FiDeSR: High-Fidelity and Detail-Preserving One-Step Diffusion Super-Resolution

Supplementary Material

This supplementary material includes: comparison with User Study (Sec. A), GAN-based Real-ISR (Sec. B), more visual comparisons (Sec. C), complexity analysis (Sec. D), ablation studies (Sec. E), implementation details of DAW (Sec. F) and LFIM (Sec. G), and comparison with Frequency-Aware Diffusion SR (Sec. H).

A. User Study

As shown in Fig. 1, we conduct a user study comparing our method with eight other diffusion-based SR approaches. A total of 20 images were randomly selected from the DIV2K [1], DRealSR [6], and RealSR [2] datasets, and 20 volunteers were asked to select the image that best balances perceptual realism and content fidelity. Although individual preferences varied, FiDeSR received the highest number of votes overall, demonstrating its superior perceptual quality compared with the competing methods.

B. Comparison with GAN-based Real-ISR

We compare our method with GAN-based Real-ISR in Table 1. As shown in the table, our FiDeSR outperforms existing GAN-based methods [4, 5, 7]. In particular, FiDeSR achieves superior results in LPIPS and DISTS, indicating improved perceptual similarity and reduced artifact formation compared to existing GAN-based methods. FiDeSR shows strong results in no-reference metrics such as CLIP-IQA, NIQE, MUSIQ, and MANIQA, suggesting that the restored images look more realistic and maintain coherent visual structure. In addition, we present visual comparisons to GAN-based Real-ISR methods in Fig. 2. As shown in the figure, GAN-based models often suffer from over-smoothed textures, inaccurate details, and unstable structure reconstruction. FiDeSR produces more faithful textures and preserves geometric structures more consistently.

C. More Visual Comparisons

As shown in Fig. 3 and Fig. 4, we provide additional visual comparisons between FiDeSR and other diffusion-based SR models. FiDeSR consistently reconstructs fine textures and intricate patterns even from heavily degraded low-quality inputs, producing results that are both highly realistic and structurally faithful compared with other methods.

D. Complexity Analysis

We compare the number of parameters and the inference time of competing diffusion model-based SR models in Table 2. The inference time is measured on the $\times 4$ SR task using 128×128 LQ images and a single NVIDIA H100 80GB GPU. Although FiDeSR incorporates both the LRRB and LFIM modules and is designed to produce perceptually faithful and realistic reconstructions, it still achieves competitive inference speed compared with existing one-step SR models. In addition, compared to FiDeSR without LRRB, introducing LRRB increases the parameter count by only 0.01B (0.8% of 1.29B) and adds a runtime overhead of 0.0063 s (8.1% of 0.078 s).

E. Ablation studies

E.1. LRRB High-Frequency Noise Prediction Refinement Visualization

By employing the LRRB, we demonstrate its effectiveness in refining noise predictions in the high-frequency domain. To visualize the spatial distribution of error improvement, we compute the high-frequency error difference map:

$$\Delta E_{\mathrm{HF}} = \|\hat{\varepsilon}_{\text{baseline}} - \varepsilon\|_{\mathrm{HF}} - \|\hat{\varepsilon}_{\text{LRRB}} - \varepsilon\|_{\mathrm{HF}} .$$

The high-frequency components are extracted using FFT-based high-pass filtering with a radial cutoff of $r_c = 0.8$ (top 20% frequencies). To improve the visibility of small differences, we apply a sign-preserving logarithmic transformation:

$$\Delta E_{\mathrm{HF}}^{(\log)} = \mathrm{sign}(\Delta E_{\mathrm{HF}}) \cdot \log(1 + |\Delta E_{\mathrm{HF}}|) .$$

Table 3 in the main paper presents the quantitative error values, while Fig. 5 in this supplement visualizes these improvements. Positive values (red) indicate regions where LRRB reduces prediction error, whereas negative values (blue) correspond to areas where the error increases. The visualization further reveals that improvements (red regions) are concentrated in perceptually important areas such as edges and fine textures, where accurate noise prediction is essential for preserving image fidelity. These results demonstrate that, by incorporating LRRB, the FiDeSR model better preserves high-frequency details from low-quality inputs, enabling more realistic and visually faithful image restoration.

E.2. DAW Module Visualization

As illustrated in Fig. 6, the Detail-aware Weighting (DAW) module generates a Detail Map using multiple spatial filters, including Sobel, Laplacian, and local variance operators.

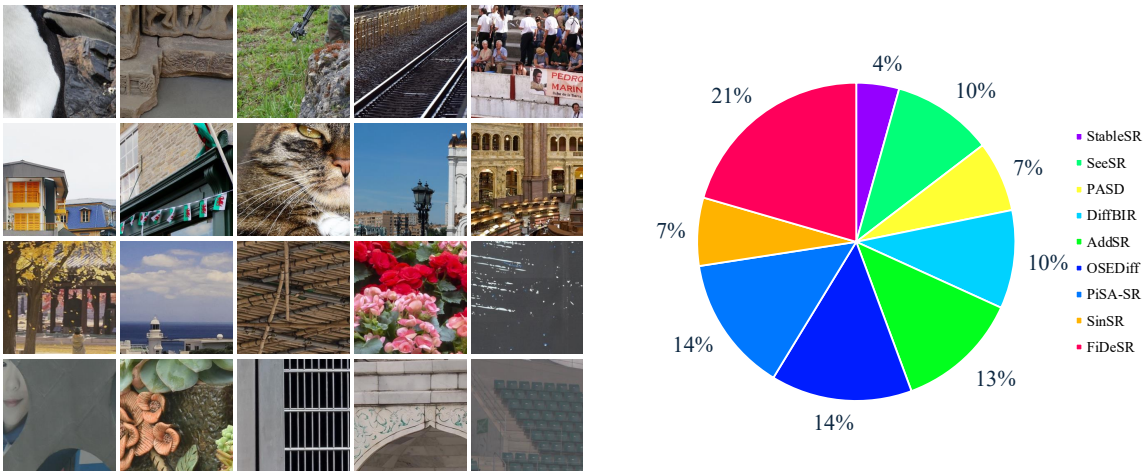

Figure 1. Ground-truth images used in the user study along with the voting results for their corresponding restored outputs.

Each filter captures different types of fine structures, enabling the model to extract complementary detail cues from the input image. The resulting Detail Map is then element-wise multiplied with the Error Map to produce the Difficulty Weight Map, which emphasizes regions where the model tends to make larger prediction errors. By guiding the network to focus more on these challenging and structurally important areas, the DAW module effectively enhances the perceptual quality of the reconstructed images.

E.3. Qualitative Analysis of LRRB and DAW Contributions

To further analyze the effectiveness of each component in FiDeSR, we conduct a qualitative ablation study, as shown in Fig. 7. By individually removing the LRRB and DAW modules, we observe clear degradation in the reconstruction quality. Excluding the LRRB reduces the model's ability to refine high-frequency structures, leading to blurry textures and loss of fine details. Without the DAW module, the model becomes less sensitive to spatially challenging regions, resulting in artifacts and reduced perceptual sharpness. In contrast, the full FiDeSR model consistently reconstructs finer textures and preserves structural fidelity even under severe degradation conditions, demonstrating the complementary contributions of both modules.

E.4. Ablations on LoRA rank

Table 3 presents the effect of varying the LoRA rank in FiDeSR. Although different ranks lead to slight variations across distortion-oriented and perceptual metrics, the overall performance remains stable and competitive. In our implementation, we adopt a LoRA rank of 8, which offers a strong balance between fidelity (PSNR/SSIM) and perceptual quality (NIQE, MUSIQ, MANIQA). The results in the table also show that other ranks (4 and 16) achieve similarly competitive performance.

Algorithm 1 Pseudo-code of Detail-Aware Weighting

Input: GT y, pred $\hat{y}$, mix p
Output: $L_{l2}, L_{lpips}, L_{csd}$
$y \leftarrow \text{to\_gray}(y)$
$(S, L, V) \leftarrow (\text{Sobel}(y), \text{Laplacian}(y), \text{Variance}(y))$
$D \leftarrow \text{box\_blur3x3}(\text{quantile\_norm}((S + L + V)/3))$
$E_{\text{pix}}, E_{\text{perc}} \leftarrow \text{L1}(\hat{y}, y),\ \text{LPIPS}(\hat{y}, y)$
$E \leftarrow (1 - p)\, E_{\text{pix}} + p\, E_{\text{perc}}$
$E \leftarrow \text{quantile\_norm}(E)$
$W \leftarrow \tanh(\text{blur}(D \odot E)/w_{\max}) \cdot w_{\max}$
$w^* \leftarrow \text{mean\_norm}(1 + \alpha \cdot W)$
$L_{l2} \leftarrow (w^* \cdot (\hat{y} - y)^2).mean()$
$L_{lpips} \leftarrow (\text{resize}(w^*) \cdot \text{LPIPS\_map}(\hat{y}, y)).mean()$
$L_{csd} \leftarrow (\text{resize}(w^*) \cdot \text{CSD}(\text{latents}, \text{prompts})).mean()$

F. Implementation Details of DAW

We summarize the implementation of DAW used during training in Alg. 1. DAW computes a detail map D from spatial operators (Sobel, Laplacian, and Variance) on the HQ target x_H, and an error map E by mixing pixel-level (L1) and perceptual (LPIPS) discrepancies between x_{SR} and x_H with coefficient p. These maps are combined to form a per-pixel difficulty weight, which is applied to both the reconstruction loss and the CSD loss.

G. Implementation Details of LFIM

Additional Details of LFIM on LF and HF. Beyond the main description in the paper, the inference code reveals several important implementation aspects of the low-

Table 1. Quantitative comparison with GAN-based Real-ISR Methods. Best results are highlighted in **red**.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	CLIPIQA ↑	NIQE ↓	MUSIQ ↑	MANIQA ↑	FID ↓
DRealSR	Real-ESRGAN	28.61	0.8051	0.2819	**0.2089**	0.4519	6.6896	54.2678	0.4904	147.68
	BSRGAN	28.70	0.8028	0.2858	0.2144	0.5093	6.5387	57.1626	0.4844	155.59
	LDL	28.20	**0.8124**	**0.2792**	0.2127	0.4475	7.1360	53.9464	0.4894	155.53
	FiDeSR	**28.90**	0.7907	0.2836	0.2112	**0.6974**	**6.2014**	**65.7820**	**0.6239**	**127.97**
RealSR	Real-ESRGAN	25.68	0.7614	0.2709	0.2060	0.4485	5.7936	60.3674	0.5505	135.20
	BSRGAN	**26.37**	**0.7651**	0.2656	0.2124	0.5119	5.6361	63.2870	0.5420	141.30
	LDL	25.28	0.7565	0.2750	0.2120	0.4554	5.9905	60.9277	0.5494	142.68
	FiDeSR	26.02	0.7457	**0.2626**	**0.1965**	**0.6896**	**5.3194**	**69.8245**	**0.6681**	**109.68**
DIV2K	Real-ESRGAN	24.29	**0.6372**	0.3112	0.2141	0.5277	4.6790	61.0621	0.5485	37.63
	BSRGAN	**24.58**	0.6269	0.3351	0.2275	0.5247	4.7510	61.1953	0.5041	44.22
	LDL	23.83	0.6344	0.3256	0.2227	0.5179	4.8549	60.0382	0.5328	42.28
	FiDeSR	24.33	0.6250	**0.2678**	**0.1845**	**0.6873**	**4.6644**	**68.8672**	**0.6384**	**23.30**

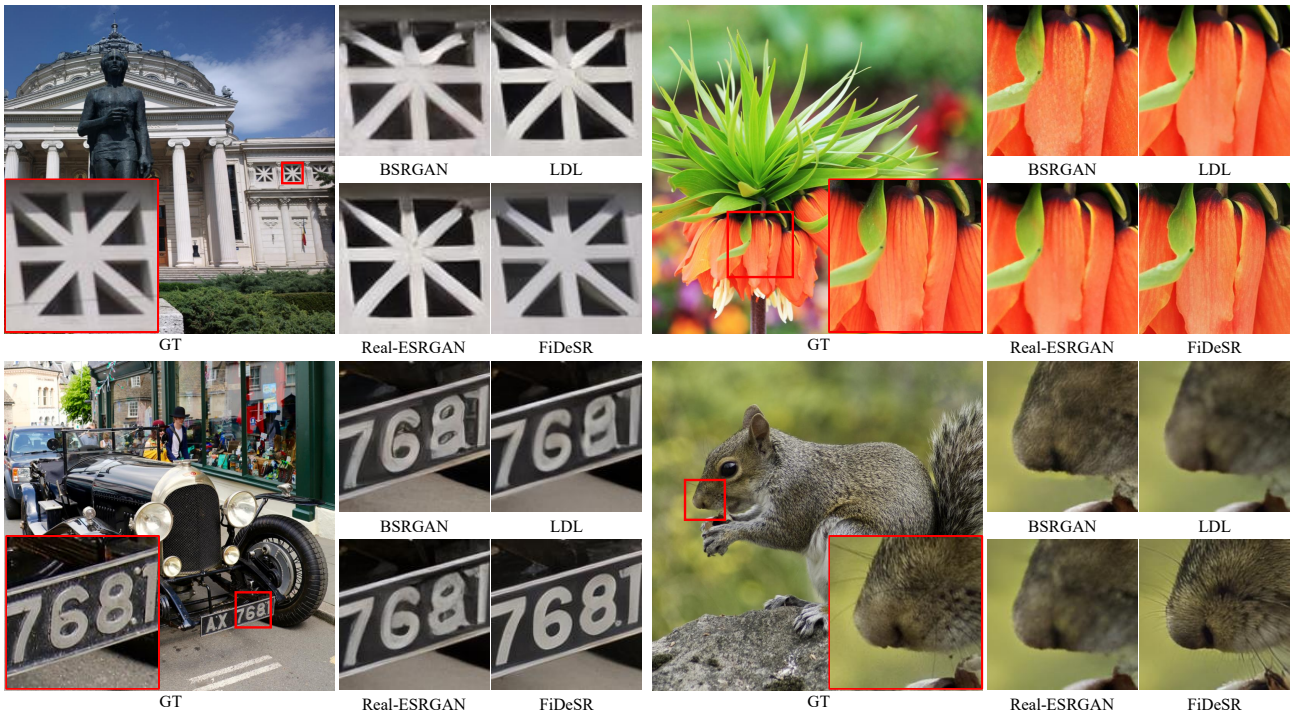

Figure 2. Qualitative comparisons between FiDeSR and GAN-based Real-ISR methods.

frequency injection (LFIM on LF) and high-frequency injection (LFIM on HF) mechanisms that are not explicitly discussed in the main manuscript.

Frequency Decomposition and Injection Intensity. Both variants of LFIM operate directly in the latent space z using FFT-based Butterworth filtering. The injection intensity is governed by two global weighting parameters: `lf_alpha` for low-frequency reinforcement and `hf_beta` for high-frequency enhancement. These coefficients determine how strongly the filtered components are injected back into the latent representation. Larger values of `lf_alpha` result in stronger stabilization of global structures, while higher `hf_beta` promote more pronounced enhancement of textures and edges. As observed in Table 4 of the main paper, increasing low-frequency injection monotonically improves PSNR and SSIM, whereas stronger high-frequency injection leads to higher MUSIQ and MANIQA scores.

As summarized in Table 4, adjusting both `lf_alpha` and `hf_beta` jointly provides a flexible way to control the balance between structural fidelity and perceptual sharpness, and both the overall injection strength and the LF/HF

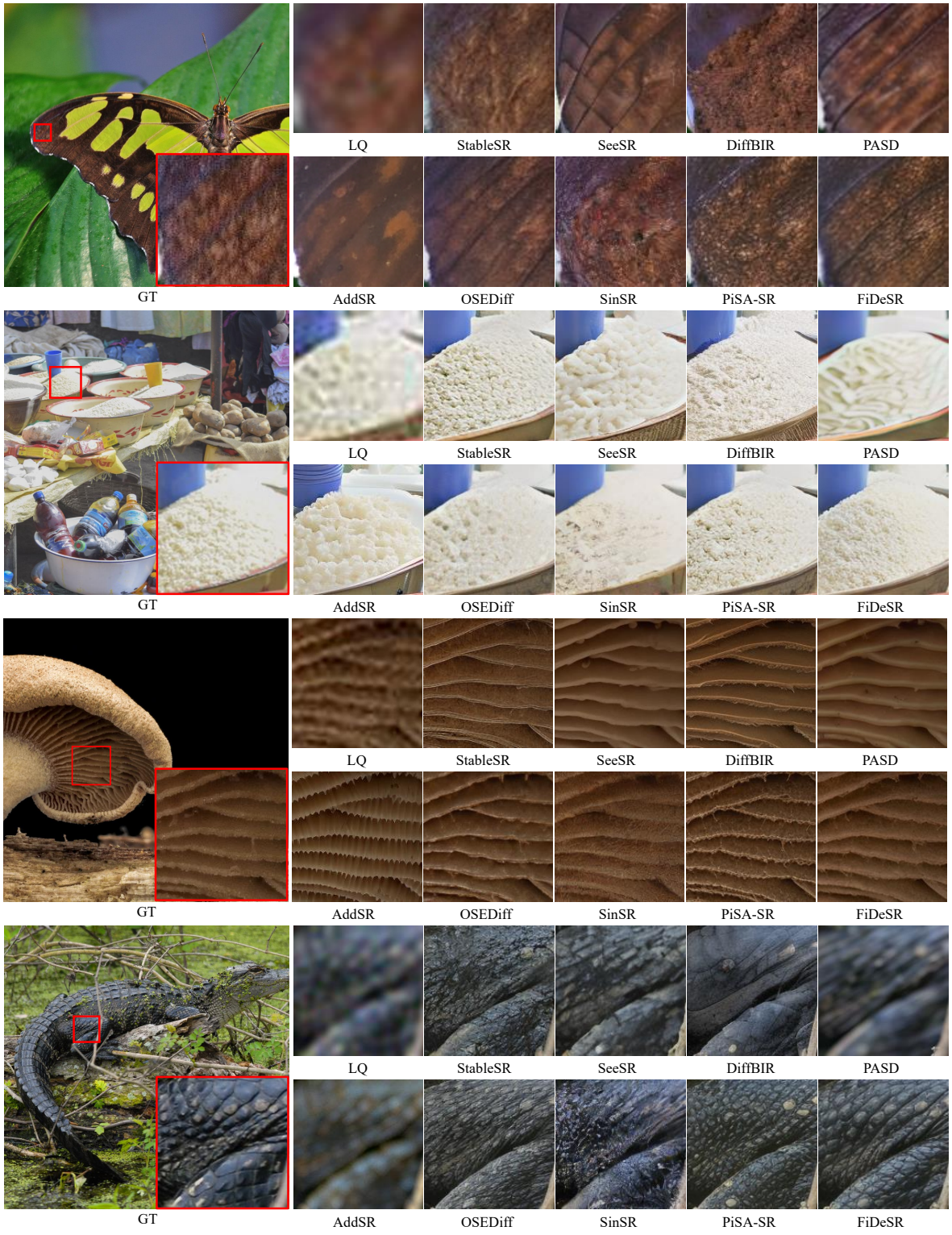

Figure 3. Qualitative comparisons between FiDeSR and different diffusion-based methods on DIV2K dataset. FiDeSR effectively reconstructs fine details while preserving overall image fidelity.

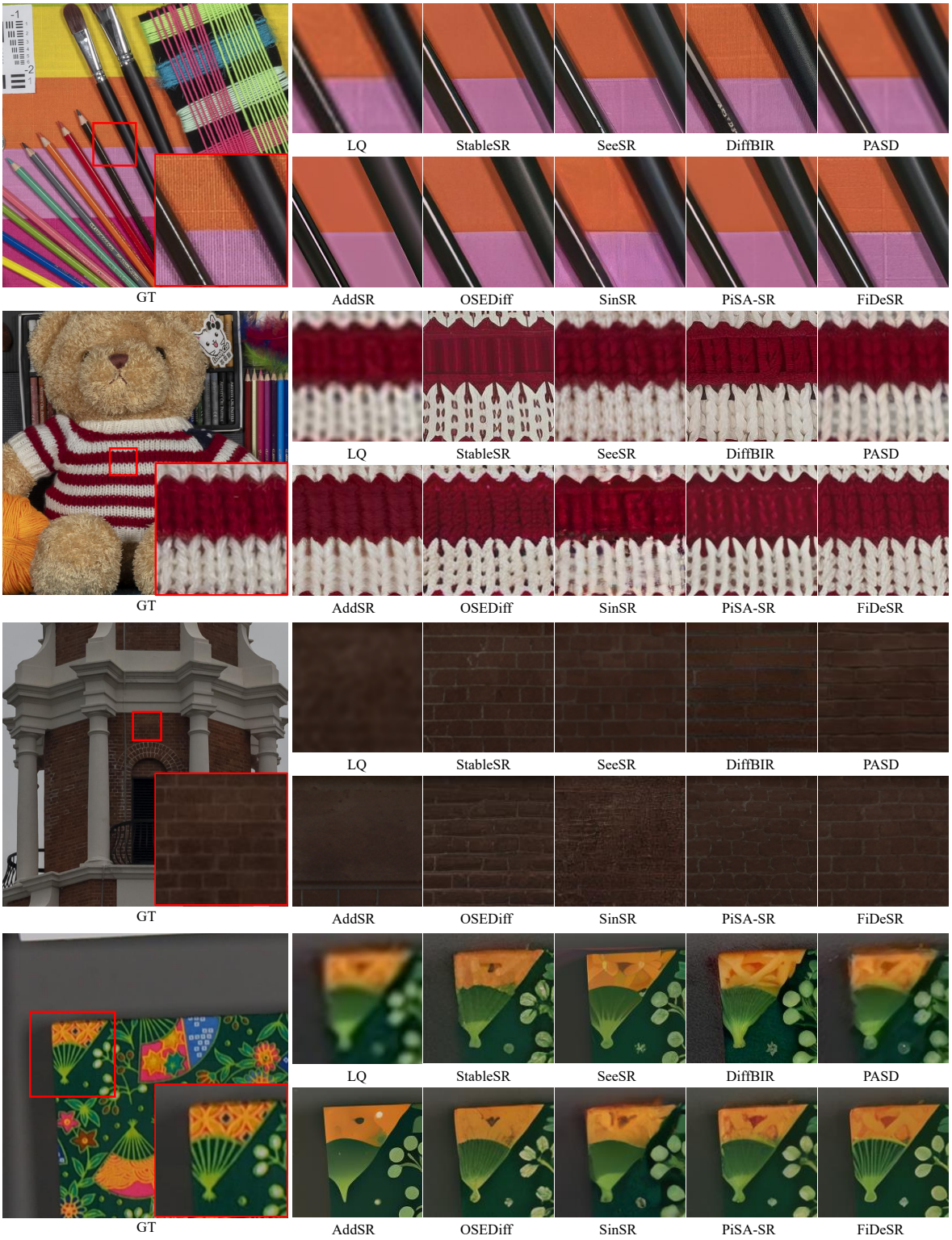

Figure 4. Qualitative comparisons between FiDeSR and different diffusion-based methods on DRealSR and RealSR dataset. FiDeSR effectively reconstructs fine details while preserving overall image fidelity.

Table 2. Comparison of inference steps, runtime, and model parameters among diffusion-based SR methods.

Metric	StableSR	DiffBIR	SeeSR	PASD	AddSR	SinSR	OSEDiff	PISA-SR	FiDeSR (Ours)
Inference Step	200	50	50	20	4	1	1	1	1
Inference Time (s)	7.52	2.04	3.30	2.10	0.76	0.097	0.087	0.057	0.078
#Params (B)	1.56	1.68	2.51	2.31	2.28	0.18	1.77	1.30	1.29

Table 3. Ablation study of LoRA rank for FiDeSR on the RealSR dataset (evaluated before applying LFIM).

LoRA Rank	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	**DISTS** ↓	**CLIPIQA** ↑	**NIQE** ↓	**MUSIQ** ↑	**MANIQA** ↑	**FID** ↓
4	26.5196	0.7577	0.2573	0.1967	0.6554	5.3488	68.4282	0.6442	112.9637
8	26.2542	0.7498	0.2604	0.1963	0.6764	5.3332	69.3610	0.6580	108.7867
16	26.4582	0.7563	0.2508	0.1920	0.6660	4.6858	68.0145	0.6525	108.6132

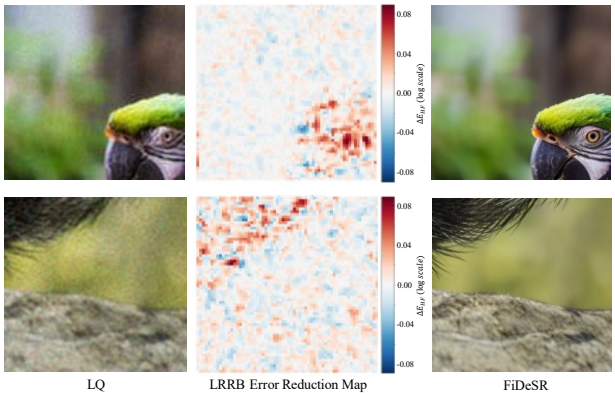

Figure 5. Spatial distribution of high-frequency noise prediction error improvement by LRRB. ΔE_{HF} represents the difference between baseline and LRRB error magnitudes in the high-frequency domain (top 20% frequencies, $r_c = 0.8$). Positive values (red) indicate regions where LRRB reduces prediction error, while negative values (blue) indicate regions where error increases. Color intensity represents the magnitude of change in log scale.

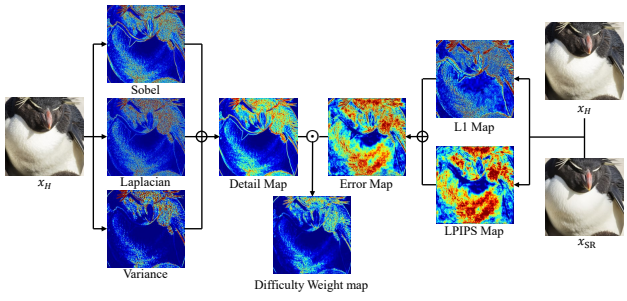

Figure 6. Visualization of the Detail-aware Weighting (DAW) module. Detail Map generated by spatial filters (Sobel, Laplacian, Variance), is element-wise multiplied by the Error Map to create the Difficulty Weight Map.

ratio can be freely manipulated depending on the desired behavior. In practice, these parameters allow users to steer the reconstruction preference, ranging from clean and stable outputs to sharper results with more pronounced texture details.

In our implementation, we adopt a balanced configuration of `lf_alpha` $= 0.2$ and `hf_beta` $= 0.2$, which we found to yield a well-rounded compromise between global structural consistency and perceptual detail enhancement. This setting avoids excessively biasing the model toward either distortion-centric or perception-centric behavior, producing stable and visually coherent results across datasets.

LFIM on Low Frequency LFIM on LF extracts the low-frequency residual Δ_{LP} through a Butterworth low-pass filter and selectively injects it back into z using spatial and channel gating. The spatial gate is derived from Sobel, Laplacian, and local variance maps, limiting LF injection in detail-rich regions to avoid oversmoothing. The channel gate evaluates Pseudo-PSD energy to identify channels dominated by structural information. The final LF injection is applied as

$$z \leftarrow z + \texttt{lf\_alpha} \cdot M_{\mathrm{sp}} \cdot M_{\mathrm{ch}} \cdot \Delta_{LP},$$

with optional morphological erosion available to refine the spatial mask boundaries. This mechanism stabilizes illumination, coarse geometry, and tone consistency, reducing structural distortion in the restored output.

LFIM on High Frequency LFIM on HF extracts the high-frequency component Δ_{HP} either directly or using the differential high-pass term $\mathrm{HPF}(z) - \mathrm{HPF}(\hat{z})$ when enabled by `hf_use_diff`. Spatial gating emphasizes edge regions through a detail-dependent exponent γ, while channel gating selects frequency-rich channels complementary to the LF gate. The final injection is computed as

$$z \leftarrow z + \texttt{hf\_beta} \cdot M_{\mathrm{sp}}^{HF} \cdot M_{\mathrm{ch}}^{HF} \cdot \Delta_{HP}.$$

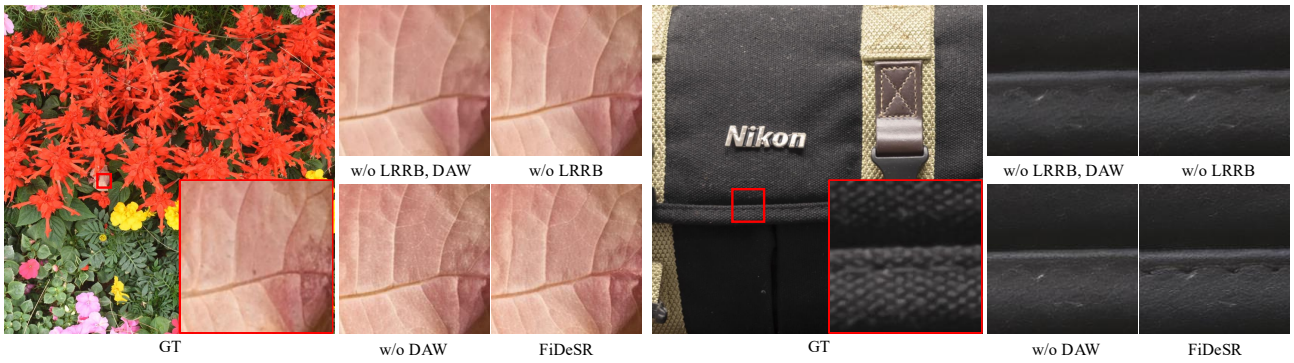

Figure 7. Qualitative ablation study illustrating the contributions of the LRRB and DAW modules. Even under challenging degradation conditions, the full FiDeSR model reconstructs finer details and preserves texture fidelity most effectively.

Table 4. Ablation study on different LF/HF injection strength ratios of LFIM (`lf_alpha`, `hf_beta`) on the RealSR dataset.

LF/HF Ratio	PSNR ↑	SSIM ↑	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑
(0.2, 0.2)	26.0249	0.7457	0.6896	69.8245	0.6681
(0.4, 0.2)	26.0658	0.7464	0.6877	69.7645	0.6666
(0.4, 0.4)	25.8630	0.7428	0.6934	70.0136	0.6737
(0.6, 0.6)	25.6519	0.7391	0.6956	70.1429	0.6789

Table 5. Comparison between FiDeSR and TFDSR on SR benchmarks.

Dataset	Method	PSNR ↑	LPIPS ↓	NIQE ↓	MANIQA ↑	FID ↓
DRealSR	TFDSR	27.88	0.3417	6.2667	0.6164	155.66
	FiDeSR	**28.90**	**0.2836**	**6.2014**	**0.6239**	**127.97**

This selective enhancement sharpens textures, edges, and micro-patterns without altering global luminance structure, and results in substantial boosts to perceptual metrics.

Summary. LFIM on LF enhances global structural fidelity, whereas LFIM on HF improves perceptual sharpness. The use of balanced injection intensities enables FiDeSR to recover both reliable structure and rich detail, ultimately yielding high-quality, frequency-aware super-resolution results.

H. Comparison with Frequency-Aware Diffusion SR

In Table 5, we provide a quantitative comparison with TFDSR [3], a frequency-aware diffusion SR method, on the DRealSR dataset. FiDeSR consistently outperforms TFDSR in both full-reference and no-reference metrics. Specifically, FiDeSR achieves a lower LPIPS score, indicating that the restored images are more perceptually and semantically consistent with the ground-truth. Furthermore, improved NIQE and MANIQA scores indicate that FiDeSR generates more naturalistic details and higher visual quality.

References

[1] 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*, pages 126–135, 2017. 1

[2] 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*, pages 3086–3095, 2019. 1

[3] Yueying Li, Hanbin Zhao, Jiaqing Zhou, Guozhi Xu, Tianlei Hu, Gang Chen, and Haobo Wang. A timestep-adaptive frequency-enhancement framework for diffusion-based image super-resolution. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 1503–1511, 2025. 7

[4] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5657–5666, 2022. 1

[5] 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*, pages 1905–1914, 2021. 1

[6] 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 *European conference on computer vision*, pages 101–117. Springer, 2020. 1

[7] 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*, pages 4791–4800, 2021. 1