

D²R²OSR: Degradation-Disentangled Representation for Real-World Omnidirectional Image Super-Resolution

Hongyu An¹, Xinfeng Zhang^{1*}(✉), Xu Fan¹, Shijie Zhao²,
Li Zhang², and Ruiqin Xiong³

¹ School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China

{anhongyu22, xufan24}@mailsucas.ac.cn, xfzhang@ucas.ac.cn

² ByteDance Inc., Shenzhen, China

{zhaoshijie.0526, lizhang.idm}@bytedance.com

³ School of Computer Science, Peking University, Beijing, China
rqxiong@pku.edu.cn

Abstract. With the growing demand for immersive visual experiences, high-quality omnidirectional images (ODIs) have become increasingly important. However, limitations in imaging devices and transmission bandwidth often lead to low-resolution ODIs, hindering the rendering of fine-grained 360° details, especially in the presence of real-world degradations and geometric distortions. Existing real-world super-resolution (Real-SR) methods are inadequate for ODIs, as their degradation models fail to account for the complex imaging pipeline involving fisheye capture and Equirectangular Projection (ERP), introducing severe aliasing and projection-specific distortions. To address these challenges, we propose **D²R²OSR**, a **D**egradation-**D**isentangled **R**epresentation framework for **R**eal-world **O**mnidirectional image **S**uper-**R**esolution. D²R²OSR explicitly models degradations arising from both fisheye imaging and ERP projection, guided by two key insights: (1) projection priors play a critical role in shaping real-world degradations, and (2) human perception in immersive environments is inherently viewpoint-centric. Accordingly, we introduce a Perspective Projection Representation (PPR) operating alongside the ERP branch to capture viewpoint-aware features, together with a Degradation-Specific Module (DSM) that jointly models ERP-induced geometric distortions and PPR-specific real-world degradations. Extensive experiments demonstrate that D²R²OSR achieves state-of-the-art performance and produces visually compelling, high-fidelity omnidirectional Real-SR results while maintaining favorable computational efficiency for low-resource deployment.

Keywords: Super-Resolution · Omnidirectional Image · Degradation Representation

* Corresponding author.

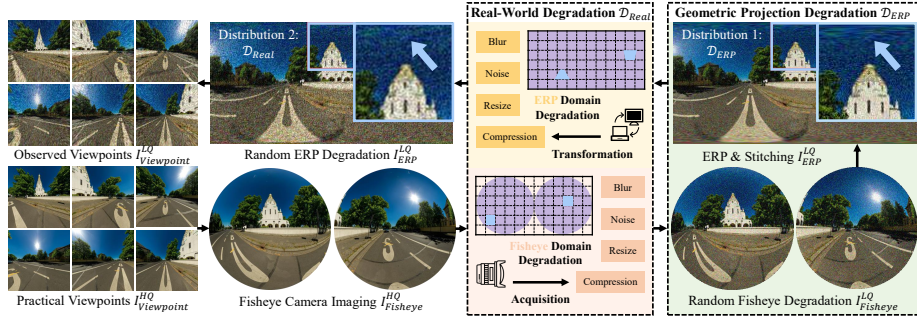


Fig. 1: Real-world degradations in the omnidirectional imaging pipeline. During fisheye capture and subsequent ERP conversion, degradations such as blur, noise, resize, and compression are introduced and further amplified by projection-induced distortions (e.g., geometric stretching and shape deformation), resulting in severe visual artifacts. The zoomed-in regions reveal distinct, domain-specific degradation characteristics in fisheye and ERP domains, motivating separate modeling for faithful restoration.

1 Introduction

Omnidirectional images (ODIs), also known as 360° or panoramic images, capture the entire 360° field of view (FoV), enabling realistic visual experiences. With the rapid development of virtual and augmented reality applications, the demand for high-quality ODIs has increased dramatically. To ensure immersion, ODIs must be rendered at ultra-high resolutions (e.g., 4K, 8K, or even 16K), while local regions of visual attention—typically spanning 100° × 100° to 120° × 120° FoV [11]—should remain sharp and detailed. However, most ODIs are constrained by limited resolution owing to the high cost of precision imaging systems and restricted transmission bandwidth. Moreover, throughout the practical imaging pipeline—including acquisition, stitching, projection, transmission, processing, and display—ODIs suffer from diverse degradations such as blur, noise, resize, compression, and projection-induced distortions, all of which severely impair visual fidelity. Addressing these degradations through faithful high-quality reconstruction remains a fundamental yet unresolved challenge.

Super-resolution (SR) has therefore emerged as an effective client-side solution for enhancing ODI quality. Nevertheless, conventional SR methods [8, 14, 17, 23, 31, 43] are typically trained under the idealized Bicubic down-sampling assumption, which deviates significantly from real-world degradation processes and limits their generalization. Recent real-world SR (Real-SR) approaches [4, 6, 21, 33–35, 42] attempt to mitigate this gap by either implicitly learning degradation priors from low-resolution data or explicitly synthesizing realistic degradation models. Despite their progress, these methods largely overlook the mixed degradations and projection-related distortions intrinsic to the ODI imaging pipeline, where multiple degradation sources jointly hinder the recovery of photo-realistic, high-fidelity omnidirectional images.

ODIs are typically captured using fisheye cameras and stitched to cover the full 360° FoV. To ensure compatibility with general image transmission and processing pipelines, the spherical content is projected onto planar representations, where pixel-wise mapping inevitably introduces information loss and interpolation distortions. Among various projection methods, Equirectangular Projection (ERP) is the most widely adopted for its computational efficiency and broad applicability, enabling direct processing as standard 2D images despite non-uniform content density. After projection, ERP ODIs are subjected to compression, transmission, and enhancement, and are finally rendered as localized perspective views for immersive visualization. As illustrated in Fig. 1, diverse real-world degradations $\mathcal{D}_{Real}$ arise throughout the ODI imaging pipeline, while ERP further introduces intrinsic geometric distortions $\mathcal{D}_{ERP}$, together yielding complex and mixed degradation distributions. Existing ODI-SR methods [1, 2, 12, 13, 16, 19, 25, 27, 32, 38, 41] primarily focus on mitigating latitude-dependent distortions by leveraging geometric priors in ERP space. Although these approaches partially alleviate projection-related degradations, they remain insufficient for addressing the full spectrum of real-world degradations encountered in practical ODI pipelines, limiting restoration quality and generalization.

To address the challenges of real-world ODI-SR, we propose a mixed fish-eye-ERP degradation model that synthesizes realistic training samples by capturing diverse degradation distributions across the imaging pipeline. Building on this, we develop **D²R²OSR** (**D**egradation-**D**isentangled **R**epresentation-Based **R**eal-World **O**mnidirectional Image **S**uper-**R**esolution), a unified framework that disentangles latitude-dependent projection distortions from blind pixel-level degradations. Specifically, a perspective-aware representation branch is introduced alongside the ERP branch to enhance local viewpoint quality, while degradation-specific encoding enables both branches to adaptively compensate for distortions with or without explicit spatial priors. A projection fusion mechanism further integrates complementary features across the two domains. The main contributions of this paper are summarized as follows:

1. We present a unified degradation modeling paradigm for real-world ODI-SR that captures the coupled effects of fisheye imaging and ERP projection throughout the omnidirectional imaging pipeline. By explicitly disentangling degradations with distinct priors, D²R²OSR enables more faithful and grounded restoration of real-world ODIs.
2. We propose a Perspective Projection Representation (PPR) that bridges spherical and planar domains, effectively decoupling projection-induced geometric distortions from pixel-level real-world degradations and facilitating accurate degradation modeling from a viewpoint-centric perspective.
3. We develop a dual-domain ODI-SR framework that jointly leverages complementary omnidirectional and perspective representations, equipped with Degradation-Specific Modules (DSMs) for adaptive degradation compensation to robustly handle diverse and mixed degradation scenarios.

2 Related Work

2.1 Real-World Image Super-Resolution

With the rapid advances in deep learning, network-based image SR methods [8, 14, 17, 23, 31, 43] have largely replaced traditional approaches. However, most existing models are trained under predefined degradation assumptions with fixed patterns, which fail to capture the complexity and diversity of real-world degradations. In practice, low-resolution (LR) images are acquired under varying devices and environments, and are commonly affected by blur, noise, resize, and compression. To bridge the gap between synthetic and real-world degradations, a series of works have explored real-world SR (Real-SR). RealSR [4] constructs a large kernel pool to cover diverse degradation patterns. BSRGAN [42] and Real-ESRGAN [34] explicitly design mathematical degradation pipelines for realistic data synthesis, while ADL [28] simulates unknown down-sampling processes via adversarial learning. DASR [21] adaptively estimates degradation from LR inputs to modulate network parameters, and LWay [7] combines supervised pre-training with self-supervised learning to mitigate the synthetic-real domain gap. DKP [39] proposes unsupervised kernel estimation via Markov chain Monte Carlo sampling, whereas AdaSR [15] leverages sample-adaptive priors learned from image self-similarity. More recently, diffusion-based models [24, 36, 40] have further improved perceptual realism by introducing powerful generative priors. Nevertheless, these methods incur substantial computational overhead due to their diffusion backbones, which limits their scalability to ultra-high-resolution scenarios such as omnidirectional images.

2.2 Omnidirectional Image Super-Resolution

In ODI applications, ERP remains the dominant representation due to its simplicity, full-sphere coverage, and strong compatibility with standard architectures. Consequently, most existing ODI-SR methods [1–3, 12, 16, 19, 25, 32, 38, 41] are developed in the ERP domain and focus on mitigating geometric distortions introduced by spherical-to-planar projection. Early works [12, 25] mitigate latitude-dependent distortions and uneven content distribution through projection-related modulation. [1] introduces a perception-oriented network to enhance high-frequency details from critical viewpoints. Subsequent studies explore more advanced distortion modeling and representation strategies. OSRT [41] aligns features using continuous and adaptive offsets to alleviate projection-induced misalignment, while GDGT-OSR [38] extends this idea with distortion-guided rectangle-window self-attention for improved texture modeling. FATO [2] addresses content density imbalance by leveraging frequency-domain characteristics and high-frequency attention. BPOSR [32] combines ERP and cubemap projections through parallel attention to exploit complementary representations, whereas ODA-SRN [16] employs multi-segment parameterized convolutions for geometric distortion compensation. OmniSSR [19] introduces octadecaplex tangent interaction and gradient decomposition for zero-shot ODI-SR. Despite these

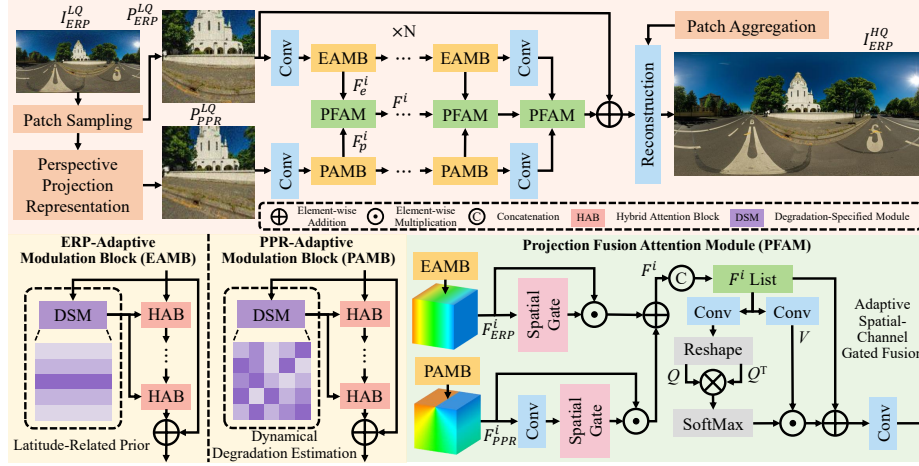


Fig. 2: Overall framework of D²R²OSR. It comprises two parallel branches, ERP and PPR, for distortion-aware restoration. PPR patches are derived from ERP patches in a nearly lossless manner, preserving consistent scene content. The PPR branch captures viewpoint-centric features to enhance perceptual quality. DSMs in the ERP branch compensate for projection-induced geometric distortions, while those in the PPR branch handle local real-world degradations. PFAMs fuse complementary features across branches to facilitate information exchange.

advances, existing methods primarily assume fixed Bicubic down-sampling and emphasize ERP geometric distortions, overlooking the complex real-world degradation encountered in ODI imaging pipelines and the non-uniform human visual perception across local viewpoints. These limitations impair practical applicability and immersive visual quality.

3 Method

3.1 Overall Framework

The D²R²OSR pipeline is demonstrated in Fig. 2. Given a low-quality (LQ) ERP image I_{ERP}^{LQ} with geometric and real-world degradations, our goal is to learn an adaptive model $\mathcal{M}$ that restores high-resolution details by modulating $\mathcal{D}_{ERP}$ and $\mathcal{D}_{Real}$. Formally, I_{ERP}^{LQ} is partitioned into ERP patches P_{ERP}^{LQ} and transformed into perspective patches P_{PPR}^{LQ} via Perspective Projection Representation (PPR), facilitating the disentanglement of ERP-induced and real-world degradations. Both patch sets are encoded into shallow features F_e^0 and F_p^0 and processed by parallel ERP and PPR branches, each consisting of N EAMBs and PAMBs with embedded DSMs. DSMs in EAMBs model latitude-related $\mathcal{D}_{ERP}$, while those in PAMBs focus on $\mathcal{D}_{Real}$. PFAMs fuse F_e^i and F_p^i at each stage, progressively refining complementary features for the final reconstruction.

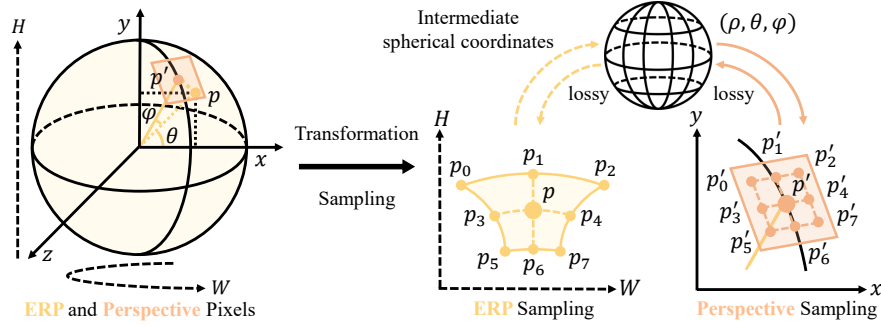


Fig. 3: Illustration of pixel distributions in the Equirectangular Projection (ERP) and Perspective Projection Representation (PPR) domains. ERP exhibits increasing latitude-dependent pixel stretching, whereas PPR maintains a uniform pixel distribution and better preserves local geometry.

3.2 Perspective Projection Representation (PPR)

Although ERP is the dominant representation for ODIs due to its simplicity and broad compatibility, projecting spherical scenes onto a planar domain inevitably introduces severe geometric distortions, particularly at high latitudes. When these distortions interact with real-world degradations throughout the ODI imaging pipeline (Fig. 1), faithful content restoration becomes substantially more challenging. To address this issue, we introduce a perspective projection representation (PPR) that processes features in a projection-independent, viewpoint-centric domain better aligned with human visual perception.

As shown in Fig. 3, the ERP-PPR transformation is formulated along the spherical cutting-plane direction. Specifically, ERP coordinates are first mapped onto the unit sphere, followed by a rotation about the sphere center o to simulate camera motion. The tangent plane at the rotated viewpoint (orange region) is then projected onto the image plane, forming an ERP→Sphere→PPR pipeline. Conventional interpolation-based resampling (e.g., Bicubic or Bilinear interpolation) relies on intermediate spherical warping, inevitably introducing information loss and blurring artifacts. To this end, we construct a more accurate perspective representation that decouples projection-included degradations from real-world distortions. Formally, for a target PPR coordinate $p'(u', v')$ corresponding to an ERP coordinate $p(u, v)$, we define a continuous mapping $p' = f(p)$. The local offset $\delta p'$ is estimated from p and its neighborhood:

$$\mathcal{P} = \{p_i | p_i = f^{-1}(p') + [\frac{m}{h}, \frac{n}{w}], [m, n] \in [-1, 0, 1]\}, \quad (1)$$

where h and w denote the height and width of the ERP representation, respectively. Under a first-order Taylor expansion, the mapping can be locally approximated in a linear form as follows:

$$\begin{aligned} \delta p'_i &= f(p) - f(p_i) \approx f(p_i) + J_f(p_i)(p - p_i) - f(p_i) \\ &= J_f(p_i)(p - p_i) = J_f(p_i)\delta p, \end{aligned} \quad (2)$$

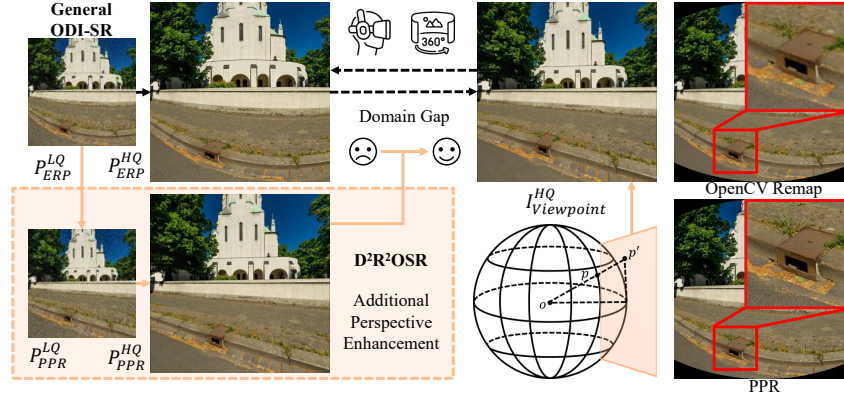


Fig. 4: PPR transforms ERP patches into perspective patches, aligning the viewing distribution with natural images and effectively decoupling projection-included degradations. Compared with the conventional ERP-to-Perspective-to-ERP OpenCV remapping, PPR yields sharper and more structurally faithful reconstructions.

where $J_f(p_i)$ denotes the Jacobian of f at p_i , and $\delta p = p - p_i$ is the local grid in the ERP domain. Since the ERP-to-PPR mapping varies with latitude, the first-order approximation cannot accurately characterize the spatially varying projection warping. To better model these non-uniform distortions, we further introduce higher-order geometric information. Following LTEW [18] and ResVR [20], we incorporate the second-order Hessian $H_f(\mathcal{P})$ together with $J_f(\mathcal{P})$ to jointly encode local orientation and curvature:

$$J_f(\mathcal{P}) = \begin{bmatrix} p'_4 - p'_3 \\ p'_1 - p'_6 \end{bmatrix}, \quad (3)$$

$$H_f(\mathcal{P}) = \begin{bmatrix} p'_3 + p'_4 - 2p' & p'_2 + p'_5 - p'_0 - p'_7 \\ p'_2 + p'_5 - p'_0 - p'_7 & p'_1 + p'_6 - 2p' \end{bmatrix}.$$

Moreover, the local displacement and its gradient in the PPR domain are approximated as $\delta p' = J_f(\mathcal{P})\delta p$, and $\delta(\nabla p') = H_f(\mathcal{P})\delta(\nabla p)$. Unlike prior works [18, 20], which primarily utilize Jacobian and Hessian matrices for pixel-level warping, we explicitly formulate the ERP→PPR conversion through $\delta p'$ and $\delta(\nabla p')$, establishing a continuous coordinate mapping framework. Concretely, the six independent elements derived from J_f and H_f are concatenated to form spatial priors $s(\mathcal{P})$, enabling direct ERP→PPR mapping without intermediate spherical resampling. Inspired by implicit neural representation methods [9, 18, 29], we further design a learnable coordinate transformation estimator $E = E_a, E_f, E_p$ conditioned on $s(\mathcal{P})$. This estimator models local textures in the Fourier domain, enabling adaptive reconstruction of high-frequency details. The estimation is formulated as follows:

$$E(\mathcal{E}(p), \delta p, s(\mathcal{P})) = E_a(\mathcal{E}(p)) \otimes \begin{bmatrix} \cos\{\pi(\langle E_f(\mathcal{E}(p)), \delta p \rangle + E_p(s(\mathcal{P})))\} \\ \sin\{\pi(\langle E_f(\mathcal{E}(p)), \delta p \rangle + E_p(s(\mathcal{P})))\} \end{bmatrix}, \quad (4)$$

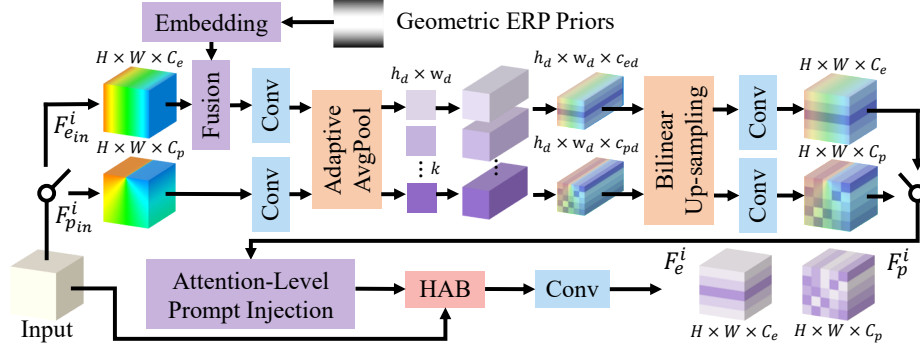


Fig. 5: Degradation-Specific Module (DSM). The DSM learns projection-aware adaptive guidance to compensate for projection-induced degradations.

where $\mathcal{E}(\cdot)$ indicates the feature encoder, while E_a , E_f , and E_p estimate the amplitude, frequency, and phase components, respectively. Here, $\langle \cdot, \cdot \rangle$ means the inner product. Based on the learned coordinate transformation estimator, the PPR-domain pixel distribution is predicted as follows:

$$P_{PPR}^{LQ} = \sum_{i \in \mathcal{P}} \omega_i \mathcal{D}(E(\mathcal{E}(p), \delta p, s(p_i))), \quad (5)$$

where $\mathcal{D}(\cdot)$ denotes a four-layer MLP, and ω_i is a local ensemble coefficient. As depicted in Fig. 4, compared with interpolation-based OpenCV Remap, the proposed formulation significantly reduces information loss during ERP $\rightarrow$ PPR conversion. Consequently, the PPR branch effectively decouples projection-induced geometric degradations, improves real-world degradation modeling, and better aligns with human visual perception.

3.3 Degradation-Specific Module (DSM)

As illustrated in Fig. 2, we introduce two complementary modules: the ERP-Adaptive Modulation Block (EAMB) and the PPR-Adaptive Modulation Block (PAMB), to address multi-degradation restoration challenges in ODIs. The EAMB operates in the ERP domain to mitigate latitude-dependent geometric distortions, while the PAMB models random real-world degradations in the perspective domain. To exploit domain-specific priors, we embed Degradation-Specific Modules (DSMs) into both branches, enabling degradation-aware modulation under heterogeneous projection distributions.

Fig. 5 shows the DSM architecture. At the i th stage, the inputs to EAMB and PAMB are $F_e^i \in \mathbb{R}^{H \times W \times C_e}$ and $F_p^i \in \mathbb{R}^{H \times W \times C_p}$, respectively. The DSM takes three inputs: branch features from EAMB/PAMB, a projection-aware conditional map, and a learnable degradation descriptor (“Prompt”). Since feature distributions vary spatially across projections, we first extract projection-aware

degradation priors using 1×1 convolutions. For the ERP branch, a latitude-related quantization map is derived to characterize spatially varying geometric distortions, while a detailed derivation is provided in the Appendix. The resulting conditional map of size $H \times W \times 1$ is fused with F_e^i to produce geometrically calibrated features $F_e^{i'}$. To achieve distribution-adaptive modulation, we further design a lightweight encoder to generate learnable degradation descriptors $D_{ERP}^i \in \mathbb{R}^{h_d \times w_d \times c_{ed}}$ and $D_{PPR}^i \in \mathbb{R}^{h_d \times w_d \times c_{pd}}$. Specifically, features are processed by 3×3 convolutions with LeakyReLU activations, followed by Adaptive Average Pooling (AAP) to obtain compact descriptor vectors. These descriptors are subsequently up-sampled and refined via convolution layers to generate feature-level modulation masks. The generation of degradation-specific features in the ERP and PPR branches is summarized as:

$$DSF_{ERP/PPR}^i = \sum_{k=1}^K \text{AAP}(\text{Conv}(F_e^{i'} / F_p^i))_k P_k, \quad (6)$$

where k represents the descriptor length. The resulting degradation-specific feature $DSF_{ERP/PPR}^i$ is concatenated with the branch input and subsequently fed into the HAB block [8] for degradation-aware feature modulation:

$$F^{i'} = \text{Cat}(F_{e/p}^i, \text{Up}(\text{Conv}(DSF_{ERP/PPR}^i))), \quad F_{e/p}^i = \text{Conv}(\text{HAB}(F^{i'})). \quad (7)$$

The entire framework is optimized end-to-end using only the reconstruction objective, without explicit degradation annotations. Built upon the representation-level disentanglement provided by PPR, the learnable descriptors implicitly organize different degradation characteristics into distinct feature subspaces. To further enhance degradation-aware modulation, we inject prompts into the value branch of window-based multi-head self-attention as $V' = V + P_v$, where P_v denotes a prompt tensor aligned with the multi-head structure. By modulating only the value component, the proposed strategy preserves query-key similarity estimation while enabling degradation-aware contextual aggregation. Compared with feature concatenation or bias-based conditioning, the proposed strategy enables head-wise contextual modulation without introducing additional attention branches or noticeable computational overhead.

3.4 Projection Fusion Attention Module (PFAM)

The aggregation of enhanced features across different distributions and stages is critical for ODI-SR performance. As demonstrated in Fig. 2, EAMBs and PAMBs operate in distinct projection domains and hierarchical levels, making effective feature integration non-trivial. Existing work [32] directly concatenates branch features followed by attention-based fusion. However, the distinct spatial distributions of ERP and PPR limit the effectiveness of naive concatenation. To address this issue, we propose a Projection Fusion Attention Module (PFAM) for adaptive cross-branch fusion.

To effectively integrate ERP and PPR features from distinct projection domains, we propose an adaptive spatial-channel gated fusion mechanism, as illustrated in Fig. 2. Given ERP and PPR features $F_e^i \in \mathbb{R}^{B \times C_e \times H \times W}$ and $F_p^i \in \mathbb{R}^{B \times C_p \times H \times W}$, we first align channel dimensions via a learnable 1×1 convolution $\mathcal{P} : \mathbb{R}^{C_p} \rightarrow \mathbb{R}^{C_e}$, yielding $\tilde{F}_p^i = \mathcal{P}(F_p^i)$.

A branch-aware spatial attention map is then computed as:

$$S_{e/p} = \text{Sigmoid} \left(W_2 * \text{LeakyReLU} \left(W_1 * F_e^i / \tilde{F}_p^i \right) \right), \quad (8)$$

where W_i denotes convolution layers, and the resulting spatial gates assign adaptive importance to different projections. The fusion is performed residually:

$$F^i = S_e \odot F_e^i + S_p \odot \tilde{F}_p^i. \quad (9)$$

The fused feature set $\mathcal{F} = \{F^i\}$ spans multiple network stages. To capture inter-stage dependencies, we further introduce a depth-wise attention mechanism over $\mathcal{F}$. Features are reshaped as $Q, V \in \mathbb{R}^{B \times D \times (CHW)}$ and depth attention is computed as $\text{Softmax}(QQ^\top)$. This design enables adaptive fusion across stages, producing more structurally coherent representations. Overall, PFAM bridges ERP and PPR branches by reducing feature misalignment and adaptively balancing geometric and real-world cues, leading to more robust real-world ODI-SR.

4 Experiments

4.1 Implementation Details

The proposed D²R²OSR presents a projection fusion representation framework for real-world ODI-SR. We train our model on Flickr360 [5], and further evaluate it on Flickr360 [5], ODI-SR [12] and SUN360 [37]. All HR ERP ODIs have a resolution of 1024×2048 . To simulate realistic degradations, We follow the degradation settings of Real-ESRGAN [34], while additionally incorporating fisheye and ERP-specific degradation distributions as indicated in Fig. 1. The network consists of $N = 6$ EAMB-PAMB pairs, each with a depth of 6. The channel dimensions are set to $c_e = 180$ and $c_p = 60$, respectively. For DSMs, we set $k = 5$, $c_{ed} = 128$, and $c_{pd} = 60$. During training, HR ODIs are randomly cropped into 256×256 patches with random flip augmented. We train the model for 500K iterations using the Adam optimizer ($lr = 1 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.99$) with an ℓ_1 loss. The batch size is set to 4. All experiments are implemented in PyTorch and conducted on NVIDIA A6000 GPUs.

4.2 Comparison Results

To validate the effectiveness of D²R²OSR for real-world ODI-SR, we compare it with representative methods from three related domains: ODI-SR, Real-SR, and general image restoration. The compared approaches include OSRT [41], BPOSr [32], Real-ESRGAN [34], SwinIR [22], HAT [8], PromptIR [26], and

Table 1: Quantitative real-world ODI-SR comparisons on Flickr360, ODI-SR, and SUN360 datasets under $\times 4$, $\times 8$, and $\times 16$ settings. Best in **red**, second-best in **blue**.

Method	×4		×8		×16	
	WS-PSNR↑ / WS-SSIM↑					
Flickr360 Dataset						
360-SS	24.20	0.6198	16.98	0.5728	15.91	0.5844
Real-ESRGAN	24.30	0.6171	21.85	0.6037	19.84	0.5934
SwinIR	25.16	0.6827	22.83	0.6347	21.19	0.6202
HAT	25.19	0.6838	22.93	0.6354	21.36	0.6206
PromptIR	25.20	0.6852	22.87	0.6376	21.40	0.6240
OSRT	25.41	0.6886	22.93	0.6354	21.41	0.6211
BPOSr	25.34	0.6864	22.88	0.6348	21.27	0.6212
AdaIR	25.49	0.6912	22.96	0.6377	21.45	0.6238
D ² R ² OSR	25.64	0.6934	23.24	0.6433	21.64	0.6249
ODI-SR Dataset						
360-SS	23.30	0.6031	17.11	0.5562	15.97	0.5667
Real-ESRGAN	23.24	0.6001	21.52	0.5904	19.61	0.5785
SwinIR	23.97	0.6537	22.32	0.6129	20.77	0.5996
HAT	23.97	0.6541	22.41	0.6139	20.92	0.6000
PromptIR	24.00	0.6555	22.37	0.6165	20.96	0.6039
OSRT	24.22	0.6569	22.39	0.6135	20.96	0.6004
BPOSr	24.14	0.6541	22.36	0.6132	20.82	0.6006
AdaIR	24.26	0.6591	22.42	0.6158	21.00	0.6036
D ² R ² OSR	24.37	0.6626	22.62	0.6210	21.18	0.6053
SUN360 Dataset						
360-SS	23.02	0.5943	16.64	0.5728	15.14	0.5875
Real-ESRGAN	23.07	0.5922	21.23	0.5926	19.34	0.5793
SwinIR	23.66	0.6439	22.04	0.6296	20.48	0.6176
HAT	23.67	0.6444	22.08	0.6300	20.55	0.6177
PromptIR	23.69	0.6464	22.07	0.6311	20.55	0.6184
OSRT	23.82	0.6484	22.13	0.6301	20.58	0.6180
BPOSr	23.75	0.6459	21.91	0.6282	20.23	0.6171
AdaIR	23.85	0.6502	22.14	0.6306	20.59	0.6185
D ² R ² OSR	24.00	0.6528	22.16	0.6328	20.69	0.6184

AdaIR [10]. Following common practice in ODI evaluation, we report PSNR and SSIM together with ERP-specific metrics WS-PSNR [30] and WS-SSIM [44], all computed on the luminance (Y) channel. Since the parts of compared approaches are not originally designed for real-world ODI-SR, we retrain them under a unified training paradigm using our simulated real-world ODI-SR dataset. All methods share identical training settings, ensuring that performance differences arise solely from architectural design differences.

Quantitative and Qualitative Comparison As displayed in Tab. 1, D²R²OSR achieves the best quantitative performance across almost all benchmarks and

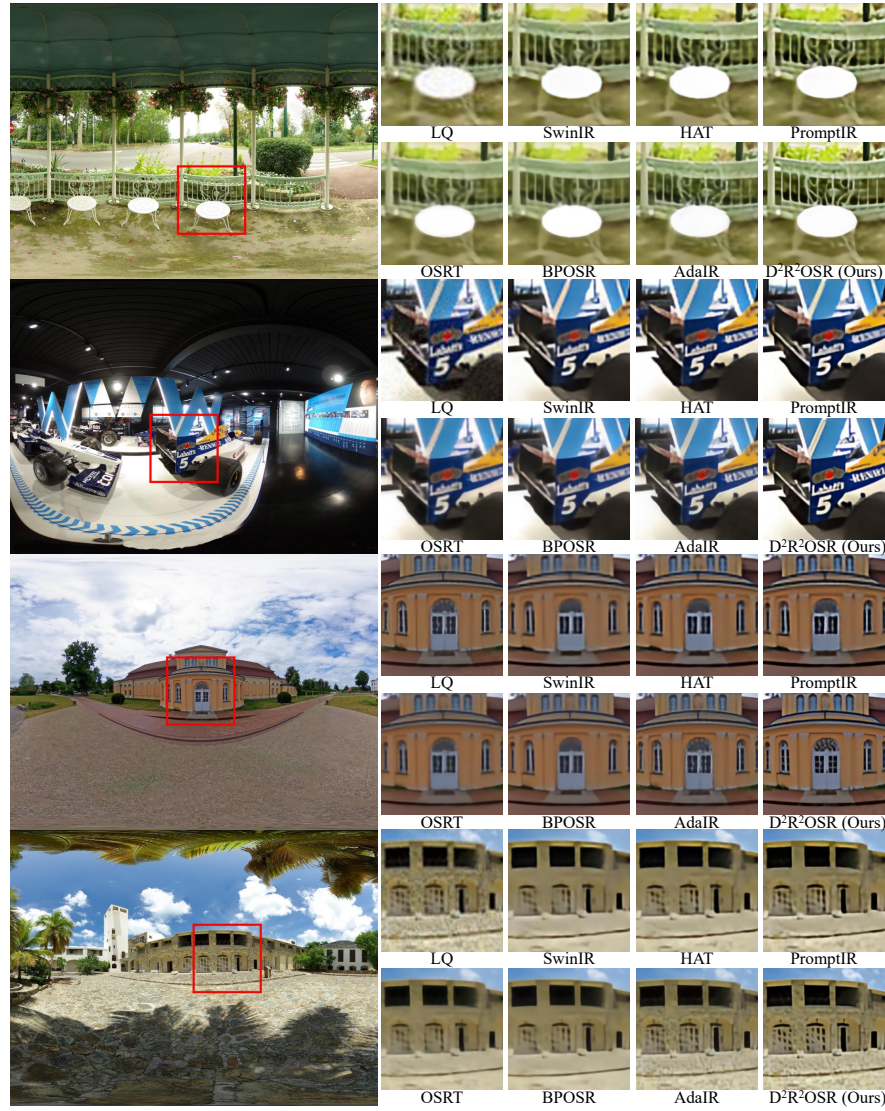


Fig. 6: Qualitative $\times 4$ real-world ODI-SR comparison on ERP ODIs.

scaling factors, validating the effectiveness of the proposed framework for real-world ODI-SR. In particular, D^2R^2OSR consistently outperforms the second-best approach AdaIR in WS-PSNR, with the largest gain of 0.28, dB achieved on Flickr360 under the $\times 8$ setting. Even under the extremely challenging $\times 16$ setting, the proposed method maintains clear advantages on all three datasets, demonstrating strong robustness under severe degradations and large upscaling factors. More quantitative comparisons of perceptual quality metrics with other methods are provided in the Appendix.

Table 2: Computational efficiency comparison on the SUN360 dataset.

Metric	360-SS	Real-ESRGAN	SwinIR	HAT	PromptIR	OSRT	BPOSR	AdaIR	D ² R ² OSR (Ours)
Params. (M)↓	0.19	16.70	11.67	20.33	48.85	6.02	<u>2.07</u>	28.90	3.50
FLOPs (T)↓	0.73	2.35	1.70	2.84	0.67	0.79	0.32	0.96	<u>0.55</u>
Runtime (s)↓	0.21	0.69	0.75	0.51	0.47	1.33	0.42	0.38	<u>0.34</u>

Table 3: Ablation studies for the parallel branch design and Perspective Projection Representation (PPR) on the real-world Flickr360 dataset.

Method	Branch		Representation			WS-PSNR↑	PSNR↑
	Single	Dual	ERP	Pers	PPR	WS-SSIM↑	SSIM↑
Baseline	✓		✓			25.20 / 0.6858	24.63 / 0.6644
Perspective (Pers)	✓			✓		25.09 / 0.6820	24.51 / 0.6602
Dual-Baseline		✓	✓			25.35 / 0.6865	24.77 / 0.6650
Baseline + Pers		✓	✓	✓		25.47 / 0.6882	24.90 / 0.6669
Baseline + PPR		✓	✓		✓	25.50 / 0.6892	24.94 / 0.6682

Fig. 6 presents qualitative comparisons with representative ODI-SR and image restoration methods. Upon inspection of zoomed-in regions, only D²R²OSR effectively suppresses severe artifacts while faithfully recovering fine details. For example, the chair structures exhibit more accurate textures and color consistency, while the picture frame in the second scene is reconstructed without introducing noticeable artifacts or noise. These observations corroborate the superior capability of D²R²OSR in restoring fine structures and preserving visual fidelity.

4.3 Computational Efficiency

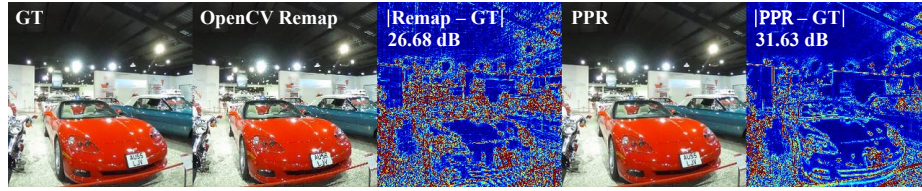
Table 2 compares the computational efficiency of different methods. Runtime is measured on a single NVIDIA A6000 GPU with a batch size of 1 and an input resolution of 256×512 under the $\times 4$ setting. The proposed D²R²OSR achieves competitive efficiency with only 3.50M parameters, 0.55T FLOPs, and an inference time of 0.34s per image. Compared with most Transformer-based methods, D²R²OSR provides a favorable balance between model complexity and inference speed. In particular, it is approximately $4\times$ faster than OSRT (1.33, s/image) while achieving superior restoration performance.

4.4 Ablation Studies

In this section, we perform ablation studies to validate the effectiveness of each developed component. Unless otherwise specified, all experiments are performed under the full D²R²OSR configuration for $\times 4$ real-world ODI-SR.

Table 4: Round-trip reconstruction quality across different latitude regions.

Viewpoint	Latitude	PSNR $\uparrow$ / SSIM $\uparrow$	
		OpenCV Remap	PPR (Ours)
Equator	0°	29.28 / 0.9056	34.89 / 0.9703
Mid-lat 30	30°	30.94 / 0.9233	36.27 / 0.9765
Mid-lat 60	60°	36.21 / 0.9404	40.07 / 0.9723
Overall	–	33.63 / 0.9295	38.17 / 0.9728

**Fig. 7:** Round-trip reconstruction visualization.

Effects of different dual-branch architectures and projection representations. Tab. 3 summarizes the impact of different branch designs and projection representations. Directly projecting ERP ODIs into the perspective domain (second row) does not improve performance. Simply introducing an additional branch by duplicating the network with reduced feature dimensions (third row) yields a 0.15 dB WS-PSNR improvement. Compared with feature expansion within the same ERP distribution, incorporating a dedicated perspective branch (fourth row) provides complementary viewpoint information and further improves WS-PSNR by 0.27 dB. Replacing conventional perspective projection (OpenCV Remap) with the proposed PPR (fifth row) further increases WS-PSNR to 25.50 dB. These results validate the effectiveness of the proposed dual-branch design and the PPR representation for real-world ODI-SR.

Geometric fidelity of PPR. To further validate the proposed PPR, we evaluate the round-trip ERP $\rightarrow$ PPR $\rightarrow$ ERP reconstruction quality on 1,200 training images across different latitude bands. As reported in Tab. 4, PPR outperforms OpenCV Remap by an average of 4.54 dB PSNR, indicating substantially reduced projection-induced information loss across all latitudes. Qualitative comparisons in Fig. 7 further show cleaner residual maps (brighter indicates larger reconstruction error) and better preservation of fine structures, such as the plate text “AUSS”. These results demonstrate the superior geometric fidelity of PPR and explain its consistent gains achieved by the dual-branch framework.

Impact of individual architecture modules. We further evaluate the effectiveness of the proposed Degradation-Specific Module (DSM) and Projection Fusion Attention Module (PFAM) on the PPR-based dual-branch framework. As shown in Tab. 5, introducing DSM_{ERP} and DSM_{PPR} improves WS-PSNR by 0.08 dB, demonstrating the benefit of projection-specific degradation model-

Table 5: Ablation studies of DSM and PFAM on the real-world Flickr360 dataset.

Method	WS-PSNR↑ WS-SSIM↑	PSNR↑ SSIM↑
Baseline + PPR (Baseline _P)	25.50 / 0.6892	24.94 / 0.6682
Baseline _P + DSM _{ERP}	25.56 / 0.6906	25.02 / 0.6702
Baseline _P + DSM _{PPR}	25.58 / 0.6909	25.03 / 0.6703
Baseline _P + DSM _{ERP&PPR} (Baseline _D)	25.58 / 0.6910	25.06 / 0.6712
Baseline _D + Attention-Level Injection (ALI)	25.57 / 0.6918	25.08 / 0.6716
Baseline _D + ALI + PFAM	25.64 / 0.6934	25.13 / 0.6726

ing. Further incorporating attention-level injection (ALI) improves WS-PSNR to 25.57 dB by enabling degradation-aware modulation within Transformer attention. Finally, PFAM brings an additional 0.07 dB gain through adaptive cross-branch fusion in the spatial-channel domain. These results demonstrate that projection-specific degradation modeling and adaptive cross-projection fusion complement each other, jointly enabling D²R²OSR to effectively handle the complex degradations encountered in real-world ODIs.

5 Conclusion

In this paper, we propose D²R²OSR, a Degradation-Disentangled Representation network for Real-world Omnidirectional image Super-Resolution in practical scenarios. First, we model a realistic two-stage Fisheye-to-ERP degradation process based on the ODI imaging pipeline. Motivated by two observations: (1) geometric projection degradations are entangled with random real-world distortions, and (2) users typically focus on limited viewpoints within the full 360° space, we develop a dual-branch architecture with Perspective Projection Representation (PPR) to explicitly separate degradations across different distributions while better aligning with human visual perception. Moreover, we integrate Degradation-Specific Modules (DSMs) into both branches to adaptively extract degradation priors. The enhanced features are further aggregated through Projection Fusion Attention Modules (PFAMs) at each stage to facilitate information interaction across projection domains. Extensive experiments on multiple datasets demonstrate that D²R²OSR achieves state-of-the-art real-world ODI-SR performance while maintaining practical efficiency with fast inference for real-world viewpoint enhancement applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62461160310, 62521007, 62431011), and the Fundamental Research Funds for the Central Universities (E2ET1104).

References

1. An, H., Zhang, X.: Perception-oriented omnidirectional image super-resolution based on transformer network. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 3583–3587. IEEE (2023)
2. An, H., Zhang, X., Zhao, S., Zhang, L.: FATO: Frequency Attention Transformer for Omnidirectional Image Super-Resolution. In: Proceedings of the 6th ACM International Conference on Multimedia in Asia. pp. 1–7 (2024)
3. An, H., Zhang, X., Zhao, S., Zhang, L., Xiong, R.: Spatio-temporal distortion aware omnidirectional video super-resolution. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 2309–2317 (2026)
4. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3086–3095 (2019)
5. Cao, M., Mou, C., Yu, F., Wang, X., Zheng, Y., Zhang, J., Dong, C., Li, G., Shan, Y., Timofte, R., et al.: NTIRE 2023 challenge on 360° omnidirectional image and video super-resolution: Datasets, methods and results. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1731–1745 (2023)
6. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Investigating tradeoffs in real-world video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5962–5971 (2022)
7. Chen, H., Li, W., Gu, J., Ren, J., Sun, H., Zou, X., Zhang, Z., Yan, Y., Zhu, L.: Low-Res Leads the Way: Improving generalization for super-resolution by self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25857–25867 (2024)
8. 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. pp. 22367–22377 (2023)
9. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8628–8638 (2021)
10. Cui, Y., Zamir, S.W., Khan, S., Knoll, A., Shah, M., Khan, F.S.: AdaIR: Adaptive all-in-one image restoration via frequency mining and modulation. In: 13th International Conference on Learning Representations, ICLR 2025. pp. 57335–57356. International Conference on Learning Representations, ICLR (2025)
11. Dasari, M., Bhattacharya, A., Vargas, S., Sahu, P., Balasubramanian, A., Das, S.R.: Streaming 360-degree videos using super-resolution. In: IEEE INFOCOM 2020-IEEE Conference on Computer Communications. pp. 1977–1986. IEEE (2020)
12. 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. pp. 9189–9198 (2021)
13. Deng, X., Wang, H., Xu, M., Li, L., Wang, Z.: Omnidirectional image super-resolution via latitude adaptive network. *IEEE Transactions on Multimedia* **25**, 4108–4120 (2022)
14. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part IV 13. pp. 184–199. Springer (2014)

15. Fan, Z., Yang, W., Guo, Z., Liu, J.: Towards real-world super resolution with adaptive self-similarity mining. *IEEE Transactions on Image Processing* (2024)
16. Ji, X., Xu, C., Zhong, L., Yang, S., Xiao, H., Muntean, G.M.: Dual enhancement in odi super-resolution: adapting convolution and upsampling to projection distortion. In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. pp. 902–910 (2024)
17. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4681–4690 (2017)
18. Lee, J., Choi, K.P., Jin, K.H.: Learning local implicit fourier representation for image warping. In: *European Conference on Computer Vision*. pp. 182–200. Springer (2022)
19. Li, R., Sheng, X., Li, W., Zhang, J.: OmniSSR: Zero-shot omnidirectional image super-resolution using stable diffusion model. In: *European Conference on Computer Vision*. pp. 198–216. Springer (2024)
20. Li, W., Zhao, S., Chen, B., Cheng, X., Li, J., Zhang, L., Zhang, J.: ResVR: Joint rescaling and viewport rendering of omnidirectional images. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. pp. 78–87 (2024)
21. Liang, J., Zeng, H., Zhang, L.: Efficient and degradation-adaptive network for real-world image super-resolution. In: *European Conference on Computer Vision*. pp. 574–591. Springer (2022)
22. 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*. pp. 1833–1844 (2021)
23. 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*. pp. 136–144 (2017)
24. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: DiffBIR: Toward blind image restoration with generative diffusion prior. In: *European conference on computer vision*. pp. 430–448. Springer (2024)
25. Ozcinar, C., Rana, A., Smolic, A.: Super-resolution of omnidirectional images using adversarial learning. In: *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. pp. 1–6. IEEE (2019)
26. Potlapalli, V., Zamir, S.W., Khan, S.H., Shahbaz Khan, F.: PromptIR: Prompting for all-in-one image restoration. *Advances in Neural Information Processing Systems* **36**, 71275–71293 (2023)
27. Shen, X., Wang, Y., Zheng, S., Xiao, K., Yang, W., Wang, X.: Fast omni-directional image super-resolution: Adapting the implicit image function with pixel and semantic-wise spherical geometric priors. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 39, pp. 6833–6841 (2025)
28. Son, S., Kim, J., Lai, W.S., Yang, M.H., Lee, K.M.: Toward real-world super-resolution via adaptive downsampling models. *IEEE transactions on pattern analysis and machine intelligence* **44**(11), 8657–8670 (2021)
29. Son, S., Lee, K.M.: SRWarp: Generalized image super-resolution under arbitrary transformation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 7782–7791 (2021)
30. Sun, Y., Lu, A., Yu, L.: Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE signal processing letters* **24**(9), 1408–1412 (2017)

31. Tian, Y., Chen, H., Xu, C., Wang, Y.: Image processing GNN: Breaking rigidity in super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24117 (2024)
32. Wang, J., Cui, Y., Li, Y., Ren, W., Cao, X.: Omnidirectional image super-resolution via bi-projection fusion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5454–5462 (2024)
33. Wang, L., Wang, Y., Dong, X., Xu, Q., Yang, J., An, W., Guo, Y.: Unsupervised degradation representation learning for blind super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10581–10590 (2021)
34. 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. pp. 1905–1914 (2021)
35. Wei, Y., Gu, S., Li, Y., Timofte, R., Jin, L., Song, H.: Unsupervised real-world image super resolution via domain-distance aware training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13385–13394 (2021)
36. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: SeeSR: Towards semantics-aware real-world image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25456–25467 (2024)
37. Xiao, J., Ehinger, K.A., Oliva, A., Torralba, A.: Recognizing scene viewpoint using panoramic place representation. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2695–2702. IEEE (2012)
38. Yang, C., Dong, R., Xiao, J., Zhang, C., Lam, K.M., Zhou, F., Qiu, G.: Geometric distortion guided transformer for omnidirectional image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology* (2025)
39. Yang, Z., Xia, J., Li, S., Huang, X., Zhang, S., Liu, Z., Fu, Y., Liu, Y.: A dynamic kernel prior model for unsupervised blind image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26046–26056 (2024)
40. Yu, F., Gu, J., Li, Z., Hu, J., Kong, X., Wang, X., He, J., Qiao, Y., Dong, C.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25669–25680 (2024)
41. 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. pp. 13283–13292 (2023)
42. 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. pp. 4791–4800 (2021)
43. 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). pp. 286–301 (2018)
44. Zhou, Y., Yu, M., Ma, H., Shao, H., Jiang, G.: Weighted-to-spherically-uniform ssim objective quality evaluation for panoramic video. In: 2018 14th IEEE International Conference on Signal Processing (ICSP). pp. 54–57. IEEE (2018)