WHEN CAPTURE IS NOT ENOUGH: SYNTHETIC DATA FOR REAL-WORLD ARBITRARY-SCALE IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Achieving high-quality arbitrary-scale super-resolution (ASSR) results is crucial for real-world multimedia applications. However, existing real-world SR datasets are primarily constructed through multi-camera synchronization with post-alignment, beam splitters for direct paired capture, or zoom lenses for sequential imaging. Yet these approaches either fail to provide fine-grained scaling factors required for arbitrary SR or struggle to preserve degradation consistency in dynamic scenes due to temporal misalignment. Therefore, these methods are not suited for capturing real-world datasets for arbitrary-scale super-resolution. Since capturing genuine arbitrary-scale datasets is infeasible, we construct a real-world SR dataset with truly arbitrary-scales through a degradation-aware synthesis approach. Specifically, we extract degradation characteristics from real low-resolution (LR) images and employ them to synthesize LR images at arbitrary scales, ensuring that synthesized data faithfully reflects real-world degradation patterns. We further propose the DP-ArbSR, a Dual-Prior Arbitrary-Scale Super-Resolution Network, which explicitly incorporates both degradation priors and texture priors to enhance feature representation and local feature fusion. Extensive experiments demonstrate that our synthesized dataset enables models to generalize robustly to real-world degradations, and DP-ArbSR achieves competitive performance. The dataset and the implementation code used in this work will be released.

1 INTRODUCTION

Arbitrary-scale image super-resolution (SR) reconstructs high-resolution (HR) images from low-resolution (LR) inputs at continuous magnifications (*e.g.*, $\times 2$, $\times 3.7$) with one model Ray et al. (2024); Gandikota & Chandramouli (2024); Chen et al. (2023b;c); Liu et al. (2023); Chen et al. (2024); Korkmaz et al. (2024); Ledig et al. (2017); Dong et al. (2014); Yue et al. (2024); Wang et al. (2024a;b); Zhang et al. (2024); Wu et al. (2024); Kim & Kim (2024); Wei et al. (2023); Cai et al. (2022). Learning these continuous representations requires fine-grained paired HR-LR images, often simulated with bicubic downsampling Hu et al. (2019); Chen et al. (2021); Lee & Jin (2022); Wei & Zhang (2023); Chen et al. (2023a); Yao et al. (2023); Cao et al. (2023); Wang et al. (2021b); models trained on such pairs struggle with real degradations Cai et al. (2019); Fu et al. (2024). Existing alternatives model degradation Wang et al. (2021a); Liu & Sun (2013); Elad & Feuer (1997); Yuan et al. (2018); Fritsche et al. (2019); Wang et al. (2021c); Liu et al. (2023); Chen et al. (2024); Zhang et al. (2021) or collect real datasets Zhang et al. (2019); Köhler et al. (2019); Chen et al. (2019); Qu et al. (2016); Wei et al. (2020); Cai et al. (2019); Fu et al. (2024); Li et al. (2024).

For simulation-based approaches, mainstream strategies include degradation with predefined physical parameter combinations Wang et al. (2021c); Zhang et al. (2021) and degradation modeling using neural networks guided by degradation features Chen et al. (2024); Liu et al. (2023). The former relies on fixed parameters, which struggle to accommodate the complex degradation characteristics at different scales. The latter offers stronger representational capability but requires that degradation features corresponding to different scaling factors be sufficiently distinguishable to be applicable to arbitrary-scale image synthesis. However, existing methods have yet to effectively resolve this issue. While recent work like InterFlow Park et al. (2023) attempts to bridge this gap via linear latent interpolation between the few discrete scales available in real-world datasets, this simplistic approach

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.

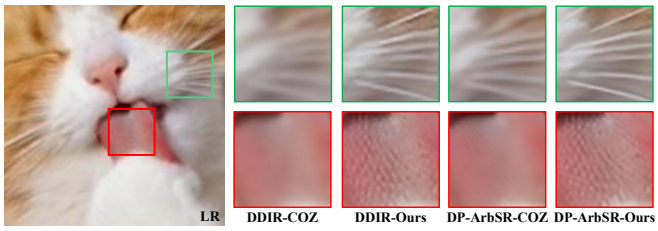

Figure 1: Visual comparison of $\times 6$ super-resolution results. From left to right: DDIR Li et al. (2024) trained on the COZ dataset Fu et al. (2024), DDIR trained on our dataset, our DP-ArbSR trained on the COZ dataset, and DP-ArbSR trained on our dataset. (The LR image is shown enlarged for clarity.)

fails to capture the highly non-linear dynamics of real-world degradation across continuous arbitrary scales.

For real-world data acquisition, major approaches include multi-camera synchronized capture followed by post-alignment Chen et al. (2019); Cai et al. (2019), direct alignment using beam splitters Köhler et al. (2019); Yue et al. (2022); Joze et al. (2020), and continuous capturing with optical zoom lenses Zhang et al. (2019); Fu et al. (2024); Li et al. (2024). Nonetheless, these methods have significant limitations. Multi-camera systems require strict hardware synchronization and spatial alignment, and in scenarios requiring data collection at numerous scales (*e.g.*, 50 different scales), system complexity grows exponentially. While beam splitters can address image registration issues, their inability to achieve multi-path light splitting makes them unsuitable for arbitrary-scale SR data collection. Optical zoom imaging theoretically enables continuous-scale acquisition, but in practice, the nonlinear motion characteristics of zoom mechanisms result in uneven scale distribution, and the delay introduced by zooming prevents accurate alignment of moving objects between consecutive images. Furthermore, naive physical re-capture methods (*e.g.*, photographing HR screens or printed postcards) inevitably introduce secondary artifacts such as moiré patterns and corrupt the natural degradation distribution.

Compared to synthetic data, real-world data exhibits more diverse degradation types, and models trained on such data typically perform better in practical applications. However, current acquisition pipelines still struggle to provide arbitrary-scale training data that is both high-quality and precisely aligned in time and space, whereas simulation-based methods can readily produce large-scale, strictly aligned samples with well-controlled scale distributions. To bridge this gap, we construct a training dataset that preserves real-world degradation characteristics while remaining practical to build. Compared with previous efforts Fu et al. (2024); Li et al. (2024); Park et al. (2023), our dataset more closely matches the practical requirements of ideal arbitrary-scale SR training data, with a more uniform and finer-grained distribution over continuous scale factors, richer scene diversity, and stronger support for high-quality visual reconstruction. Concretely, we dynamically generate degradation kernels based on the degradation features of a reference LR image and apply them to HR images to synthesize LR images whose degradation characteristics are consistent with the reference. To ensure reliable extraction of degradation features, we develop a degradation feature extractor using a rank-based learning strategy, enabling it to capture subtle degradation differences among LR images with nearby scaling factors in the same scene. Our comparison with other methods is shown in Tab. 1.

To enhance the SR stability and arbitrary-scale performance, we propose the Dual-Prior Arbitrary-scale SR Network (DP-ArbSR) comprising: a Diffusion-based Texture Prior Generation module, a plug-and-play module integrating degradation and texture priors, and a Dual-Attention Feature Fusion module (spatial and channel) combining texture priors with enriched encoder features. By integrating the texture prior from the Diffusion model with features extracted by the SR model itself and employing implicit neural function representations, our method produces SR results that are of higher quality and exhibit more natural details.

The primary contributions of this work are as follows:

- To the best of our knowledge, we construct the closest approximation to ideal arbitrary-scale SR training data to date, enabling fine-grained, well-aligned continuous-scale supervision with realistic degradations and broader scene diversity.

Table 1: Comparison with previous super-resolution datasets.

Dataset	Conference	Real-World	Arbitrary-Scale	Fine-Grained	Comprehensive	Paired LR
DIV2K Agustsson & Timofte (2017)	CVPRW 2017	✗	✓	✓	✓	-
RealSR Cai et al. (2019)	ICCV 2019	✓	✗	✗	✓	-
SR-RAW Zhang et al. (2019)	CVPR 2019	✓	✗	✗	✓	-
DRealSR Wei et al. (2020)	ECCV 2020	✓	✗	✗	✗	-
COZ Fu et al. (2024)	CVPR 2024	✓	✓	Long-tail	✗	-
RealArbiSR Li et al. (2024)	ECCV 2024	✓	✓	✗	✗	-
DegAE Liu et al. (2023)	CVPR 2023	✓	✓	Not Completely	✓	Not required
InterFlow Park et al. (2023)	ICML 2023	✓	✓	Genuine	✓	Required
Ours	Required	✓	✓	Genuine	✓	Not required

- We introduce the Dual-Prior Arbitrary-Scale Super-Resolution Network (DP-ArbSR), which leverages degradation prior and texture prior to enhance robustness in real-world scenarios.
- Extensive experiments validate the effectiveness of our dataset and DP-ArbSR by comparing them to state-of-the-art methods.

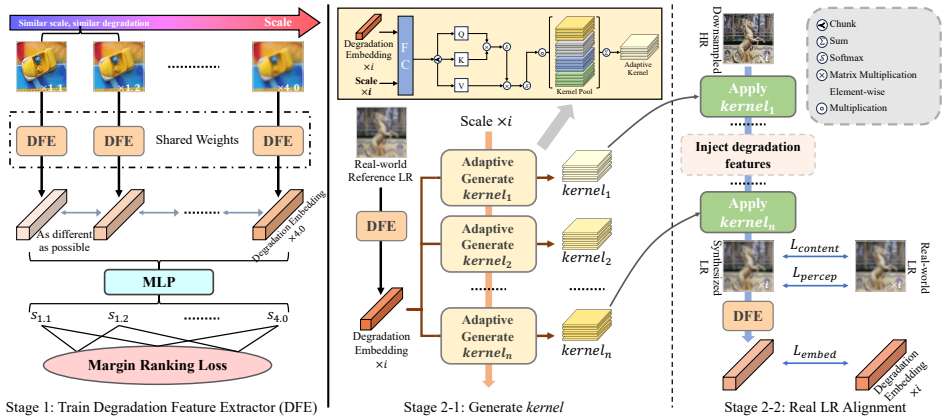

Figure 2: The proposed Degraded Image Synthesis Framework. Stage 1: Degradation feature extractor (DFE) learns to discriminate degradation patterns through rank-based learning. Stage 2-1: Adaptive kernel generate conditioned on scale factors and target degradation embeddings. Stage 2-2: Kernels are applied to degrade images.

2 Degraded Image Synthesis

To simulate realistic degradation processes, we propose a novel degraded image synthesis method, as illustrated in Fig. 2. Specifically, a degradation feature extractor (DFE) is first trained via ranking-based learning to better capture fine-grained degradation differences among images from adjacent scales. The resulting degradation feature embeddings are then used as priors to guide the adaptive construction of degradation kernels, which are applied to a downsampled high-resolution image $HR_{\downarrow}$ to synthesize LR images with controllable degradation characteristics. This design enables the synthesized LR images to adapt to diverse scale factors and degradation patterns while preserving precise control over the degradation properties.

2.1 Stage 1: Train Degradation Feature Extractor

Prior work has shown that degradation features can be extracted from a single image. Liu *et al*. Liu et al. (2023) use a pretrained SRGAN generator with several downsampling layers, a design later adopted by subsequent work Chen et al. (2024). CDFormer Liu et al. (2024) instead uses paired HR and LR images for degradation extraction and employs diffusion to generate degradation features.

In ASSR, LR images of the same scene at adjacent scales exhibit only subtle degradation differences, making them difficult to distinguish without additional constraints. Motivated by No-Reference Image Quality Assessment, we train a ResNet-50-based DFE with ranking learning to produce a 512-dimensional embedding. During training, a removable auxiliary MLP with parameters θ_M maps the DFE $\mathcal{E}$ outputs for a same-scene image pair to relative degradation scores:

$$\begin{aligned} s_1 &= \text{MLP}(\mathcal{E}(i_1;\ \theta_{\mathcal{E}});\ \theta_M) \\ s_2 &= \text{MLP}(\mathcal{E}(i_2;\ \theta_{\mathcal{E}});\ \theta_M) \end{aligned} \tag{1}$$

For each scene, we form pairs at different resolutions so that scene content remains fixed while degradation severity changes. This provides an explicit ordering constraint: the lower-resolution

observation should receive a higher degradation score. We jointly optimize θ_M and $\theta_{\mathcal{E}}$ with margin-ranking loss Liu et al. (2017):

$$L(s_1,\ s_2,\ y) = \max(0, -y * (s_1 - s_2) + \text{margin}),$$
$$y = \begin{cases} 1, & \text{if } r_1 < r_2, \\ -1, & \text{if } r_1 > r_2 \end{cases} \tag{2}$$

where r_1 and r_2 denote the image resolutions. The scores encode relative rather than absolute degradation severity, allowing the DFE to focus on fine-grained differences between nearby scales without requiring a calibrated quality label. The auxiliary MLP is needed only to impose this ordering during training. At inference, it is removed and only the DFE is retained to obtain the degradation embedding:

$$E_i = \mathcal{E}(i;\ \theta_{\mathcal{E}}) \tag{3}$$

2.2 Stage 2-1: Kernel Generation

In real-world scenarios, degradation kernels are highly complex and exhibit significant variations across different magnification levels. To more accurately simulate this intricate degradation process and ensure that the degradation characteristics of the synthesized images align with actual conditions, this study introduces scale factors and fine-grained degradation embedding extracted by the DFE trained in Stage 1 as prior information to guide the generation of degradation kernels. Specifically, we employ an attention mechanism to dynamically combine different kernels from a kernel pool to generate degradation kernels that align with the specified scale and degradation embedding. The mathematical formulation is as follows:

$$q, k, v = \text{Chunk}(\text{FC}(\text{Concat}\{s,\ E\})),$$
$$v' = \text{Softmax}(qk^T) * v,$$
$$w = \text{Softmax}(MLP(v')), \tag{4}$$
$$kernel = \sum_{i=1}^{n} kernel_pool_i \cdot w_i$$

where s is the scaling factors, and n is the size of the kernel pool. The adaptive degradation kernel $kernel$ and pool elements $kernel_pool_i$ are in $\mathbb{R}^{out \times in \times kH \times kW}$, where out and in denote output and input channels, and kH, kW specify the kernel size. The term FC refers to a fully connected layer.

To enhance the representation of degradation, as illustrated in Fig. 2, we generate a set of degradation kernels for subsequent image synthesis. Notably, each generated kernel is associated with its own corresponding kernel pool. The kernels within the pool are a set of learnable parameters, whose values are adjusted during the training process by the loss function in Stage 2-2.

2.3 Stage 2-2: Real LR Alignment

This stage mainly trains the synthesis network to effectively exploit degradation embeddings; once training is complete, paired LR-HR images are no longer required for actual image synthesis.

Once the degradation kernels are prepared, they are applied for the synthesis of degraded images. Specifically, the input first passes through a convolutional layer followed by a ReLU function. The generated degradation kernels are then used for the second convolution, and the resulting output is added to the input. For the first convolution, each degradation kernel has a dedicated convolutional layer. To synthesize the degraded image, we first map the channel number of the input image $I_{\downarrow}^{HR}$ to the desired quantity using a convolutional layer, then sequentially apply the degradation kernels generated in Stage 2-1. Finally, a ConvBlock is applied to restore the channel count to 3. To improve efficiency, the generation and application of degradation kernels are performed synchronously; each degradation kernel is immediately applied to the input image as it is generated.

To ensure that the synthesized images closely align with real-world degradation characteristics, we incorporate an additional degradation-aware feature injection step during the synthesis process. This is achieved by using the degradation embedding to modulate the weights of different channels in a channel-attention manner. To further strengthen the influence of the degradation embedding, an additional bias is added to each channel. The operation can be formulated as:

$$X' = X \odot \text{Sigmoid}(\text{FC}(E)) + \text{FC}(E) \tag{5}$$

where $X, X' \in \mathbb{R}^{HW \times C}$ represent the input and output feature respectively, while $\odot$ denotes element-wise multiplication along the channel dimension.

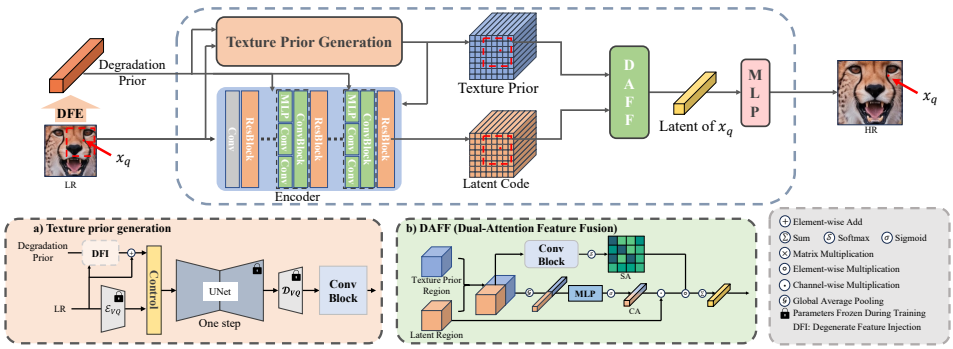

Figure 3: The proposed Dual-Prior Arbitrary-scale Super-Resolution Network.

We train all modules in Stage 2 using three loss functions: content reconstruction loss L_{content}, perceptual loss L_{per}, and degradation embedding loss L_{embed}.

Content Reconstruction Loss: To enhance degradation diversity, the generated image $I_{\downarrow}^{HR'}$ is allowed slight detail variations compared to the reference I^{LR}. Specifically, a Gaussian blur kernel k is applied to both images, and their L_1 loss is computed as:

$$L_{content} = \|I_{\downarrow}^{HR'} \otimes k - I^{LR} \otimes k\|_1 \tag{6}$$

Perceptual Loss: We adopt a VGG$_L$based perceptual loss Johnson et al. (2016):

$$L_{per} = \sum_{l} w_l \cdot \|\text{VGG}_l(I_{\downarrow}^{HR'}) - \text{VGG}_l(I^{LR})\|_2^2 \tag{7}$$

where $\text{VGG}_l(I)$ is the feature map of image I at layer l, and w_l is the weight for that layer, with L being the selected VGG layers.

Degradation Embedding Loss: To align the synthesized image's degradation embedding $E_{HR'}$ with the target E, we use cosine similarity, defining the embedding loss as:

$$L_{embed} = 1 - \frac{E_{HR'} \cdot E}{\|E_{HR'}\| \times \|E\|} \tag{8}$$

The overall loss is defined as $L = L_{\text{content}} + \lambda_1 L_{\text{per}} + \lambda_2 L_{\text{embed}}$, where $\lambda_1 = 0.001$ and $\lambda_2 = 0.005$.

2.4 Image Synthesis

During the image synthesis stage, no LR-HR paired images are required. Given an HR image and a reference LR image LR_R, we first downsample the HR image by a target scale factor i to obtain $HR\downarrow$, and then extract the degradation embedding E_R from LR_R using the DFE trained in Stage 1. Conditioned on E_R and the scale factor, Stage 2-1 generates adaptive degradation kernels, which are applied in Stage 2-2 to synthesize the target LR image LR_i. To support training and evaluation, we collect 588 CC0-licensed photographic images, using 535 scenes for training and the rest for testing. For each training scene, we synthesize 100 images with scale factors uniformly sampled from $\times 1.0$ to $\times 4.0$, while for testing we only use integer scale factors. Our dataset additionally includes scene types that are not present in existing datasets such as COZ Fu et al. (2024) and RealArbiSR Li et al. (2024). It therefore covers diverse scene categories and provides a more uniform scale distribution. Complete visualizations of the per-scene synthesis sequence and dataset statistics are provided in the supplementary material. Additional implementation details are also provided there.

3 Method

Recent ASSR methods represent an image as a continuous implicit function rather than discrete pixels Chen et al. (2021); Lee & Jin (2022); Wei & Zhang (2023); Chen et al. (2023a); Cao et al. (2023); Fu et al. (2024); Li et al. (2024). Given a coordinate x and latent code Z extracted by an encoder such as EDSR Lim et al. (2017) or RDN Zhang et al. (2018b), the corresponding RGB value is modeled as $I(x) = f(Z, x)$. Because a latent code is not directly available at an arbitrary query coordinate x_q, existing methods interpolate predictions from its neighboring codes:

$$I(x_q) = \sum_{(i,j)\in S} w_{i,j} f(Z_{i,j}, x_{i,j} - x_q), \tag{9}$$

where S denotes the local neighborhood, $x_{i,j}$ and $Z_{i,j}$ are its coordinates and latent codes, and $w_{i,j}$ is the interpolation weight. However, a single encoder struggles with diverse degradations and scale factors, while the small local neighborhood limits the effective receptive field for RGB prediction.

3.1 PROPOSED SUPER-RESOLUTION NETWORK

Although the synthesized training data alleviate the data scarcity issue for underrepresented scenes, real-world arbitrary-scale super-resolution remains challenging due to complex degradations and diverse texture patterns that cannot be fully addressed by data augmentation alone. To further improve robustness and generalization in real-world scenarios, we propose a **D**ual-**P**rior **Arb**itrary-scale **S**uper-**R**esolution Network (DP-ArbSR), as shown in Fig. 3. Specifically, DP-ArbSR incorporates both a degradation prior, extracted by the DFE trained in Stage 1, and a texture prior, generated by a diffusion model pre-trained on large-scale real-world images, to enhance feature extraction and local feature fusion. By jointly exploiting these two priors, the proposed network better handles diverse real-world degradation conditions and scene contents.

3.2 DEGRADATION-TEXTURE PRIOR GENERATION

By utilizing the DFE model trained in the previous section, we directly use the extracted degradation embeddings as the degradation prior. During the process of obtaining the texture prior, as shown in Fig. 3(a), the degradation prior first modulates the LR image according to Eq. (5). Subsequently, the LR image serves as the control input for the diffusion model to guide the generation of the texture prior. ResShift's distilled SinSRYue et al. (2023); Wang et al. (2024b) serves as the backbone of the diffusion model, along with its original control layer consisting of a single convolutional layer. Benefiting from this distilled design, the diffusion module accomplishes texture-prior generation with only a single denoising step during inference, eliminating expensive iterative sampling and further improving efficiency. VQGAN Esser et al. (2021) is further employed to encode images into the latent space, reducing computational cost. To preserve the pre-trained knowledge of the diffusion model on data distribution, all relevant modules (indicated by the lock icon in the figure) are frozen during training. Only the control layer is trained to generate an effective texture prior.

3.3 PRIOR FOR ENCODER

To incorporate prior knowledge during the encoding phase, we design a plug-and-play module integrated into the EDSR encoder Lim et al. (2017). The degradation prior P_D undergoes initial processing through a MLP followed by sigmoid function to generate channel attention weights. Simultaneously, the texture prior P_T and intermediate features F_E from the encoder output undergo channel compression via separate 1×1 convolutional layers. These compressed features are concatenated to form composite feature F. The computed channel attention from P_D is then applied to F through element-wise multiplication along the channel dimension, producing refined features F'. A convolutional block then integrates prior knowledge before residual fusion with the original encoder features F_E, ultimately yielding the output features F_o. The detailed calculation process is as follows:

$$\begin{aligned} F &= \text{Concat}\{\text{Conv}_{1\times1}(F_E),\ \text{Conv}_{1\times1}(P_T)\}, \\ F' &= F \odot \text{Sigmoid}(\text{MLP}(P_D)), \\ F_o &= F_E + \text{ConvBlock}(F') \end{aligned} \tag{10}$$

3.4 PRIOR FOR LOCAL DUAL-ATTENTION FEATURE FUSION

To expand the receptive field of the neural network when predicting the RGB value at x_q, we choose to capture a larger surrounding region of x_q for prediction, instead of the conventional method of using only four neighboring pixels. Specifically, we take x^* as the center and extract a region R^Z of size $r \times r$ (with default $r = 5$) from the latent code Z. At the same time, we extract the corresponding region R^T from the texture prior. Due to the expanded feature extraction range, we propose the Dual-Attention Feature Fusion module (DAFF) to effectively integrate these features, as shown in Fig. 3(b). We first concatenate the extracted regions R^Z and R^T to obtain R, then perform global average pooling on R, followed by an MLP and a Sigmoid layer to generate channel attention ca. The formula is as follows:

$$\begin{aligned} R &= \text{Concat}\{R^Z,\ R^T\} \\ ca &= \text{Sigmoid}(\text{MLP}(\mathcal{G}(R))) \end{aligned} \tag{11}$$

where $\mathcal{G}(\cdot)$ denotes the global average pooling operation. Next, we calculate a spatial attention weight sa to quantify the contribution of each pixel within the region to the final result. To reduce computational cost, we generate the attention map using convolution instead of the self-attention mechanism used in LIT Chen et al. (2023a). The concatenated feature R is passed through a convolutional block, which reduces the channels to 1, and a Softmax function is applied to obtain the

Table 2: Quantitative comparison (PSNR (dB)) with SOTA methods on COZ, RealArbiSR, and our synthesized test scenarios.

Methods	COZ Fu et al. (2024)					RealArbiSR Li et al. (2024)			Our Synthesized Test Scenarios				
	In-scale			Out-of-scale		In-scale			In-scale			Out-of-scale	
	×2	×3	×4	×5	×6	×2	×3	×4	×2	×3	×4	×5	×6
EDSR-LIIF Chen et al. (2021)	28.81	26.68	25.24	24.37	23.28	29.85	27.03	25.42	31.91	28.10	25.78	24.05	23.76
EDSR-LTE Lee & Jin (2022)	28.86	26.70	25.26	24.40	23.27	29.96	27.12	25.46	31.96	28.14	25.83	24.09	23.81
EDSR-SRNO Wei & Zhang (2023)	28.82	26.67	25.23	24.43	23.31	29.94	27.13	25.44	31.97	28.08	25.80	24.12	23.76
EDSR-LIT Chen et al. (2023a)	28.84	26.73	25.26	24.39	23.33	29.97	27.16	25.47	32.00	28.11	25.78	24.15	23.80
EDSR-CiaoSR Cao et al. (2023)	28.85	26.72	25.27	24.44	23.33	29.99	27.15	25.52	31.97	28.13	25.83	24.04	23.75
EDSR-LMI Fu et al. (2024)	28.87	26.74	25.33	24.48	23.36	29.98	27.13	25.61	32.01	28.10	25.85	24.08	23.77
EDSR-DDIR Li et al. (2024)	28.87	26.75	25.31	24.43	23.36	30.02	27.17	25.64	32.05	28.14	25.84	24.13	23.82
RDN-LIIF Chen et al. (2021)	28.87	26.73	25.28	24.40	23.33	29.90	27.11	25.47	31.98	28.17	25.84	24.12	23.84
RDN-LTE Lee & Jin (2022)	28.90	26.75	25.31	24.42	23.31	29.99	27.19	25.51	32.02	28.18	25.89	24.14	23.89
RDN-SRNO Wei & Zhang (2023)	28.87	26.74	25.26	24.46	23.32	29.97	27.21	25.50	32.02	28.13	25.85	24.20	23.83
RDN-LIT Chen et al. (2023a)	28.89	26.81	25.31	24.44	23.37	29.99	27.21	25.55	32.07	28.15	25.85	24.23	23.86
RDN-CiaoSR Cao et al. (2023)	28.89	26.79	25.32	24.47	23.38	30.02	27.22	25.57	32.02	28.17	25.89	24.13	23.83
RDN-LMI Fu et al. (2024)	28.91	26.84	25.44	24.56	23.42	30.04	27.23	25.69	32.06	28.19	25.92	24.17	23.86
RDN-DDIR Li et al. (2024)	28.92	26.87	25.45	24.52	23.40	30.08	27.26	25.70	32.09	28.22	25.91	24.19	23.90
DP-ArbSR (ours)	**29.05**	**26.98**	**25.52**	**24.65**	**23.50**	**30.15**	**27.33**	**25.77**	**32.19**	**28.29**	**25.98**	**24.30**	**23.99**

spatial attention sa. The calculation is given by:

$$sa = \text{Softmax}(\text{ConvBlock}(R)) \tag{12}$$

Once the attention weights for each channel and the attention of each pixel to the result are obtained, the final RGB value can be computed using the following equation:

$$\begin{aligned} R' &= R^Z \odot ca, \\ I(x_q) &= f(\sum_{i=0}^{r}\sum_{j=0}^{r} sa_{ij} \cdot R'_{ij},\ x^* - x_q) \end{aligned} \tag{13}$$

4 Experiments

4.1 Experimental Settings

Datasets. We train the degradation feature extractor (DFE) and synthesis network on COZ Fu et al. (2024). The SR network uses three training configurations: fine-grained COZ augmentation, our synthetic data alone, or mixed COZ, RealArbiSR Li et al. (2024), and synthetic data. We test on COZ, RealArbiSR, and our synthesized scenarios. In Tab. 2, $\times 2$–$\times 4$ are labeled in-scale and $\times 5$–$\times 6$ out-of-scale, relative to the $\times 1$–$\times 4$ sampling range of our synthetic training images.

Implementation Details. The DFE learns from sampled LR-HR patch pairs via ranking and is evaluated with SROCC. The synthesis network incorporates degradation embeddings and is evaluated by cosine similarity (Eq. (8)); paired images are required for training but not synthesis inference. We optimize SR with Adam and L1 loss and report PSNR. Further details are in the supplementary material.

4.2 Quantitative Experiments

The Ability to Generate Fine-Grained LR Images. We use DegAE Liu et al. (2023) as the baseline and augment COZ separately with LR images generated by DegAE and our degradation model. Both generators are applied to the same source scenes so that the comparison isolates the synthesized degradation rather than scene content. To test whether the resulting supervision benefits different architectures, we train LIIF Chen et al. (2021), LTE Lee & Jin (2022), LMI Fu et al. (2024), DDIR Li et al. (2024), and DP-ArbSR on both augmented datasets, then evaluate them on COZ and RealArbiSR. Averaging across the five models avoids attributing a gain to a particular SR architecture. The results in Tab. 3(a) show consistent improvements at the challenging $\times 4$–$\times 6$ factors. Compared with DegAE data, ours improves mean PSNR by 0.068 dB on COZ $\times 4$, 0.048 dB on COZ $\times 6$, and 0.036 dB on RealArbiSR $\times 4$, supporting its ability to synthesize fine-grained degradations.

Effectiveness of Degradation Types. We further collect 535 internet scenes that are underrepresented or absent in existing datasets Fu et al. (2024); Li et al. (2024), and synthesize separate training sets with DegAE and our model. Unlike the preceding COZ augmentation, this setting tests whether a degradation generator can turn new scene content into useful supervision without relying on paired real LR observations. We train the same five ASSR methods and test them on COZ and RealArbiSR. Even when trained only on synthetic observations of these scenes, our degradation model achieves higher mean PSNR across representative scales (Tab. 3(a)), indicating a better approximation of complex real-world degradations.

Table 3: PSNR (dB): (a) mean across LIIF, LTE, LMI, DDIR, and DP-ArbSR for fine-grained COZ augmentation and collected-scene synthesis; (b) DP-ArbSR ablations on the COZ test set. Complete per-model and per-scale synthesis results are in the supplementary material.

(a) Data synthesis

Setting	Data	COZ ×4	COZ ×6	RealArbiSR ×4
Fine-gr.	Original	25.174	23.188	25.478
	DegAE	25.212	23.236	25.520
	Ours	**25.280**	**23.284**	**25.556**
Collected	Bicubic	24.614	22.674	24.772
	DegAE	24.732	22.818	24.872
	Ours	**25.086**	**23.128**	**25.460**

(b) Ablation

Methods	Scale				
	×2	×3	×4	×5	×6
DP-ArbSR	29.05	26.98	25.52	24.65	23.50
DP-ArbSR-RDN	29.15	27.06	25.59	24.77	23.59
DP-ArbSR-K2	28.98	26.88	25.46	24.59	23.41
DP-ArbSR-K8	29.08	26.99	25.50	24.67	23.53
DP-ArbSR-a	28.91	26.84	25.37	24.48	23.32
DP-ArbSR-b	28.95	26.89	25.44	24.57	23.38
DP-ArbSR-c	29.00	26.94	25.47	24.60	23.47
DP-ArbSR-d	28.98	26.90	25.44	24.59	23.39
DP-ArbSR-e	28.90	26.84	25.39	24.49	23.37
DP-ArbSR-f	28.96	26.93	25.46	24.58	23.45
DP-ArbSR-g	29.04	26.98	25.52	24.63	23.50
DP-ArbSR-h	29.01	26.92	25.46	24.59	23.44

Comparison with Bicubic Downsampled Data. To distinguish learned degradation synthesis from conventional resizing, we also generate the collected-scene training data using bicubic downsampling. The same models are trained with bicubic, DegAE, and our data and evaluated using PSNR and SSIM on COZ and RealArbiSR. Relative to bicubic data, ours improves mean PSNR by 0.472, 0.454, and 0.688 dB on the three representative settings in Tab. 3(a); the gains over DegAE are 0.354, 0.310, and 0.588 dB. Complete per-model PSNR matrices are provided in the supplementary material, together with additional SSIM and LPIPS evaluations.

Models Trained on Mixed Data. Finally, we train eight methods—LIIF Chen et al. (2021), LTE Lee & Jin (2022), SRNO Wei & Zhang (2023), LIT Chen et al. (2023a), CiaoSR Cao et al. (2023), LMI Fu et al. (2024), DDIR Li et al. (2024), and DP-ArbSR—on mixed data comprising COZ, RealArbiSR, and our synthetic dataset. We evaluate all models on COZ, RealArbiSR, and additional synthesized scenes to cover both seen and complementary degradation domains. As shown in Tab. 2, DP-ArbSR outperforms methods not designed for real-world ASSR and remains competitive with LMI and DDIR. These results show that our network can exploit mixed training data without sacrificing robustness across evaluation domains.

4.3 Qualitative Experiments

Testing set. We compare LMI Fu et al. (2024), DDIR Li et al. (2024), and DP-ArbSR trained on COZ, DegAE (DAE) Liu et al. (2023), or our synthetic data across three scenes: synthetic ($\times 3$), COZ ($\times 4$), and RealArbiSR ($\times 4$). Across these cases, models trained exclusively on COZ data struggle with texture recovery in synthetic animal scenes, while DAE-trained models restore less detail on the RealArbiSR and COZ test sets. Overall, our synthetic method surpasses DegAE in visual quality and efficiently generates scenes absent from current datasets. The supplementary material provides complete comparisons across datasets, models, and training configurations, together with additional qualitative examples.

Generalizability. We test DDIR Li et al. (2024) trained on COZ, DegAE (DAE), or our data against DP-ArbSR trained on our data, using Sony A7R2, Panasonic S5M2, and iPhone 12Pro scenes. Results in Fig. 4 show that the model trained on our synthetic data generalizes well to real-world scenarios, outperforming DegAE's approach. Notably, our method surpasses COZ-trained models on scenes rare in the real dataset, such as animals and dynamic water surfaces, while achieving similar visual quality to COZ on common architectural scenes.

User study. We conduct two user studies with 20 participants on 20 real-world scenes (10 static and 10 dynamic) to evaluate real-world adaptability and perceptual quality. In the first study, participants choose their preferred $\times 4$ and $\times 6$ results between models trained on our synthetic data and those trained on COZ. The models trained on our data are preferred more often at both magnifications, suggesting that our synthesized data better matches real-world degradation. In the second study, participants rank the visual quality of $\times 4$ outputs from representative methods. Our DP-ArbSR achieves the highest ranking scores, demonstrating superior perceptual quality. Complete plots and additional results are provided in the supplementary material.

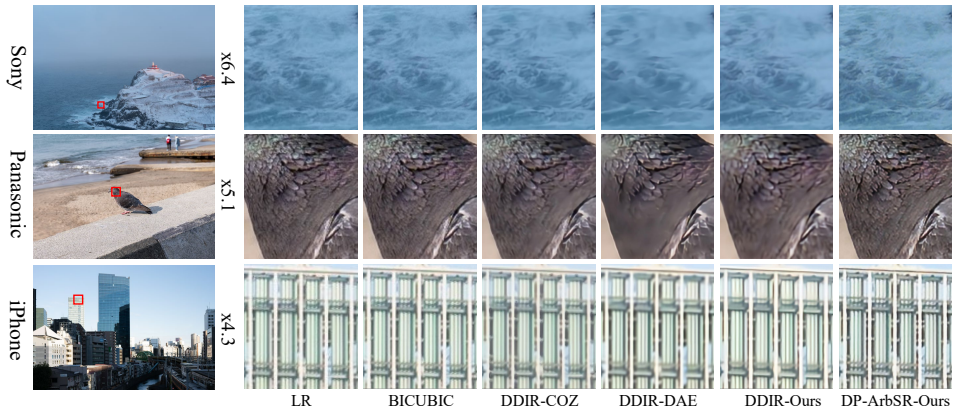

Figure 4: Real-world visual comparison of DDIR Li et al. (2024) trained on COZ Fu et al. (2024), DegAE-synthesized (DAE) Liu et al. (2023), or our data, and DP-ArbSR trained on our data. More examples are provided in the supplementary material.

4.4 Ablation Experiments

Rank-Based Learning. We use the pretrained degradation extractor Liu et al. (2023); Chen et al. (2024) as DegAE-O and retrain only its downsampling layers with our ranking objective (DegAE-RL), isolating the learning objective. On COZ, the mean cosine similarity between same-scene, adjacent-scale embeddings falls from 0.99 to 0.89. The lower similarity shows improved discrimination of fine-grained degradation rather than scene content. Scale-granularity results are provided in the supplementary material.

Kernel Generation. We set the kernel pool size to 2 and 8 to obtain DP-ArbSR-K2 and DP-ArbSR-K8. As shown in Tab. 3(b), K2 is consistently below the default at every scale, indicating insufficient degradation diversity. K8 changes PSNR only marginally, from -0.02 to $+0.03$ dB across the five scales, while adding computation. We therefore retain the default as a balanced choice; further pool analysis is provided in the supplementary material.

Plug-and-Play Module. Variant (a) removes the Plug-and-Play module to measure its overall contribution, while (b) retains only one instance to test whether repeated prior injection is necessary. Variant (c) removes channel attention from the module, and (d) removes degradation injection (Eq. (5)) during texture-prior generation. Their lower results at every reported scale in Tab. 3(b) show that both the priors and their repeated integration contribute to reconstruction quality. We also integrate the module into RDN as DP-ArbSR-RDN; its consistent gains over the default backbone results further demonstrate portability across ASSR backbones.

DAFF. Variant (e) removes DAFF and predicts as in LIIF Chen et al. (2021), testing whether conventional local interpolation is sufficient. Variants (f) and (g) set the local region to $r = 3$ and $r = 7$, while (h) removes channel attention. Removing DAFF or channel attention degrades performance, confirming that both spatial aggregation and channel reweighting contribute to local feature fusion. The smaller $r = 3$ limits contextual information, whereas $r = 7$ gives no clear gain because a larger region may dilute key local cues and make relevant pixels harder to distinguish. We therefore use $r = 5$ by default (Tab. 3(b)).

Qualitative Ablation Study. We additionally inspect a synthesized test scene at $\times 4$. Removing the texture prior weakens texture details (-a), excluding degradation injection introduces blur (-d), and removing DAFF impairs local texture reconstruction (-e). These visual changes agree with the quantitative ablation; the complete comparison is provided in the supplementary material.

5 Conclusion

We propose a degradation-aware synthesis method that constructs the first fine-grained, label-aligned dataset for real-world arbitrary-scale SR. We also introduce DP-ArbSR, integrating degradation and texture priors for feature extraction and local fusion. Experiments and user studies validate the contributions and demonstrate state-of-the-art performance.

AI USE STATEMENT

Generative AI tools assisted with retrieval and discovery of related literature, migration to the ICLR 2027 LaTeX template, compilation diagnostics, and editorial condensation of prose. The authors are responsible for verifying the literature, wording, scientific claims, experimental results, data, and conclusions in the final submission.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 126–135, 2017.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. 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.

Qing Cai, Jinxing Li, Huafeng Li, Yee-Hong Yang, Feng Wu, and David Zhang. TDPN: Texture and Detail-Preserving Network for Single Image Super-Resolution. *IEEE Transactions on Image Processing*, 31:2375–2389, 2022.

Jiezhang Cao, Qin Wang, Yongqin Xian, Yawei Li, Bingbing Ni, Zhiming Pi, Kai Zhang, Yulun Zhang, Radu Timofte, and Luc Van Gool. Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1796–1807, 2023.

Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. Camera lens super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1652–1660, 2019.

Hao-Wei Chen, Yu-Syuan Xu, Min-Fong Hong, Yi-Min Tsai, Hsien-Kai Kuo, and Chun-Yi Lee. Cascaded local implicit transformer for arbitrary-scale super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18257–18267, 2023a.

Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Haoze Sun, Xueyi Zou, Zhensong Zhang, Youliang Yan, and Lei Zhu. 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.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22367–22377, 2023b.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. 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.

Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12312–12321, 2023c.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 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.

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, 2015.

Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3599–3608. IEEE, 2019.

Huiyuan Fu, Fei Peng, Xianwei Li, Yejun Li, Xin Wang, and Huadong Ma. Continuous Optical Zooming: A Benchmark for Arbitrary-Scale Image Super-Resolution in Real World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3035–3044, 2024.

Kanchana Vaishnavi Gandikota and Paramanand Chandramouli. Text-guided Explorable Image Super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25900–25911, 2024.

Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-SR: A magnification-arbitrary network for super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1575–1584, 2019.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016.

Hamid Reza Vaezi Joze, Ilya Zharkov, Karlton Powell, Carl Ringler, Luming Liang, Andy Roulston, Moshe Lutz, and Vivek Pradeep. Imagepairs: Realistic super resolution dataset via beam splitter camera rig. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 518–519, 2020.

Jinseok Kim and Tae-Kyun Kim. Arbitrary-Scale Image Generation and Upsampling using Latent Diffusion Model and Implicit Neural Decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9202–9211, 2024.

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.

Thomas Köhler, Michel Bätz, Farzad Naderi, André Kaup, Andreas Maier, and Christian Riess. Toward bridging the simulated-to-real gap: Benchmarking super-resolution on real data. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2944–2959, 2019.

Cansu Korkmaz, A Murat Tekalp, and Zafer Dogan. Training Generative Image Super-Resolution Models by Wavelet-Domain Losses Enables Better Control of Artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5926–5936, 2024.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, 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.

Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1929–1938, 2022.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

Zhiheng Li, Muheng Li, Jixuan Fan, Lei Chen, Yansong Tang, Jiwen Lu, and Jie Zhou. Learning Dual-Level Deformable Implicit Representation for Real-World Scale Arbitrary Super-Resolution. In *European Conference on Computer Vision*, pp. 352–368. Springer, 2024.

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*, 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 workshops*, pp. 136–144, 2017.

Ce Liu and Deqing Sun. On Bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):346–360, 2013.

Qingguo Liu, Chenyi Zhuang, Pan Gao, and Jie Qin. CDFormer: When Degradation Prediction Embraces Diffusion Model for Blind Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7455–7464, June 2024.

Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pp. 1040–1049, 2017.

Yihao Liu, Jingwen He, Jinjin Gu, Xiangtao Kong, Yu Qiao, and Chao Dong. Degae: A new pretraining paradigm for low-level vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23292–23303, 2023.

Seobin Park, Dongjin Kim, Sungyong Baik, and Tae Hyun Kim. Learning controllable degradation for real-world super-resolution via constrained flows. In *International Conference on Machine Learning*, pp. 27188–27203. PMLR, 2023.

Chengchao Qu, Ding Luo, Eduardo Monari, Tobias Schuchert, and Jürgen Beyerer. Capturing ground truth super-resolution data. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2812–2816. IEEE, 2016.

Abhisek Ray, Gaurav Kumar, and Maheshkumar H Kolekar. CFAT: Unleashing Triangular Windows for Image Super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26120–26129, 2024.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Haoze Sun, Wenbo Li, Jianzhuang Liu, Haoyu Chen, Renjing Pei, Xueyi Zou, Youliang Yan, and Yujiu Yang. Coser: Bridging image and language for cognitive super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25868–25878, 2024.

Wei Sun, Dong Gong, Qinfeng Shi, Anton van den Hengel, and Yanning Zhang. Learning to Zoom-In via Learning to Zoom-Out: Real-World Super-Resolution by Generating and Adapting Degradation. *IEEE Transactions on Image Processing*, 30:2947–2962, 2021.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pp. 1–21, 2024a.

Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10581–10590, 2021a.

Longguang Wang, Yingqian Wang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning a single network for scale-arbitrary super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4801–4810, 2021b.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

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*, pp. 1905–1914, 2021c.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. SinSR: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25796–25805, 2024b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Min Wei and Xuesong Zhang. Super-resolution neural operator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18247–18256, 2023.

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pp. 101–117. Springer, 2020.

Pengxu Wei, Ziwei Xie, Guanbin Li, and Liang Lin. Taylor Neural Network for Real-World Image Super-Resolution. *IEEE Transactions on Image Processing*, 32:1942–1951, 2023.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. 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.

Jie-En Yao, Li-Yuan Tsao, Yi-Chen Lo, Roy Tseng, Chia-Che Chang, and Chun-Yi Lee. Local implicit normalizing flow for arbitrary-scale image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1776–1785, 2023.

Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 701–710, 2018.

Huanjing Yue, Zhiming Zhang, and Jingyu Yang. Real-rawvsr: Real-world raw video super-resolution with a benchmark dataset. In *European Conference on Computer Vision*, pp. 608–624. Springer, 2022.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36: 13294–13307, 2023.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36, 2024.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. 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.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018a.

Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3762–3770, 2019.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2472–2481, 2018b.

Yuzhe Zhang, Jiawei Zhang, Hao Li, Zhouxia Wang, Luwei Hou, Dongqing Zou, and Liheng Bian. Diffusion-based Blind Text Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25827–25836, 2024.

APPENDIX

A THE CHALLENGES OF ASISR IN REAL-WORLD

A.1 REQUIREMENTS FOR IDEAL DATASETS

Arbitrary-scale super-resolution (SR) in real-world settings necessitates datasets with continuous magnification coverage across a specified range (e.g., 1.0× to 4.0×) at sub-pixel-level granularity (e.g., 0.01× intervals). Such fine-grained sampling is critical for modeling nonlinear texture variations, as even a narrow range like 3.0×-3.5× requires ≥50 distinct magnification levels. Simultaneously, all multi-scale images must exhibit strict spatiotemporal alignment: spatially, identical scene geometry, illumination, and optical centers must be preserved across magnifications to avoid misregistration artifacts; temporally, dynamic scenes demand microsecond-level synchronization to eliminate motion-induced misalignment, as even minor temporal offsets (e.g., 10 ms) can cause pixel-level displacements in fast-moving objects.

A.2 PHYSICAL LIMITATIONS IN DATA ACQUISITION

Single-lens optical zoom systems face inherent trade-offs: mechanical focal adjustments introduce time delays (100 ms–1 s per magnification change), causing content mismatch for dynamic scenes, while nonlinear distortion (e.g., barrel / pincushion effects) and discrete zoom steps (e.g., 0.5× increments) undermine sub-pixel alignment precision. Multi-camera arrays, though theoretically capable of parallel capture, become impractical due to hardware complexity—synchronizing 50 cameras with varying magnifications requires customized optics and triggers. Furthermore, minute installation errors ($< 0.1^\circ$ tilt) propagate as parallax distortions over distance, making cross-camera alignment infeasible. Dynamic scenes exacerbate these issues through conflicting exposure requirements: high magnifications (e.g., 4×) necessitate short exposure to mitigate motion blur but sacrifice Signal-to-Noise Ratio, while low magnifications tolerate longer exposure, creating inconsistent noise-motion coupling across scales.

A.3 THE LIMITATIONS OF EXISTING DATASETS

Currently, only two real-world datasets are specifically designed for arbitrary-scale image super-resolution (ASISR): COZ Fu et al. (2024) (CVPR 2024) and RealArbiSR Li et al. (2024) (ECCV 2024). While these datasets have advanced ASISR research in real-world scenarios, they exhibit limitations that hinder model generalization and practical applicability in complex environments.

COZ. COZ is the first real-world dataset for ASISR, constructed based on data generation strategies commonly used in synthetic settings Hu et al. (2019); Chen et al. (2021); Lee & Jin (2022); Chen et al. (2023a); Wei & Zhang (2023); Yao et al. (2023); Cao et al. (2023). These methods generate high and low resolution (HR-LR) image pairs of the same scene across fine-grained scaling factors within a given range. For example, given an HR image, a random scale factor s ($1 \leq s \leq 4$) is selected, and an HR patch of size $48s \times 48s$ is cropped and downsampled to 48×48, forming an LR-HR training pair.

To achieve this in real-world settings, COZ captures images of the same scene at different zoom levels using an optical zoom lens. A motor-driven mechanism precisely controls focal length adjustments to ensure fine-grained scaling. However, two key limitations remain. First, due to the nonlinear nature of focal length changes, the dataset exhibits a long-tailed scale distribution, with an overrepresentation of low-scale data and a scarcity of high-scale samples. This imbalance reduces model generalization at high scales, leading to unstable reconstruction performance. Simply discarding low-scale data is not an ideal solution, as these images are valuable for training, and their absence degrades overall model performance. Second, COZ is restricted to static scenes. Since multiple images at different focal lengths must be captured for the same scene, the content must remain unchanged. Any motion during zooming misaligns images, preventing the formation of valid LR-HR pairs. Consequently, dynamic scenes such as water waves and flying birds cannot be captured, limiting scene diversity and reducing real-world applicability.

RealArbiSR. RealArbiSR employs a data acquisition approach similar to COZ, using an optical zoom lens to capture the same scene at different scales. However, it suffers from two major draw-



Figure 5: The diverse and intricate scenes examples in our dataset.

backs. First, like COZ, it is limited to static scenes due to its reliance on optical zoom for HR-LR pair collection, restricting model generalization to real-world dynamic environments. Second, its scale distribution is sparse, containing only fixed magnifications ($1.5\times, 2\times, 2.5\times, 3\times, 3.5\times, 4\times$), in contrast to COZ's fine-grained scale variations.

Existing capture methods struggle to accommodate dynamic scenes, and increasing the number of cameras does not resolve this issue. Therefore, a more efficient data synthesis approach is necessary to overcome these limitations and improve ASISR model adaptability to real-world scenarios.

B Related Work

Single-Image Super-Resolution (SISR) Early approaches to SISR typically relied on handcrafted interpolation methods such as bilinear or bicubic interpolation, which often produced blurry edges and smoothed textures due to their inability to recover high-frequency details. SRCNN Dong et al. (2015) was one of the earliest methods to introduce convolutional neural networks (CNNs) into the SISR task, achieving an end-to-end mapping between low-resolution and high-resolution images. Subsequently, a series of CNN-based models, including VDSR Kim et al. (2016), EDSR Lim et al. (2017), and RDN Zhang et al. (2018b), further advanced the field by improving reconstruction accuracy.

With the recent success of transformer architectures in computer vision tasks, transformer-based SISR models such as SwinIR Liang et al. (2021) and HAT Chen et al. (2023b) have demonstrated remarkable performance on various benchmarks. In parallel, to enhance the perceptual quality of super-resolved images, generative adversarial network (GAN)-based models have also been explored. SRGAN Ledig et al. (2017) was the first GAN-based SR method, achieving visually pleasing results at the cost of a drop in PSNR. ESRGAN Wang et al. (2018) and Real-ESRGAN Wang et al. (2021c) subsequently improved perceptual quality and real-world performance through enhanced network designs and training strategies.

However, these methods are typically trained for a fixed set of upscaling factors (*e.g.*, $\times 2$, $\times 4$), and most of them adopt pixel-shuffle-based upsampling, making it difficult to handle arbitrary, especially fractional, scaling factors. To address these limitations, arbitrary-scale super-resolution (ASSR) has been proposed. ASSR aims to reconstruct HR images at arbitrary scales, including non-integer magnification factors, using a single unified model. MetaSR Hu et al. (2019) was the first neural network-based method to tackle this problem by employing meta-learning techniques to dynamically

predict the weights of the upsampling layers. LIIF Chen et al. (2021) introduced an implicit neural representation framework, treating images as continuous functions and predicting pixel values at arbitrary coordinates, which enables flexible scaling. Building upon this idea, several follow-up works such as LTE Lee & Jin (2022), CLIT Chen et al. (2023a), SRNO Wei & Zhang (2023), CiaoSR Cao et al. (2023), LMI Fu et al. (2024), and DDIR Li et al. (2024) have demonstrated the great potential of implicit representations for ASSR.

Real-World SISR Methods. Although remarkable progress has been made in the field of SR, most existing SR models are trained on synthetic datasets. Typically, LR images are generated by simply downsampling HR images using bicubic interpolation at predefined scaling factors. However, when these models are deployed in real-world scenarios, the super-resolved images often exhibit undesirable artifacts and ghosting effects. This performance degradation is primarily due to the fact that real-world degradations are far more complex than simple downsampling operations, making the learned SR models insufficient for handling practical conditions Fu et al. (2024); Li et al. (2024). To mitigate this gap, several works Wang et al. (2021c); Liu et al. (2023); Park et al. (2023) have attempted to design more sophisticated degradation models to enhance the model's ability to simulate realistic degradation processes. Additionally, some studies Liu et al. (2024); Sun et al. (2021) enhance the stability of SR results by enabling the model to identify degradation features. Meanwhile, other researchers Cai et al. (2019); Chen et al. (2019) collect real-world images captured by cameras to construct real-world SR datasets.

In the context of ASSR, several attempts have been made to address the data limitations. InterFlow Park et al. (2023), for instance, generates fine-grained LR training images by blending latent features of LR and HR images. However, this method relies on temporally and spatially aligned LR-HR image pairs, and due to its use of weighted blending for latent features, the generated results fail to fully align with the degradation trends of real-world data. The COZ dataset Fu et al. (2024) is the first work to collect real-world ASSR data by capturing the same scene at different zoom levels with a variable focal length lens and aligning them through cropping. While effective to some extent, the COZ dataset suffers from a long-tail distribution, with a surplus of low-scale images and a scarcity of high-scale samples. Directly discarding low-scale images is undesirable, as they still contain valuable degradation patterns. Similarly, the RealArbiSR dataset Li et al. (2024) provides LR images at several fixed scales but lacks fine-grained, continuous-scale coverage.

In addition, these real-world datasets face several practical limitations. Due to the time required for lens zooming, dynamic objects such as flowing water, animals, and pedestrians cannot be captured consistently across scales. Furthermore, optical distortions caused by lens zooming make it impossible to achieve perfect spatial alignment between images captured at different focal lengths. This misalignment accumulates over varying scales, often causing SR models to produce overly smoothed results, leading to detail loss. Moreover, whether using multiple cameras or beam-splitters, existing dataset collection strategies still fall short of providing temporally and spatially aligned image sets comparable to those generated by synthetic pipelines.

Diffusion Models in SISR. Recently, diffusion models have shown impressive capabilities in various image generation tasks, including image SR. These models gradually add noise to an image and learn to reverse this process through iterative denoising, enabling the progressive recovery of high-frequency details. Early works such as SR3 Saharia et al. (2022) and SRDiff Li et al. (2022) leverage iterative denoising processes to progressively recover high-frequency details from noisy low-resolution inputs, achieving state-of-the-art perceptual quality. Subsequent methods integrate conditional information, like degradation priors Wu et al. (2024) or text prompts Sun et al. (2024), to guide the reverse diffusion process, improving both fidelity and controllability. Despite their promising results, diffusion-based SR models often suffer from high computational costs due to their iterative nature, motivating recent studies Yue et al. (2023); Wang et al. (2024b) to explore accelerated sampling strategies.

C Details of Data Synthesis

C.1 Dataset Introduction

We collect 588 copyright-free photographic images from the internet, all licensed under CC0, and will publicly disclose the data sources to ensure reproducibility and compliance. Compared with



Figure 6: Examples of training data from our synthesized dataset. The eleven images are evenly selected from 100 images in a training scene. The first row simulates scene variations during the data acquisition process, while the second row shows the actual training data.

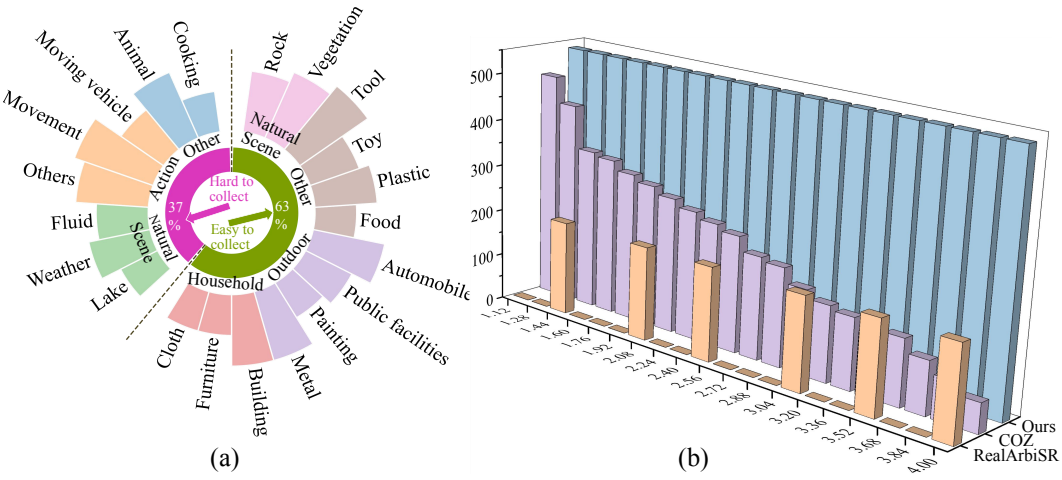

Figure 7: Dataset statistics. (a) Scene-category distribution of our synthesized dataset. (b) Distribution of magnification scales ($\times 1.0$–$\times 4.0$) in our dataset, COZ Fu et al. (2024), and RealArbiSR Li et al. (2024), where our dataset shows a more uniform scale distribution.

existing ASSR datasets such as COZ Fu et al. (2024) and RealArbiSR Li et al. (2024), our dataset additionally includes scene types that are absent from these datasets. Representative examples are shown in Fig. 5.

For data partitioning, 535 scenes are randomly selected for training, while the remaining 53 scenes are used for testing. For each training scene, we synthesize 100 LR images with scale factors uniformly sampled from $\times 1.0$ to $\times 4.0$. This dense sampling strategy allows the model to observe continuous degradation variations across scales during training. In contrast, for the test set, we only synthesize images at five discrete scale factors, i.e., $\times 2$, $\times 3$, $\times 4$, $\times 5$, and $\times 6$, following a standard evaluation protocol.

Our synthesized dataset also covers diverse scene categories and provides a more uniform scale distribution than existing datasets. These properties make it suitable for training and evaluating arbitrary-scale super-resolution models under broader scene and scale settings. Figure 6 illustrates the per-scene synthesis sequence, while Fig. 7 summarizes the scene categories and scale distributions.

C.2 Degradation Embedding

Before synthesizing high-quality data, we construct a Degradation Embedding, which serves as auxiliary information alongside the original image for generating LR images with specific degradation patterns. To build this embedding, we design a degradation feature extractor inspired by previous studies Liu et al. (2023); Chen et al. (2024). Specifically, we adopt ResNet-50 as the feature extraction backbone and employ an MLP to reduce the dimensionality to 512.

Unlike existing methods, we observe that degradation features of similar scale factors exhibit strong similarities, making models less sensitive to minor variations. To address this, a Rank-based Learning strategy is introduced to enhance the model's ability to distinguish subtle degradation patterns.

We have validated the effectiveness of this training strategy in the main text. Here, we additionally provide experimental results showing that when DFE is trained with sampling granularities of 10, 25, and 35, and used for dataset synthesis, the PSNR of DP-ArbSR on the COZ test set for $\times 2$, $\times 3$, and $\times 4$ decreases by 0.33/0.42/0.41, 0.19/0.23/0.16, and 0.07/0.08/0.10, respectively.

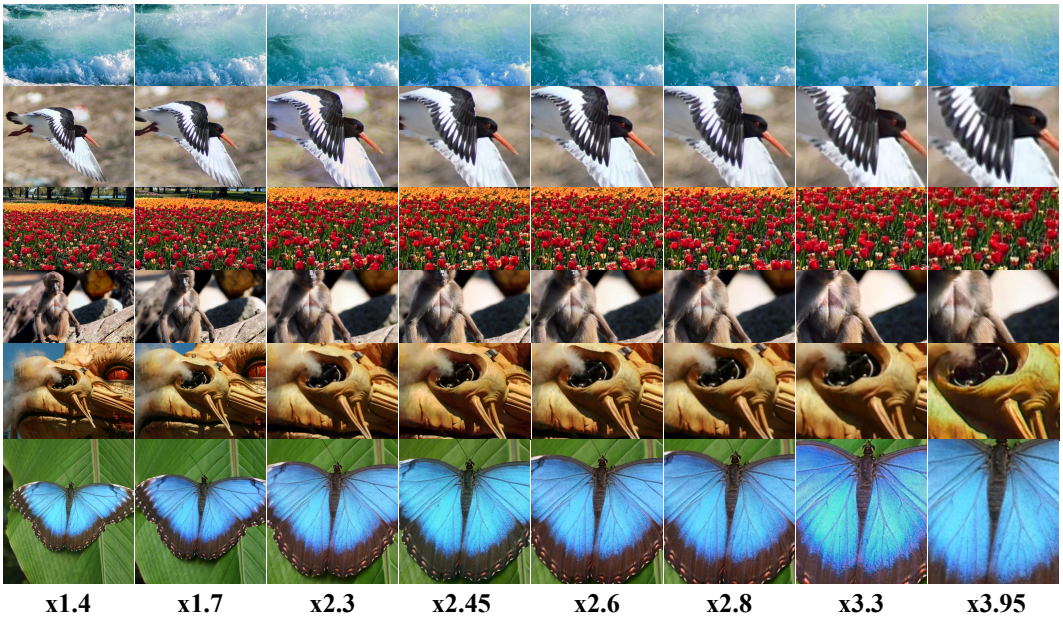

Figure 8: Additional examples of training data sequences from our dataset. To provide a more intuitive understanding, the images illustrate the degradation process of image quality as the simulated scene is progressively enlarged from left to right.

D Experiment

D.1 Details of Degraded Image Synthesis

Datasets. We use the V2 version of the COZ dataset Fu et al. (2024), which contains 264 training scenes (14,794 images) and 37 test scenes, to train the degradation feature extractor and the degraded image synthesis network. For training the DFE, 90% of the training scenes are used, with the remaining scenes reserved for testing.

Implementation Details.

Degradation Feature Extractor. To train the degradation feature extractor, we employ a ranking-based learning strategy. For each scene, we randomly select two images, denoted as LR and HR. A 128×128 patch is cropped from the LR image, and a corresponding patch is extracted from the HR image. Both patches are then input into the DFE with a MLP for ranking prediction. The model is evaluated using the Spearman rank correlation coefficient (SROCC), calculated as:

$$SROCC = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{14}$$

where d_i is the ranking difference for each sample pair, and n is the total number of samples. The pre-trained SRGAN weights are frozen, and the model is trained for 500 epochs with an Adam optimizer and a learning rate of 1×10^{-4}. The margin in the margin-ranking loss is set to 0.5, and the MLP is removed after training.

Image Synthesis Network. For the image synthesis network, a pair of patches is extracted from the HR and LR images, with the LR patch of size 48×48 serving as the label. The model is trained

for 300 epochs with the Adam optimizer, at a learning rate of 1×10^{-4}, and evaluated using cosine similarity defined in Eq. (8). It should be noted that this step is intended to enable the image synthesis network to learn how to utilize degradation embeddings. After training, paired LR-HR images are no longer required during the actual image synthesis process.

D.2 Details of SR Network

Datasets. We conduct experiments using three datasets: a modified fine-grained COZ dataset Fu et al. (2024) corrected by the image synthesis network, a synthetic dataset generated entirely by this network, and a mixed dataset that combines the original COZ dataset, the RealArbiSR dataset Li et al. (2024), and the synthetic dataset. Performance is evaluated on two real-world arbitrary-scale super-resolution datasets—the COZ test set and the RealArbiSR dataset, both from the V2 version—as well as on our synthesized test scenarios.

Implementation Details. We follow the experimental setup from previous works Chen et al. (2021); Fu et al. (2024); Li et al. (2024), training the network with the L1 loss function and optimizing with the Adam optimizer, with an initial learning rate of 1×10^{-4}. During training, a 48×48 patch is cropped from the LR image and the corresponding HR patch serves as the label. Models are trained for 400 epochs, with the learning rate halved at epoch 200. For other models Chen et al. (2021); Lee & Jin (2022); Cao et al. (2023); Chen et al. (2023a); Fu et al. (2024); Li et al. (2024); Wei & Zhang (2023), we adhere to their original experimental configurations. Peak Signal-to-Noise Ratio (PSNR) is used as the evaluation metric for all models.

D.3 More Quantitative Experiments

Complete Per-Model Results. The main paper reports mean PSNR over five methods for the fine-grained COZ augmentation and collected-scene synthesis experiments. Tables 6 and 7 provide the complete per-model and per-scale results used to compute those summaries.

Impact of Removing Low-Magnification Data. Low magnifications are removed to assess the impact on model performance. Specifically, LR images below $\times 1.5$ are excluded. LIIF Chen et al. (2021), LTE Lee & Jin (2022), LMI Fu et al. (2024), DDIR Li et al. (2024), and our DP-ArbSR model are trained on both the original COZ dataset Fu et al. (2024) and the filtered version, then evaluated on the COZ and RealArbiSR Li et al. (2024) test sets. As shown in Tab. 5, while removing low magnifications balanced the data distribution, performance decreased at $\times 2$ and $\times 3$, with minimal improvement at higher magnifications.

More Evaluation Metrics. We utilize additional metrics, Structural Similarity Index (SSIM) Wang et al. (2004) and Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. (2018a), to assess the overall effectiveness of our approach. We train eight models, including LIIF Chen et al. (2021), LTE Lee & Jin (2022), SRNO Wei & Zhang (2023), LIT Chen et al. (2023a), CiaoSR Cao et al. (2023), LMI Fu et al. (2024), DDIR Li et al. (2024), and our model, on a mixed dataset that combines the COZ training set, the RealArbiSR dataset, and our synthesized data from the previous section. We then test these models on the COZ dataset and our synthesized scenes. The experimental results are shown in Tab. 4, demonstrating that our method outperforms others on two additional evaluation metrics.

Comparison with InterFlow. InterFlow Park et al. (2023) is a representative real-world synthesis method that generates intermediate-scale training samples by interpolating latent codes between two captured images at different magnifications. To make the comparison as fair as possible, we construct two synthesized training sets on the RealArbiSR Li et al. (2024) training split. For InterFlow, we use the $\times 2$ and $\times 4$ images in each scene as the two endpoints and synthesize training samples covering the scale range from $\times 2$ to $\times 4$. For our method, we use the same $\times 2$ images as inputs and synthesize the same number of samples over the identical scale range, while using degradation embeddings extracted from COZ Fu et al. (2024) as priors. We then train LIIF Chen et al. (2021), LMI Fu et al. (2024), DDIR Li et al. (2024), and DP-ArbSR on these two synthesized datasets and evaluate them on the RealArbiSR test set using PSNR. As shown in Tab. 8, models trained on our synthesized data consistently outperform those trained on InterFlow-synthesized data by about 0.9 dB on average. We further inspect the RealArbiSR training pairs and find that some $\times 2$/$\times 4$ image pairs are not strictly aligned. Since InterFlow blends latent codes from both images,

Table 4: Quantitative comparison (SSIM ↑ Wang et al. (2004) / LPIPS ↓ Zhang et al. (2018a)) with SOTA methods on COZ and our synthesized test scenarios.

Methods	COZ Fu et al. (2024)				
	In-scale			Out-of-scale	
	×2	×3	×4	×5	×6
EDSR-LIIF Chen et al. (2021)	0.814/0.364	0.767/0.474	0.741/0.549	0.720/0.581	0.705/0.611
EDSR-LTE Lee & Jin (2022)	0.811/0.370	0.763/0.480	0.739/0.552	0.721/0.594	0.704/0.634
EDSR-SRNO Wei & Zhang (2023)	0.816/0.371	0.761/0.483	0.732/0.551	0.726/0.588	0.705/0.618
EDSR-LIT Chen et al. (2023a)	0.818/0.374	0.770/0.486	0.733/0.560	0.725/0.597	0.703/0.629
EDSR-CiaoSR Cao et al. (2023)	0.818/0.373	0.769/0.496	0.736/0.572	0.727/0.606	0.704/0.637
EDSR-LMI Fu et al. (2024)	0.819/0.382	0.771/0.494	0.738/0.572	0.728/0.607	0.708/0.635
EDSR-DDIR Li et al. (2024)	0.820/0.378	0.772/0.506	0.740/0.584	0.729/0.629	0.708/0.663
RDN-LIIF Chen et al. (2021)	0.816/0.337	0.770/0.453	0.746/0.525	0.725/0.547	0.708/0.581
RDN-LTE Lee & Jin (2022)	0.816/0.343	0.768/0.461	0.744/0.533	0.724/0.556	0.706/0.590
RDN-SRNO Wei & Zhang (2023)	0.821/0.342	0.771/0.465	0.735/0.537	0.728/0.560	0.708/0.594
RDN-LIT Chen et al. (2023a)	0.823/0.346	0.765/0.469	0.736/0.539	0.728/0.559	0.709/0.594
RDN-CiaoSR Cao et al. (2023)	0.820/0.352	0.767/0.467	0.744/0.541	0.731/0.563	0.708/0.601
RDN-LMI Fu et al. (2024)	0.823/0.361	0.774/0.478	0.746/0.549	0.733/0.578	0.713/0.604
RDN-DDIR Li et al. (2024)	0.824/0.370	0.775/0.491	0.745/0.562	0.734/0.591	0.714/0.633
DP-ArbSR (ours)	**0.828/0.313**	**0.779/0.421**	**0.749/0.493**	**0.736/0.519**	**0.716/0.566**

Methods	Our Synthesized Test Scenarios				
	In-scale			Out-of-scale	
	×2	×3	×4	×5	×6
EDSR-LIIF Chen et al. (2021)	0.856/0.216	0.760/0.322	0.694/0.433	0.639/0.517	0.622/0.543
EDSR-LTE Lee & Jin (2022)	0.856/0.217	0.762/0.323	0.696/0.435	0.641/0.519	0.623/0.544
EDSR-SRNO Wei & Zhang (2023)	0.857/0.217	0.765/0.324	0.698/0.437	0.645/0.521	0.624/0.547
EDSR-LIT Chen et al. (2023a)	0.860/0.221	0.770/0.325	0.701/0.440	0.649/0.525	0.628/0.548
EDSR-CiaoSR Cao et al. (2023)	0.862/0.223	0.774/0.325	0.706/0.441	0.653/0.527	0.632/0.551
EDSR-LMI Fu et al. (2024)	0.866/0.226	0.779/0.328	0.709/0.443	0.658/0.528	0.635/0.554
EDSR-DDIR Li et al. (2024)	0.867/0.227	0.778/0.328	0.710/0.445	0.658/0.527	0.636/0.556
RDN-LIIF Chen et al. (2021)	0.866/0.192	0.774/0.296	0.712/0.416	0.647/0.498	0.635/0.527
RDN-LTE Lee & Jin (2022)	0.867/0.194	0.775/0.299	0.711/0.417	0.656/0.497	0.637/0.530
RDN-SRNO Wei & Zhang (2023)	0.870/0.195	0.778/0.302	0.714/0.421	0.660/0.501	0.640/0.532
RDN-LIT Chen et al. (2023a)	0.872/0.197	0.780/0.302	0.712/0.420	0.661/0.499	0.639/0.530
RDN-CiaoSR Cao et al. (2023)	0.872/0.198	0.783/0.307	0.714/0.422	0.663/0.500	0.640/0.532
RDN-LMI Fu et al. (2024)	0.875/0.200	0.788/0.312	0.718/0.427	0.665/0.503	0.643/0.538
RDN-DDIR Li et al. (2024)	0.877/0.201	0.785/0.310	0.717/0.423	0.666/0.501	0.644/0.538
DP-ArbSR (ours)	**0.880/0.172**	**0.789/0.281**	**0.720/0.394**	**0.671/0.472**	**0.648/0.508**

this misalignment is propagated to the synthesized samples and becomes more noticeable at larger scales, as illustrated in Fig. 9. In contrast, our method only requires a single reference LR image to synthesize new training samples, avoiding the sensitivity to pairwise alignment and leading to more stable degradation modeling.

Model Parameters and Inference Speed. As shown in Tab. 9, we present the model parameters of several representative EDSR-based ASSR methods used in previous sections, along with those of our proposed model. In addition, we measure the inference time of each method on the COZ X4 test set using a single NVIDIA RTX 3090 GPU. Although our model incorporates a diffusion module, resulting in a relatively larger number of parameters, it avoids the use of computationally expensive self-attention mechanisms in the upsampling stage. Consequently, our method still achieves competitive inference speed.

Table 5: Quantitative comparison of PSNR (dB) results, using LIIF Chen et al. (2021), LTE Lee & Jin (2022), LMI Fu et al. (2024), DDIR Li et al. (2024), and our DP-ArbSR model trained on the original COZ data and COZ data with LR images below $\times 1.5$ removed.

Dataset	Methods	COZ Fu et al. (2024)					RealArbiSR Li et al. (2024)		
		Scale							
		$\times 2$	$\times 3$	$\times 4$	$\times 5$	$\times 6$	$\times 2$	$\times 3$	$\times 4$
Original COZ	LIIF	28.72	26.52	25.14	24.23	23.12	29.84	27.02	25.35
	LTE	28.73	26.52	25.12	24.23	23.11	29.93	27.04	25.39
	LMI	28.77	26.57	25.17	24.23	23.22	29.87	27.08	25.43
	DDIR	28.76	26.60	25.15	24.25	23.14	29.96	27.13	25.57
	DP-ArbSR (ours)	28.92	26.74	25.29	24.45	23.35	30.02	27.19	25.65
Remove Low-Magnification COZ Data	LIIF	28.65	26.51	25.12	24.23	23.13	29.80	27.03	25.35
	LTE	28.68	26.53	25.12	24.25	23.12	29.91	27.03	25.39
	LMI	28.70	26.52	25.14	24.33	23.23	29.83	27.04	25.42
	DDIR	28.65	26.65	25.15	24.26	23.17	29.94	27.11	25.46
	DP-ArbSR (ours)	28.89	26.73	25.30	24.47	23.35	29.98	27.17	25.66

Table 6: Complete PSNR (dB) comparison of LIIF Chen et al. (2021), LTE Lee & Jin (2022), LMI Fu et al. (2024), DDIR Li et al. (2024), and DP-ArbSR trained on the original COZ dataset, DegAE-synthesized fine-grained COZ data, and our method-synthesized fine-grained COZ data, evaluated on COZ and RealArbiSR.

Synthesis method	Methods	COZ Fu et al. (2024)					RealArbiSR Li et al. (2024)		
		Scale							
		$\times 2$	$\times 3$	$\times 4$	$\times 5$	$\times 6$	$\times 2$	$\times 3$	$\times 4$
Original COZ	LIIF	28.72	26.52	25.14	24.23	23.12	29.84	27.02	25.35
	LTE	28.73	26.52	25.12	24.23	23.11	29.93	27.04	25.39
	LMI	28.77	26.57	25.17	24.23	23.22	29.87	27.08	25.43
	DDIR	28.76	26.60	25.15	24.25	23.14	29.96	27.13	25.57
	DP-ArbSR (ours)	28.92	26.74	25.29	24.45	23.35	30.02	27.19	25.65
DegAE	LIIF	28.74	26.61	25.17	24.23	23.16	29.88	27.08	25.40
	LTE	28.75	26.59	25.16	24.23	23.16	29.98	27.07	25.43
	LMI	28.79	26.61	25.18	24.24	23.27	29.91	27.11	25.47
	DDIR	28.79	26.65	25.20	24.31	23.20	30.00	27.16	25.61
	DP-ArbSR (ours)	28.94	26.78	25.36	24.51	23.39	30.04	27.24	25.69
Ours	LIIF	28.77	26.66	25.22	24.38	23.21	29.91	27.12	25.45
	LTE	28.81	26.66	25.24	24.36	23.20	30.02	27.12	25.47
	LMI	28.82	26.69	25.26	24.43	23.33	29.97	27.14	25.49
	DDIR	28.85	26.70	25.27	24.38	23.26	30.04	27.19	25.65
	DP-ArbSR (ours)	28.98	26.83	25.41	24.54	23.42	30.09	27.29	25.72

D.4 More Ablation Experiments

Rank-Based Learning. The main paper reports a compact ablation to verify the benefit of our rank-based learning strategy for degradation representation learning. Here we provide a more detailed comparison. We consider the original pretrained degradation extractor DegAE-O, its rank-based retrained variant DegAE-RL, our final DFE, and three DFE variants trained with different scale sampling granularities. Specifically, DFE-10, DFE-25, and DFE-35 are trained using 10, 25, and 35 LR images sampled as uniformly as possible over the scale range within each scene, respectively.

Following the protocol in the main paper, we compute the average cosine similarity between degradation embeddings extracted from adjacent scales within the same scene on the COZ test set. Lower similarity indicates that the learned embeddings are more sensitive to subtle degradation differences

Table 7: Complete PSNR (dB) comparison of LIIF Chen et al. (2021), LTE Lee & Jin (2022), LMI Fu et al. (2024), DDIR Li et al. (2024), and DP-ArbSR trained on data synthesized by bicubic downsampling, DegAE Liu et al. (2023), and our method on the collected scenes, evaluated on COZ and RealArbiSR.

Synthesis method	Methods	COZ Fu et al. (2024)					RealArbiSR Li et al. (2024)		
		Scale							
		×2	×3	×4	×5	×6	×2	×3	×4
Bicubic	LIIF	28.01	25.98	24.58	23.63	22.56	28.67	26.15	24.80
	LTE	28.03	26.01	24.59	23.86	22.74	28.65	26.14	24.77
	LMI	28.21	26.14	24.60	23.82	22.67	28.73	26.24	24.74
	DDIR	28.04	26.08	24.60	23.75	22.68	28.58	26.10	24.76
	DP-ArbSR (ours)	28.22	26.15	24.70	23.86	22.72	28.85	26.18	24.79
DegAE	LIIF	28.15	26.12	24.67	23.75	22.71	28.80	26.29	24.89
	LTE	28.15	26.13	24.70	23.95	22.88	28.75	26.25	24.87
	LMI	28.30	26.26	24.76	23.94	22.81	28.99	26.31	24.82
	DDIR	28.20	26.24	24.71	23.90	22.80	28.70	26.21	24.85
	DP-ArbSR (ours)	28.36	26.33	24.82	24.02	22.89	29.06	26.37	24.93
Ours	LIIF	28.72	26.51	25.01	24.15	23.02	29.85	26.92	25.34
	LTE	28.72	26.53	24.98	24.24	23.04	29.94	27.03	25.38
	LMI	28.75	26.59	25.15	24.31	23.18	29.85	27.10	25.40
	DDIR	28.74	26.51	25.07	24.24	23.10	29.97	27.12	25.55
	DP-ArbSR (ours)	28.87	26.69	25.22	24.38	23.30	30.01	27.19	25.63

Table 8: Quantitative comparison (PSNR (dB)) of models trained on InterFlow-synthesized and our synthesized datasets constructed from the RealArbiSR training split, evaluated on the RealArbiSR test set.

Synthesis method	Methods	RealArbiSR Li et al. (2024)		
		×2	×3	×4
InterFlow Park et al. (2023)	LIIF Chen et al. (2021)	28.98	26.08	24.40
	LMI Fu et al. (2024)	29.06	26.15	24.50
	DDIR Li et al. (2024)	29.14	26.23	24.60
	DP-ArbSR (ours)	29.19	26.27	24.66
Ours	LIIF Chen et al. (2021)	29.83	27.00	25.38
	LMI Fu et al. (2024)	29.89	27.08	25.47
	DDIR Li et al. (2024)	29.96	27.15	25.58
	DP-ArbSR (ours)	**30.02**	**27.21**	**25.65**

across scales. As shown in Tab. 10, introducing rank-based learning reduces the similarity from 0.99 to 0.89, indicating improved discrimination between closely related degradations.

Moreover, increasing the sampling granularity progressively improves feature discrimination, with the similarity decreasing from 0.92 to 0.86 and 0.79 as the number of sampled LR images increases from 10 to 25 and 35. Our final DFE achieves the lowest similarity of 0.73, highlighting the benefit of the proposed training strategy.

Qualitative Ablation. Figure 10 complements the quantitative ablation in the main paper. Removing the texture prior weakens fine textures, removing degradation injection introduces visible blur, and removing DAFF impairs local texture reconstruction.

D.5 More Qualitative Experiments

Testing set. We further compare the visual results of two additional state-of-the-art methods—LIT Chen et al. (2023a) and CiaoSR Cao et al. (2023)—across three evaluation benchmarks: the COZ

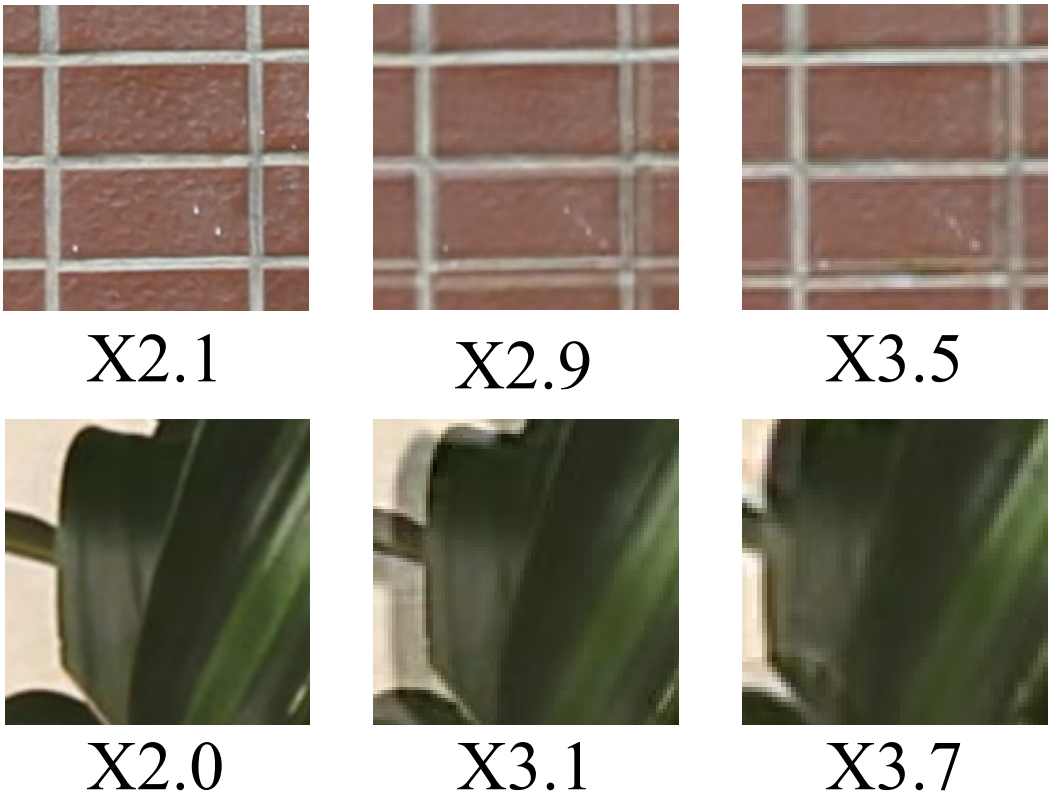

Figure 9: Examples of the alignment issue in RealArbiSR Li et al. (2024). Slight misregistration between the $\times 2$ and $\times 4$ images of the same scene is amplified when InterFlow interpolates their latent codes, leading to inconsistent structures in the synthesized samples.

Table 9: Model parameters and inference time, measured on the COZ $\times 4$ test set using a single NVIDIA RTX 3090 GPU. LMI, DDIR, and LIT use the EDSR backbone.

Model	LMI Fu et al. (2024)	DDIR Li et al. (2024)	LIT Chen et al. (2023a)	DP-ArbSR
Parameters	1.31M	3.15M	6.55M	181.2M
Inference Time	4m25s	7m25s	10m32s	5m49s

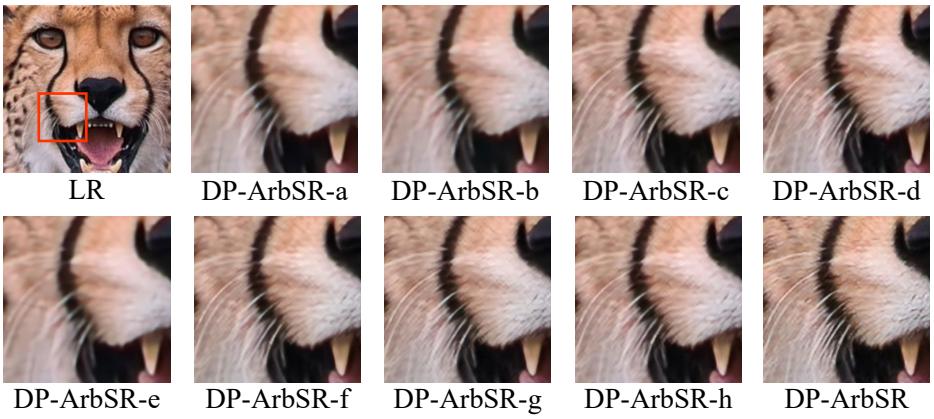

Figure 10: Qualitative ablation of DP-ArbSR on our synthesized test set at $\times 4$.

Table 10: Detailed ablation on rank-based learning and scale sampling granularity. Similarity denotes the average cosine similarity between adjacent-scale degradation embeddings on the COZ test set, where lower values indicate better discrimination.

Metric	Rank-based learning			Sampling granularity		
	DegAE-O	DegAE-RL	DFE	DFE-10	DFE-25	DFE-35
Sim. ↓	0.99	0.89	0.73	0.92	0.86	0.79

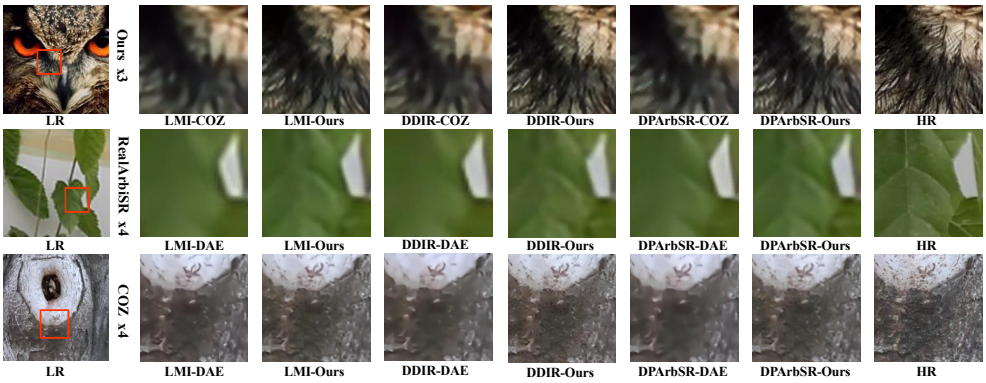

Figure 11: Qualitative comparison of LMI Fu et al. (2024), DDIR Li et al. (2024), and our DP-ArbSR across the synthesized, RealArbiSR, and COZ test settings.

test set, RealArbiSR test set, and our synthesized test set. Beyond the two test cases discussed in the main text, we extend the analysis to four additional experimental configurations: (1) models trained on DegAE-generated dataLiu et al. (2023) and our synthesized data evaluated on our synthesized test set; (2) models trained on COZ and our synthesized data tested on the RealArbiSR test set; (3) models trained on COZ and synthesized data assessed on the COZ test set; and (4) models trained on DegAE and synthesized data evaluated on the COZ test set. An overview is provided in Fig. 11, with the extended comparisons shown in Fig. 18.

Experimental results demonstrate that models trained on our synthesized data significantly outperform those trained exclusively on COZ data when tested on our synthesized test set, exhibiting richer texture details and superior sharpness. These models also achieve marginally better performance on the RealArbiSR test set. Notably, even when evaluated on the original COZ test set, our synthesized data-trained models surpass COZ-trained counterparts in scenarios underrepresented in the COZ training data, while maintaining comparable performance in scenarios where COZ training samples are abundant. This advantage stems from the larger volume of training data in our synthesized dataset compared to COZ.

Crucially, models trained on our synthesized data consistently outperform DegAE-trained models across all test sets. This validates that our degradation synthesis method more effectively captures real-world degradation characteristics than DegAE, enabling visually discernible improvements in output quality. Visual comparisons confirm that our approach better preserves fine structures while suppressing artifacts, with differences readily apparent in side-by-side evaluations.

Generalizability. To evaluate the generalization of our synthetic data trained model, we capture images with the Sony A7R2, Panasonic S5M2, and iPhone 12 Pro. We train SRNO Wei & Zhang (2023), LMI Fu et al. (2024) and DDIR Li et al. (2024) with real data, data from the DegAE degradation method, and our synthetic data. Then we compare them to DP-ArbSR trained on our synthetic data. We provide four additional test scenarios for comparison using the previously mentioned models. The results are shown in Fig. 19, Fig. 20 and Fig. 21.

Findings further demonstrate that models trained on our synthesized data generalize well to real-world scenarios and outperform those trained on DegAE-generated data. In some cases, our models achieve visual quality identical to real data and perform better in scenarios where real data is absent.

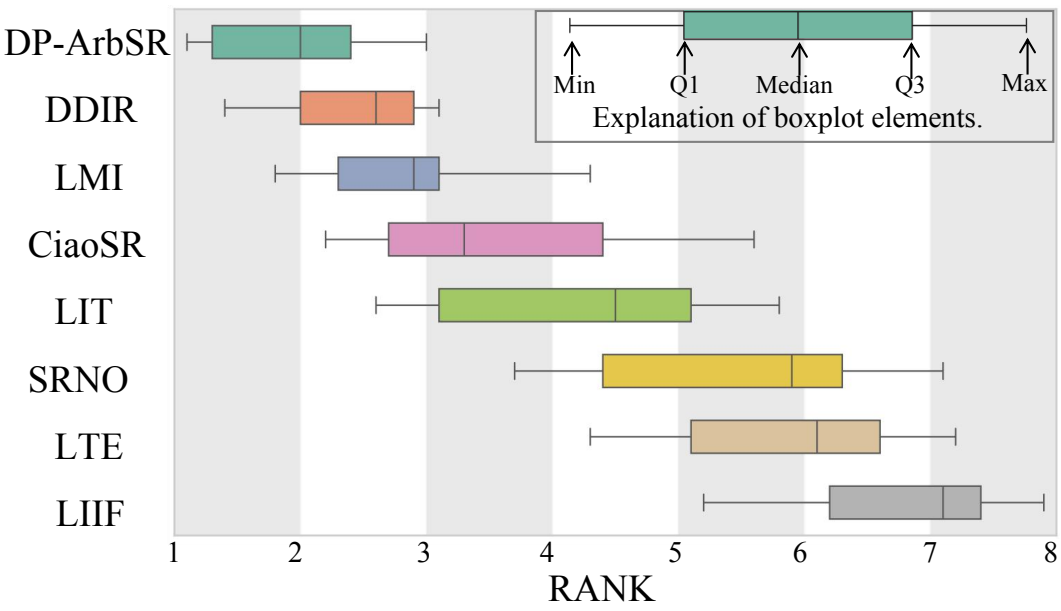

Figure 12: User Study 2: Subjective rankings of seven existing models and our proposed method at ×4 magnification. Users ranked the models from 1 (best) to 8 (worst) based on image quality, considering texture preservation, artifact suppression, and content fidelity.

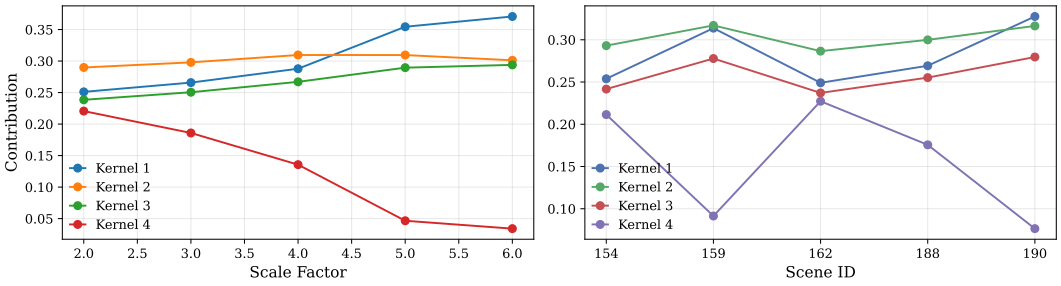

Figure 13: Effect of the scaling factor and degradation embedding on the contribution of different kernels in the kernel pool.

D.6 User Study

We conduct two user studies to further validate the real-world generalization of our synthesized data and the perceptual quality of DP-ArbSR. A total of 20 participants are invited. To build the test set, we capture 20 real-world scenes using three devices, namely Sony A7R2, Panasonic S5M2, and iPhone 12 Pro, including 10 static scenes and 10 dynamic scenes. These self-captured scenes are used solely as stimuli for perceptual evaluation and are not part of our contributed synthesized dataset. Because everyday scenes may contain private or identifiable content, this user-study stimulus set will not be publicly redistributed.

In the first study, we evaluate whether our synthesized data can better support real-world arbitrary-scale SR than the original COZ training data. For each method, we compare the results of the model trained on COZ with those of the same model trained on our synthesized data at $\times 4$ and $\times 6$ magnifications. Participants are asked to select their preferred result for each image pair separately at the two magnifications. The results in Fig. 17 show that models trained on our synthesized data receive more preference votes at both $\times 4$ and $\times 6$. We further separate the results on dynamic scenes, where the advantage of our synthesized data becomes more evident, indicating that our data better covers real-world content that is difficult to collect in existing datasets.



Figure 14: Visualization of intermediate results, including the texture prior in the encoder and the local spatial attention weight map. The orientation of the spatial attention map closely aligns with the actual texture structure.

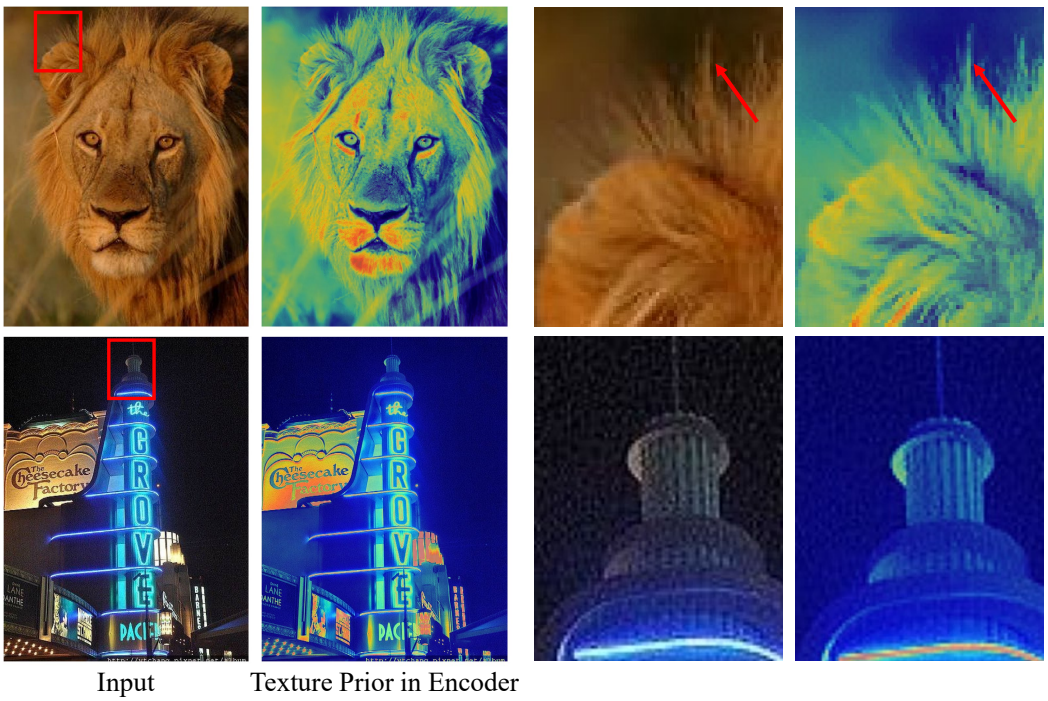

Figure 15: Visualization results of the texture prior in the encoder, which can provide clearer texture features and features with less noise for the encoder.

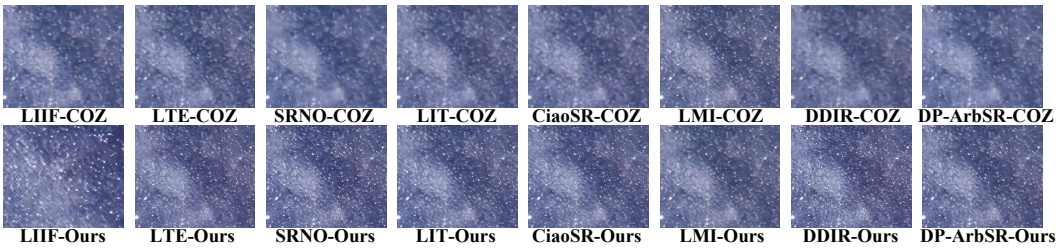

Figure 16: One of the samples in the user study comparison chart.

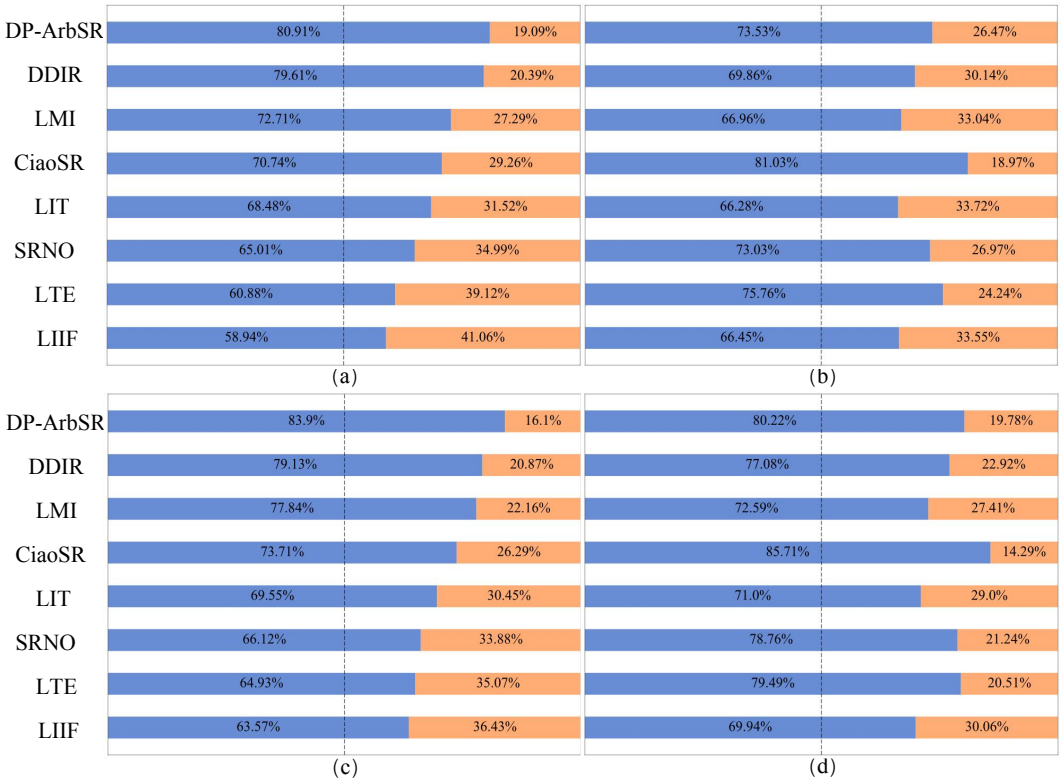

Figure 17: User Study 1: preference comparison between models trained on our synthesized data and those trained on the COZ dataset. The blue bars on the left denote the votes for models trained on our synthesized data. The top row shows the results at $\times 4$, and the bottom row shows the results at $\times 6$. Figures (a) and (c) report the preference votes on all 20 real-world scenes, while Figures (b) and (d) report the preference votes on the 10 dynamic scenes. Models trained on our synthesized data are preferred more often at both magnifications, with a clearer advantage on dynamic scenes.

In the second study, we compare seven baseline methods Chen et al. (2021); Lee & Jin (2022); Wei & Zhang (2023); Chen et al. (2023a); Cao et al. (2023); Fu et al. (2024); Li et al. (2024) together with our DP-ArbSR at $\times 4$ magnification. Participants rank the visual quality of the outputs from 1 (best) to 8 (worst) according to texture fidelity, artifact suppression, and content fidelity. As shown in Fig. 12, our method achieves the best overall ranking, demonstrating superior perceptual quality in real-world scenarios.

D.7 Analysis of Kernel Pool Contributions

To further investigate the effectiveness of the kernel pool design, we analyze how the contributions of individual kernels vary with scale and scene content. First, we extract degradation embeddings from the same scene image in the COZ test set under different scaling factors and visualize the

corresponding kernel weights in the image synthesis network. As shown in Fig. 13, the contributions of different kernels change smoothly as the scale increases, indicating that the kernel pool can adapt continuously to scale variation rather than relying on a fixed degradation pattern. In particular, some kernels, such as $kernel_1$ and $kernel_4$, are more sensitive to scale changes, suggesting that different kernels specialize in different degradation characteristics.

We further analyze the kernel weights of five scenes with different color, brightness, and texture characteristics at a fixed $\times 4$ scale. As shown in Fig. 13, although several kernels exhibit relatively consistent trends across scenes, the contributions of some kernels vary noticeably with scene content. Through visual inspection, we observe that scenes No.159, No.188, and No.190 contain relatively higher noise levels compared with the other scenes. Correspondingly, the weight of $kernel_4$ decreases more sharply in these scenes. This phenomenon suggests that $kernel_4$ may be related to noise-related degradation, and its contribution is adjusted according to the noise level in the scene. These observations indicate that the proposed kernel pool is influenced by both scaling factors and degradation characteristics, enabling more adaptive synthesis of degraded images across different scenes.

D.8 Intermediate Result Visualization

We visualize the texture priors processed by the encoder, as well as the weight maps used in local feature fusion. The texture prior is averaged along the channel dimension for visualization, while the local weight map employs color intensity to indicate each pixel's contribution. As shown in Fig. 14 and Fig. 15, the model effectively identifies texture orientation and dynamically adjusts fusion weights during local feature integration. Additionally, compared to the original input, the introduced texture prior exhibits reduced noise and clearer structural details, which facilitates more effective feature extraction by the encoder.

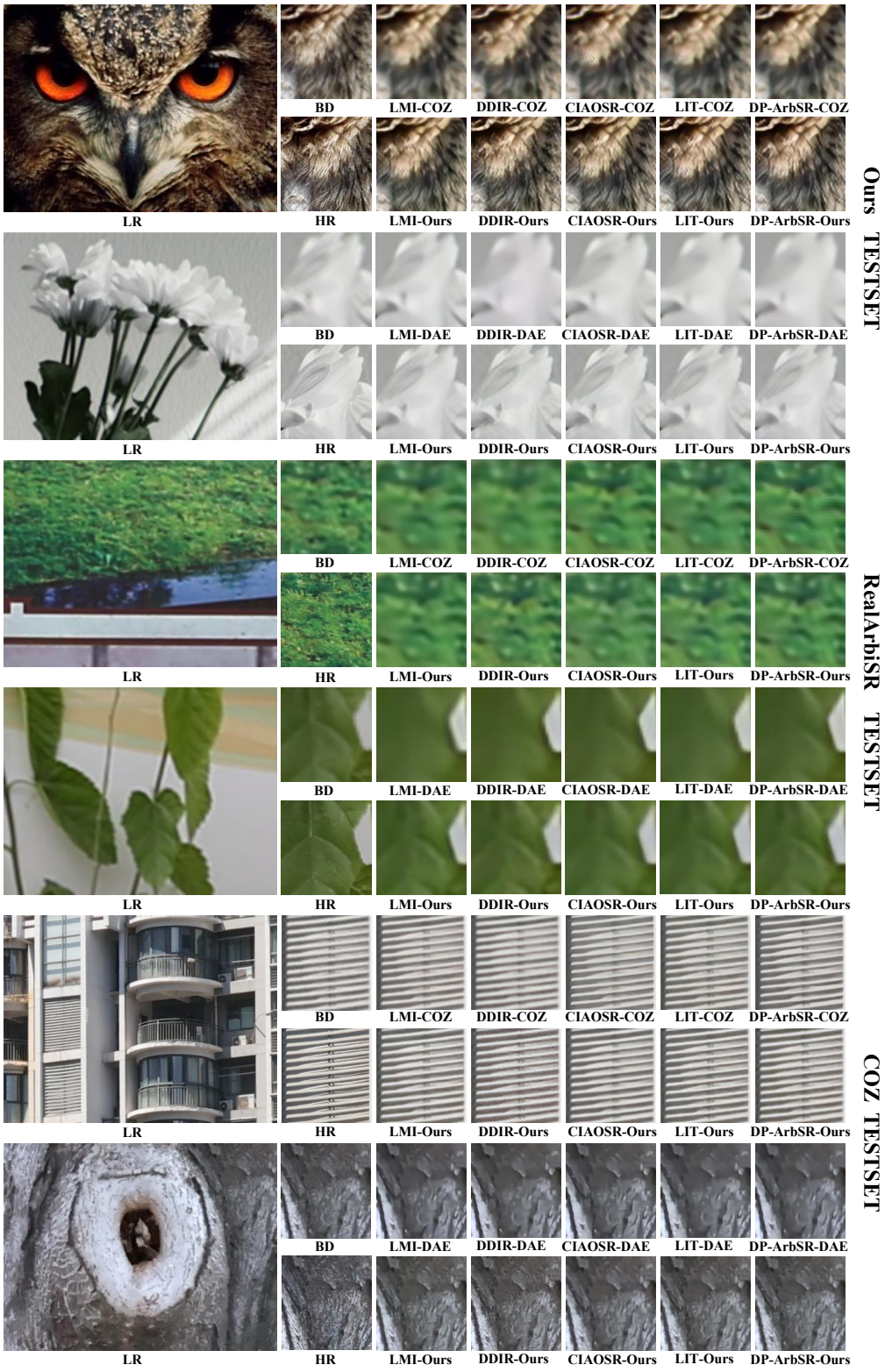

Figure 18: Visual comparison on synthetic scenes, RealArbiSR Li et al. (2024), and COZ Fu et al. (2024). Models including LMI Fu et al. (2024), DDIR Li et al. (2024), CiaoSR Cao et al. (2023), LIT Chen et al. (2023a), and our proposed DP-ArbSR are trained on different datasets: the COZ dataset, synthetic training data generated by DegAE Liu et al. (2023), and our self-constructed synthetic dataset.

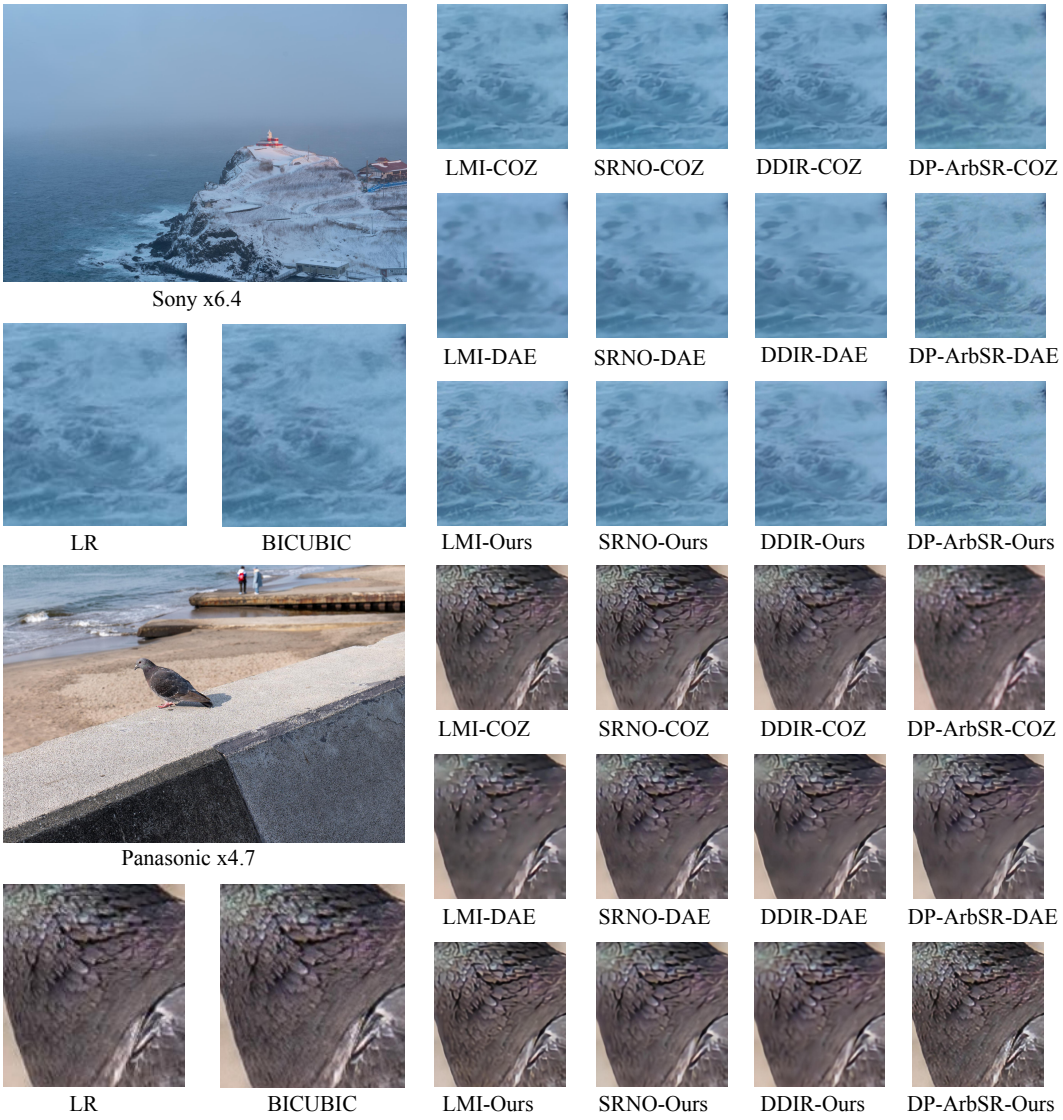

Figure 19: Visual comparison of LMIFu et al. (2024) trained on COZ, DegAE-synthesized, and our synthesized datasets, SRNOWei & Zhang (2023) trained on COZ, DegAE-synthesized, and our synthesized datasets, DDIRLi et al. (2024) trained on COZ, DegAE-synthesized, and our synthesized datasets, and our DP-ArbSR trained on our data, using real-world LR from Sony A7R2, Panasonic S5M2. Our method outperforms COZ-trained models on dynamic water surfaces and animal scenes.

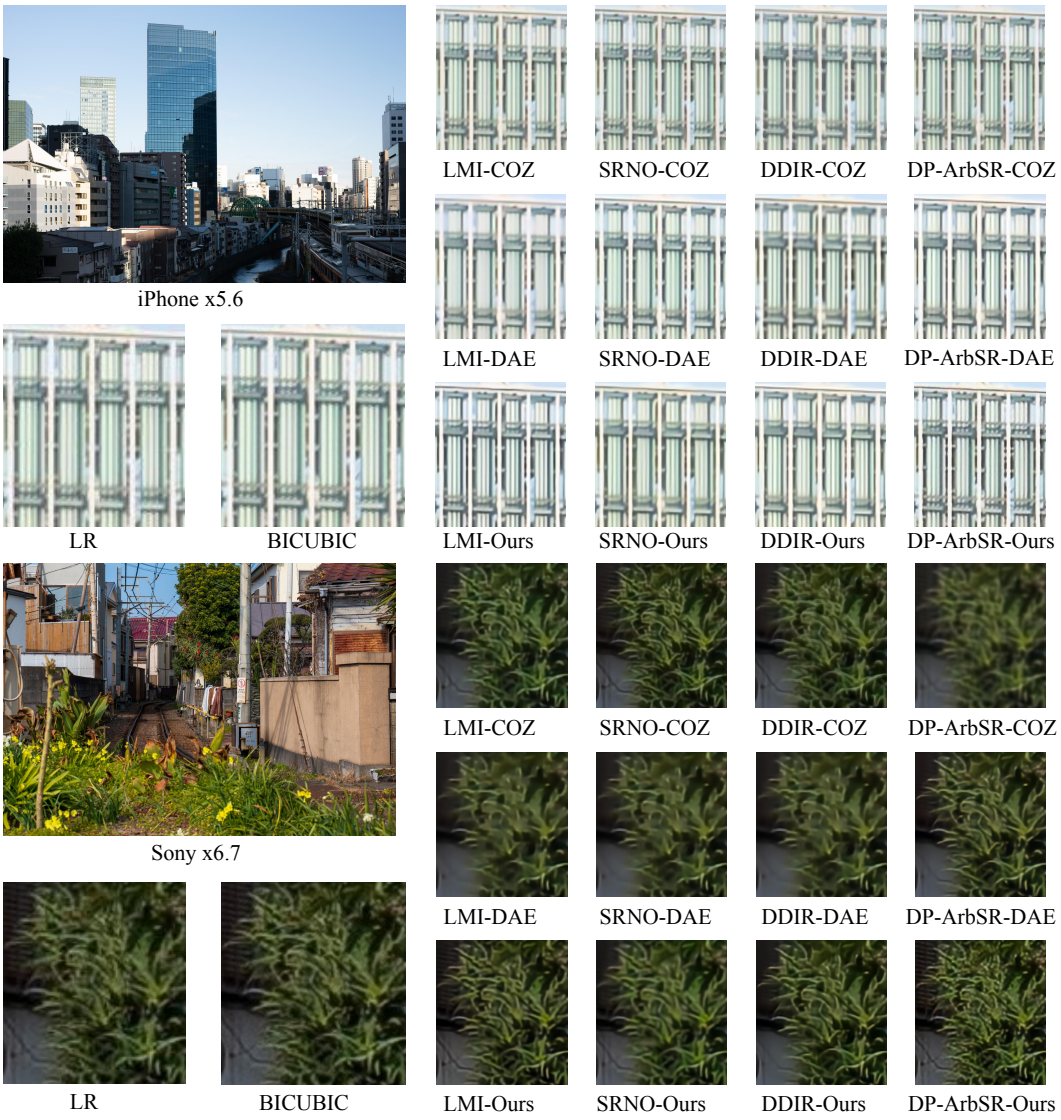

Figure 20: Visual comparison of LMIFu et al. (2024) trained on COZ, DegAE-synthesized, and our synthesized datasets, SRNOWei & Zhang (2023) trained on COZ, DegAE-synthesized, and our synthesized datasets, DDIRLi et al. (2024) trained on COZ, DegAE-synthesized, and our synthesized datasets, and our DP-ArbSR trained on our data, using real-world LR from iPhone 12 Pro, Sony A7R2. Our method outperforms COZ-trained models on Vegetation scenes, while achieving similar visual results to COZ on its common architectural scene.



Figure 21: Visual comparison of LMIFu et al. (2024) trained on COZ, DegAE-synthesized, and our synthesized datasets, SRNOWei & Zhang (2023) trained on COZ, DegAE-synthesized, and our synthesized datasets, DDIRLi et al. (2024) trained on COZ, DegAE-synthesized, and our synthesized datasets, and our DP-ArbSR trained on our data, using real-world LR from Panasonic S5M2, iPhone 12 Pro. Our method outperforms COZ-trained models on Rooftop scenes, while achieving similar visual results to COZ on its common architectural scene.