information

*Article*

# An Omnidirectional Image Super-Resolution Method Based on Enhanced SwinIR

**Xiang Yao [1,2], Yun Pan [1,2,*] and Jingtao Wang [1,2]**

[1] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China; yaoxiang@cuc.edu.cn (X.Y.); jingtao0621@cuc.edu.cn (J.W.)
[2] School of Computer and Cyberspace Security, Communication University of China, Beijing 100024, China
* Correspondence: pany@cuc.edu.cn

**Abstract:** For the significant distortion problem caused by the special projection method of equi-rectangular projection (ERP) images, this paper proposes an omnidirectional image super-resolution algorithm model based on position information transformation, taking SwinIR as the base. By introducing a space position transformation module that supports deformable convolution, the image preprocessing process is optimized to reduce the distortion effects in the polar regions of the ERP image. Meanwhile, by introducing deformable convolution in the deep feature extraction process, the model's adaptability to local deformations of images is enhanced. Experimental results on publicly available datasets have shown that our method outperforms SwinIR, with an average improvement of over 0.2 dB in WS-PSNR and over 0.030 in WS-SSIM for $\times 4$ pixel upscaling.

**Keywords:** omnidirectional image; image super-resolution; equi-rectangular projection; SwinIR; shift windows; deep learning



**Citation:** Yao, X.; Pan, Y.; Wang, J. An Omnidirectional Image Super-Resolution Method Based on Enhanced SwinIR. *Information* **2024**, *15*, 248. https://doi.org/10.3390/info15050248

Academic Editor: Vincenzo Moscato

Received: 8 April 2024
Revised: 25 April 2024
Accepted: 26 April 2024
Published: 28 April 2024



## 1. Introduction

With the breakthrough growth of augmented reality (AR) and virtual reality (VR) applications, omnidirectional image processing has gradually received attention from scientific research, and omnidirectional image super-resolution technology has also seen development. Compared to single-frame images, omnidirectional images have larger dimensions. In practical applications, to conserve memory and bandwidth, lower-resolution equi-rectangular projection (ERP) images are often stored and transmitted. However, the resolution of omnidirectional images is closely related to users' immersion and visual experience. Therefore, it is necessary to find a method to reconstruct high-definition images from low-resolution omnidirectional images.

Initially, people used regularization methods, algorithms based on neighborhood embedding, and a strategy to reconstruct high-resolution images by exploiting redundant similar blocks within low-resolution images [1–3]. However, such traditional algorithms often perform poorly when dealing with large-scale upscaled images, making it difficult to achieve the desired task goals. Recently, methods based on deep learning have developed rapidly and made significant contributions to the single-image super-resolution (SISR) field. After convolutional neural network (CNN)-based super-resolution methods [4–10] were first applied, algorithms based on generative adversarial networks (GAN) [11–18], visual transformers (ViTs) [19–23], and diffusion models [24,25] further advanced this technology. It is worth noting that SwinIR [26], as a Swin Transformer-based image super-resolution method, utilizes the hierarchical structure and local attention mechanism of the Swin Transformer to effectively handle the long-range dependency problem of images, demonstrating excellent performance in single-image super-resolution tasks. However, due to uneven pixel density and texture complexity across dimensions, these methods cannot be directly applied to ERP images.

Research on image super-resolution initially focused on stitching and processing multiple low-resolution omnidirectional images, but such methods were inefficient and produced subpar results. Subsequently, the LAU-NET [27] method segmented ERP images into multiple strips based on different latitudes and treated distortion within each latitude range separately, yet this approach overlooked the connection between adjacent patches. SphereSR [28] addressed the distortion issues caused by complex projections from a projection perspective, but the computational cost involved was too high, making it impractical for real-world applications. Recently, OSRT [29] proposed that downsampling from fisheye images better reveals the geometric properties of ERP images but lacks explicit positional information.

To address the above issues, this paper builds upon SwinIR [26] and proposes an omnidirectional image super-resolution method based on positional information transformation. By introducing a Positional Information Transformation module supporting deformable convolution, the preprocessing process of images is optimized, reducing the distortion effect of ERP images in polar regions. Meanwhile, by introducing deformable convolution during deep feature extraction, the model's adaptability to local image deformations is enhanced.

## 2. Related Work

### *2.1. Single-Image Super-Resolution (SISR)*

Since SRCNN [4] introduced deep learning into single-image super-resolution tasks, various CNN architectures have been widely studied to further improve the performance of image super-resolution algorithms. For example, DRCN [5] adopts an innovative approach by implementing weight sharing between convolutional layers within recursive blocks to reduce the complexity of the model. DenseNet [6] proposes the first dense structure, where the output of each convolutional layer is connected to the outputs of all subsequent convolutional layers, ensuring direct connections between every layer and all subsequent layers. EDSR [7] initially uses residual blocks without batch normalization as the basic building blocks, forming deeper super-resolution networks.

SwinIR [26] is an image super-resolution model based on the Swin Transformer architecture, which incorporates a shift-window mechanism in its design. This mechanism helps simulate long-range dependencies, enhancing the model's understanding of distant image information. Compared to traditional methods, SwinIR achieves performance enhancement with fewer parameters, implying that it can handle larger-scale image data and produce clearer, higher fidelity results with similar computational resources. Additionally, by leveraging the advantages of the Swin Transformer architecture, SwinIR can simultaneously capture both local and global information, leading to superior performance in super-resolution tasks.

### *2.2. Omnidirectional Image Super-Resolution (ODISR)*

In order to qualitatively assess the generation quality of omnidirectional image super-resolution, researchers proposed a method called weighted spherical-to-spherical peak signal-to-noise ratio (WS-PSNR) [30], which aims to evenly distribute pixel points on a sphere and assign different weights to pixels in different regions to comprehensively evaluate and reflect the generation quality of omnidirectional image super-resolution. Subsequently, attempts were made to apply GAN to omnidirectional image super-resolution. Due to the maturity of single-frame image super-resolution methods, researchers tried to adapt single-frame image super-resolution methods to ERP images and adjusted existing models using loss functions such as L1 to adapt to the special characteristics of omnidirectional images. Considering the importance of different regions in the image, researchers introduced the weighted structural similarity index (WS-SSIM) [31] to evaluate model performance.

LAU-NET [27] addressed the issue of uneven pixel density caused by distortion at different latitudes in ERP images by segmenting ERP images into multiple strips based on

different latitudes and handling distortion within each latitude range separately through separate learning. Subsequent optimizations of this model considered the issue of area stretching ratio, using it as an additional condition input to help the network better understand the distortion of the image and achieve more precise correction, yielding good results. SphereSR [28] addressed the problem of distortion in planar projection images caused by complex projections. This method introduced a feature extraction module capable of extracting features from different types of projections (such as central projection, equidistant cylindrical projection, etc.) on the sphere. These extracted spherical features were used as parameters that were input into the spherical local implicit image function (SLIIF) to predict the RGB values corresponding to the spherical coordinates, thus achieving high-resolution reconstruction results for arbitrary projection types. Although this method achieved good results, the computational cost involved was too high, making it impractical for real-world applications. Currently, the mainstream approach is still to process ERP images because ERP projection calculation is simple, easy to implement, and has higher efficiency and reliability in practical applications. The OSRT [29] method approached the problem from the perspective of image sampling, suggesting that downsampling from fisheye images better captures the geometric properties of ERP images and is more consistent with the real imaging process. OSRT also designed an image distortion-aware transformer, which can adjust the distortion of images in real-time based on the distortion of ERP images, without the need for manual intervention or complex preprocessing steps.

## 3. Architectural Details

In this section, we first describe the network structure of the proposed omnidirectional image super-resolution model based on SwinIR [26] and then provide detailed introductions to each module.

### *3.1. The Entire Network Architecture*

The entire network architecture is based on improvements made to SwinIR, as illustrated in Figure 1, and mainly consists of the following key components: firstly, the location transformation module preprocesses ERP images to reduce distortion in polar regions, providing optimized input for super-resolution reconstruction. Next is the enhanced SwinIR framework, which effectively handles long-distance dependencies by leveraging the hierarchical structure and local attention mechanism of the Swin Transformer to enhance super-resolution reconstruction performance. Deformable convolutional layers are introduced into the model, replacing standard convolutional layers to better adapt to local deformations in the images, particularly achieving better results in handling distortion in ERP images. The entire model's operation proceeds as follows: firstly, the input low-resolution ERP images undergo preprocessing and distortion correction through the location transformation module, then through the improved SwinIR framework, shallow feature extraction, deep feature extraction, and, finally, image reconstruction for super-resolution is sequentially performed, ultimately outputting high-quality high-resolution ERP images.

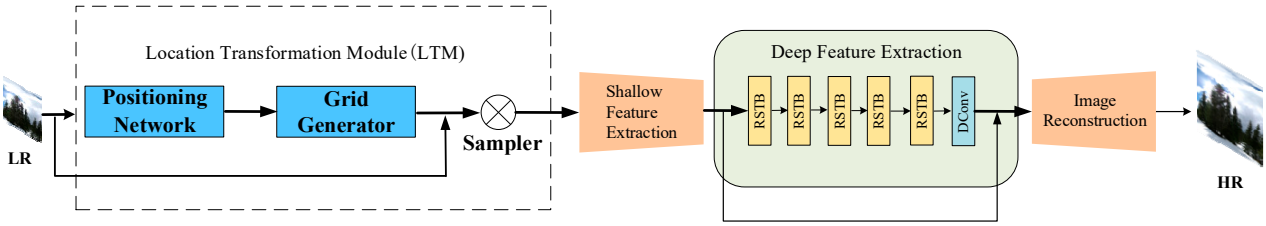


**Figure 1.** The Network Architecture of Our Proposed Model (This model aims to convert low resolution (LR) images into high resolution (HR) images).

### *3.2. The Location Transformation Module*

ERP images exhibit significant distortion in the polar regions due to their unique projection method, posing a considerable challenge for super-resolution reconstruction algorithms. To overcome this challenge, this paper introduces a location transformation module (LTM) with support for deformable convolutions. The LTM aims to optimize the preprocessing of images, reduce distortion in the polar regions of ERP images, and enhance the network's ability to perform spatial transformations [32]. The LTM enables the network to learn how to adaptively adjust images for different spatial transformation tasks, thereby improving subsequent processing.

As shown in Figure 2, the LTM applies spatial transformations to feature maps during a single forward pass. This transformation depends on the specific input, generating a single output feature map. It allows the neural network to dynamically modify the spatial layout of its input feature maps, enabling the network to automatically perform spatial transformations on images without any additional supervision. This enhances the model's adaptability and robustness to geometric deformations in the input data. When dealing with ERP images, the spatial transformation module can be particularly effective as ERP images often exhibit significant distortion at the poles, which can result in poor image quality during super-resolution tasks. By introducing the LTM, the spatial content of images can be dynamically adjusted during the network's forward pass, thereby preprocessing ERP images and improving the performance of subsequent tasks. The LTM can automatically learn spatial transformations to correct distortions or perspective biases in images, providing SwinIR with more standardized inputs. Detailed descriptions of each module will be provided below.

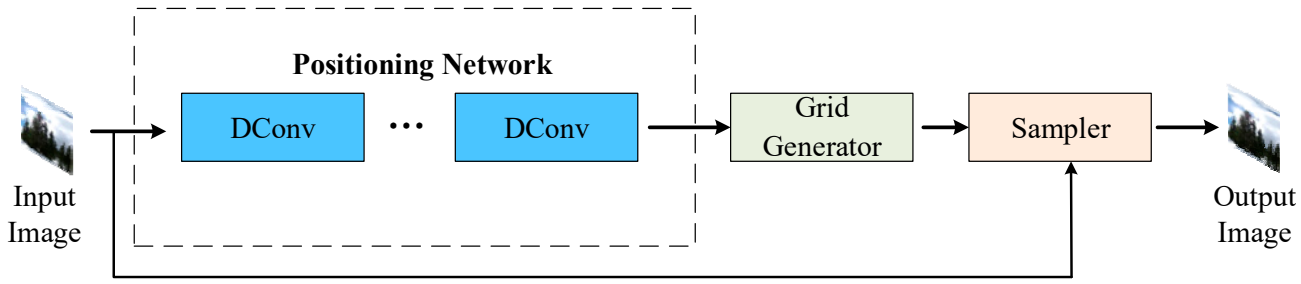


**Figure 2.** LTM Module.

#### 3.2.1. Positioning Network

The positioning network is the first part of the location information module, responsible for learning the spatial transformation parameters of the input feature map. The core function of this network is to automatically determine the optimal adjustment of the spatial layout of the input data, facilitating the subsequent super-resolution tasks.

The positioning network receives an input feature map $U \in \mathbb{R}^{H \times W \times C}$, where $W$ is the width, $H$ is the height, and $C$ is the number of channels. Its task is to output a set of transformation parameters, $\theta$, which define the spatial transformation, $T_\theta$, applied to the feature map. The size of the transformation parameters, $\theta$, depends on the type of transformation being parameterized. For example, for affine transformations, $\theta$ is six-dimensional, as affine transformations can be fully defined by six parameters (translation, scaling, rotation, and skew), as illustrated in Figure 3.

The function, $f_{\text{loc}}(U)$, of the positioning network can take any form, such as a fully connected network or a convolutional network, but the key is that it should include a final regression layer to generate the transformation parameters, $\theta$. This means that regardless of the internal structure of the network, its output is a set of numerical values that can be directly used to define spatial transformations. The positioning network enables the LTM to adaptively perform spatial preprocessing based on the specific content of the input feature

map. This means that the network can automatically perform operations such as cropping, rotating, and scaling, which are necessary for optimizing subsequent task performance.

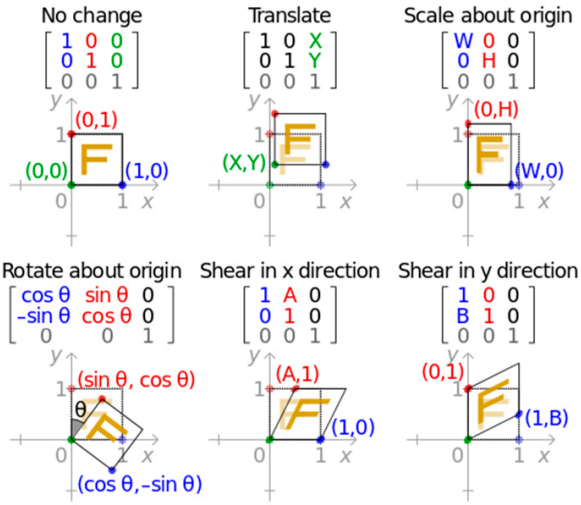


**Figure 3.** Affine Transformation Diagram.

### 3.2.2. Grid Generator

The grid generator is the second key component of the LTM, following the positioning network. Its main responsibility is to create a sampling grid based on the transformation parameters, $\theta$, output by the positioning network. This grid determines how to sample from the input feature map to generate the transformed output feature map. This process is a direct means of implementing spatial positional transformations, allowing the model to dynamically adjust the spatial layout of its input.

The grid generator computes the corresponding input pixel positions for each output pixel position based on the transformation parameters, $\theta$. This means that for each pixel position in the output feature map, the grid generator specifies a sampling point location from the input feature map. This process can be described by the following Formula (1), taking affine transformation as an example:

$$T_\theta(x,y) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \theta_{11}x + \theta_{12}y + \theta_{13} \\ \theta_{21}x + \theta_{22}y + \theta_{23} \end{bmatrix} \quad (1)$$

where $(x, y)$ represents the pixel position in the output feature map and $\theta$ represents the affine transformation parameters learned by the positioning network. In this way, each output position is mapped back to a specific position on the input feature map. The coordinates of each point obtained by the transformation formula constitute the sampling grid. This grid directly guides the sampling positions in the subsequent sampling steps, determining from which positions in the input feature map to extract pixel values to construct the transformed output feature map.

### 3.2.3. Sampler

The sampler is the third and final major component of the LTM. Its function is to sample pixel values from the input feature map according to the sampling grid provided by the grid generator in order to generate the transformed output feature map. This step is the actual execution stage of spatial transformation, ensuring that the model can dynamically adjust the spatial layout of its input data as needed.

The sampler first receives the transformation parameters, $\theta$, output by the positioning network, which define how the spatial transformation from the input feature map, $U$, to the output feature map, $U$, is performed. Based on these parameters, the sampler computes

a set of sampling points, $T_\theta(G)$, which represents the continuous spatial positions that need to be sampled in the input feature map, $U$. Each sampling point, $(x_s, y_s)$, defines a continuous spatial position in the input feature map, $U$, and the sampler applies a sampling kernel, $k(\cdot, \Phi_x, \Phi_y)$, at these positions to compute the values of the corresponding pixels in the output feature map, $V$. The sampling kernel determines how discrete output values are interpolated from the continuous spatial positions of the input feature map. The sampling process can be mathematically represented by the following Formula (2):

$$V_c^i = \sum_{n=1}^{H} \sum_{m=1}^{W} U_c^{nm} k\left(x_s^i - m; \Phi_x\right) k\left(y_s^i - n; \Phi_y\right) \tag{2}$$

where $U_c^{nm}$ is the value at channel $c$ and position $(n, m)$ in the input feature map. $V_c^i$ is the sampled value at channel $c$ and position $\left(x_t^i, y_t^i\right)$ in the output feature map. $k(\cdot; \Phi_x)$ and $k\left(\cdot; \Phi_y\right)$ are the sampling kernel functions in the $x$ and $y$ directions, respectively.

### *3.3. SwinIR with Deformable Convolution*

SwinIR [26] is an image restoration model based on the Swin Transformer, which achieves high-quality image reconstruction through three key modules: shallow feature extraction, deep feature extraction, and high-quality (HQ) image reconstruction. These modules work together to process low-quality (LQ) image inputs and produce high-quality output images. This part of this project used the code (online available at https://github.com/JingyunLiang/SwinIR (accessed on 26 August 2021)) from Liang et al. [26].

First, the model transforms the input low-quality image into shallow features using a 3 × 3 convolutional layer, preparing for subsequent deep feature extraction. Then, in the deep feature extraction stage, the model utilizes a module consisting of K residual Swin Transformer [33] blocks and a 3 × 3 convolutional layer to extract deep features from the shallow features, capturing richer image information. Finally, by aggregating shallow and deep features, the model reconstructs the high-quality image. Shallow features primarily contain low-frequency information, while deep features focus on recovering lost high-frequency information. Through long skip connections, the model effectively transmits low-frequency information and helps the deep feature extraction module focus on recovering high-frequency information, thereby improving the training stability and image reconstruction quality. The final image reconstruction module typically employs sub-pixel convolutional layers for upsampling and utilizes residual learning to recover the residual between low-quality and high-quality images, thus achieving high-quality image reconstruction.

To enhance the performance of super-resolution (SR) on ERP images, we replaced the standard convolutional layers in SwinIR with deformable convolutions [34]. This substitution allows for the dynamic adjustment of the shape of the convolutional kernel to adapt to specific geometric variations in the input data, thereby improving the model's adaptability to changes in image shape and the flexibility of its receptive field. As illustrated in Figure 4, deformable convolutions introduce additional learnable parameters (i.e., offsets) during the convolution process, dynamically adjusting the sampling points of the convolutional kernel to accommodate local deformations in the image content. These offsets, learned by the network, enable the convolutional kernel to adaptively align with key features in the image rather than sampling solely at fixed, regular grid points. The entire process is expressed as Formula (3), as follows:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{3}$$

A regular grid, $R$, is used to sample on the input feature map, $x$. Then, the sampled values are multiplied by the weights, $w$, and summed. $p_0$ represents the position on the output feature map, $y$, $p_n$ is the relative position within the grid $R$, and $\Delta p_n$ is the offset

corresponding to the position $p_n$, allowing the sampling position $p_0 + p_n$ to be appropriately adjusted according to the content of the input feature map.

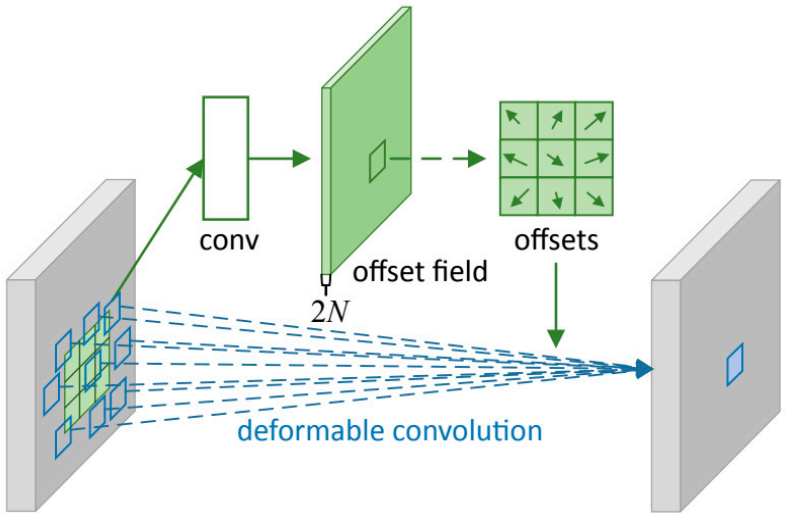


**Figure 4.** The Process of Deformable Convolution.

## 4. Experiments

### *4.1. Datasets*

ODI-SR [8]. The dataset is specifically designed for research on omnidirectional image super-resolution. "ODI" stands for omnidirectional image, representing a 360-degree field of view using ERP images. The ODI-SR dataset comprises a series of high-resolution omnidirectional images along with their corresponding low-resolution versions, covering diverse scenes such as indoor, urban landscapes, and natural environments. It aims to provide a comprehensive benchmark platform for evaluating and comparing the performance of various omnidirectional image super-resolution techniques.

SUN360 [35]. This is a large-scale omnidirectional image dataset, part of the SUN (Scene Understanding) project aimed at creating a comprehensive benchmark dataset for visual scene understanding. The SUN360 dataset contains tens of thousands of omnidirectional images covering various indoor and outdoor environments. Each omnidirectional image provides a 360-degree field of view represented in ERP format.

### *4.2. Implementation Detail*

The GPU used in this experiment is A800*4. During the training phase, we followed the data partitioning settings of the ODI-SR dataset, conducting model training on the ODI-SR training set. The resolution of ERP HR images is set to 1024 $\times$ 2048, with upscaling factors of $\times$2 and $\times$4. Predefined downsampling kernels are directly applied to the ERP images for downsampling. The loss function used is L1, and optimization is performed using the Adam optimizer with an initial learning rate of $2 \times 10^{-4}$. The total batch size is set to 32, and the input patch size is 64. The model is trained for 300 k iterations, with the learning rate halved at 150 k, 240 k, and 270 k iterations. During the evaluation phase, the model is tested on the ODI-SR test set and the SUN360 dataset. PSNR, SSIM, LPIPS as well as their WS-PSNR and WS-SSIM are used as evaluation metrics to assess performance.

### *4.3. Quantitative Results*

During the experimental process, we trained and compared the performance of mainstream super-resolution methods including Bicubic, SRCNN [4], EDSR [7], VDSR [8], ESRGAN [12], RCAN [13], BSRGAN [14], Real-ESRGAN [15] and SwinIR [26]. The experimental results are shown in Tables 1–4. The best results are represented in red colors.

**Table 1.** Experimental Results for ×2 Upscaling Comparison.

| **Scale** | **×2** | | | | | |
|---|---|---|---|---|---|---|
| **Method** | **ODI-SR** | | | **SUN 360 Panorama** | | |
| | **PSNR** | **SSIM** | **LPIPS** | **PSNR** | **SSIM** | **LPIPS** |
| **Bicubic** | 28.26 | 0.8216 | 0.353 | 28.54 | 0.8279 | 0.398 |
| SRCNN [4] | 29.03 | 0.8452 | 0.342 | 29.26 | 0.8426 | 0.364 |
| VDSR [8] | 30.12 | 0.8703 | 0.265 | 30.11 | 0.8733 | 0.302 |
| RCAN [13] | 30.15 | 0.8725 | 0.226 | 30.52 | 0.8745 | 0.264 |
| EDSR [7] | 30.32 | 0.8711 | 0.269 | 30.65 | 0.8720 | 0.279 |
| ESRGAN [12] | 30.36 | 0.8769 | 0.205 | 30.85 | 0.8812 | 0.198 |
| BSRGAN [14] | 30.32 | 0.8795 | 0.187 | 30.98 | 0.8839 | 0.175 |
| Real-ESRGAN [15] | 30.59 | 0.8819 | 0.155 | 31.19 | 0.8825 | 0.196 |
| SwinIR [26] | 30.54 | 0.8825 | 0.115 | 31.25 | 0.8846 | 0.132 |
| LTM-SwinIR (ours) | 30.67 | 0.8836 | 0.102 | 31.39 | 0.8857 | 0.118 |

**Table 2.** Experimental Results for ×2 Upscaling Comparison.

| **Scale** | **×2** | | | |
|---|---|---|---|---|
| **Method** | **ODI-SR** | | **SUN 360 Panorama** | |
| | **WS-PSNR** | **WS-SSIM** | **WS-PSNR** | **WS-SSIM** |
| **Bicubic** | 27.32 | 0.8059 | 28.50 | 0.8356 |
| SRCNN [4] | 28.20 | 0.8312 | 28.96 | 0.8402 |
| VDSR [8] | 29.56 | 0.8716 | 29.65 | 0.8736 |
| RCAN [13] | 29.63 | 0.8669 | 29.41 | 0.8749 |
| EDSR [7] | 29.65 | 0.8772 | 29.66 | 0.8768 |
| ESRGAN [12] | 29.86 | 0.8769 | 29.79 | 0.8775 |
| BSRGAN [14] | 30.25 | 0.8698 | 30.25 | 0.8799 |
| Real-ESRGAN [15] | 30.36 | 0.8716 | 30.19 | 0.8859 |
| SwinIR [26] | 30.32 | 0.8720 | 30.39 | 0.8886 |
| LTM-SwinIR (ours) | 30.54 | 0.8799 | 30.62 | 0.8870 |

From the experimental results, it can be observed that the LTM demonstrates strong potential performance compared to many mainstream single-image super-resolution methods. The presence of the sliding window mechanism enables SwinIR and our proposed LTM-SwinIR to excel in capturing long-range dependencies (i.e., global information). Consequently, they generally outperform traditional SR methods relying on convolutional dependencies, such as RCAN and ESRGAN, in terms of recovering image details. Although RCAN and ESRGAN introduce attention mechanisms and residual learning to enhance feature extraction capabilities, the experimental results indicate that they are less efficient in multi-scale feature fusion and utilization compared to LTM-SwinIR. In comparison to SwinIR, LTM-SwinIR demonstrates a more noticeable advantage in distortion perception. The application of the LTM module and the deformable convolution blocks offers significant superiority in handling ERP image processing.

**Table 3.** Experimental Results for ×4 Upscaling Comparison.

| **Scale** | **×4** | | | | | |
|---|---|---|---|---|---|---|
| **Method** | **ODI-SR** | | | **SUN 360 Panorama** | | |
| | **PSNR** | **SSIM** | **LPIPS** | **PSNR** | **SSIM** | **LPIPS** |
| **Bicubic** | 25.39 | 0.7089 | 0.574 | 25.29 | 0.7069 | 0.608 |
| SRCNN [4] | 25.69 | 0.7319 | 0.428 | 26.16 | 0.7365 | 0.526 |
| VDSR [8] | 26.75 | 0.7622 | 0.399 | 27.13 | 0.7639 | 0.422 |
| RCAN [13] | 26.89 | 0.7599 | 0.352 | 27.22 | 0.7659 | 0.395 |
| EDSR [7] | 27.08 | 0.7624 | 0.403 | 27.35 | 0.7709 | 0.355 |
| ESRGAN [12] | 26.99 | 0.7689 | 0.326 | 27.39 | 0.7738 | 0.386 |
| BSRGAN [14] | 27.26 | 0.7695 | 0.295 | 27.29 | 0.7729 | 0.308 |
| Real-ESRGAN [15] | 27.32 | 0.7702 | 0.302 | 27.50 | 0.7755 | 0.226 |
| SwinIR [26] | 27.36 | 0.7708 | 0.282 | 27.56 | 0.7795 | 0.256 |
| LTM-SwinIR (ours) | 27.41 | 0.7726 | 0.203 | 27.99 | 0.7820 | 0.199 |

**Table 4.** Experimental Results for ×4 Upscaling Comparison.

| **Scale** | **×4** | | | |
|---|---|---|---|---|
| **Method** | **ODI-SR** | | **SUN 360 Panorama** | |
| | **WS-PSNR** | **WS-SSIM** | **WS-PSNR** | **WS-SSIM** |
| **Bicubic** | 24.96 | 0.6985 | 25.38 | 0.7059 |
| SRCNN [4] | 25.13 | 0.7256 | 26.02 | 0.7423 |
| VDSR [8] | 26.16 | 0.7459 | 26.98 | 0.7812 |
| RCAN [13] | 26.23 | 0.7449 | 27.12 | 0.7859 |
| EDSR [7] | 26.44 | 0.7478 | 27.30 | 0.7860 |
| ESRGAN [12] | 26.39 | 0.7502 | 27.35 | 0.7895 |
| BSRGAN [14] | 26.41 | 0.7519 | 27.46 | 0.7899 |
| Real-ESRGAN [15] | 26.49 | 0.7522 | 27.52 | 0.7906 |
| SwinIR [26] | 26.61 | 0.7546 | 27.60 | 0.7915 |
| LTM-SwinIR (ours) | 26.69 | 0.7553 | 27.82 | 0.7966 |

### *4.4. Ablation Study*

To demonstrate the effectiveness of the LTM module and the deformable convolution layers in feature extraction of LTM-SwinIR, ablation experiments were conducted in this section. The experimental parameters were consistent with those described earlier, and testing was performed at a scale factor of ×4 on the dataset. The results are presented in Table 5. The best results are represented in red colors.

### *4.5. Qualitative Results*

In this section, a comparative analysis was conducted between several mainstream methods and LTM-SwinIR. The proposed method in this chapter demonstrates relatively superior results in both quantitative comparative experiments and qualitative presentations. This is particularly notable for addressing images like ERP that exhibit distortion at specific locations. The experimental results are shown in Figures 5–13. The best results are represented in red colors.

**Table 5.** Experimental Results for ×4 Upscaling Ablation Study.

| Scale | Component | | ×4 | | | |
|---|---|---|---|---|---|---|
| Model | D-Conv | LTM | ODI-SR | | SUN 360 Panorama | |
| | | | WS-PSNR | WS-SSIM | WS-PSNR | WS-SSIM |
| **1** | x | x | 26.49 | 0.7509 | 27.09 | 0.7895 |
| **2** | x | √ | 26.68 | 0.7544 | 27.32 | 0.7933 |
| **3** | √ | √ | 26.76 | 0.7558 | 27.40 | 0.7956 |

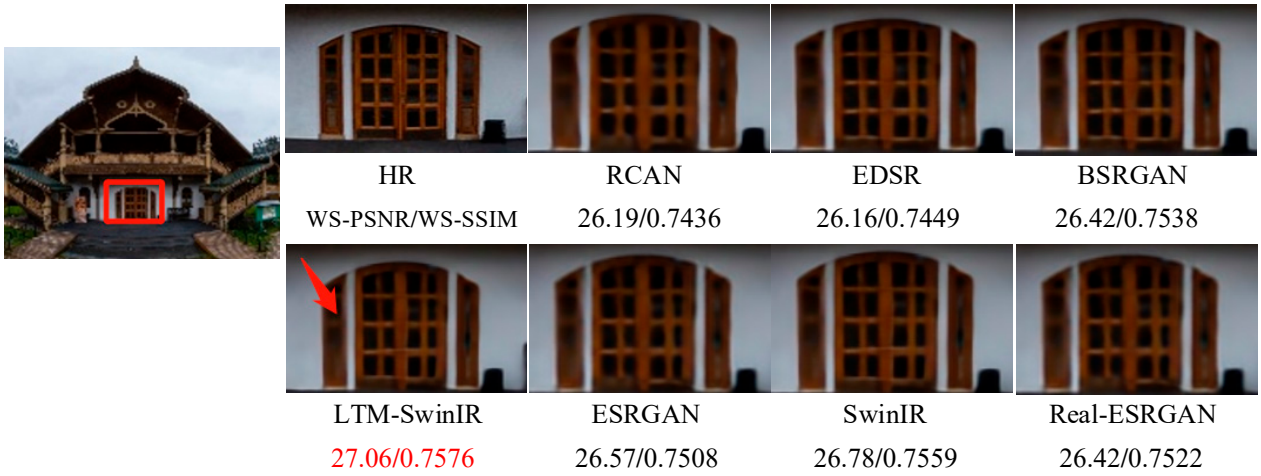


**Figure 5.** Qualitative Comparison of ×4 Pixel Upsampling Results.


**Figure 6.** Qualitative Comparison of ×4 Pixel Upsampling Results.

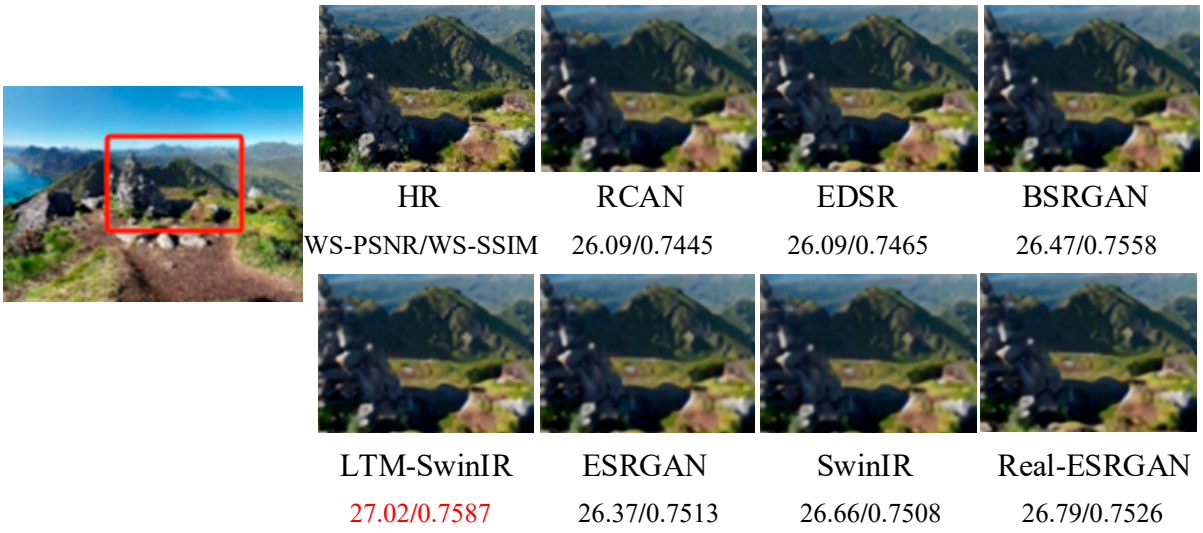


**Figure 7.** Qualitative Comparison of ×4 Pixel Upsampling Results.


**Figure 8.** Qualitative Comparison of ×4 Pixel Upsampling Results.

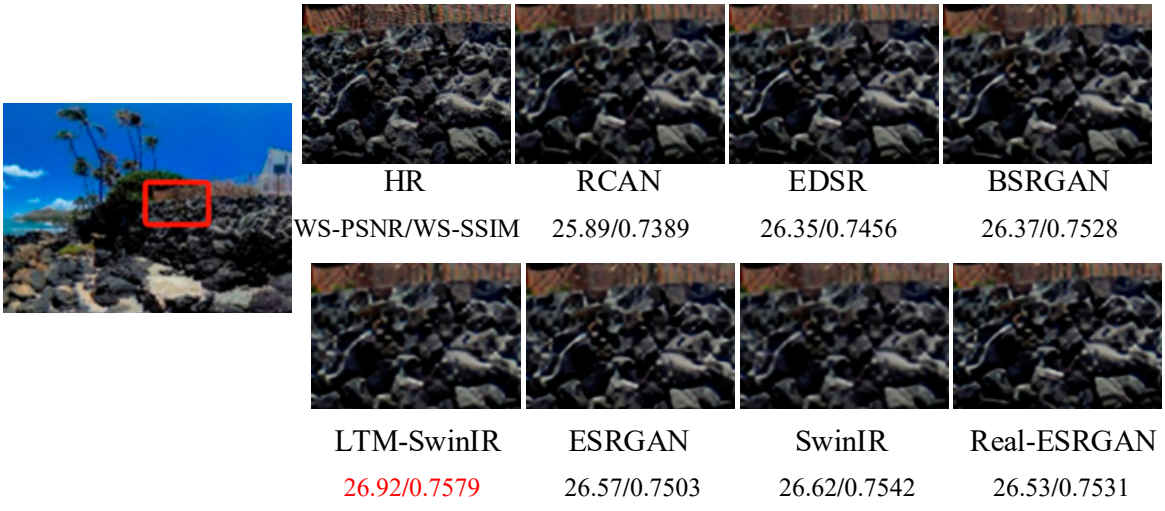


**Figure 9.** Qualitative Comparison of ×4 Pixel Upsampling Results.


**Figure 10.** Qualitative Comparison of ×4 Pixel Upsampling Results.

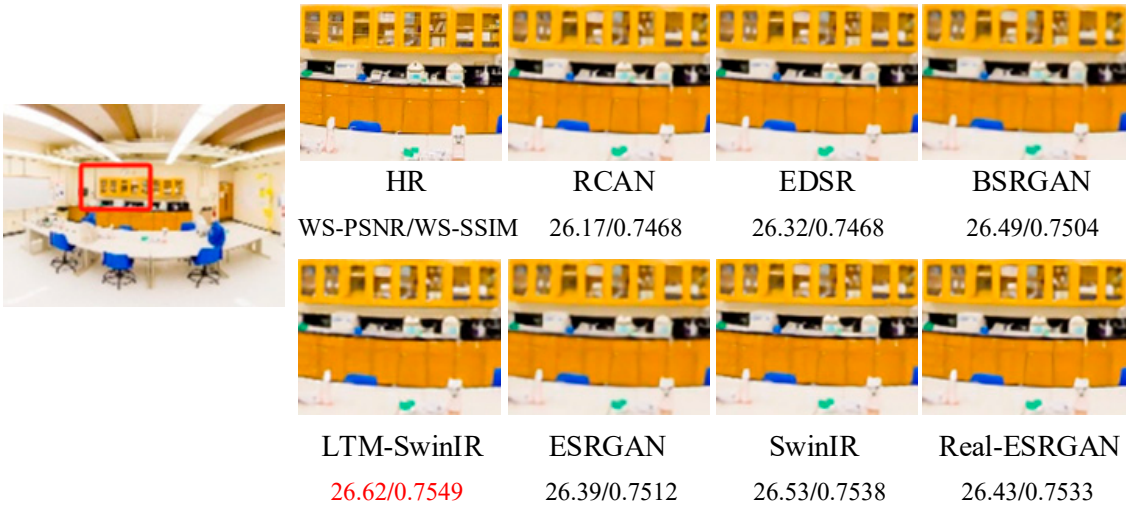


**Figure 11.** Qualitative Comparison of ×4 Pixel Upsampling Results.


**Figure 12.** Qualitative Comparison of ×4 Pixel Upsampling Results.

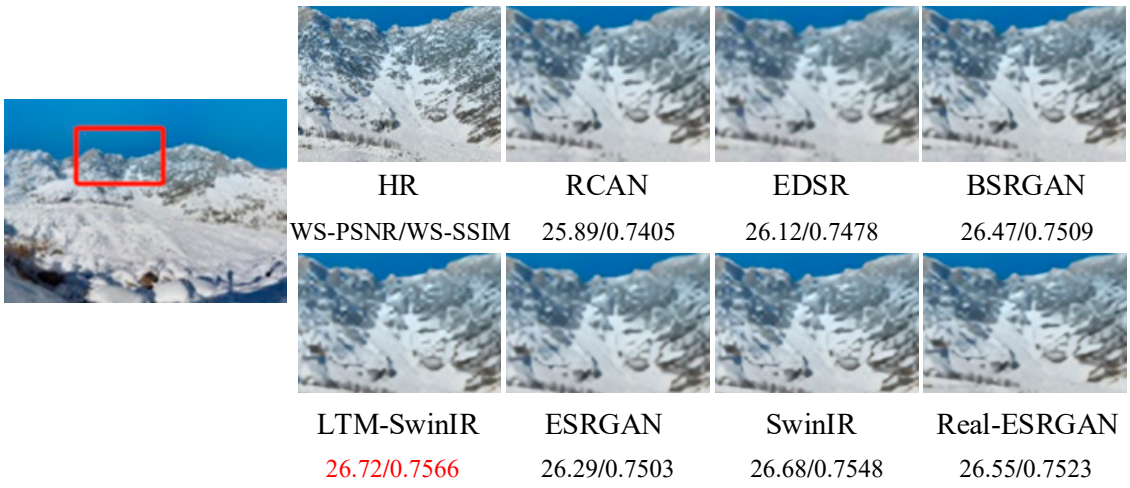


**Figure 13.** Qualitative Comparison of ×4 Pixel Upsampling Results.

### *4.6. Training and Validation Results*

The ODI-SR dataset was used for training at a ×4 pixel scale, with experimental settings as described in Section 4.2. It iterated for a total of 250 epochs. The changes in evaluation metrics WS-PSNR and loss function are shown in Figures 14 and 15. The experiments generally met the expectations, validating the rigor of the approach.

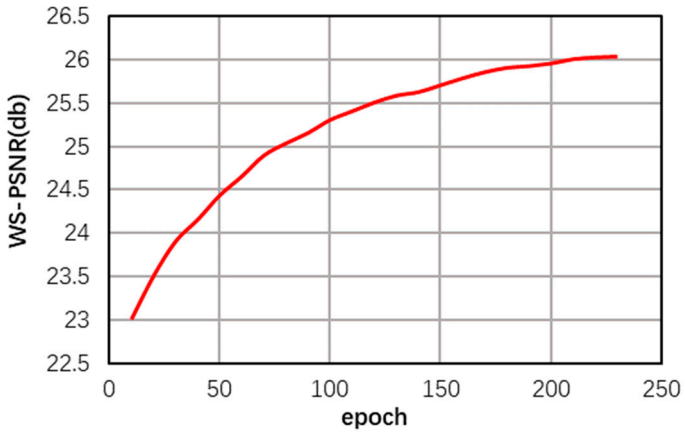


**Figure 14.** Changes in Results as Training Progresses.

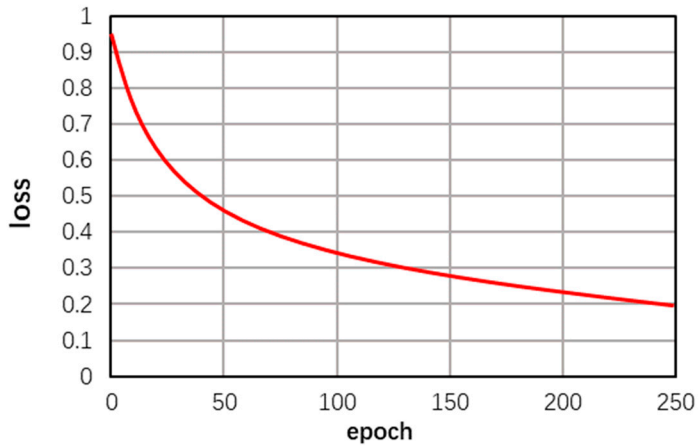


**Figure 15.** Changes in Loss Function.

## 5. Conclusions

This paper proposes an omnidirectional image super-resolution method based on enhanced SwinIR, leveraging the spatial transformation capability of the location transformation module (LTM) and deformable convolution layers in feature extraction to enhance SwinIR's ability to handle spatial deformations and different perspectives in images. This fusion significantly enhances the model's capability to handle complex spatial relationships and details in omnidirectional images, thereby achieving higher-quality super-resolution reconstruction.

The LTM supports deformable convolution, enabling spatial transformations of feature maps in a single forward pass. This transformation depends on specific inputs, generating a single output feature map. It allows neural networks to dynamically modify the spatial layout of their input feature maps, enabling automatic spatial transformations without requiring any additional supervision. This enhances the model's adaptability and robustness to geometric deformations in the input data. When dealing with ERP images, the LTM can be particularly effective in addressing distortions present at the poles. By introducing the LTM, the spatial content of images can be dynamically adjusted during the forward propagation of the network, thereby preprocessing ERP images to improve the performance of subsequent tasks. The replacement of standard convolutional layers with deformable convolutional layers in the feature extraction layer of SwinIR is primarily motivated by the need to enhance the adaptability and flexibility of convolutional neural networks to geometric deformations in images. This replacement strengthens the network's ability to handle non-rigid deformations in images, particularly for omnidirectional images and other applications requiring high geometric flexibility.

In the future, research should focus on developing more efficient model architectures and algorithms to optimize and lightweight the models, reducing the demand for computational resources, enhancing the processing speed, and achieving real-time or near real-time omnidirectional image super-resolution processing. Additionally, future efforts can explore the introduction of more diverse training data, advanced data augmentation techniques, and new training strategies to further improve the model's generalization and robustness across different scenarios and conditions. Currently, the application scope of omnidirectional image super-resolution methods is limited. In the future, integrating image super-resolution technology with the latest research findings from other disciplines can broaden its applicability.

**Author Contributions:** Conceptualization, X.Y.; methodology, X.Y.; investigation, Y.P. and J.W.; writing—original draft preparation, X.Y.; writing—review and editing, Y.P. and J.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Natural Science Foundation of China (61972050, 62201525, 62272040, 62172005), research on the strategic project of the Science and Technology Commission of the Ministry of Education of China (JYB2022-01).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data are available in a publicly accessible repository.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Bevilacqua, M.; Roumy, A.; Guillemot, C.; Alberi Morel, M.-L. Lowcomplexity single-image super-resolution based on nonnegative neighbor embedding. In Proceedings of the British Machine Vision Conference, London, UK, 3–7 September 2012; pp. 135.1–135.10. [CrossRef]
2. Zhang, K.; Gao, X.; Tao, D.; Li, X. Single image super-resolution with non-local means and steering kernel regression. *IEEE Trans. Image Process.* **2012**, *21*, 4544–4556. [CrossRef] [PubMed]
3. Gao, X.; Zhang, K.; Tao, D.; Li, X. Image super-resolution with sparse neighbor embedding. *IEEE Trans. Image Process.* **2012**, *21*, 3194–3205. [PubMed]

4. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef] [PubMed]
5. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-recursive convolutional network for image super-resolution. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1637–1645.
6. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
7. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced deep residual networks for single image super-resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 136–144.
8. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
9. Wang, H.; Chen, X.; Ni, B.; Liu, Y.; Liu, J. Omni aggregation networks for lightweight image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
10. Lin, J.; Luo, X.; Hong, M.; Qu, Y.; Xie, Y.; Wu, Z. Memory-Friendly Scalable Super-Resolution via Rewinding Lottery Ticket Hypothesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
11. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photorealistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
12. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, C. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018; pp. 1–16.
13. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 286–301.
14. Zhang, K.; Liang, J.; Van Gool, L.; Timofte, R. Designing a practical degradation model for deep blind image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021.
15. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021.
16. Mou, C.; Wu, Y.; Wang, X.; Dong, C.; Zhang, J.; Shan, Y. Metric learning based interactive modulation for real-world super-resolution. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022.
17. Park, J.; Son, S.; Lee, K.M. Content-aware local gan for photo-realistic super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023.
18. Lee, M.; Heo, J.-P. Noise-free optimization in early training steps for image super-resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38. No. 4.
19. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; Gao, W. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, DC, USA, 14–19 June 2020; pp. 12299–12310.
20. Zhou, Y.; Li, Z.; Guo, C.L.; Bai, S.; Cheng, M.M.; Hou, Q. Srformer: Permuted self-attention for single image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023.
21. Chen, Z.; Zhang, Y.; Gu, J.; Kong, L.; Yang, X.; Yu, F. Dual aggregation transformer for image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023.
22. Zhou, X.; Huang, H.; He, R.; Wang, Z.; Hu, J.; Tan, T. Msra-sr: Image super-resolution transformer with multi-scale shared representation acquisition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023.
23. Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; Dong, C. Activating more pixels in image super-resolution transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
24. Luo, X.; Xie, Y.; Qu, Y.; Fu, Y. SkipDiff: Adaptive Skip Diffusion Model for High-Fidelity Perceptual Image Super-resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38. No. 5.
25. Yuan, Y.; Yuan, C. Efficient Conditional Diffusion Model with Probability Flow Sampling for Image Super-resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38. No. 7.
26. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 1833–1844.
27. Deng, X.; Wang, H.; Xu, M.; Guo, Y.; Song, Y.; Yang, L. Lau-net: Latitude adaptive upscaling network for omnidirectional image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 9189–9198.
28. Yoon, Y.; Chung, I.; Wang, L.; Yoon, K.J. 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, New Orleans, LA, USA, 18–24 June 2022; pp. 5677–5686.

29. Yu, F.; Wang, X.; Cao, M.; Li, G.; Shan, Y.; Dong, C. Osrt: Omnidirectional image super resolution with distortion-aware transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.
30. Sun, Y.; Lu, A.; Yu, L. Weighted-to-spherically uniform quality evaluation for omnidirectional video. *IEEE Signal Process. Lett.* **2017**, *24*, 1408–1412. [CrossRef]
31. Zhou, Y.; Yu, M.; Ma, H.; Shao, H.; Jiang, G. Weighted-to-spherically-uniform ssim objective quality evaluation for panoramic video. In Proceedings of the 2018 14th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 12–16 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 54–57.
32. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial Transformer Networks. *arXiv* **2015**, arXiv:1506.02025.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 764–773.
35. Xiao, J.; Ehinger, K.A.; Oliva, A.; Torralba, A. Recognizing scene viewpoint using panoramic place representation. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 2695–2702.