DISENTANGLED PANORAMA SUPER-RESOLUTION WITH LATITUDE-WISE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Panorama super-resolution suffers from entangled content and degradation features, as well as latitude-dependent distortion caused by equirectangular projection. Existing end-to-end methods jointly learn mixed features, limiting their generalization and amplifying artifacts in high-latitude regions. We propose a decoupled dual-branch framework that explicitly separates sample-specific content features from universal degradation patterns. Two dedicated encoders are trained with a cross-sample feature fusion strategy, forcing the degradation branch to capture only content-agnostic information while the content branch focuses on semantic and geometric structures. To handle non-uniform stretching, we design a latitude-wise block with adaptive convolution whose dilation rate is dynamically modulated by the spherical polar angle. Edge-guided spatial attention and a composite loss with feature consistency further enhance detail recovery and training stability. Our method achieves state-of-the-art performance on standard benchmarks.

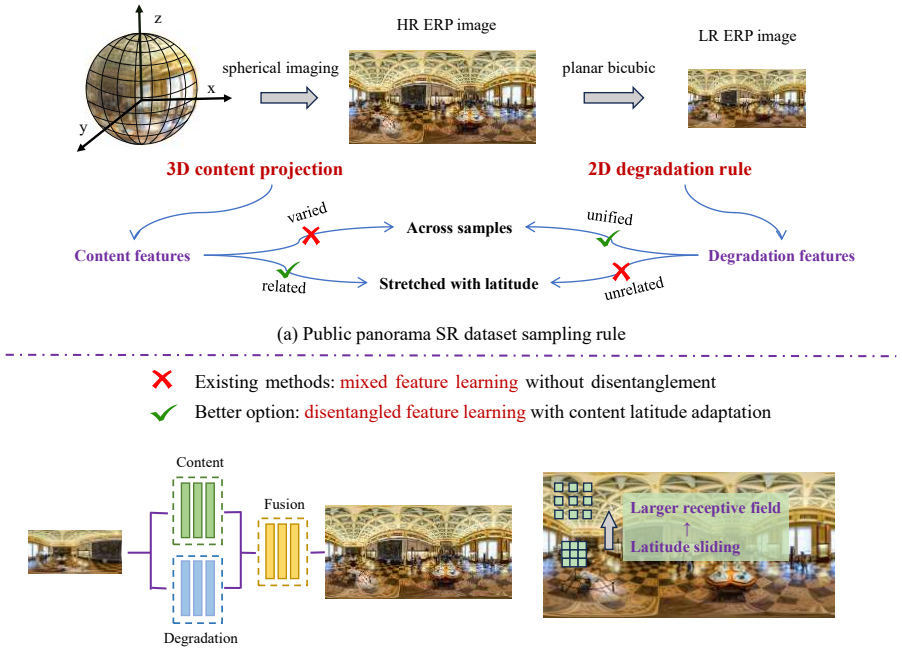

Figure 1: (a) Public panorama SR dataset sampling rule; (b) Decoupled network and latitude-related receptive field design.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Panorama super-resolution (SR) aims to recover sharp, high-quality images from low-resolution (LR) data. High-quality panoramic imagery is increasingly essential for real-world applications such as self-driving cars (Yang et al., 2021; Zhang et al., 2024), virtual reality (Meng et al., 2025; Cai et al., 2024; Sun et al., 2023; Cao et al., 2023; Yoon et al., 2022; Deng et al., 2021; Fang et al., 2022), and robotics (Wang et al., 2018; Liu et al., 2013; Meng et al., 2025). Panoramic cameras capture dense 360° scenes at high resolution to preserve visual fidelity across the entire viewing sphere. However, such high-resolution (HR) panoramic images create heavy burdens on data storage and transmission (Wang et al., 2025; Zhang et al., 2021; Deng et al., 2021; Wang et al., 2024b; An et al., 2024). To alleviate these burdens, a common strategy is to store and transmit downsampled low-resolution (LR) versions and later recover the original resolution via SR. In such a pipeline, downsampling is naturally implemented through bicubic interpolation, and the panoramic SR community follows this practice to construct LR-HR training pairs by downsampling HR ground-truth images (Deng et al., 2021; Wang et al., 2026). As shown in Figure 1(a), this downsampling pipeline serves as the standard construction rule for public panoramic SR datasets.

Specifically, these datasets (Deng et al., 2021; Wang et al., 2026) follow a two-step sampling rule. First, raw 3D scene content is projected from a spherical model onto a 2D equirectangular projection (ERP) image. This spherical-to-planar projection causes non-uniform stretching, where image content becomes progressively stretched toward higher latitudes. Second, the HR ERP images are degraded to LR versions (typically $\times 8$ or $\times 16$ downsampling) using 2D bicubic interpolation, and the resulting LR images are stored in standard compressed formats (e.g., PNG) for disk-based access, as is common in public datasets. This pipeline inherently produces two distinct types of features, which are conflated in existing datasets. Content features are directly shaped by the 3D projection, varying significantly across samples and exhibiting clear latitude-dependent stretching. In contrast, degradation features originate from the 2D downsampling operation itself, a spatially uniform process whose parameters are independent of the 3D projection geometry. While its local artifacts may vary with underlying image structure, the degradation rule remains consistent across all samples and fundamentally separable from the content distortions caused by spherical projection. This distinction between sample-specific content and sample-agnostic degradation is often overlooked in current design. Consequently, networks trained end-to-end on such data inevitably entangle these two factors, impairing their ability to generalize across diverse scenes and exacerbating reconstruction artifacts in high-latitude regions.

As highlighted in Figure 1(b), existing panorama SR methods adopt an end-to-end training paradigm that learns mixed content and degradation features without recognizing or exploiting their inherent separability (Yu et al., 2023; Wang et al., 2024a; Cai et al., 2024; Yang et al., 2025; Shen et al., 2025; Wang et al., 2026). By conflating the two, this holistic approach prevents the network from learning invariance to degradation variations and limits its capacity to adapt to latitude-dependent content distortions. A more principled approach is to explicitly disentangle these two feature types while equipping the content branch with latitude adaptation. To achieve this, we design a decoupled dual-branch network composed of a content encoder that isolates sample-specific content features and a degradation encoder that captures universal, sample-agnostic degradation patterns. Crucially, we enforce this disentanglement through a cross-sample feature fusion strategy during training, ensuring that the degradation branch learns only content-independent information. Complementing this decoupling, we further design a latitude-adaptive receptive field mechanism. As shown in Figure 1(b), this mechanism allows the network's receptive field to expand dynamically as latitude increases, ensuring robust feature extraction that effectively compensates for non-uniform ERP stretching caused by 3D projection.

The contributions of this paper are as follows.

- A novel decoupled dual-branch pipeline that explicitly separates sample-specific content features from universal degradation features, enforced through cross-sample feature fusion to address the feature entanglement inherent in existing methods.
- A latitude-wise block with adaptive convolution that dynamically adjusts dilation rates according to spherical geometry, compensating for non-uniform ERP stretching and achieving latitude-adaptive receptive fields.

- Tailored feature enhancement and training strategies that complement the decoupled architecture, including layer normalization for stabilizing ERP feature learning, edge-guided spatial attention for detail recovery, and a composite loss with feature consistency, together achieving state-of-the-art performance.

2 Related Work

Single Image Super-Resolution (SISR). Benefiting from deep learning, single-image super-resolution has achieved remarkable progress and outperforms traditional interpolation and sparse coding methods. For CNN-based methods, SRCNN (Dong et al., 2016) pioneered the end-to-end SISR paradigm, but its shallow structure limits contextual modeling. VDSR (Kim et al., 2016) improves reconstruction performance with deep residual learning. EDSR (Lim et al., 2017) further enhances performance by stacking residual blocks and removing batch normalization. RCAN (Zhang et al., 2018) adopts channel attention to adaptively weight feature channels and refine texture recovery. For GAN-based methods, SRGAN (Ledig et al., 2017) first introduces adversarial loss to replace MSE and uses VGG perceptual loss to improve visual quality. Inspired by SRGAN, Real-ESRGAN (Wang et al., 2021) proposes flexible techniques for real-world degradation. For ViT-based methods, IPT (Chen et al., 2021) strengthens unified representation via large-scale pre-training and multi-task learning; SwinIR (Liang et al., 2021) models long-range dependencies with sliding windows and achieves efficient high-performance reconstruction. For diffusion-based methods, SR3 (Sahak et al., 2023) models complex distributions via conditional diffusion with strong robustness; ResShift (Yue et al., 2023) realizes efficient reconstruction through residual shifting, balancing accuracy and computational efficiency. These advances, particularly in residual learning, attention mechanisms, and efficient reconstruction, serve as important technical foundations for developing panoramic SR methods.

Omnidirectional Image Super-Resolution (ODISR). Panoramic super-resolution methods focus on correcting ERP projection distortion and modeling spherical geometry to reduce information loss during projection transformation. Early work such as LAU-Net (Deng et al., 2021) divides ERP images into latitude-wise blocks to handle pixel stretching, but its non-overlapping partitions disrupt global coherence and fragment spatial information across block boundaries. OSRT (Yu et al., 2023) introduces a distortion-aware Transformer that adapts attention to non-uniform ERP sampling, yet its end-to-end design jointly encodes content and degradation without explicit separation. SphereSR (Yoon et al., 2022) lifts ERP images to a continuous spherical representation via SLIIF, achieving projection-consistent super-resolution at the cost of expensive spherical convolutions that limit inference speed. FATO (An et al., 2024) strengthens panoramic feature extraction with attention mechanisms but focuses on local feature enhancement without addressing the underlying entanglement between projection distortion and degradation patterns. FAOR (Shen et al., 2025) further optimizes feature modeling and achieves improved performance on several scenes, but still operates within a joint learning paradigm that couples content and degradation features, limiting generalization under varying geometric conditions. More recent methods, including BPOSR (Wang et al., 2024a), LAPR (Cai et al., 2024), GDGT-OSR (Yang et al., 2025), and EAM (Wang et al., 2026), advance the field with stronger backbones or more efficient designs. However, none of these approaches explicitly recognizes that the synthetic data pipeline produces two inherently distinct feature types: sample-specific content features shaped by 3D projection and sample-agnostic degradation features governed by the uniform 2D downsampling rule. This conflation leads to suboptimal high-latitude reconstruction and limits generalization.

In contrast, our method explicitly decouples these two feature spaces through dual encoders with cross-sample feature fusion, and further compensates for latitude-dependent stretching via adaptive convolution modulated by spherical geometry. This design directly targets the entangled representation that constrains existing ODISR methods.

3 Method

In this section, we introduce the overall architecture of our proposed network. Our model employs a dual-branch decoupled structure to separately extract degradation features and content features. To alleviate distortion and non-uniform sampling in equirectangular panoramic images, we design a

latitude-wise block with latitude-aware dilated convolutions. We further integrate channel attention and edge-guided spatial attention to enhance feature representation, and employ a composite loss for optimization.

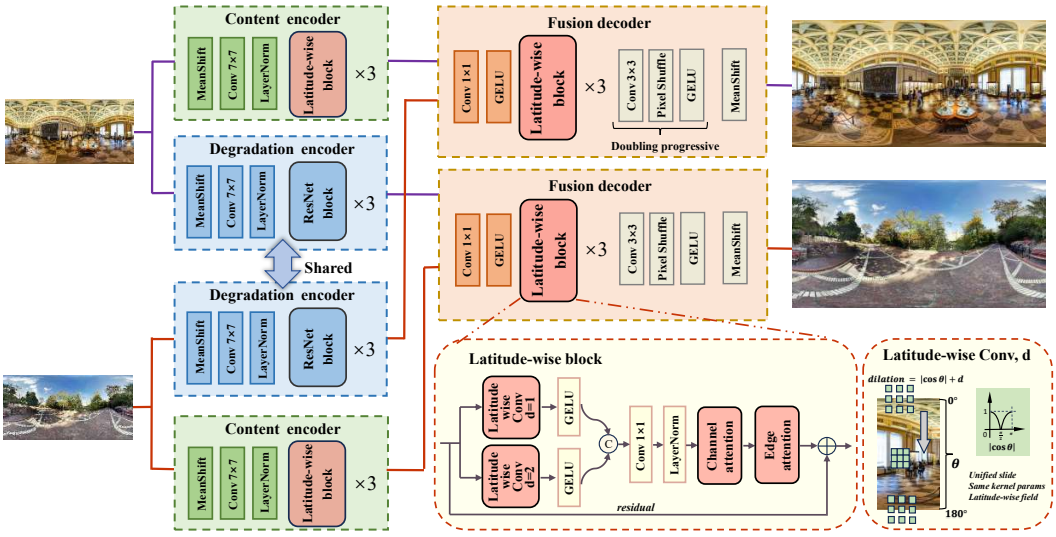

Figure 2: Overall architecture of the proposed decoupled panorama super-resolution framework.

3.1 Decoupled Pipeline

Existing panoramic SR methods predominantly adopt an end-to-end paradigm that jointly models content features (which vary with latitude and scene semantics) and degradation features (which are content-independent and globally shared), failing to exploit the inherent separability between the two.

As shown in Figure 2, our decoupled framework consists of a degradation encoder and a content encoder that separately extract the two types of features and map them to 64-dimensional channel-wise features. Given two different low-resolution (LR) panoramic inputs $I_{\text{LR},1}$ and $I_{\text{LR},2}$, the feature extraction process is formulated as:

$$\begin{cases} F_{C,1} = \mathcal{E}_C(I_{\text{LR},1}),\ F_{D,1} = \mathcal{E}_D(I_{\text{LR},1}) \\ F_{C,2} = \mathcal{E}_C(I_{\text{LR},2}),\ F_{D,2} = \mathcal{E}_D(I_{\text{LR},2}) \end{cases} \tag{1}$$

where $\mathcal{E}_C(\cdot)$ denotes the content encoder that extracts content features F_C correlated with image semantics and latitude, and $\mathcal{E}_D(\cdot)$ denotes the degradation encoder that extracts global, content-independent degradation features F_D.

Although the local intensity of downsampling artifacts varies with image gradients, this variation is purely a function of low-level structures and remains semantic-agnostic. A perfectly constant response would be undesirable, as it would provide no useful information for reconstruction. The goal of the degradation encoder is to capture these spatially varying but semantic-agnostic artifact patterns, not to trivially output a uniform feature. Degradation features extracted from one sample therefore remain valid for reconstructing another. To enforce this disentanglement, we employ a cross feature fusion strategy during training. The degradation features from one image are paired and fused with the content features from the other image, then fed into the fusion decoder:

$$\begin{aligned} I_{\text{SR},1} &= \mathcal{D}\left(\text{Concat}(F_{C,1}, F_{D,2})\right), \\ I_{\text{SR},2} &= \mathcal{D}\left(\text{Concat}(F_{C,2}, F_{D,1})\right). \end{aligned} \tag{2}$$

where $\mathcal{D}(\cdot)$ denotes the fusion decoder and $\text{Concat}(\cdot)$ denotes channel-wise concatenation. This cross-sample fusion forces the degradation encoder to discard content-dependent information, as any content leakage would impair the reconstruction of the paired sample. In practice, we simply

randomly pair samples within each training batch; we find this straightforward strategy sufficient for the degradation encoder to learn content-invariant features, without requiring any specialized pairing mechanism.

The equirectangular projection (ERP) causes severe horizontal stretching in polar regions, which degrades content feature extraction when using fixed convolutional kernels. To address this, we adopt latitude-wise blocks in the content encoder for adaptive latitude-aware feature extraction. In contrast, degradation patterns are content-agnostic, so polar pixel stretching does not interfere with their extraction. We therefore use standard residual blocks with shared weights in the degradation encoder for deep degradation modeling.

After the decoupled extraction, content features and degradation features from the two input images are cross-paired, concatenated along the channel dimension, and fed into the fusion decoder. The decoder consists of three latitude-wise blocks with progressive upsampling, finally outputting the super-resolved image.

3.2 LATITUDE-WISE BLOCK

ERP projection suffers from two critical issues. Pixels in polar regions are severely stretched horizontally, causing the physical pixel scale to vary drastically with latitude, which violates the uniform pixel scale assumption of standard planar images. Moreover, applying fixed convolutional kernels globally leads to highly inconsistent physical receptive fields across different latitudes, resulting in unstable feature extraction and loss of high-frequency details. To address these latitude-related distortions while explicitly exploiting multi-scale context for panoramic super-resolution, we propose a latitude-wise residual block, whose detailed structure is illustrated in the bottom-right corner of Figure 2.

Given an input feature map F_{in}, the block first feeds it into two parallel branches with latitude-wise convolution (Sec. 3.3), where the scales are set to $s = 1$ and $s = 2$ respectively:

$$\begin{cases} F_{\text{branch1}} = \text{LatConv}_{s=1}(F_{\text{in}}), \\ F_{\text{branch2}} = \text{LatConv}_{s=2}(F_{\text{in}}) \end{cases} \tag{3}$$

The two branches employ different dilation rates, enabling them to capture complementary context at varying scales. This multi-scale design follows the common practice in image restoration (Wang et al., 2026), where parallel branches with different receptive fields are used to extract both fine and coarse spatial features. Each branch is followed by a GELU activation, and the resulting features are concatenated along the channel dimension for fusion. The fused features are then processed by a 1×1 convolution, followed by layer normalization (LN). While most panoramic SR methods employ batch normalization (BN) (Ozcinar et al., 2019; Coors et al., 2018; Fakour-Sevom et al., 2018; Wang et al., 2026), BN relies on batch-level statistics and can introduce instability given the large cross-sample feature variations in ERP data. We therefore adopt LN, which normalizes per-sample and better preserves image-specific characteristics crucial for SR. The normalized features are further enhanced by channel attention and edge-guided spatial attention, before being added back to the original input via a residual connection to preserve low-frequency base features.

3.3 LATITUDE-WISE CONVOLUTION

To address the receptive field stretching that intensifies with latitude in panoramic images, we design a latitude-adaptive convolutional receptive field by dynamically adjusting the kernel dilation rate. ERP images exhibit no stretching near the equator but extreme horizontal stretching at the poles. Our design strictly adheres to the spherical mathematical model to align the receptive field with the actual spherical geometry. We leverage the polar angle θ (zenith angle), which has a one-to-one correspondence with the vertical coordinate y of the ERP image, as follows:

$$\theta(y) = \frac{y}{H} \times \pi \tag{4}$$

where H is the image height and $y \in [0, H-1]$. $\theta = 0^\circ$ corresponds to the North Pole, $\theta = 90^\circ$ to the equator, and $\theta = 180^\circ$ to the South Pole.

In an ERP projection, the horizontal pixel stretching ratio at latitude θ is proportional to $|\cos\theta|$. To maintain a consistent physical receptive field on the sphere, the receptive field in the projected image should expand accordingly. We therefore adopt $|\cos\theta|$ as the latitude weight, which provides a smooth and physically grounded modulation. Its curve, illustrated in the bottom-right corner of Figure 2, naturally transitions from 1 at the poles to 0 at the equator and back to 1 at the opposite pole. Based on this weight, we define the dynamic dilation rate as:

$$\text{dilation} = |\cos\theta| + s \tag{5}$$

where s is the input scale parameter. Taking $s = 1$ as an example, at the equator where $\theta = 90°$, $|\cos\theta| = 0$ and the convolution dilation rate is 1. As the convolutional kernel slides toward either pole, $|\cos\theta|$ gradually increases from 0 to 1, and the dilation rate dynamically expands from 1 to 2. Since the resulting dilation is generally non-integer, we implement it via weighted interpolation of two convolutions with integer dilation rates (**details are provided in Appendix A.7, including a discretized integer-dilation variant**). This ensures that high-latitude regions with severe pixel stretching are equipped with a larger receptive field through increased kernel spacing to compensate for the sparse spherical pixel mapping. The equator maintains an optimal compact receptive field for precise feature extraction.

3.4 Dual attentions

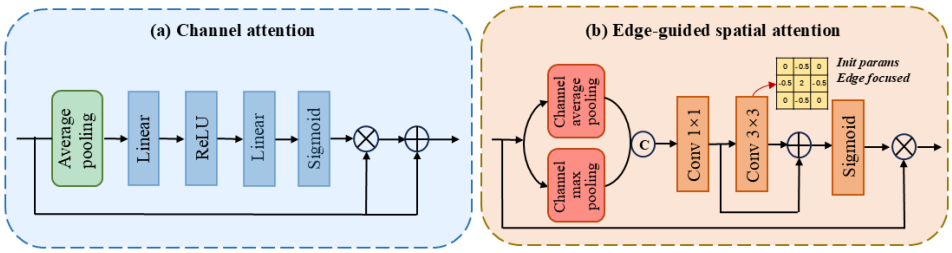

Figure 3: (a) Channel attention; (b) Edge-guided spatial attention.

For panoramic super-resolution, attention is critical for edge restoration. ERP projection and unified 2D downsampling cause edge blurring and distortion. We introduce channel attention and edge-guided spatial attention to enhance discriminative channels and prioritize edge features for restoration.

Channel attention. As illustrated in Figure 3(a), the channel attention adaptively recalibrates the importance of features across different channels. It first applies global average pooling to the input feature to aggregate spatial information into a global descriptor. Two fully connected layers with ReLU activation are then used to model inter-channel dependencies and generate attention weights. The learned weights are multiplied with the input features to emphasize informative channels and suppress redundant ones. A residual connection is added to ensure training stability.

Edge-guided spatial attention. Conventional spatial attention relies on global statistical features and fails to explicitly focus on edge regions distorted by stretching. We propose an edge-guided spatial attention (Figure 3(b)), which leverages explicit edge priors to guide the network toward high-frequency details and improve structural clarity. Specifically, we first perform channel-wise average and max pooling on the input features, concatenate the results, and fuse them via a 1×1 convolution to generate a preliminary spatial attention map. To inject explicit edge prior, we apply a trainable convolutional layer that is initialized with a cross-shaped edge detection kernel. This initialization provides a stable starting point for extracting edge features, while the trainable parameters allow the network to adaptively refine edge patterns and suppress background noise. The resulting edge-enhanced feature is fused with the spatial attention map via element-wise addition. Finally, a Sigmoid function maps the fused features to attention weights, which are multiplied element-wise with the original input to highlight edge regions and restore high-frequency details.

3.5 LOSS FUNCTION

To match our decoupled architecture, we propose a three-component joint loss that imposes constraints from reconstruction accuracy, feature consistency, and structural similarity. Together, these terms ensure the network stably learns disentangled representations.

The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{L1} + 0.1 \cdot \mathcal{L}_{\text{feat}} + 0.01 \cdot \mathcal{L}_{\text{SSIM}} \tag{6}$$

- **Reconstruction loss.** We adopt the L_1 loss to enforce pixel-level reconstruction accuracy between the super-resolved output and the ground-truth HR panorama.
- **Feature consistency loss** (weight = 0.1). This term enforces degradation features to be content-invariant by penalizing the cosine dissimilarity between degradation features extracted from in-batch paired samples.
- **SSIM loss** (weight = 0.01). This term minimizes the structural dissimilarity between the SR output and the HR ground truth, helping to alleviate distortion caused by non-uniform ERP stretching and improve perceptual quality.

4 EXPERIMENTS

4.1 IMPLEMENTATION

We train our model on the ODI-SR dataset (Deng et al., 2021) with spherical augmentation following (Wang et al., 2026), resulting in 12,000 training samples. This expanded training set follows the protocol introduced by EAM (Wang et al., 2026), and yields stronger baselines than those reported in earlier works trained on the original dataset. For fair comparison, we therefore primarily benchmark against EAM, which shares the same training data scale. The network is implemented in PyTorch and trained on a single NVIDIA TITAN RTX (24GB memory). We use the Adam optimizer with an initial learning rate of 0.001 and a batch size of 4, and train for 170 epochs. We evaluate on the ODI-SR and SUN360 test sets (100 images each) using the standard panorama-specific metrics WS-PSNR and WS-SSIM (Deng et al., 2021), following the established evaluation protocol.

Table 1: Quantitative results on ODI-SR and SUN360 datasets.

Method	×16				×8			
	ODI-SR		SUN360		ODI-SR		SUN360	
	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM	WS-PSNR	WS-SSIM
Bicubic	17.12	0.4332	17.56	0.4638	19.64	0.5908	19.72	0.5403
EDSR (Lim et al., 2017) (CVPR'17)	22.24	0.6090	21.83	0.5974	23.97	0.6483	23.79	0.6472
RCAN (Zhang et al., 2018) (ECCV'18)	22.49	0.6176	21.86	0.5938	24.26	0.6554	23.88	0.6542
360-SS (Ozcinar et al., 2019) (MMSP'19)	19.65	0.5431	19.62	0.5308	21.65	0.6417	21.48	0.6352
LAU-Net (Deng et al., 2021) (CVPR'21)	22.52	0.6284	22.05	0.6058	24.36	0.6602	24.24	0.6708
SphereSR (Yoon et al., 2022) (CVPR'22)	22.51	0.6370	21.95	0.6342	24.37	0.6777	24.17	0.6820
OSRT (Yu et al., 2023) (CVPR'23)	22.69	0.6261	22.13	0.6388	24.53	0.6780	24.38	0.7072
BPOSR (Wang et al., 2024a) (AAAI'24)	22.72	0.6285	22.16	0.6433	24.61	0.6782	24.47	0.7084
GDGT-OSR (Yang et al., 2025) (TCSVT'25)	22.78	0.6087	22.60	0.6303	24.60	0.6687	25.00	0.7068
FAOR (Shen et al., 2025) (AAAI'25)	22.78	0.6008	22.53	0.6211	24.58	0.6581	24.77	0.6914
EAM (Wang et al., 2026) (CVPR'26)	23.86	0.6290	23.49	0.6415	25.69	0.6839	25.81	0.7102
Ours	**24.07**	**0.6370**	**23.74**	**0.6523**	**25.78**	**0.6874**	**25.95**	**0.7152**

Table 2: Computational efficiency comparison across methods.

Model	FLOPs	Params	Time
SwinIR (Liang et al., 2021)	900 G	11.5 M	0.982 s
360-SS (Ozcinar et al., 2019)	**15 G**	1.6 M	0.025 s
LAU-Net (Deng et al., 2021)	685 G	9.4 M	0.443 s
SphereSR (Yoon et al., 2022)	587 G	8.7 M	0.401 s
LAPR (Cai et al., 2024)	372 G	7.8 M	0.312 s
EAM (Wang et al., 2026)	38 G	2.0 M	**0.022 s**
Ours	106 G	**1.3 M**	0.030 s

Table 3: Ablation study on network components.

LConv	SA	CA	DFeat	WS-PSNR	WS-SSIM
×	✓	✓	✓	25.54	0.6817
✓	×	✓	✓	25.44	0.6766
✓	✓	×	✓	25.45	0.6792
✓	✓	✓	×	25.46	0.6772
✓	✓	✓	✓	**25.78**	**0.6874**

LConv: Latitude-wise convolution; SA: Spatial attention; CA: Channel attention; DFeat: Degradation Feature.

4.2 QUANTITATIVE RESULTS

Table 1 reports quantitative comparisons on ODI-SR and SUN360, and all data are sourced from the original paper. Our method achieves the best results. Compared to EAM (Wang et al., 2026), the strongest open-source baseline, our method matches or surpasses its performance across all settings and both datasets, despite using only 1.3M parameters against EAM's 2.0M. The advantage is particularly evident at ×16, where we outperform EAM by 0.21 dB on ODI-SR and 0.25 dB on SUN360. These results demonstrate that the proposed disentanglement strategy achieves strong reconstruction quality without relying on architectural complexity.

Table 2 compares FLOPs, parameters, and inference time on ODI-SR with $\times 8$ upscaling. Our method achieves the smallest parameter count across all compared methods (1.3 M), with competitive FLOPs (106 G) and fast inference (0.030 s per image). This compact profile validates the lightweight design of the proposed architecture.

4.3 ABLATION STUDY AND ANALYSIS

We conduct ablation experiments on ODI-SR under the $\times 8$ setting to verify the contribution of each core component. Table 3 reports the results. The loss ablation is provided in Appendix A.1.

Latitude-wise convolution (LConv). Replacing the latitude-adaptive convolution with standard residual blocks reduces WS-PSNR by 0.24 dB, confirming that the dynamically adjusted receptive field effectively suppresses polar distortion.

Spatial attention (SA). Removing the edge-guided spatial attention causes a notable drop in both WS-PSNR and WS-SSIM, as the network loses the ability to explicitly enhance edge regions distorted by ERP stretching. We also perform experiments on substituting our attention module, as detailed in Appendix A.2.

Channel attention (CA). Disabling channel attention weakens the adaptive recalibration of feature channels, leading to a clear decline in detail recovery.

Degradation feature extraction (DFeat). Removing the cross-sample fusion constraint keeps the dual-branch structure and parameter count unchanged, but eliminates the mechanism that forces one branch to learn degradation-specific features. This causes a clear performance drop, confirming the effectiveness of the degradation modeling strategy.

Latitude-wise region analysis. The latitude-adaptive convolution is motivated by spherical projection geometry: pixel stretching intensifies toward the poles, requiring larger receptive fields. To verify this alignment, we evaluate WS-PSNR separately in equatorial (central one-third) and high-latitude (upper and lower thirds combined) regions. Removing the latitude-wise convolution causes a 0.22 dB drop at the equator versus 0.33 dB at high latitudes. The performance degradation in high-latitude regions is markedly greater than that in equatorial regions, demonstrating that the latitude-adaptive dilation can effectively compensate for pixel stretching distortion caused by latitudinal variations, which is fully consistent with the original design intention.

Table 4: Latitude-wise ablation: WS-PSNR in equatorial and high-latitude regions.

Model	Equatorial Region	High-Latitude Region
w/o LConv	24.47	27.57
Full Model	24.69	27.90
Gain	+0.22	+0.33

4.4 QUALITATIVE RESULTS

Figure 4 provides visual comparisons on ODI-SR. Our method consistently produces sharper results that are closer to the HR ground truth. More $\times 8/\times 16$ examples are provided in Appendix A.8.

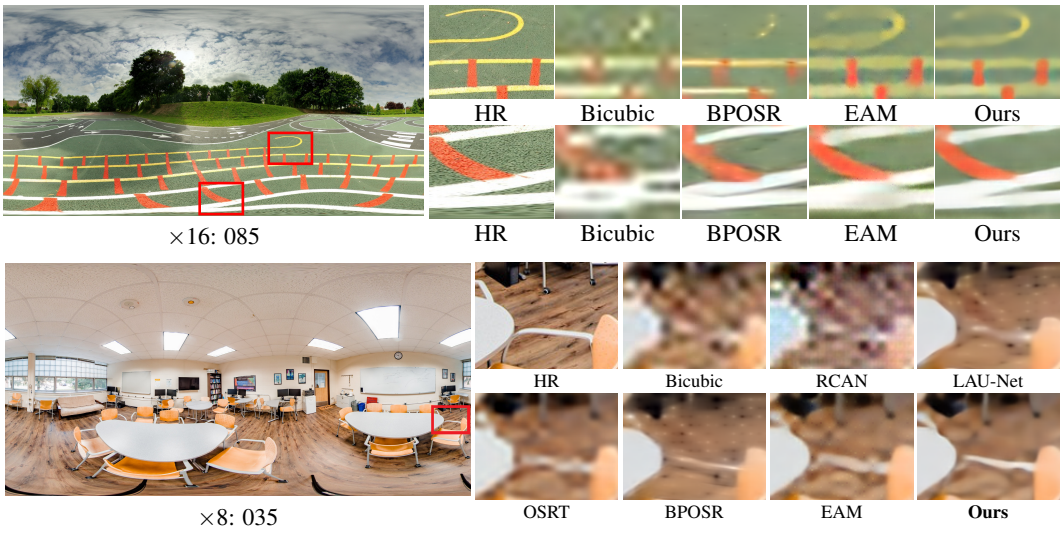

Figure 4: Visual comparisons of $\times 8$ and $\times 16$ super-resolution results on ODI-SR and SUN360.

4.5 Degradation features visualization and analysis

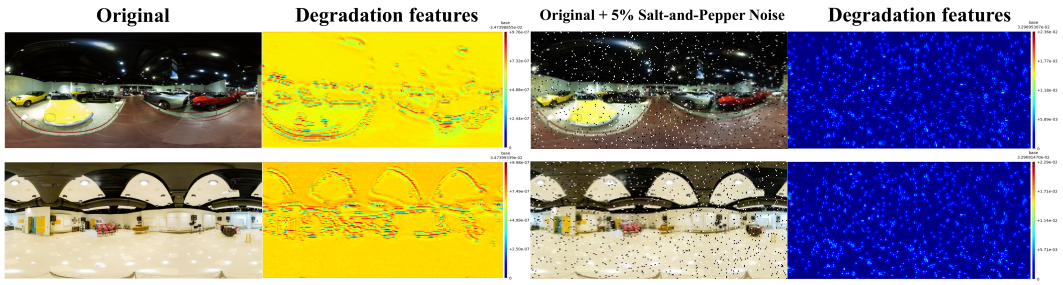

Figure 5: Visualization of degradation features under different degradation patterns.

Figure 5 visualizes channel-averaged degradation encoder features under two different degradations. For bicubic downsampling (second column), the features show a warm tone with ringing-like patterns around edges, consistent with bicubic-specific artifacts. With 5% salt-and-pepper noise (fourth column), the features shift to a cool tone, with sparse high responses exactly matching the noise distribution. The clear shift confirms the encoder captures degradation-specific signals rather than generic edges. Further analysis and decoupling evidence are given in Appendices A.4 and A.5.

5 Conclusion

This work presents a disentangled dual-branch framework for panorama super-resolution under equirectangular projection. By separating sample-dependent content features from universal degradation patterns via cross-sample feature fusion, together with latitude-wise adaptive convolution, our method mitigates latitude-dependent geometric distortion and yields competitive reconstruction performance. Compared with the strong baseline EAM (Wang et al., 2026), our model obtains clear gains especially under the challenging $\times 16$ setting with fewer parameters (1.3 M vs. 2.0 M). However, our dual-branch and latitude-adaptive design brings moderately higher FLOPs than the most lightweight EAM. Our current training mainly focuses on synthetic bicubic degradation following standard panorama SR dataset pipelines. For future work, we plan to investigate more diverse real-world degradation models, to further boost generalization toward practical in-the-wild panoramic imagery.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with code debugging, implementation of auxiliary scripts, and LaTeX-related troubleshooting. We have not used generative AI tools for proposing or refining the core research hypothesis, designing the proposed method, designing the experimental methodology, generating or modifying datasets, formulating mathematical claims, assisting in the writing of proofs, or interpreting experimental results, and qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for information searching, grammar correction, language polishing, improving readability and refining figure captions. We have reviewed all AI-assisted work. All AI-assisted code was inspected, modified when necessary, and verified by the authors through running the implementation and checking the experimental pipeline; all AI-assisted text was manually reviewed to ensure that it accurately reflected the authors' original method, experiments, results, and claims, and that no unsupported claims, fabricated citations, or incorrect technical statements were introduced. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our paper provides the details necessary to reproduce the proposed method and experiments. Section 3 describes the decoupled dual-branch architecture, latitude-wise convolution, channel and edge-guided spatial attention, feature fusion strategy, and training objectives. Section 4.1 specifies the datasets, data augmentation protocol, model implementation, training hardware, optimizer, learning rate, batch size, number of training epochs, evaluation datasets, and evaluation metrics. The quantitative comparisons, computational-efficiency measurements, and component and latitude-wise ablation studies are reported in Tables 1-4. Additional loss and attention ablations, analyses of the degradation encoder and feature disentanglement, the alternative disentanglement constraint, non-integer dilation implementation details, and further qualitative comparisons are provided in Appendix A.1-A.8. All experiments use the public ODI-SR and SUN360 benchmark datasets and follow the established evaluation protocol described in the paper.

REFERENCES

Hongyu An, Xinfeng Zhang, Shijie Zhao, and Li Zhang. Fato: Frequency attention transformer for omnidirectional image super-resolution. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, MMAsia '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400712739. doi: 10.1145/3696409.3700232. URL https://doi.org/10.1145/3696409.3700232.

Qing Cai, Mu Li, Dongwei Ren, Jun Lyu, Haiyong Zheng, Junyu Dong, and Yee-Hong Yang. Spherical pseudo-cylindrical representation for omnidirectional image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 873–881, 2024.

Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Gen Li, Ying Shan, Radu Timofte, Xiaopeng Sun, Weiqi Li, Zhenyu Zhang, Xuhan Sheng, Bin Chen, Haoyu Ma, Ming Cheng, Shijie Zhao, Wanwan Cui, Tianyu Xu, Chunyang Li, Long Bao, Heng Sun, Huaibo Huang, Xiaoqiang Zhou, Yuang Ai, Ran He, Renlong Wu, Yi Yang, Zhilu Zhang, Shuohao Zhang, Junyi Li, Yunjin Chen, Dongwei Ren, Wangmeng Zuo, Qian Wang, Hao-Hsiang Yang, Yi-Chung Chen, Zhi-Kai Huang, Wei-Ting Chen, Yuan-Chun Chiang, Hua-En Chang, I-Hsiang Chen, Chia-Hsuan Hsieh, Sy-Yen Kuo, Zebin Zhang, Jiaqi Zhang, Yuhui Wang, Shuhao Cui, Junshi Huang, Li Zhu, Shuman Tian, Wei Yu, and Bingchun Luo. Ntire 2023 challenge on 360deg omnidirectional image and video super-resolution: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1731–1745, June 2023.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12299–12310, 2021.

Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 525–541, 2018.

Xin Deng, Hao Wang, Mai Xu, Yichen Guo, Yuhang Song, and Li Yang. Lau-net: Latitude adaptive upscaling network for omnidirectional image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9189–9198, June 2021.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295–307, 2016. doi: 10.1109/TPAMI.2015.2439281.

Vida Fakour-Sevom, Esin Guldogan, and Joni-Kristian Kämäräinen. 360 panorama super-resolution using deep convolutional networks. In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 4: VISAPP*, pp. 159–165. INSTICC, SciTePress, 2018. ISBN 978-989-758-290-5. doi: 10.5220/0006618901590165.

Yuming Fang, Liping Huang, Jiebin Yan, Xuelin Liu, and Yang Liu. Perceptual quality assessment of omnidirectional images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 580–588, 2022. doi: 10.1609/aaai.v36i1.19937.

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1646–1654, 2016.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1833–1844, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

Ming Liu, Cédric Pradalier, and Roland Siegwart. Visual homing from scale with an uncalibrated omnidirectional camera. *IEEE Transactions on Robotics*, 29(6):1353–1365, 2013. doi: 10.1109/TRO.2013.2272251.

Ming Meng, Yonggui Zhu, Yufei Zhao, Zhaoxin Li, and Zhe Zhu. 3d indoor scene geometry estimation from a single omnidirectional image: A comprehensive survey. *Computational Visual Media*, 11(3):431–464, 2025. doi: 10.26599/CVM.2025.9450438.

Cagri Ozcinar, Aakanksha Rana, and Aljosa Smolic. Super-resolution of omnidirectional images using adversarial learning. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6. IEEE, 2019.

Hshmat Sahak, Daniel Watson, Chitwan Saharia, and David Fleet. Denoising diffusion probabilistic models for robust image super-resolution in the wild. *arXiv preprint arXiv:2302.07864*, 2023.

Xuelin Shen, Yitong Wang, Silin Zheng, Kang Xiao, Wenhan Yang, and Xu Wang. Fast omnidirectional 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*, volume 39, pp. 6833–6841. AAAI, 2025.

Xiaopeng Sun, Weiqi Li, Zhenyu Zhang, Qiufang Ma, Xuhan Sheng, Ming Cheng, Haoyu Ma, Shijie Zhao, Jian Zhang, Junlin Li, and Li Zhang. Opdn: Omnidirectional position-aware deformable network for omnidirectional image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1293–1301, June 2023.

Jiangang Wang, Yuning Cui, Yawen Li, Wenqi Ren, and Xiaochun Cao. Omnidirectional image super-resolution via biprojection fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5454–5462, 2024a.

Shaolin Wang, Yuying Li, Lei Zhong, Shigang Li, and Jianfeng Li. Edge-focused super-resolution for omnidirectional images with spherical geometric augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Tsun-Hsuan Wang, Hung-Jui Huang, Juan-Ting Lin, Chan-Wei Hu, Kuo-Hao Zeng, and Min Sun. Omnidirectional cnn for visual place recognition and navigation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2341–2348, 2018. doi: 10.1109/ICRA.2018.8463173.

Xin Wang, Shiqi Wang, Jinxing Li, Mu Li, Jinkai Li, and Yong Xu. Omnidirectional image super-resolution via position attention network. *Neural Networks*, 178:106464, 2024b. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2024.106464. URL https://www.sciencedirect.com/science/article/pii/S0893608024003885.

Xingtao Wang, Kaixin Wu, Jinyu Zhang, Yuxuan Wang, and Wenrui Li. Panoextend: An omnidirectional image super-resolution method based on spherical expansion. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, MMAsia '25 Workshops, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400722479. doi: 10.1145/3769748.3773344. URL https://doi.org/10.1145/3769748.3773344.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1905–1914, 2021.

Cuixin Yang, Rongkang Dong, Jun Xiao, Cong Zhang, Kin-Man Lam, Fei Zhou, and Guoping Qiu. Geometric distortion guided transformer for omnidirectional image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Kailun Yang, Jiaming Zhang, Simon Reiss, Xinxin Hu, and Rainer Stiefelhagen. Capturing omni-range context for omnidirectional segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1376–1386, June 2021.

Youngho Yoon, Inchul Chung, Lin Wang, and Kuk-Jin Yoon. Spheresr: 360deg image super-resolution with arbitrary projection via continuous spherical image representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5677–5686, June 2022.

Fanghua Yu, Xintao Wang, Mingdeng Cao, Gen Li, Ying Shan, and Chao Dong. Osrt: Omnidirectional image super-resolution with distortion-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13283–13292, June 2023.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *Advances in Neural Information Processing Systems*, 2023.

Xue Zhang, Gene Cheung, Yao Zhao, Patrick Le Callet, Chunyu Lin, and Jack Z. G. Tan. Graph learning based head movement prediction for interactive 360 video streaming. *IEEE Transactions on Image Processing*, 30:4622–4636, 2021. doi: 10.1109/TIP.2021.3073283.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 286–301, 2018.

Yuqi Zhang, Liang Chu, Zixu Wang, He Tong, Jincheng Hu, and Jihao Li. A novel panorama depth estimation framework for autonomous driving scenarios based on a vision transformer. *Sensors*, 24(21), 2024. ISSN 1424-8220. doi: 10.3390/s24217013. URL https://www.mdpi.com/1424-8220/24/21/7013.

A APPENDIX

A.1 LOSS ABLATION

Table 5 reports the ablation study on the three loss terms. Removing either the feature consistency loss or the SSIM loss leads to a clear performance drop, confirming that each term contributes to the final result. The full combination of L1 loss, feature consistency loss, and SSIM loss achieves the best performance, verifying that the three terms are complementary and jointly benefit panoramic SR.

Table 5: Ablation study on loss.

L1 Loss	Feature consistency Loss	SSIM Loss	ODI-SR	
			WS-PSNR	WS-SSIM
✓	×	✓	25.56	0.6787
✓	✓	×	25.50	0.6782
✓	✓	✓	**25.78**	**0.6874**

A.2 ON THE ABLATION STUDY OF SPATIAL ATTENTION REPLACEMENT

We replaced our edge-guided spatial attention with the distortion-aware attention module DAAB from OSRT (Yu et al., 2023). Under ×8, the replaced result was (25.38, 0.6716), compared to our full method at (25.78, 0.6874). This is likely because DAAB mainly boosts distortion-aware feature interactions, while our network already models latitude-dependent geometric distortions of ERP images via LConv; introducing DAAB further may lead to functional redundancy. By contrast, edge-guided spatial attention focuses on high-frequency edge regions and provides complementary detail enhancement for geometric modeling, making it better suited for our panoramic super-resolution task.

A.3 ON THE CHOICE OF A LEARNABLE DEGRADATION ENCODER

We choose a learnable degradation encoder over a fixed, non-learnable alternative because the actual data contains subtle degradation signals beyond the analytic bicubic model. Specifically, after HR images are bicubic-downsampled to LR, they are typically compressed for storage. The encoding and decoding process may introduce subtle, image-dependent variations beyond what a fixed bicubic-aware filter can capture. These compression-specific artifacts vary across images and cannot be captured by a fixed bicubic-aware filter. The learned encoder naturally absorbs these additional signals at the cost of only a few hundred extra parameters, making it a simple and practical choice.

A.4 ON THE OBSERVED DEGRADATION FEATURE PATTERNS AND THEIR IMPLICATIONS

Figure 5 shows channel-averaged feature maps from the degradation encoder.

Under standard bicubic degradation, the degradation features exhibit a warm tone with local variations near image gradients. These variations form diffuse, ringing-like envelopes around affected edges, consistent with bicubic-specific artifacts rather than generic edge responses. They do not appear at every contour and are driven by gradient strength rather than object identity. A perfectly uniform map would collapse to a trivial scalar, providing no useful information.

When 5% fixed salt-and-pepper noise is added to the input images (fourth column), the degradation features shift dramatically to a cool overall tone, with sparse, localized high responses that precisely match the distribution of the noise pixels. The edge-associated ringing patterns are no longer visually dominant, as the strong, isolated noise responses compress the feature range and dominate the intensity distribution. The feature maps directly reflect the spatial distribution of the salt-and-pepper noise, with bright spots appearing exactly where the noise is present in the degraded image.

If the encoder produced generic edge responses, the features would remain similar across both degradation types, and removing the degradation branch would not affect performance. Yet the observed patterns differ markedly, and the ablation study shows a clear performance drop when the branch is

removed. The degradation encoder thus captures content-agnostic, degradation-specific signals that provide meaningful guidance for reconstruction.

A.5 Additional Evidence for Disentanglement

As shown in Figure 5, our degradation encoder extracts content-agnostic representations under different degradation types. Although response patterns differ markedly, both respond to low-level structures rather than semantic content. We further conducted two additional experiments to strengthen the evidence.

cross-swapping test (reconstruction-level): we swapped degradation features across different samples under both bicubic and salt-and-pepper degradation settings (i.e., reconstructing image 'B' using content features from 'B' and degradation features from 'A'). In both cases, WS-PSNR remained unchanged. This demonstrates that under the cross-sample constraint, the degradation encoder learns a unified content-agnostic pattern for each degradation type, confirming that our disentanglement strategy generalizes across different degradations.

feature similarity analysis (feature-level): for each test image, we extracted content and degradation features using both the bicubic-trained and salt-and-pepper-trained models, and computed cross-degradation cosine similarities. Over the entire test set, the average content feature similarity was 0.66, which strongly indicates that content features remain stable across degradations, even though the two models were trained independently and their feature spaces are not explicitly aligned. In contrast, the average degradation feature similarity was only 0.10, confirming that the degradation encoder captures fundamentally different signals under different degradation types. This quantitatively validates the effectiveness of disentanglement at the feature level. Notably, the reconstruction-level cross-swapping experiment directly proves disentanglement, while the feature-level similarity analysis serves as a complementary observation; both are consistent.

Together, these experiments confirm that our disentanglement strategy successfully enables the degradation encoder to learn degradation-related features that are free of content information.

A.6 On alternative disentanglement constraint

To further validate the necessity of our cross-sample fusion strategy, we designed a baseline that removes the cross-sample constraint and instead applies an orthogonal regularization between the content and degradation features within the same sample, without any cross-sample feature swapping. This baseline achieved (25.53, 0.6799), which is notably lower than our full method at (25.78, 0.6874). The results confirm that the cross-sample constraint, as a stronger form of disentanglement, is empirically superior to the same-sample counterpart.

A.7 Implementation of non-integer dilation

The latitude-wise convolution produces a continuous dilation value $|\cos\theta| + s$ that is generally non-integer, while standard convolutional layers require integer dilation rates. To achieve smooth latitude-dependent modulation, we compute the convolution at the two nearest integer dilation rates and blend their outputs:

$$\mathrm{LatConv}(F) = (1 - |\cos\theta|) \cdot \mathrm{Conv}_s(F) + |\cos\theta| \cdot \mathrm{Conv}_{s+1}(F) \tag{7}$$

where Conv_s and Conv_{s+1} are standard convolutions with integer dilation s and $s+1$. At the equator ($|\cos\theta| = 0$), only Conv_s contributes; toward the poles ($|\cos\theta| \to 1$), Conv_{s+1} dominates. This matches the two-branch design in Sec. 3.2 and achieves a continuously varying receptive field without discretization artifacts.

To validate the benefit of our continuous weighting strategy, we construct a discrete baseline using nearest-integer dilation. Specifically, we partition the feature map into latitude segments according to the rounded dilation value: a dilation rate of 1 is assigned to the central one-third latitude band, and a rate of 2 to the two outer polar bands. Regional outputs are concatenated to form the final feature map. Replacing our smooth continuous weighting with this integer-truncated discrete scheme yields clear performance degradation: the overall WS-PSNR falls from 25.78 to 25.55 and WS-SSIM drops from 0.6874 to 0.6796. This demonstrates that smooth latitude-dependent interpolation outperforms discontinuous dilation transitions.

A.8 MORE COMPARISONS OF QUALITATIVE RESULTS



Figure 6: More visual comparisons of $\times 8$ and $\times 16$ super-resolution results on ODI-SR and SUN360.