Efficient Adaptation for Real-World Omnidirectional Image Super-Resolution

Cuixin Yang*, Rongkang Dong* and Kin-Man Lam*
* The Hong Kong Polytechnic University, Hong Kong
E-mail: cuixin.yang@connect.polyu.hk, rongkang97.dong@connect.polyu.hk, enkmlam@polyu.edu.hk

***Abstract*—With the increasing popularity of virtual techniques, such as virtual reality (VR) and augmented reality (AR), super-resolution (SR) of omnidirectional images has been crucial for more immersive and realistic experiences. This advancement also enhances the quality of images for various visual applications. Researchers have started exploring omnidirectional image super-resolution (ODISR). However, existing methods primarily address the problem using synthetic data pairs, where low-resolution (LR) images are generated using fixed, predefined kernels, such as bicubic downsampling. Consequently, the performance of these methods drops significantly when applied to real-world data. To address this issue, in this paper, we propose exploring the rich image priors from existing SR models designed for 2D planar images and adapting them for real-world ODISR. Specifically, we employ low-rank adaptation (LoRA) to adapt a large-scale model from the 2D planar image domain to the omnidirectional image domain by training only the decomposed matrices. This approach significantly reduces the number of parameters and computational resources required. Experimental results demonstrate that the proposed method outperforms other state-of-the-art methods both quantitatively and qualitatively.**

I. Introduction

With the development of virtual reality (VR) and augmented reality (AR), omnidirectional images (ODIs) have become increasingly prevalent in our daily lives, providing immersive virtual scenes. High-resolution (HR) ODIs are essential for these virtual scenarios, because they provide richer details and textures. However, it is costly to capture, store, and transmit HR ODIs [1]. Image super-resolution (SR) is a promising technique to address this challenge, aiming to reconstruct HR images from their low-resolution (LR) counterparts [2].

Recently, with the development of deep-learning techniques, significant progress has been made in many computer vision tasks [3], [4], including the field of omnidirectional image super-resolution (ODISR) [1], [5]–[10]. Despite the remarkable advancements in ODISR, existing methods still have several limitations. A major drawback is their focus on synthetic images [11], where LR images are generated from the HR images using fixed or predefined degradation methods, such as bicubic downsampling. However, real-world omnidirectional images suffer from complex and often unknown degradations, such as noise, blurriness, and downsampling [12]. Consequently, methods trained on synthetic images perform poorly in real-world scenarios, due to the substantial domain gap between synthetic training images and real-world test images. In the field of 2D planar SR, numerous works [13], [14] have addressed this problem by combining and mixing several degradations to better simulate real-world conditions. BSRGAN [13] and Real-ESRGAN [14] employ GAN losses to make the super-resolved output more realistic in the training process. However, GAN-based methods can be unstable, leading to undesirable artifacts in the generated images [15]. Recently, generative diffusion models, which utilize a Markov chain with hundreds of inference steps, have demonstrated strong representation abilities and have dominated many computer vision tasks, including real-world image SR. Some methods [16]–[19] have proposed leveraging the powerful capabilities of pretrained large-scale text-to-image (T2I) models to provide image priors. However, these diffusion-based methods have primarily been developed for 2D planar images. To date, limited research has been conducted in the field of ODISR to utilize large-scale pretrained generative models for the reconstruction of ODIs.

To address these issues, we propose leveraging a well-trained diffusion-based model, originally trained on a real-world 2D planar dataset, to exploit the rich image priors for the real-world ODISR problem. Specifically, to bridge the gap between 2D image SR and ODISR, we adopt an advanced parameter-efficient fine-tuning adapter, e.g., LoRA [20], to adapt the large-scale pretrained 2D planar image SR model to the ODISR model. LoRA is a low-rank decomposition method that freezes the weights of pretrained model and decomposes the weights of Transformer layers into trainable low-rank matrices, significantly reducing the number of trainable parameters. This allows us to efficiently transfer the well-trained knowledge from the 2D image domain to the ODI domain with a minimal number of trainable parameters.

The main contributions of this paper are as follows:

- Unlike previous works that focus on synthetic omnidirectional image super-resolution (ODISR), we investigate the challenge of real-world ODISR, where images are subject to multiple degradations, such as noise, blur, jpeg compression, and downsampling.
- We propose adopting the LoRA adapter to adapt a large-scale pretrained model from the 2D image domain to the ODI domain. By fine-tuning only the decomposed weight matrices of the Transformer layers, this approach significantly reduces the number of trainable parameters and training time compared to the original large-scale pretrained model.
- Experimental results demonstrate the effectiveness and superiority of the proposed method quantitatively and qualitatively.

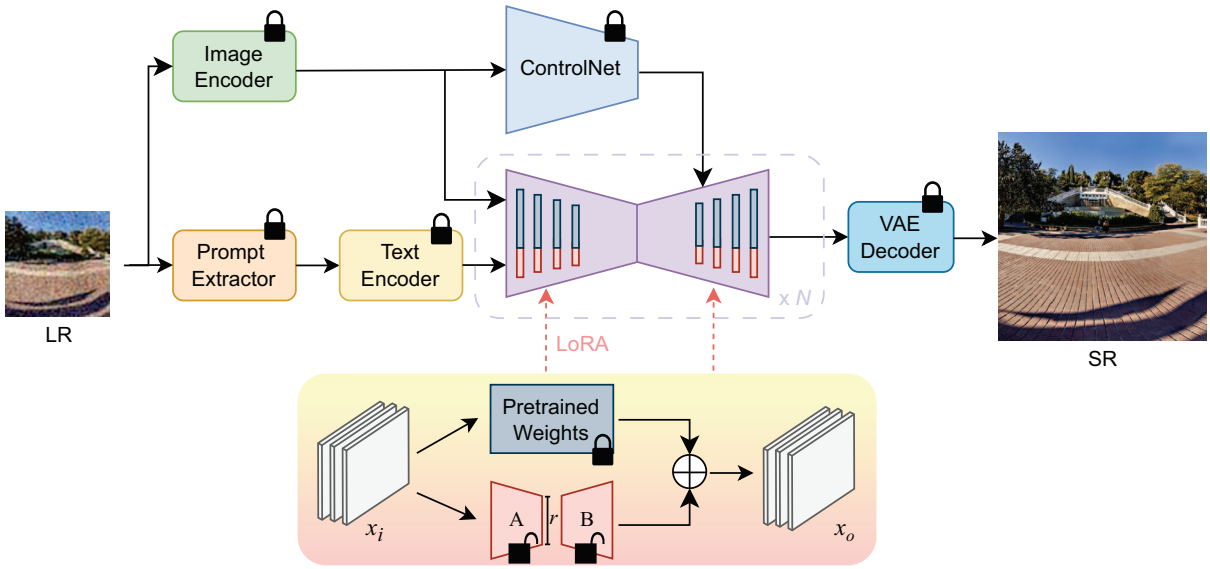

Fig. 1. The overview pipeline of the proposed method for real-world omnidirectional image super-resolution. Only A and B, which are low-rank decomposed matrices of the pretrained weight matrix, are trainable.

II. Related Works

A. Omnidirectional Image Super-Resolution (ODISR)

Omnidirectional images are projected as 2D equirectangular (ERP) images for image processing. LAU-Net [7] is a progressive pyramid network to highlight the non-uniform pixel density across latitudes in ERP images. However, training multiple network levels for different latitude bands is computationally expensive and can result in inconsistencies between the bands. Inspired by LIIF [21], SphereSR [9] creates a continuous spherical image representation to predict RGB values across various projection types. Although SphereSR can flexibly resolve ODIs with arbitrary projection types, it requires training multiple network branches for different projections. OSRT [47] utilizes Fisheye downsampling, applying uniform bicubic downsampling on the original ODIs. OSRT [1] was proposed to apply fisheye downsampling on the ODIs to perform uniform bicubic downsampling on the original ODIs. In addition, OSRT also proposes to adopt a deformable attention mechanism to make full use of the distortion map in the reconstruction. GDGT-OSR [10] proposes to utilize the rectangle-window-based transformer [22] for better adapting to the distortion of ERP images, and introduce the distortion-guided mechanism to modulate the attention area.

However, the abovementioned methods primarily focus on synthetic ODIs, which struggle to perform well on real-world data. In this paper, we explore real-world ODISR to address this problem.

B. Generative Models for Real-World Image Super-Resolution

In recent years, researchers have employed generative models, such as GAN [23] and diffusion networks [24], in the field of real-world image SR. SRGAN [25] was the first to adopt the GAN loss [23] in the SR training process to generate photo-realistic images. BSRGAN [13] and Real-ESRGAN [14] randomly shuffle the degradation of an image to generate realistic training pairs. However, training GAN models is unstable, and GAN models can easily generate unnatural visual artifacts. Recently, researchers have started exploring the potential of more powerful pretrained text-to-image models, such as Stable Diffusion [26], for solving real-world image SR problems. StableSR [16] employed a time-aware encoder and a feature warping module to balance quality and fidelity. PASD [17] feeds both low-level and high-level features into the pretrained Stable Diffusion model with a pixel-aware cross-attention module. SeeSR [19] takes both text prompts and image embedding as input to improve generation performance.

Despite this advancement, limited research has been conducted in real-world ODISR. In this paper, we aim to utilize the rich image priors provided by large-scale pretrained models to bridge the gap between 2D planar images and ODIs.

III. Methodology

A. Framework Overview

In this paper, we leverage a pretrained large-scale model that exploits both text and image representations, e.g., SeeSR [19], as our backbone. The overview pipeline, depicted in Fig. 1, consists of a prompt extractor, text encoder, image encoder, stable diffusion (SD) with ControlNet, and the LoRA adapter. During the training phase, all modules except the LoRA adapter are frozen. Specifically, we only train the decomposed matrices, e.g., A and B in the LoRA module, which are the low-rank matrices of the pretrained weight matrix. We applied LoRA to query, key, value and output projection matrices in the self-attention module, i.e., W_q, W_k, W_v, W_o.

B. LoRA for Real-World Omnidirectional Image Super-Resolution

Low-rank adaptation (LoRA) [20] is a parameter-efficient training technique that inserts a small number of trainable weights instead of fine-tuning all of the model's parameters. Based on the theory that pretrained models have a low "intrinsic rank" when adapting to a downstream task [27], LoRA posits that the updates to the model weights, i.e., ΔW, still have a low "intrinsic rank" and can be projected into a small space. The rank of a matrix represents the maximum number of linearly independent vectors (rows or columns) in it. When adapting a pretrained model to the downstream task, ΔW can be approximated by a small number of linearly independent vectors.

The bottom of Fig. 1 illustrates the mechanism of weight parametrization using LoRA. When fine-tuning the pretrained large-scale model, a large weight matrix $W \in \mathcal{R}^{m\times n}$ are decomposed into two trainable low-rank matrices, i.e., $A \in \mathcal{R}^{r\times n}$ and $B \in \mathcal{R}^{m\times r}$, where the rank $r \ll min(m, n)$. The pretrained weights are kept unchanged. Then, the dot product of the two low-rank matrices is added to the freezed pretrained weight matrix. The forward pass with LoRA can be expressed as follows:

$$x_o = Wx_i + \Delta W x_i = Wx_i + BAx_i, \tag{1}$$

where x_i and x_o are input and output features, respectively.

Fig. 2. An example of an equirectangular projection image.

In this paper, we address the problem of real-world ODISR, which is still a relatively unexplored area. In SR, omnidirectional images (ODIs) are typically projected as equirectangular projection (ERP) images for easier processing. Fig. 2 shows an example of an ERP image, which exhibits various geometric distortions due to the projection from a sphere surface (ODI) to a 2D plane (ERP image). These distortions vary across different latitudes, making ERP images distinct from common 2D planar images.

Significant advancements have been made in the field of real-world 2D image SR [30]–[33], particularly with algorithms utilizing recent large-scale models [16]–[19]. While these advanced real-world SR models provide rich image priors, directly applying 2D image SR models to ODIs is not advisable due to the domain gap between these two image types. One potential solution is to fine-tune large-scale real-world 2D image SR models. Considering the vast number of parameters in state-of-the-art (SOTA) large-scale models, fully fine-tuning these models from scratch is inefficient and requires substantial computational resources. To address these issues, we propose using the parameter-efficient adapter LoRA to adapt the 2D planar image SR model for ODIs. This approach not only maintains well-trained features for real-world image SR, but also facilitates adaptation for the ODI domain by training only a small number of parameters. Notably, the LoRA adapter is applied to the query, key, value, and output projection matrices in the self-attention module.

IV. Experiments

A. Datasets

We adopt the ODI-SR dataset [7], which consists of 800 HR images with a resolution of 1024×2048, for training. For testing, we use the ODI-SR testing dataset and the SUN 360 Panorama dataset [34]. Both testing datasets contain 100 HR images with a resolution of 1024×2048. Due to the lack of real-world ODI datasets, we follow the degradation pipeline of Real-ESRGAN [14] to synthesize LR-HR training pairs from these public datasets. The degradation involves Gaussian noise, Poisson noise, blur, jpeg compression, and downsampling. The scaling factor used in our experiments is 4. During training, the training pairs are cropped into patches, with the HR image patch size being 512×512.

B. Experimental Settings

We conducted all experiments using PyTorch [35]. For the LoRA adapter, we set the rank of the A and B matrices to 4. The learning rate was set to 5×10^{-5}, and the batch size was 4. We use the constant learning rate schedule as SeeSR [19] during the training process. The total number of training iterations 50,000. We used AdamW [36] as the optimizer with $\beta 1 = 0.9$ and $\beta 2 = 0.999$, and the weight decay is 0.01. We ran the experiments on an Nvidia GeForce RTX 3090 GPU, and the model was trained for approximately 2 days.

C. Experimental Results

1) Quantitative Comparison with State-of-the-Art Methods: Table I shows the quantitative results of different methods on the ODI-SR and SUN 360 Panorama test datasets. We compare our method with other SOTA SR methods, such as ESRGAN [28], SwinIR [29], PASD [17], and OSRT [1]. Notably, ESRGAN [28], SwinIR [29] and PASD [17] are SR methods for 2D planar images, while OSRT [1] is a SOTA method for ODISR. In addition, SwinIR [29] and OSRT [1] are transformer-based methods, while ESRGAN [28] and PASD [17] are photo-realistic SR methods based on generative models. The backbone of our method is the model proposed in SeeSR [19], which is also based on the generative diffusion model. We use multiple evaluation metrics to evaluate the performance of different methods. PSNR and SSIM are common metrics for evaluating SR methods. NIQE [37] is a no-reference image quality evaluation metric. FID [38] is

TABLE I

QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE ODI-SR AND SUN 360 PANORAMA DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE SCALING FACTOR IS 4.

Methods	ODI-SR					SUN 360 Panorama				
	PSNR↑	SSIM↑	NIQE↓	LPIPS↓	FID↓	PSNR↑	SSIM↑	NIQE↓	LPIPS↓	FID↓
Bicubic	21.22	0.5128	9.4243	0.6264	-	21.2177	0.5128	9.4243	0.6218	-
ESRGAN [28]	20.92	0.4871	8.8908	0.6299	116.97	20.78	0.4744	8.8908	0.6262	143.78
SwinIR [29]	20.87	0.4852	8.7825	0.6307	-	20.74	0.4722	8.789	0.6268	-
PASD [17]	21.18	0.5727	**4.6861**	0.4602	77.43	21.27	0.5783	**4.6471**	0.4593	75.52
OSRT [1]	19.70	0.4389	7.9834	0.6345	-	19.12	0.4213	7.7438	0.8153	-
Ours	**21.51**	**0.5903**	5.4967	**0.4182**	**70.29**	**21.77**	**0.5989**	5.3447	**0.4307**	**66.02**

TABLE II

COMPARISON RESULTS ON DIFFERENT RANKS IN LORA. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE SCALING FACTOR IS 4.

Rank	Params	ODI-SR					SUN 360 Panorama				
		PSNR↑	SSIM↑	NIQE↓	LPIPS↓	FID↓	PSNR↑	SSIM↑	NIQE↓	LPIPS↓	FID↓
r=4	1.59M	**21.51**	**0.5903**	**5.4967**	**0.4185**	70.29	**21.77**	**0.5989**	**5.3447**	0.4307	66.02
r=8	3.18M	**21.51**	**0.5903**	**5.4967**	**0.4185**	**70.27**	21.69	0.5965	5.3657	**0.4181**	**65.4257**

TABLE III

QUANTITATIVE RESULTS WITH DIFFERENT TRAINING DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE SCALING FACTOR IS 4.

Finetune Dataset	ODI-SR					SUN 360 Panorama				
	PSNR↑	SSIM↑	NIQE↓	LPIPS↓	FID↓	PSNR↑	SSIM↑	NIQE↓	LPIPS↓	FID↓
DIV2K	21.49	**0.5903**	5.53	**0.4174**	**68.98**	21.74	**0.6009**	5.3904	0.4313	67.19
ODI-SR	**21.51**	**0.5903**	**5.4967**	0.4182	70.29	**21.77**	0.5989	**5.3447**	**0.4307**	**66.02**

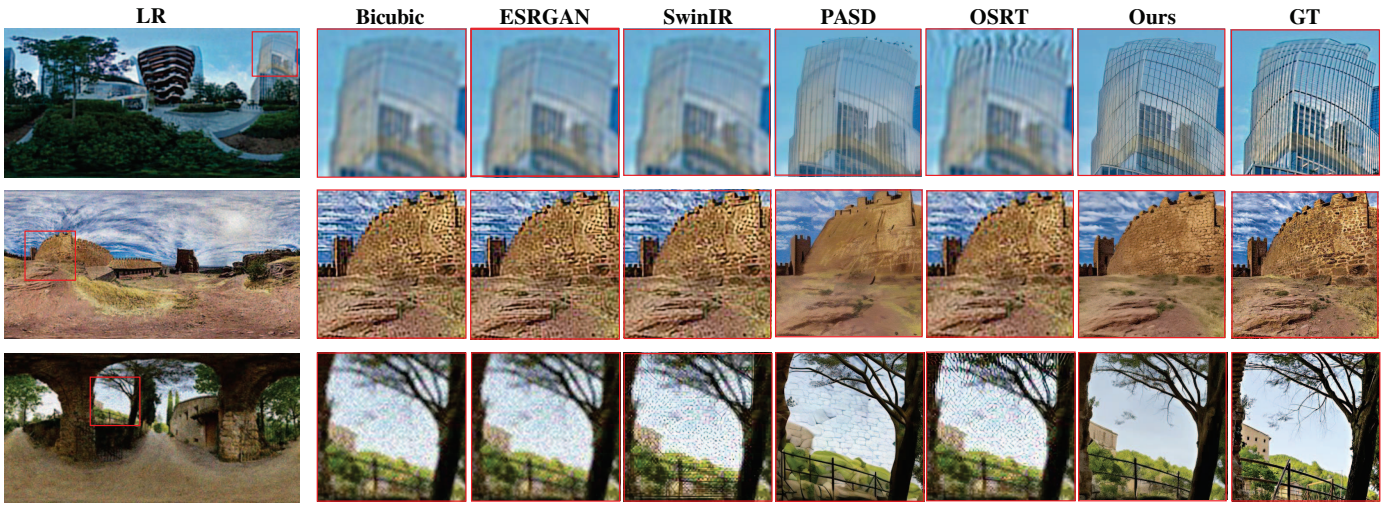

Fig. 3. Qualitative comparisons among different methods on SUN 360 Panorama testing dataset. The scaling factor is 4.

an evaluation metric that measures the distance between the distribution of images generated by a generative model and the distribution of real images (ground truth). LPIPS [39] is used to measure the perceptual similarity between two images. The higher PSNR and SSIM represent the better performance, while the lower NIQE, FID, and LPIPS mean the better results. As can be seen from Table I, our method outperforms other SR methods in most evaluation metrics on the ODI-SR and SUN 360 Panorama test datasets. In terms of the NIQE metric, the proposed method is the closest to PASD [17], which achieves the best result of NIQE. This demonstrates the effectiveness and superiority of our method.

2) Qualitative Comparison with State-of-the-Art Methods: Fig. 3 shows the visual results of different methods on the SUN 360 Panorama test dataset. As we can see, although SwinIR [29] and OSRT [1] perform well in restoring synthetic images, their performance degrades significantly when tested on real-world images which contain multiple and more complex degradation. ESRGAN [28] also struggles to reconstruct images with complex and severe degradation and introduces many unpleasant artifacts, which is a drawback of the GAN-based method. PASD [17] produces images that are too smooth and cannot restore details. Furthermore, from the third row in Fig. 3, we can see that PASD [17] restores the sky as the wall. The possible reason is that PASD is based on a generative

model, so it will generate some wrong content in some situations. Our method can produce more visually pleasing results with richer details and textures, which are closer to ground truths. Qualitative comparisons between different types of SR methods also demonstrate the superiority of the proposed methods.

3) The Rank in LoRA: The rank r determines the dimension of the trainable matrices, i.e., $A \in \mathcal{R}^{r\times n}$ and $B \in \mathcal{R}^{m\times r}$, in LoRA. A larger r represents more trainable parameters and more computational resources. We compare the impact and number of trainable parameters of different r on various evaluation metrics. Table II shows the comparisons between $r = 4$ and $r = 8$ on the two test datasets. Although the number of trainable parameters of the model with $r = 8$ is 3.18M, which is twice as large as that of the model with $r = 4$, i.e., 1.59M, the model with $r = 4$ performs equally well, or even better than the model with $r = 8$. One possible reason is that the domain gap between 2D planar images and omnidirectional images is not very large, so there is no need to fine-tune a large number of parameters. Therefore, we set $r = 4$ as the default value of the rank in our experiments. It is worth mentioning that the number of trainable parameters is 1.13B if the model is fully fine-tuned, which requires significantly more computational resources compared to using LoRA.

4) Fine-tuning Using Different Datasets: After applying LoRA, we fine-tune the model using different datasets, e.g., DIV2K and ODI-SR. DIV2K is a common SR dataset that involves 800 HR 2D planar images for training. DIV2K and ODI-SR represent two different image distributions. When fine-tuning using these two different datasets, we keep the architecture of the model and training configuration the same. As shown in Table III, the model trained using ODI-SR performs better than or on par with the model trained using DIV2K. This illustrates that it is important to keep the distributions of training data and test data consistent.

V. Conclusions

In this paper, we propose an efficient adaptation method for real-world omnidirectional image super-resolution (ODISR). Unlike previous real-world ODISR methods, we focus on SR of real-world ODIs with more complex degradations. Based on the advanced backbone for real-world 2D planar images, we propose to adopt the low-rank adaptation (LoRA) technique to adapt the SR domain from 2D planar images to omnidirectional images (ODIs). On one hand, training the low-rank decomposed matrices of the pretrained model is both time and resource efficient. On the other hand, the knowledge of pretrained model can be efficiently adapted to downstream taskS. Quantitative and qualitative results demonstrate the effectiveness and superiority of the proposed method.

References

[1] 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.

[2] C. Yang, J. Xiao, Y. Ju, G. Qiu, and K.-M. Lam, "Improving robustness of single image super-resolution models with monte carlo method," in *Proceedings of the IEEE International Conference on Image Processing*, 2023, pp. 2135–2139.

[3] R. Dong and K.-M. Lam, "Bi-center loss for compound facial expression recognition," *IEEE Signal Processing Letters*, vol. 31, pp. 641–645, 2024.

[4] Z. Lyu, J. Xiao, C. Zhang, and K.-M. Lam, "Ai-generated image detection with wasserstein distance compression and dynamic aggregation," in *2024 IEEE International Conference on Image Processing*, IEEE, 2024, pp. 3827–3833.

[5] H. Nagahara, Y. Yagi, and M. Yachida, "Super-resolution from an omnidirectional image sequence," in *Proceedings of the IEEE International Conference on Industrial Electronics, Control and Instrumentation*, vol. 4, 2000, pp. 2559–2564.

[6] 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.

[7] 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.

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

[9] 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.

[10] C. Yang, R. Dong, J. Xiao, *et al.*, "Geometric distortion guided transformer for omnidirectional image super-resolution," *arXiv preprint arXiv:2406.10869*, 2024.

[11] J. Gu, H. Lu, W. Zuo, and C. Dong, "Blind super-resolution with iterative kernel correction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1604–1613.

[12] K. Zhang, L. V. Gool, and R. Timofte, "Deep unfolding network for image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3217–3226.

[13] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4791–4800.

[14] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1905–1914.

[15] 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.

[16] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution," *International Journal of Computer Vision*, pp. 1–21, 2024.

[17] T. Yang, P. Ren, X. Xie, and L. Zhang, "Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization," *arXiv preprint arXiv:2308.14469*, 2023.

[18] X. Lin, J. He, Z. Chen, *et al.*, "Diffbir: Towards blind image restoration with generative diffusion prior," *arXiv preprint arXiv:2308.15070*, 2023.

[19] 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, pp. 25 456–25 467.

[20] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[21] 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.

[22] 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.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[24] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 23 593–23 606.

[25] C. Ledig, L. Theis, F. Huszár, *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.

[26] 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.

[27] A. Aghajanyan, L. Zettlemoyer, and S. Gupta, "Intrinsic dimensionality explains the effectiveness of language model fine-tuning," *arXiv preprint arXiv:2012.13255*, 2020.

[28] X. Wang, K. Yu, S. Wu, *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision Workshops*, 2018, pp. 1–16.

[29] 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.

[30] S. Bell-Kligler, A. Shocher, and M. Irani, "Blind super-resolution kernel estimation using an internal-gan," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 32, 2019.

[31] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin, "Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 701–710.

[32] Y. Huang, S. Li, L. Wang, T. Tan, *et al.*, "Unfolding the alternating optimization for blind super resolution," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 5632–5643.

[33] A. Shocher, N. Cohen, and M. Irani, ""zero-shot" super-resolution using deep internal learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3118–3126.

[34] 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*, 2012, pp. 2695–2702.

[35] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 32, 2019.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[37] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[38] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of Multivariate Analysis*, vol. 12, no. 3, pp. 450–455, 1982.

[39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.