:2309.05239*, 2023.

[32] Z. Chen, Y. Zhang, J. Gu, L. Kong, and X. Yang, "Recursive generalization transformer for image super-resolution," in *Proceedings of the International Conference on Learning Representations*, 2024.

[33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[34] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 2256–2265.

[35] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.

[36] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proceedings of the International Conference on Learning Representations*, 2021.

[37] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, and L. Zhang, "Seesr: Towards semantics-aware real-world image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[38] Y. Wang, W. Yang, X. Chen, Y. Wang, L. Guo, L.-P. Chau, Z. Liu, Y. Qiao, A. C. Kot, and B. Wen, "Sinsr: Diffusion-based image super-resolution in a single step," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[39] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[40] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 095–13 105.

[41] Z. Yue, J. Wang, and C. C. Loy, "Resshift: Efficient diffusion model for image super-resolution by residual shifting," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 36, 2024.

[42] H. Nagahara, Y. Yagi, and M. Yachida, "Super-resolution from an omnidirectional image sequence," in *Proceedings of the Annual Conference of Industrial Electronics Society*, vol. 4, 2000, pp. 2559–2564.

[43] Z. Arican and P. Frossard, "Joint registration and super-resolution with omnidirectional images," *IEEE Transactions on Image Processing*, vol. 20, no. 11, pp. 3151–3162, 2011.

[44] L. Bagnato, Y. Boursier, P. Frossard, and P. Vandergheynst, "Plenoptic based super-resolution for omnidirectional image sequences," in *Proceedings of the IEEE International Conference on Image Processing*, 2010, pp. 2829–2832.

[45] V. Fakour-Sevom, E. Guldogan, and J.-K. Kämäräinen, "360 panorama super-resolution using deep convolutional networks," in *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 1, 2018, p. 1.

[46] C. Ozcinar, A. Rana, and A. Smolic, "Super-resolution of omnidirectional images using adversarial learning," in *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, 2019, pp. 1–6.

[47] Y. Zhang, H. Zhang, D. Li, L. Liu, H. Yi, W. Wang, H. Suitoh, and M. Odamaki, "Toward real-world panoramic image enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 628–629.

[48] A. Nishiyama, S. Ikehata, and K. Aizawa, "360 single image super resolution via distortion-aware network and distorted perspective images," in *IEEE International Conference on Image Processing*, 2021, pp. 1829–1833.

[49] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8628–8638.

[50] X. Chai, F. Shao, Q. Jiang, and H. Ying, "Tccl-net: Transformer-convolution collaborative learning network for omnidirectional image super-resolution," *Knowledge-Based Systems*, vol. 274, no. 110625, pp. 1–15, 2023.

[51] X. Chai, F. Shao, H. Chen, B. Mu, and Y.-S. Ho, "Super-resolution reconstruction for stereoscopic omnidirectional display systems via dynamic convolutions and cross-view transformer," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, no. 5025012, pp. 1–12, 2023.

[52] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 10 347–10 357.

[53] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 213–229.

[54] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 568–578.

[55] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

[56] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 683–17 693.

[57] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4794–4803.

[58] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1408–1412, 2017.

[59] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[60] W. Wang, W. Chen, Q. Qiu, L. Chen, B. Wu, B. Lin, X. He, and W. Liu, "Crossformer++: A versatile vision transformer hinging on cross-scale attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3123–3136, 2023.

[61] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 510–519.

[62] A. A. Baniya, T.-K. Lee, P. W. Eklund, and S. Aryal, "Omnidirectional video super-resolution using deep learning," *IEEE Transactions on Multimedia*, 2023.

[63] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba, "Recognizing scene viewpoint using panoramic place representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2695–2702.

[64] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 286–301.

[65] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[66] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong, "Activating more pixels in image super-resolution transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 367–22 377.

[67] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 32, 2019.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, 2015.

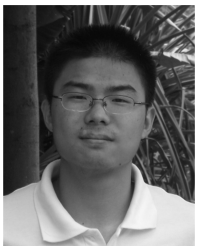
Cuixin Yang received the B.Sc. degree and M.Sc. degree from the College of Electronic and Information Engineering, Shenzhen University, in 2019 and 2022, respectively. She is currently pursuing the Ph.D. degree from the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University. Her research interests include image processing, restoration, enhancement and deep generative models.

Rongkang Dong received the B.Sc. degree from the South China University of Technology (SCUT), Guangzhou, China, in 2020, the M.Sc. degree from the Hong Kong Polytechnic University (PolyU), Hong Kong, in 2022. He is currently pursuing the Ph.D. degree from the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong. His research interests include facial expression recognition, computer vision, and deep learning.

Jun Xiao received his B.Sc. degree in Telecommunication Engineering from the Guangdong University of Technology, Guangzhou, China, in 2016. He completed his M.Sc. degree with distinction and Ph.D. degree from the Department of Electronic and Information Engineering, the Hong Kong Polytechnic University, Hong Kong, in 2018 and 2022, respectively. He is currently a postdoctoral fellow in the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University. His research focuses on image and video processing, restoration, enhancement, multimedia computing, and deep generative models.

Cong Zhang received the B.E. degree from the School of Electronics and Information, Northwestern Polytechnical University, in 2018, and the M.S. degree from the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, in 2021. He is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University. His current research interests include computer vision, machine learning, and remote sensing.

Kin-Man Lam received his Associateship in Electronic Engineering with distinction from The Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic) in 1986, his M.Sc. degree in communication engineering from the Department of Electrical Engineering, Imperial College, U.K., in 1987, and his Ph.D. degree from the Department of Electrical Engineering, University of Sydney, Australia, in 1996. From 1990 to 1993, he was a lecturer at the Department of Electronic Engineering of The Hong Kong Polytechnic University. He joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University again as an Assistant Professor in October 1996. He became an Associate Professor in 1999, and has been a Professor since 2010. He was actively involved in professional activities. He has been a member of the organizing committee or program committee of many international conferences. Prof. Lam was the Chairman of the IEEE Hong Kong Chapter of Signal Processing between 2006 and 2008, and was the Director-Student Services and the Director-Membership Services of the IEEE SPS between 2012 and 2014, and between 2015 and 2017, respectively. He was also the VP-Member Relations and Development and VP-Publications of the Asia-Pacific Signal and Information Processing Association (APSIPA) between 2014 and 2017, and between 2017 and 2021, respectively. He was an Associate Editor of IEEE Trans. on Image Processing between 2009 and 2014, and Digital Signal Processing between 2014 and 2018. He was also an Editor of HKIE Transactions between 2013 and 2018, and an Area Editor of the IEEE Signal Processing Magazine between 2015 and 2017. Currently, he is the the IEEE SPS VP-Membership. Prof. Lam also serves as a Senior Editorial Board member of APSIPA Trans. on Signal and Information Processing, and an Associate editor of EURASIP International Journal on Image and Video Processing. His current research interests include image and video processing, computer vision, and human face analysis and recognition.



Fei Zhou received the B.Eng. degree in electronics and information engineering from the Huazhong University of Science and Technology, in 2007, and the Ph.D. degree in electronic engineering, Tsinghua University in 2013, where he was a Postdoctoral Fellow with the Graduate School at Shenzhen from 2013 to 2016. From 2017 to 2018, he was a Visiting Scholar with the Department of Statistical Science, University College London. He is currently an Associate Professor with the College of Electronic and Information Engineering, Shenzhen University. He has authored over 60 papers internationally. His research interests include image super-resolution, image decomposition, image quality assessment, and inverse tone mapping. He is a Reviewer of many well-known journals, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and Information Sciences.

Guoping Qiu is a Professor of Visual Information Processing at the University of Nottingham, and the Chief Scientist of Everimaging Ltd, an AI-powered visual media creation platform. He researches neural networks and their applications in image processing and has developed some of the earliest image processing applications of neural networks including image coding/compression, learning image resolution enhancement/super-resolution, compression artifact removal and high dynamic range (HDR) image tone mapping. He also developed one of the earliest forms of representation learning (aka deep learning) where he designed self-organized competitive learning algorithms to learn image representation features/patterns. He has taught in universities in the UK and Hong Kong, and consulted for multinational companies in Europe, Hong Kong, and China. He is particularly known for his pioneering research in high dynamic range imaging and machine learning-based image processing technologies. He has published widely and holds several European and U.S. patents.