ScaleResfusion: Residual Rectified Flow based on Residual Vector Field

Zhenning Shi[1,*] Chen Xu[2,*] Junhao Zhang[3] Kefei Zhang[1] Linjie Liu[4] Zhedong Zheng[2,†] Tao Li[1,†]

[1]Nankai University, Tianjin, China

[2]University of Macau, Macau, China

[3]Csiro Data 61, Australia

[4]Beihang University, Beijing, China

[*]Equal Contribution. [†]Corresponding Authors: litao@nankai.edu.cn, zhedongzheng@um.edu.mo.

arXiv:2607.25275v1 [cs.CV] 28 Jul 2026

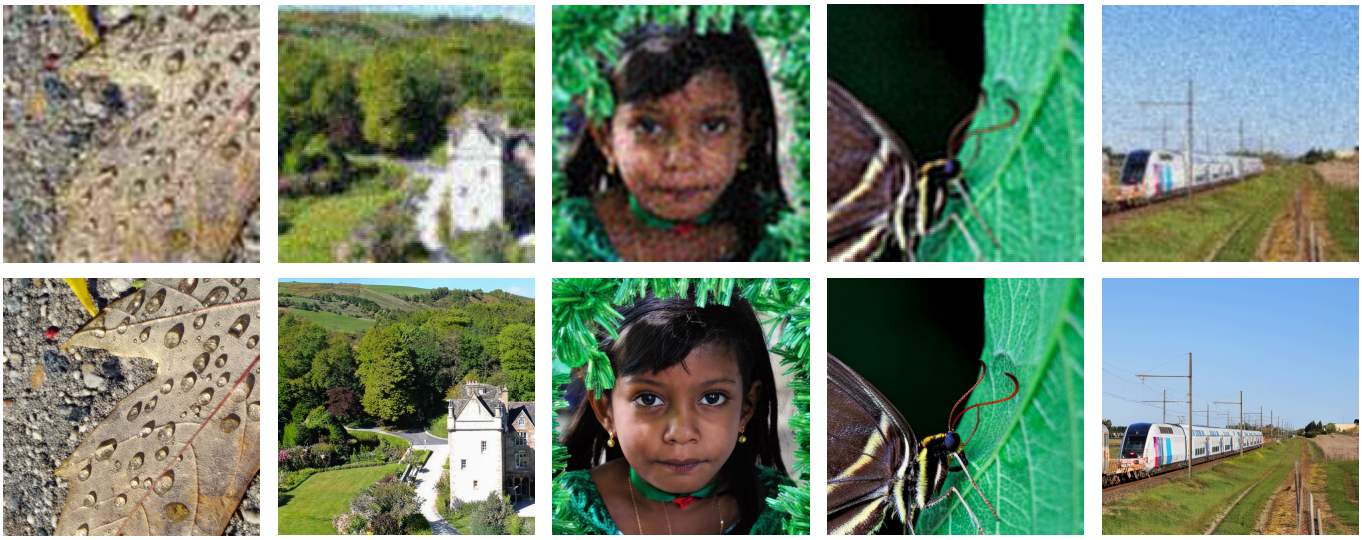


Fig. 1. Visual examples of **ScaleResfusion** for Real-IR. The top row shows LQ inputs and the bottom row shows our restored results with only **4 sampling steps**. ScaleResfusion preserves the structure, identity, and content of the input images while synthesizing photo-realistic details.

Real-world Image Restoration (Real-IR) aims to recover high-quality (HQ) images from complex and unknown degradations. Although recent diffusion-based methods have substantially improved perceptual quality, their current designs leave two key challenges unresolved. Methods that start from Gaussian noise are slow and often less faithful to the degraded input. Residual-based methods usually train from scratch, which makes it hard to exploit modern pre-trained generative priors. In this paper, we present **ScaleResfusion**, a scalable diffusion framework for real-world image restoration built on pre-trained text-to-image rectified-flow models. The core of our method is **Residual Rectified Flow**, which introduces the residual term R into the Standard Rectified Flow. Instead of starting from pure noise, it uses a residual transport path that starts from noisy low-quality (LQ) images and admits an exact acceleration point. By learning the **residual vector field**, Residual Rectified Flow keeps the output distribution and linear diffusion process consistent with the pre-trained rectified-flow models. This makes parameter-efficient fine-tuning possible at scale. We further introduce a knowledge-distillation pipeline to reduce sampling cost while maintaining restoration quality. Extensive experiments on multiple real-world restoration tasks show that ScaleResfusion achieves state-of-the-art performance with much higher efficiency. These results suggest a practical and scalable way to adapt large pre-trained diffusion models to real-world image restoration. Our code and model are available at `github.com/YukinoshitaLove/ScaleResfusion`.

Additional Key Words and Phrases: Diffusion-based Models, Residual Diffusion Models, Real World Image Restoration, Image Super-Resolution.

1 Introduction

Image restoration (IR) aims to recover a high-quality (HQ) image from low-quality (LQ) observations [Chen et al. 2021, 2023; Dai et al. 2019; Dong et al. 2014; Ledig et al. 2017; Liang et al. 2021; Wang et al. 2021b, 2018; Zhang et al. 2017, 2022b, 2018b]. In real-world image restoration (Real-IR), the degradation process is usually complex, unknown, and compound, making the inverse problem highly ill-posed in practical scenarios [Wang et al. 2021b; Zhang et al. 2021]. The central challenge is to recover realistic details while preserving fidelity to the degraded input.

Recently, diffusion-based models [Ho et al. 2020; Lipman et al. 2022; Liu et al. 2022; Song et al. 2020a; Song and Ermon 2019; Song et al. 2020b] have shown remarkable ability in synthesizing high-fidelity restored images [Guo et al. 2023; He et al. 2025; Özdenizci and Legenstein 2023; Saharia et al. 2022; Wang et al. 2025c; Whang et al. 2022]. To leverage the powerful prior knowledge of pre-trained text-to-image (T2I) diffusion models [Esser et al. 2024; Rombach et al. 2022], recent works [Lin et al. 2024; Wang et al. 2024; Wu et al. 2024b; Yang et al. 2024; Yu et al. 2024] introduce the LQ image as a conditional input [Zhang et al. 2023], producing more realistic images than GAN-based methods with more stable training. However, these methods usually consider only LQ images as the condition and start diffusion from Gaussian noise, resulting in lengthy inference and weak consistency between the generated output and the LQ

input. Although some works [Dong et al. 2025; Li et al. 2025; Wu et al. 2024a] reduce diffusion steps through knowledge distillation to achieve single-step inference, they tend to learn a LQ-HQ mapping while abandoning the multi-step sampling characteristic of diffusion models, leading to reduced diversity and fidelity.

Another line of diffusion-based restoration methods [Delbracio and Milanfar 2023; Kawar et al. 2022; Liu et al. 2023; Luo et al. 2023a,b; Shi et al. 2024; Wang et al. 2025b; Yue et al. 2023] starts the diffusion process from noisy LQ images rather than Gaussian noise, reducing inference steps while preserving the sampling capability of diffusion models. However, these works typically train a new diffusion model from scratch. Their training objectives and customized noise schedulers often deviate from the output distribution and diffusion process of modern pre-trained models, making direct reuse difficult. Resfusion [Shi et al. 2024] addresses this issue in DDPM [Ho et al. 2020] by learning the resnoise and unifying training and inference through a smooth equivalence transformation. Nevertheless, modern pre-trained models [Cai et al. 2025; Esser et al. 2024; Labs 2025] are typically based on Rectified Flow [Liu et al. 2022], preventing vanilla Resfusion from scaling further.

To solve the problems mentioned above, we propose **ScaleResfusion**, a scalable diffusion framework that adapts pre-trained T2I Rectified Flow models to Real-IR tasks. Its core is **Residual Rectified Flow**, which incorporates the image residual into the Rectified Flow trajectory while preserving its linear transport form. This design allows sampling to start from a noisy LQ image instead of pure Gaussian noise, reducing redundant generation steps without discarding the stochastic sampling behavior of diffusion models. Since the optimization target **residual vector field** only differs from the standard Rectified Flow target by a residual offset, ScaleResfusion can efficiently adapt large pre-trained backbones with only **Low-Rank Adaptation (LoRA)** [Hu et al. 2022]. We further introduce a knowledge-distillation based parameter-efficient fine-tuning (PEFT) pipeline to reduce training and inference cost. Experiments show that ScaleResfusion achieves state-of-the-art performance on multiple Real-IR tasks with significantly faster inference.

Our contributions can be summarized as follows:

(1) We formulate **Residual Rectified Flow** for Real-IR by incorporating image residuals into the Rectified Flow trajectory, enabling the diffusion process to start directly from noisy LQ images within the Rectified Flow framework.

(2) We derive the **residual vector field** as the optimization target, which differs from the standard Rectified Flow target only by a residual offset and thus enables LoRA-only PEFT for pre-trained models based on Rectified Flow.

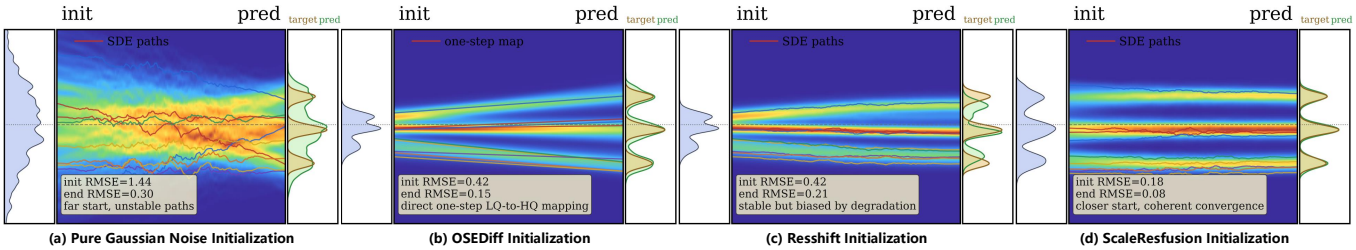
(3) Based on ScaleResfusion, we construct an efficient training pipeline with knowledge distillation, demonstrating that residual-based diffusion models can be scaled to billion-scale pre-trained models for Real-IR **with only 4 sampling steps**.

2 Related Work

Traditional Real-IR methods typically learn a direct LQ-to-HQ mapping under simplified or synthetic degradations [Chen et al. 2021, 2023; Dong et al. 2014; Liang et al. 2021; Wang et al. 2021b; Zhang et al. 2017, 2022b, 2018b], while GAN-based methods improve perceptual realism by matching the natural image distribution [Goodfellow et al. 2014; Ledig et al. 2017; Liang et al. 2022; Wang et al. 2018; Zhang et al. 2021]. However, these methods often suffer from over-smoothed details, unstable adversarial training, or uncontrollable artifacts. These limitations have shifted recent Real-IR research toward diffusion-based restoration, which offers stronger generative priors and a more stable training paradigm. In this context, we summarize current diffusion-based methods from two perspectives: diffusion models initialized from Gaussian noise and those initialized from noisy LQ images.

Diffusion models initialized from Gaussian noise. Recent Real-IR methods often leverage pre-trained T2I diffusion priors by injecting the LQ image as a condition while keeping the original Gaussian noise initialization. StableSR [Wang et al. 2024] balances fidelity and perceptual quality by fine-tuning a time-aware encoder and using controllable feature wrapping. DiffBIR [Lin et al. 2024] restores the LR images with a reconstruction network and then uses Stable Diffusion [Rombach et al. 2022] to synthesize realistic details. SeeSR [Wu et al. 2024b] extracts semantic features from the input image to better activate the generative prior of Stable Diffusion. PASD [Yang et al. 2024] combines pixel-aware cross-attention with degradation-robust guidance to preserve local structure during generation. SUPIR [Yu et al. 2024] further improves generative fidelity through negative-quality prompts, restoration-guided sampling, and larger-scale pre-training. Overall, despite promising perceptual quality, these methods still sample from Gaussian white noise with the LQ image only as a condition, leading to 20–50 step inference and weak input consistency. This motivates our Residual Rectified Flow's noisy-LQ initialization.

Diffusion models initialized from noisy LQ images. To reduce inference cost, OSEDiff [Wu et al. 2024a], TSD-SR [Dong et al. 2025], and FluxSR [Li et al. 2025] distill pre-trained diffusion priors into a direct LQ-to-HQ mapping for one-step restoration. While efficient, they largely replace iterative sampling with a deterministic mapping, which may reduce diversity and fidelity. Other works preserve iterative sampling by initializing diffusion from noisy LQ images. IR-SDE [Luo et al. 2023a] models degradation and restoration with a mean-reverting stochastic differential equation. Refusion [Luo et al. 2023b] performs realistic large-size restoration through latent-space diffusion. ResShift [Yue et al. 2023] accelerates super-resolution by shifting the residual between LQ and HQ images. Residual Denoising Diffusion Models (RDDM) [Liu et al. 2024a] jointly model the residual and noise terms. Resfusion [Shi et al. 2024] learns resnoise to unify training and inference through a smooth equivalence transformation, preserving consistency with the original diffusion process. Overall, noisy-LQ initialization shortens the sampling trajectory while preserving the diffusion sampling nature, but most existing methods train task-specific diffusion models from scratch and use customized objectives or schedulers, making them difficult to reuse with pre-trained T2I priors. This motivates our ScaleResfusion, which starts from noisy LQ images while remaining consistent with the pre-trained Rectified Flow process.



(a) Pure Gaussian Noise Initialization (b) OSEDiff Initialization (c) Resshift Initialization (d) ScaleResfusion Initialization

Fig. 2. Initialization comparison from a transport perspective. (a) Gaussian-noise initialization starts far from the restoration manifold and induces unstable paths. (b) One-step restoration starts from the LQ observation but collapses the generative restoration process into a deterministic LQ-to-HQ mapping, losing the sampling capability of diffusion models. (c) Noisy-LQ-initialized DDPM sampling injects stochasticity, yet it biases the reverse path toward the input degradation without an explicit residual correction direction. (d) ScaleResfusion starts from a noisy LQ state and follows a residual-oriented transport, producing a closer initialization and more coherent convergence toward the target distribution.

3 Method

3.1 Preliminaries

Denoising Diffusion Probabilistic Models (DDPM) [Ho et al. 2020] aim to approximate the real data distribution $p(x_0)$ with the learned model distribution $p_\theta(x_0)$ through a discrete Markov chain. Flow-based generative models [Lipman et al. 2022; Liu et al. 2022] improve sampling efficiency by replacing the discrete Markovian process with a continuous transport path between a data sample $x_0 \sim p(x_0)$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, formulated as

$$x_t = a_t x_0 + b_t \epsilon, t \in [0, 1]. \tag{1}$$

The goal is to learn a vector field $v_\theta(x_t, t)$ that approximates the target vector field along this path. Rectified Flow (RF) [Liu et al. 2022] adopts the linear path $x_t = (1-t)x_0 + t\epsilon$ and optimizes

$$\mathcal{L}_{RF} = \int_0^1 \mathbb{E}_{x_0,\epsilon}\left[\|v_\theta(x_t, t) - (\epsilon - x_0)\|^2\right] dt. \tag{2}$$

Sampling is performed by solving the reverse ODE

$$x_0 = x_1 + \int_1^0 v_\theta(x_\tau, \tau) d\tau, \tag{3}$$

This Gaussian-to-image transport is effective for generic generation, but it is redundant for image restoration since the LQ image itself already carries effective information.

3.2 Residual Modeling for Image Restoration

Given an LQ–HQ image pair, we denote the LQ observation and the HQ target as $\hat{x}_0$ and x_0 respectively, and define the residual term as

$$R = \hat{x}_0 - x_0. \tag{4}$$

Conditioned on the observed LQ image $\hat{x}_0$, recovering x_0 is therefore equivalent to estimating the correction R, since $x_0 = \hat{x}_0 - R$. This residual formulation is particularly suitable for restoration: the LQ image already preserves most semantic layout and coarse structures, while the residual mainly captures local degradation artifacts, high-frequency details, and texture corrections.

As shown in Fig. 3, the residual is much more compact than either the LQ or HQ image. In the spatial domain, R appears as a low-amplitude and localized correction around edges, textures, and degradation artifacts. After the **Scaled Fourier Transform**, it also exhibits substantially lower spectral energy than the image signals.

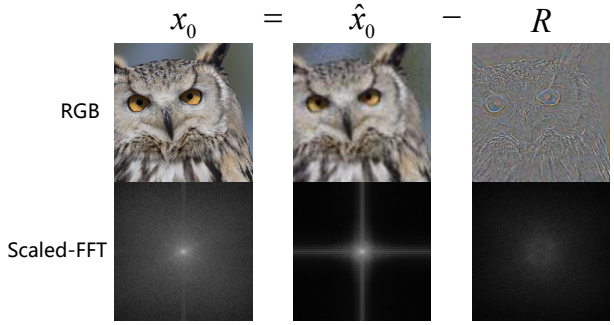


Fig. 3. The residual term R exhibits consistent low-energy, edged patterns in both the spatial and frequency domains, suggesting that RRF only needs to learn a compact residual update around the pre-trained transport.

This observation suggests that the task-specific transformation for restoration is concentrated in a compact residual subspace, rather than being distributed across the full image-generation process. Such a localized and energy-limited update is naturally compatible with **LoRA** [Hu et al. 2022], which is designed to capture low-rank adaptations around a pre-trained model.

The residual formulation is also preferable from a transport perspective. The reverse ODE is initialization-sensitive: under standard stability assumptions, an initial state closer to the task-relevant restoration manifold is more likely to be transported to a terminal state close to the desired HQ distribution. Pure Gaussian initialization is task-agnostic and far from the observed content, whereas residual modeling preserves the structural prior in $\hat{x}_0$ and makes the transport correction-oriented through R. Fig. 2 compares different strategies: Gaussian initialization suffers from large initial discrepancy; one-step restoration (OSEDiff) loses the sampling capability of diffusion models by using a deterministic LQ-to-HQ mapping; and Noisy-LQ-initialized DDPM sampling (ResShift) remains degradation-biased without an explicit residual direction. In contrast, ScaleResfusion starts from a noisy LQ state and follows a residual-oriented path, yielding more coherent trajectories and lower endpoint error. A formal measure-transport analysis is provided in Appendix A.

Overall, the above analysis shows that residual modeling provides both a compact learning target and a better-aligned transport perspective. This motivates a residual-oriented flow that starts from a noise LQ state and transports along an explicit correction direction.

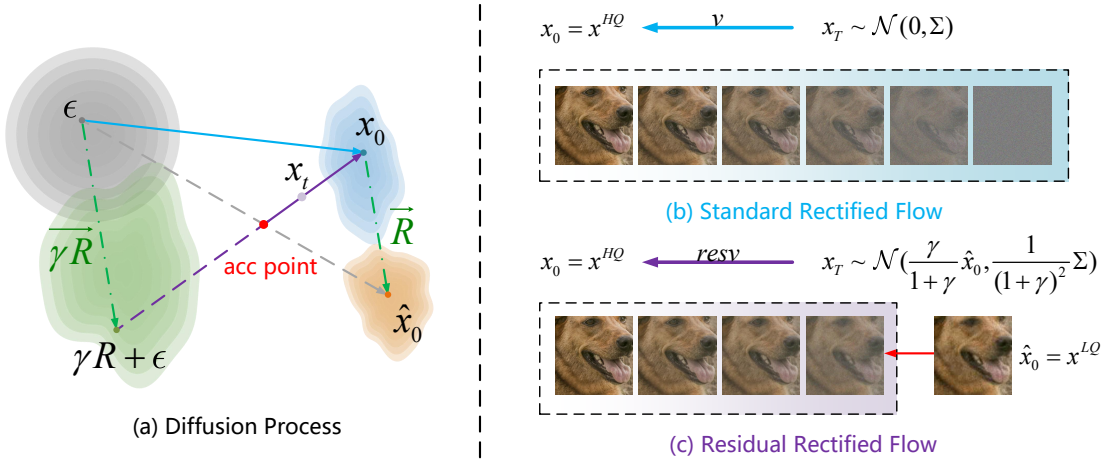


Fig. 4. Geometric Interpretation of Residual Rectified Flow (RRF). (a) RRF introduces the weighted residual term γR into the diffusion path, yielding a straight trajectory that intersects the implicit noise-to-LQ path at the acceleration point. (b) Standard Rectified Flow transports Gaussian noise to the HQ image distribution through a linear ODE. (c) Starting from the noisy-LQ state $x_{t^\star} \sim \mathcal{N}\left(\frac{\gamma}{1+\gamma}\hat{x}_0, \frac{1}{(1+\gamma)^2}\Sigma\right)$, RRF learns the residual vector field $resv$ to recover the HQ image x_0 over a shortened sampling interval.

The next subsection formalizes this idea as a continuous Residual Rectified Flow trajectory.

3.3 Residual Rectified Flow

Following the residual definition in Eq. (4), the proposed **Residual Rectified Flow** (RRF) introduces the weighted residual term into the RF trajectory in Eq. (1). Given a residual ratio $\gamma > 0$, $t \in [0, 1]$, and $\epsilon \sim \mathcal{N}(0, I)$, RRF constructs

$$x_t = (1-t)x_0 + t\gamma R + t\epsilon. \tag{5}$$

The trajectory is a straight line connecting the HQ image x_0 and the endpoint $\gamma R + \epsilon$, whose target velocity is constant with respect to t:

$$u_t(x_t \mid x_0, \epsilon, R) = \frac{dx_t}{dt} = resv = \gamma R + \epsilon - x_0. \tag{6}$$

As shown in Fig. 4, the blue arrow represents the linear diffusion process of Standard Rectified Flow (RF), which is an ODE from the Gaussian distribution to the HQ image distribution. While the purple arrow represents the proposed Residual Rectified Flow (RRF), which constructs a continuous path between x_0 and a noisy weighted residual term $\gamma R + \epsilon$. In the diffusion process of RRF, it implicitly reduces the residual term R between x_0 and $\hat{x}_0$ with the residual vector field. By the definition of similar triangles, we can obtain the expression for x_t as in Eq. (5), where the coefficient of γR is t.

When we define $\hat{x}_t = (1-t)\hat{x}_0 + t\epsilon$ with the same Gaussian noise ϵ as in x_t, the gray arrow represents an implicit Rectified Flow diffusion process from the Gaussian distribution to the LQ image $\hat{x}_0$ distribution. Due to the residual term R, the trajectories of RRF and the implicit Rectified Flow are constrained on the same hyperplane, and their **intersection** is the acceleration point [Shi et al. 2024]. The proposed Residual Rectified Flow admits an exact acceleration point due to its continuous ODE trajectory. Substituting $R = \hat{x}_0 - x_0$ into Eq. (5) gives:

$$x_t = (1-t)x_0 + t\gamma(\hat{x}_0 - x_0) + t\epsilon = (1-(1+\gamma)t)x_0 + t\gamma\hat{x}_0 + t\epsilon. \tag{7}$$

The coefficient of x_0 vanishes when $1-(1+\gamma)t^\star = 0$, which leads to the exact acceleration point $t^\star = \frac{1}{1+\gamma}$. At this point, the state of RRF can be completely represented by the LQ image $\hat{x}_0$ and noise ϵ, no longer depending on the unknown HQ image x_0, thereby enabling accelerated sampling. Accordingly, RRF trains a neural **residual velocity field** $resv_\theta(x_t, t)$ only over the effective interval $[0, t^\star]$:

$$\mathcal{L}_{RRF} = \int_0^{t^\star} \mathbb{E}\left[\|resv_\theta(x_t, t) - (\gamma R + \epsilon - x_0)\|^2\right] dt. \tag{8}$$

In the diffusion process of RRF, it actually starts from the acceleration point $t^\star$ with the initial distribution being $\mathcal{N}\left(\frac{\gamma}{1+\gamma}\hat{x}_0, \frac{1}{(1+\gamma)^2}\Sigma\right)$. We can obtain its signal-to-noise ratio (SNR) as

$$\mathrm{SNR}(x_{t^\star}) \propto \frac{\frac{\gamma}{1+\gamma}}{\frac{1}{1+\gamma}} = \gamma. \tag{9}$$

Thus, the residual ratio γ directly controls the SNR level of the starting point: larger γ yields an initialization closer to the LQ image, while smaller γ injects stronger Gaussian randomness. During sampling, RRF starts from $x_{t^\star}$ and integrates the learned ODE $x_0 = x_{t^\star} + \int_{t^\star}^0 resv_\theta(x_\tau, \tau)d\tau$ to the data endpoint. As formalized in Eq. (5) and Eq. (6), and detailed in Algorithm 1, RRF preserves the Standard Rectified Flow structure and only introduces residual offsets: $x_t = x_t^{RF} + t\gamma R$ and $resv = v^{RF} + \gamma R$. Therefore, adapting a pre-trained diffusion model to RRF does not require relearning the image-noise transport, but only fitting a compact residual correction. We thus adopt **pure LoRA** [Hu et al. 2022] as the PEFT strategy, keeping the pre-trained backbone fixed while learning low-rank updates for the residual vector field. The resulting parameter-efficient implementation is described in the following training pipeline.

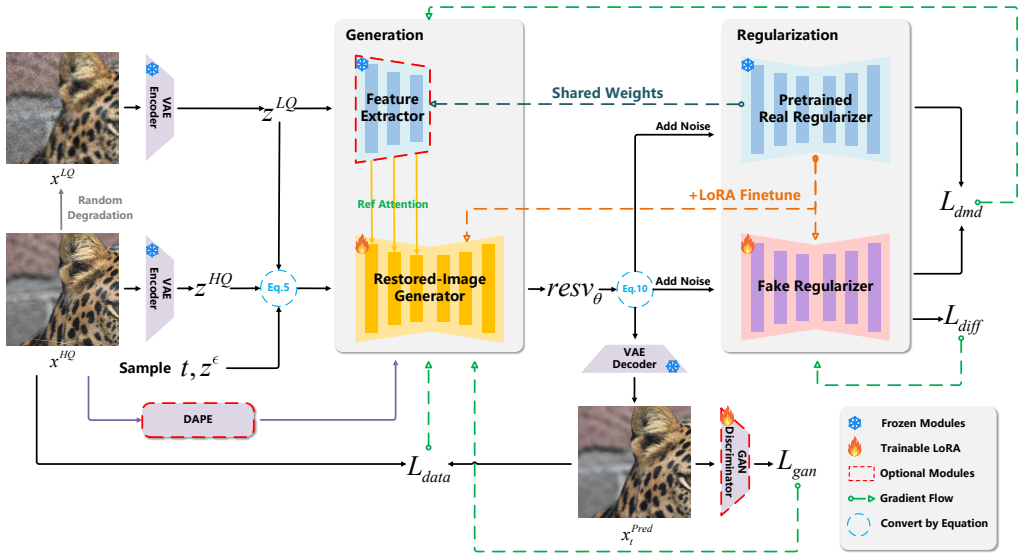


Fig. 5. Overview of our Training Pipeline, which consists of a generation module and a regularization module. In the generation module, the restored-image generator predicts the residual vector field under Residual Rectified Flow, with LQ features injected through Ref Attention and DAPE providing textual constraints. The generation and regularization modules share the same pre-trained frozen backbone weights, while LoRA adapters enable efficient adaptation. In the regularization module, DMD and optional GAN losses regularize the restored distribution toward natural HQ images.

3.4 Knowledge-Distilled Parameter-Efficient Training Pipeline

Based on the residual theory above, we design **ScaleResfusion**, a parameter-efficient training pipeline that distills pre-trained diffusion priors into an RRF generator. As shown in Fig. 5, ScaleResfusion contains a generation branch and a regularization branch. Both branches share the same frozen pre-trained backbone weights, while trainable LoRA adapters learn the residual task update.

Given an LQ–HQ pair $(\hat{x}_0, x_0)$, we encode images into the latent space with a frozen VAE and train a LoRA-adapted generator G_θ to predict the residual velocity $resv_\theta = G_\theta(x_t, t)$. The LQ features are injected into the restored-image generator through Ref Attention [Hu 2024; Shi et al. 2025]. We further use DAPE [Wu et al. 2024b] to provide additional textual constraints. For notation simplicity, x_t denotes the latent RRF state, and image-space losses are computed after VAE decoding. We convert the residual velocity output into an equivalent prediction to compute losses on the predicted clean restoration, where the VAE decoder is omitted in notation for image-space losses:

$$x_0^{pred} = x_t - t\, resv_\theta(x_t, t). \tag{10}$$

Data term. The data objective enforces pairwise fidelity and perceptual consistency:

$$\mathcal{L}_{data} = \lambda_{mse}\mathcal{L}_{MSE}(x_0^{pred}, x_0) + \lambda_{lpips}\mathcal{L}_{LPIPS}(x_0^{pred}, x_0). \tag{11}$$

Here $\mathcal{L}_{MSE}$ penalizes pixel-wise reconstruction errors and preserves the pairwise fidelity to the HQ target, while $\mathcal{L}_{LPIPS}$ measures perceptual similarity in a deep feature space [Zhang et al. 2018a] to encourage visually consistent structures and textures.

Regularization term. To align the restored distribution with natural HQ images, we introduce a regularization term that combines a DMD loss and an optional GAN loss:

$$\mathcal{L}_{reg} = \lambda_{dmd}\mathcal{L}_{DMD}(x_0^{pred}) + \lambda_{gan}\mathcal{L}_{GAN}(x_0^{pred}). \tag{12}$$

The DMD term serves as a knowledge-distillation objective following the real/fake regularizer design [Yin et al. 2024b]. Specifically, the frozen pre-trained regularizer acts as the teacher that provides the real diffusion prior, and the RRF generator is the student that learns to produce samples aligned with this prior. The LoRA-adapted fake regularizer is trained with $\mathcal{L}_{diff}$ on generated samples to estimate the generator-induced distribution, providing the fake score needed by DMD. Through this teacher–student distillation, $\mathcal{L}_{DMD}$ transfers the natural-image prior of the pre-trained diffusion model to the RRF generator and encourages restored images to stay on the HQ image manifold. The optional GAN term further sharpens local details by adversarially matching the restored distribution to real HQ images.

4 Experiment

Datasets. Following Dong et al. [2025]; Wang et al. [2024]; Wu et al. [2024a], we use LSDIR [Li et al. 2023] and the first 10K face images from FFHQ [Karras et al. 2019] for training. LR-HR pairs are synthesized with the Real-ESRGAN degradation pipeline [Wang et al. 2021b]. We use RealSR [Cai et al. 2019], DRealSR [Wei et al. 2020], and DIV2K-Val [Agustsson and Timofte 2017] for evaluation, following the same setup as StableSR [Wang et al. 2024]. Additionally, we employ a more challenging benchmark constructed by center-cropping 512×512 patches from the original LSDIR-Val split [Li et al. 2023] to further compare the performance across models.

Evaluation Metrics. To verify the performance of different models, we employ structural metrics, perceptual metrics, distribution consistency metrics, and no-reference metrics. Specifically, we report PSNR and SSIM [Wang et al. 2004] (computed on the Y channel of

Table 1. Quantitative comparison across real-world and synthetic benchmarks. Higher values indicate better performance for PSNR, SSIM, MUSIQ, and MANIQA, while lower values indicate better performance for LPIPS, DISTS, FID, and NIQE. Methods include diffusion-based baselines and our FLUX2-4B variant. The best and second-best results of each metric are highlighted in red and blue. We also report NFE and inference time to compare efficiency. **See the Appendix for the full comparison and more datasets.**

Datasets	Method	NFE	Inf. Time (ms)	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑
DRealSR	ResShift	15	848	27.05	0.74	0.39	0.26	159.90	8.65	51.24	0.47
	StableSR	200	12,036	28.04	0.75	0.33	0.23	144.15	6.60	58.53	0.56
	CCSR	6	516	28.24	0.78	0.32	0.23	157.30	6.81	66.28	0.61
	SeeSR	50	4,445	28.07	0.77	0.32	0.23	147.37	6.41	65.09	0.61
	SUPIR	50	25,252	25.09	0.65	0.42	0.28	169.48	7.39	58.79	0.55
	OSEDiff	1	266	27.92	0.78	0.30	0.22	135.41	6.46	64.69	0.59
	Ours (w/o GAN)	**4**	646	**29.77**	**0.82**	**0.25**	**0.20**	**118.18**	6.95	62.42	0.61
	Ours (w/ GAN)	**4**	646	28.26	0.78	0.29	0.21	124.03	**6.21**	**65.16**	**0.64**
RealSR	ResShift	15	848	25.66	0.74	0.33	0.25	128.03	8.07	56.89	0.51
	StableSR	200	12,036	24.62	0.70	0.31	0.22	128.54	5.78	65.48	0.62
	CCSR	6	516	25.92	0.75	0.28	0.21	122.84	5.73	69.18	0.64
	SeeSR	50	4,445	25.15	0.72	0.30	0.22	125.30	5.40	69.81	0.65
	SUPIR	50	25,252	23.65	0.66	0.35	0.25	130.38	6.11	62.09	0.58
	OSEDiff	1	266	25.15	0.73	0.29	0.21	123.53	5.65	69.08	0.63
	Ours (w/o GAN)	**4**	646	**27.05**	**0.78**	**0.24**	**0.20**	**104.74**	6.21	67.25	0.64
	Ours (w/ GAN)	**4**	646	25.78	0.75	0.26	**0.20**	106.22	**5.29**	**69.86**	**0.68**
DIV2K-Val	ResShift	15	848	24.59	0.62	0.31	0.21	30.81	6.92	58.90	0.53
	StableSR	200	12,036	23.27	0.57	0.31	0.20	24.95	4.77	65.78	0.62
	CCSR	6	516	24.30	0.63	0.30	0.20	30.84	5.34	69.53	0.61
	SeeSR	50	4,445	23.68	0.60	0.32	0.20	25.89	4.81	68.66	0.62
	SUPIR	50	25,252	22.13	0.53	0.39	0.23	31.40	5.68	63.86	0.59
	OSEDiff	1	266	23.72	0.61	0.29	0.20	26.34	4.71	67.96	0.61
	Ours (w/o GAN)	**4**	646	**24.90**	**0.64**	0.27	0.18	22.90	5.09	66.10	0.63
	Ours (w/ GAN)	**4**	646	23.63	0.60	**0.26**	**0.17**	**19.47**	**4.44**	**69.30**	**0.66**

the YCbCr color space) for distortion fidelity, LPIPS [Zhang et al. 2018a] and DISTS [Ding et al. 2020] for perceptual similarity, FID [Heusel et al. 2017] for distribution alignment, and NIQE [Zhang et al. 2015], MUSIQ [Ke et al. 2021], and MANIQA [Yang et al. 2022] as no-reference quality indicators.

Implementation Details. We train ScaleResfusion models with SD3 (2B) [Esser et al. 2024], FLUX2-Klein (4B) [Labs 2025], Z-Image (6B) [Cai et al. 2025], and FLUX2-Klein (9B) as backbones. By default, we use 4-step sampling (we discuss the effects of different sampling steps in the Appendix F.2) and the residual ratio $\gamma = 1$ (we discuss the impact of different noise-to-signal ratio initializations in Appendix F.3). We employ RAM-based [Zhang et al. 2024] DAPE following OSEDiff [Wu et al. 2024a] (we discuss the effects of the DAPE in the Appendix F.5). We use a DINOv2-based [Oquab et al. 2023] GAN discriminator identical to AddSR [Xie et al. 2024]. Detailed experimental settings are provided in Appendix A.4.

Comparison with Existing Methods. Table 1 reports comparisons on RealSR, DRealSR, and DIV2K-Val, with LSDIR results and the full method comparison provided in Appendix D. We compare ScaleResfusion with StableSR [Wang et al. 2024], SUPIR [Yu et al. 2024], CCSR [Sun et al. 2024], OSEDiff [Wu et al. 2024a], ResShift [Yue et al. 2023], and SeeSR [Wu et al. 2024b]. These methods cover iterative diffusion models, variants that start diffusion from noisy LQ images, and efficient one-step or few-step generators, providing a representative comparison for Real-IR. Unless otherwise specified, we use the ScaleResfusion built on FLUX2-Klein (4B) as the default model. On DRealSR and RealSR, the *w/o GAN* variant achieves the best overall performance across PSNR, SSIM, LPIPS, DISTS, and FID; it improves PSNR from 28.24 to 29.77 on DRealSR and from 25.92 to 27.05 on RealSR, while reducing FID from 157.30 to 118.18 on DRealSR and from 122.84 to 104.74 on RealSR. The *w/ GAN* variant further improves perceptual quality, obtaining the best LPIPS, DISTS, FID, NIQE, and MANIQA on DIV2K-Val and the best NIQE, MUSIQ, and MANIQA on DRealSR and RealSR. This perception–distortion trade-off [Blau and Michaeli 2018; Luo et al. 2024; Zhang et al. 2022a] is consistent with prior findings [Xie et al. 2024; Yu et al. 2024]. Therefore, *w/o GAN* is preferable for reconstruction fidelity, whereas *w/ GAN* is better suited for perceptual realism.

User Study. User preference evaluation adopts a pairwise Arena design: in each evaluation session, users are presented with results generated by two different models under identical input conditions and make selections based on subjective visual quality and overall preference. The study includes 48 valid independent evaluators and 1,756 comparisons. As shown in Fig. 7, in cross valid comparisons, ScaleResfusion is preferred in 71.8% of cases, compared with 18.4% for baselines and 9.8% for Tie/BothBad. It also achieves a 75.8% overall score in the Random Arena ranking, clearly outperforming all baselines. These results show that ScaleResfusion consistently delivers stronger subjective visual quality and user preference under mixed pairwise comparisons.

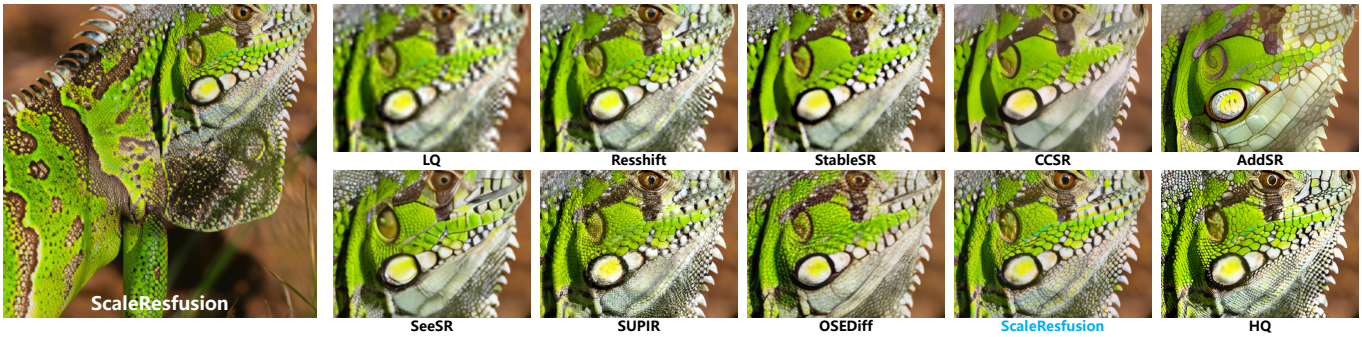


Fig. 6. Visual comparison with existing diffusion-based restoration methods. The left side shows the result of ScaleResfusion, while the right side compares local crops. Overall, ScaleResfusion better preserves the global structure while recovering more natural details without the over-smoothing or hallucinated patterns observed in competing methods.

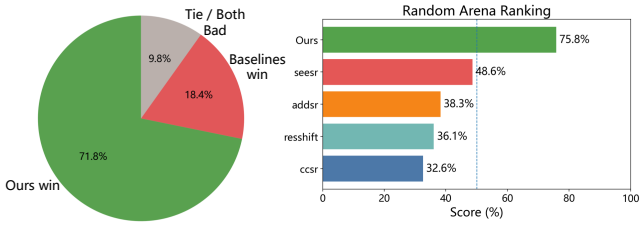


Fig. 7. Pairwise Arena user study results. Left: preference distribution when comparing ScaleResfusion against other methods, where ScaleResfusion wins 71.8% of comparisons. Right: Random Arena ranking, where ScaleResfusion achieves the highest overall score of 75.8%, indicating consistently stronger user preference.

Table 2. Ablation results of ScaleResfusion (w/ GAN) on backbone scaling and module design, where the module ablation is conducted on FLUX2-4B. Table (a) presents results on DRealSR, and Table (b) presents results on DIV2K-Val. See the Appendix for the full ablation results.

(a) Backbone Scaling

Backbone	Parameters	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓
SD3	2B	27.85	0.76	0.32	0.22	146.92
FLUX2	4B	28.26	0.77	0.29	**0.20**	**124.02**
ZImage	6B	**28.34**	**0.78**	**0.28**	**0.20**	127.39
FLUX2	9B	**28.77**	**0.79**	**0.27**	**0.19**	**116.69**

(b) Module Ablation (on FLUX2-4B)

Module				PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓
Res-init	RefNet	DMD	GAN					
×	✓	✓	×	22.60	0.55	0.35	0.23	37.58
✓	×	✓	×	23.52	0.60	0.28	0.21	23.93
✓	✓	×	×	21.90	0.51	0.38	0.24	51.25
✓	✓	✓	×	24.90	0.64	0.27	0.18	22.90
✓	✓	✓	✓	23.63	0.60	0.26	0.17	19.47

5 Ablation Studies

Backbone Scaling. Table 2 (a) and Fig. 8 show that ScaleResfusion scales favorably with stronger rectified-flow backbones, consistently improving fidelity, perceptual similarity, and distribution alignment from SD3 (2B) to FLUX2 (9B). This indicates that the proposed residual rectified-flow formulation can effectively inherit richer generative priors from larger pre-trained models rather than being tied to a specific backbone. FLUX2 (9B) achieves the best reconstruction and distribution alignment, while ZImage (6B) obtains the best MUSIQ score, suggesting that different pre-trained priors emphasize different aspects of perceptual quality. FLUX2 (4B) offers the best practical quality-efficiency trade-off, achieving competitive DISTS and FID with much lower inference cost, and is therefore used as our default backbone. More ablation studies are provided in Appendix F.

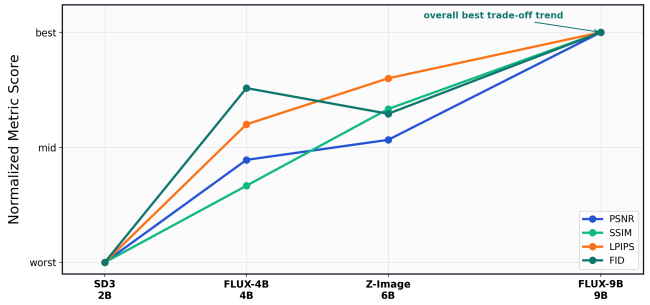


Fig. 8. Backbone scaling results of ScaleResfusion in Table 2 (a). Increasing the capacity of the pre-trained rectified-flow backbone consistently improves reconstruction fidelity and distribution alignment, while FLUX2-4B provides the best quality-efficiency trade-off.

Module Ablation. Table 2 (b) evaluates the key modules on FLUX2 (4B). Removing residual initialization causes the largest fidelity drop, confirming that starting from a residual-aware noisy LQ state is crucial for preserving image structure. Removing RefNet also degrades PSNR, SSIM, and perceptual metrics, showing the importance of injecting LQ reference features during restoration. With Res-init, RefNet, and DMD together, ScaleResfusion achieves the best distortion-oriented performance. Adding GAN further improves perceptual and distribution quality, reducing LPIPS, DISTS, and FID, while introducing the expected perception–distortion trade-off in PSNR and SSIM.

6 Conclusion

We presented ScaleResfusion, a scalable Real-IR framework that adapts pre-trained rectified-flow diffusion models through Residual Rectified Flow. By starting from a noisy LQ state and learning a residual vector field, it preserves input structure, retains diffusion sampling, and enables LoRA-based efficient training with knowledge distillation. Experiments, user studies, and ablations show strong restoration quality, clear subjective preference, favorable backbone scaling, and validate residual-aware design for scalable generative restoration.

References

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

Yochai Blau and Tomer Michaeli. 2018. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6228–6237.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. 2025. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699* (2025).

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. 2019. 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*. 3086–3095.

Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. 2022. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the 30th ACM international conference on multimedia*. 1329–1338.

Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. 2021. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12299–12310.

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

Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11065–11074.

Mauricio Delbracio and Peyman Milanfar. 2023. Inversion by direct iteration: An alternative to denoising diffusion for image restoration. *arXiv preprint arXiv:2303.11435* (2023).

Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. 2024. TurboEdit: Text-Based Image Editing Using Few-Step Diffusion Models. In *SIGGRAPH Asia 2024 Conference Papers*. 1–12.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence* 44, 5 (2020), 2567–2581.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2014. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*. Springer, 184–199.

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. 2025. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23174–23184.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. 2023. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14049–14058.

Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. 2025. Diffusion models in low-level vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Bingwen Hu, Heng Liu, Zhedong Zheng, and Ping Liu. 2025. CLIP-SR: Collaborative Linguistic and Image Processing for Super-Resolution. *IEEE Transactions on Multimedia* (2025).

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Iclr* 1, 2 (2022), 3.

Li Hu. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8153–8163.

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, David Liu, Changsheng Lu, Zhen Li, et al. 2025. Distribution matching distillation meets reinforcement learning. *arXiv preprint arXiv:2511.13649* (2025).

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. 2022. Denoising diffusion restoration models. *Advances in neural information processing systems* 35 (2022), 23593–23606.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5148–5157.

Black Forest Labs. 2025. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4681–4690.

Jianze Li, Jiezhang Cao, Yong Guo, Wenbo Li, and Yulun Zhang. 2025. One diffusion step to real-world super-resolution via flow trajectory distillation. *arXiv preprint arXiv:2502.01993* (2025).

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. 2023. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1775–1787.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1833–1844.

Jie Liang, Hui Zeng, and Lei Zhang. 2022. 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*. 5657–5666.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. 2024. Diffbir: Toward blind image restoration with generative diffusion prior. In *European conference on computer vision*. Springer, 430–448.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. 2023. I^2SB: Image-to-Image Schrödinger Bridge. *arXiv preprint arXiv:2302.05872* (2023).

Jiawei Liu, Qiang Wang, Huijie Fan, Yinong Wang, Yandong Tang, and Liangqiong Qu. 2024a. Residual denoising diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2773–2783.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003* (2022).

Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. 2024b. Text-guided texturing by synchronized multi-view diffusion. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

Xiaotong Luo, Yuan Xie, Yanyun Qu, and Yun Fu. 2024. SkipDiff: Adaptive Skip Diffusion Model for High-Fidelity Perceptual Image Super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4017–4025.

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. 2023a. Image restoration with mean-reverting stochastic differential equations. *arXiv preprint arXiv:2301.11699* (2023).

Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. 2023b. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1680–1691.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

Ozan Özdenizci and Robert Legenstein. 2023. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE transactions on pattern analysis and machine intelligence* 45, 8 (2023), 10346–10357.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

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

Zhenning Shi, Zizheng Yan, Yuhang Yu, Clara Xue, Jingyu Zhuang, Qi Zhang, Jinwei Chen, Tao Li, and Qingnan Fan. 2025. Ultra-High-Definition Reference-Based Landmark Image Super-Resolution with Generative Diffusion Prior. *arXiv preprint arXiv:2508.10779* (2025).

Zhenning Shi, Haoshuai Zheng, Chen Xu, Changsheng Dong, Bin Pan, Xueshuo Xie, Along He, Tao Li, and Huazhu Fu. 2024. Resfusion: Denoising diffusion probabilistic models for image restoration based on prior residual noise. *Advances in Neural Information Processing Systems* 37 (2024), 130664–130693.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020a. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020b. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

Lingchen Sun, Rongyuan Wu, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. 2024. Improving the Stability of Diffusion Models for Content Consistent Super-Resolution. *arXiv preprint arXiv:2401.00877* (2024).

Hebaixu Wang, Jing Zhang, Haoyang Chen, Haonan Guo, Di Wang, Jiayi Ma, and Bo Du. 2025b. Residual Diffusion Bridge Model for Image Restoration. *arXiv preprint arXiv:2510.23116* (2025).

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. 2024. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision* 132, 12 (2024), 5929–5949.

Tao Wang, Kaihao Zhang, Yong Zhang, Wenhan Luo, Björn Stenger, Tong Lu, Tae-Kyun Kim, and Wei Liu. 2025c. LLDiffusion: Learning degradation representations in diffusion models for low-light image enhancement. *Pattern Recognition* 166 (2025), 111628.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021a. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9168–9178.

Xin Wang, Di Lin, Wanchao Su, Ji Du, Renjie Zhang, Jie Zhang, Haotian Dong, Ke Xu, Qing Guo, and Ping Li. 2025a. HRC-Net: Learning Visual Hypothesis, Representative, and Collaboration for Multi-Domain Image Inpainting. *ACM Transactions on Graphics (TOG)* 44, 6 (2025), 1–13.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021b. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1905–1914.

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

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

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. 2020. 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*. Springer, 101–117.

Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. 2022. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16293–16303.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. 2024a. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems* 37 (2024), 92529–92553.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. 2024b. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 25456–25467.

Rui Xie, Ying Tai, Kai Zhang, Zhenyu Zhang, Jun Zhou, and Jian Yang. 2024. AddSR: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *arXiv preprint arXiv:2404.01717* (2024).

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1191–1200.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. 2024. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European conference on computer vision*. Springer, 74–91.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. 2024a. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems* 37 (2024), 47455–47487.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. 2024b. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6613–6623.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. 2024. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 25669–25680.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. 2023. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in neural information processing systems* 36 (2023), 13294–13307.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. 2021. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4791–4800.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. 2017. Learning deep CNN denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3929–3938.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3836–3847.

Lin Zhang, Lei Zhang, and Alan C Bovik. 2015. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing* 24, 8 (2015), 2579–2591.

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

Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. 2022b. Efficient long-range attention network for image super-resolution. In *European conference on computer vision*. Springer, 649–667.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2024. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1724–1732.

Yuehan Zhang, Bo Ji, Jia Hao, and Angela Yao. 2022a. Perception-distortion balanced ADMM optimization for single-image super-resolution. In *European Conference on Computer Vision*. Springer, 108–125.

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

Yuqing Zhang, Hao Xu, Yiqian Wu, Sirui Chen, Sirui Lin, Xiang Li, Xifeng Gao, and Xiaogang Jin. 2025. AlignTex: Pixel-Precise Texture Generation from Multi-View Artwork. *ACM Transactions on Graphics (TOG)* 44, 4 (2025), 1–12.

Zirui Zhang and Kevin Zhang. 2025. Boundary-Conditioned Inpainting for Constraint-Consistent Procedural Generation Completion. In *Proceedings of the SIGGRAPH Asia 2025 Technical Communications*. 1–4.

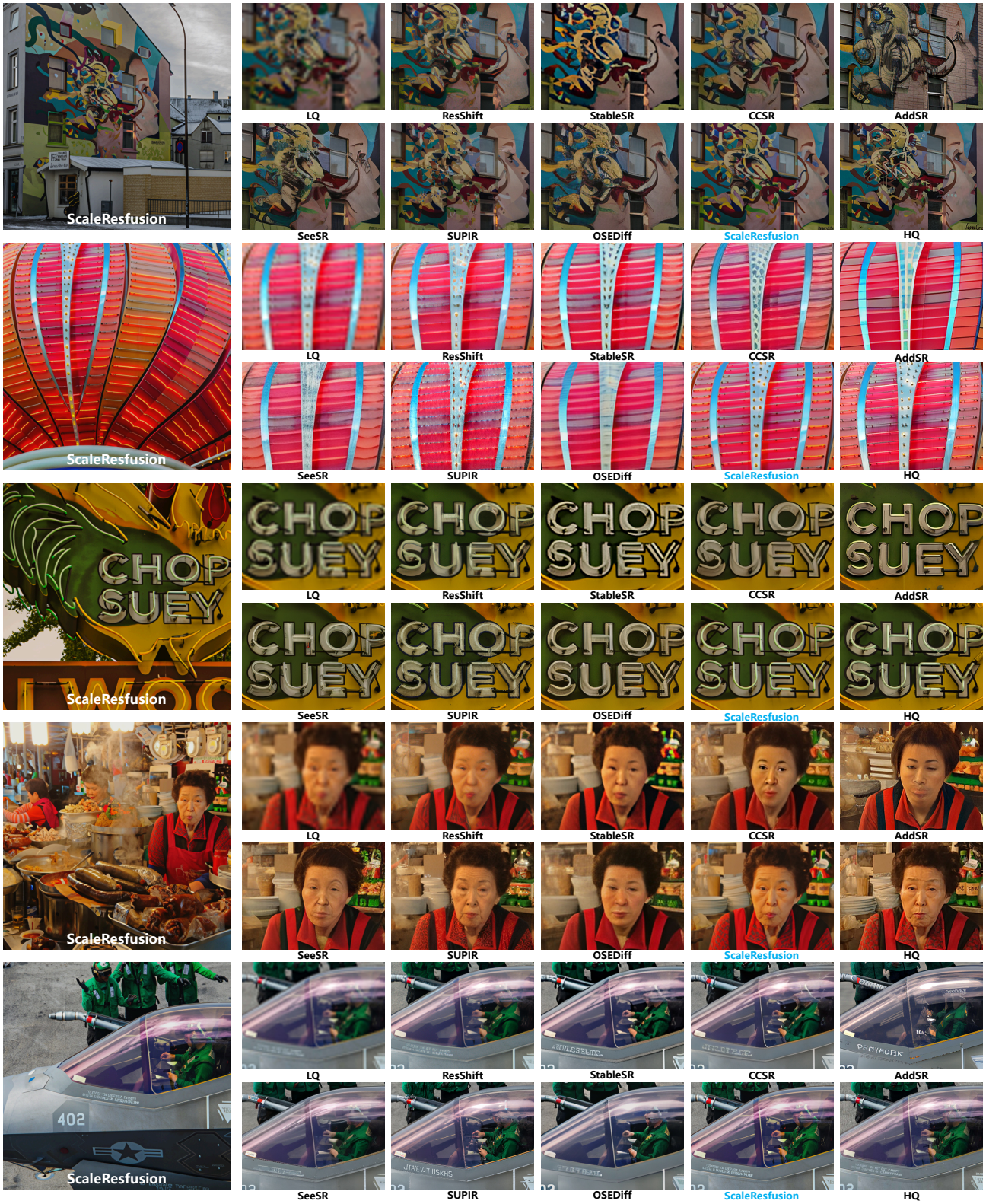


Fig. 9. More visual comparison with existing diffusion-based restoration methods. The left side shows the results of ScaleResfusion, while the right side compares local crops. Overall, ScaleResfusion better preserves the global structure while recovering more natural details without the over-smoothing or hallucinated patterns observed in competing methods.

Fig. 10. Visual comparison on WebPhoto-Test [Wang et al. 2021a], an in-the-wild face restoration benchmark without paired ground truth. In this zero-shot setting, ScaleResfusion better handles out-of-distribution degradations and restores more natural facial details than competing methods.

ScaleResfusion: Residual Rectified Flow based on Residual Vector Field

Supplementary Materials

Contents

A Residual Initialization Leads to a Better Transport Path **13**

B Algorithm **15**

C Experimental Setting Details **15**

D Comparison with Other Methods **15**

E Caution on No-Reference Metrics **17**

F Extra Ablation Study **18**

F.1 Ablation on Res-init 18

F.2 Ablation on Number of Function Evaluations (NFE) 20

F.3 Ablation on Residual Ratio 20

F.4 Ablation on DMD Regularization 20

F.5 Ablation on DAPE 21

F.6 Ablation on Feature Extractor 22

F.7 Ablation on LoRA Rank 23

G Theoretical Justification of Residual Rectified Flow **23**

H More Visual Results **25**

A Residual Initialization Leads to a Better Transport Path

In this subsection, we provide a theoretical justification for residual initialization from the perspective of measure differential equations. Our key goal is to show that the reverse process induced by the learned residual velocity field is stable with respect to its initial measure. Consequently, in image restoration, an initialization that is closer to the restoration manifold leads to a terminal distribution that is closer to the target HQ distribution. We further show that, conditioned on the observed LQ image, the residual variable R is the most natural task-aware quantity for constructing such a structured initialization.

From sample trajectories to measure flows. After the acceleration point $t^\star = \frac{1}{1+\gamma}$, the reverse process of Residual Rectified Flow is governed by the ODE

$$\frac{dX_t}{dt} = v_\theta(X_t, t), \qquad t \in [0, t^\star], \tag{13}$$

where v_θ denotes the learned residual velocity field and $X_{t^\star}$ is a random initial state. Let

$$\mu_t := \mathrm{Law}(X_t)$$

be the probability distribution of X_t at time t. Under standard regularity assumptions, the family of measures $\{\mu_t\}_{t\in[0,t^\star]}$ satisfies the continuity equation

$$\partial_t \mu_t + \nabla \cdot \big(v_\theta(\cdot, t)\mu_t\big) = 0. \tag{14}$$

Let $\Phi_{t,s}$ denote the flow map induced by Eq. (13), namely

$$X_t = \Phi_{t,s}(X_s).$$

Then the corresponding measure solution can be written in push-forward form as

$$\mu_t = (\Phi_{t,s})_\# \mu_s. \tag{15}$$

Therefore, the reverse generation process can be interpreted as a measure transport operator

$$\mathcal{T}_{t\leftarrow s}(\mu_s) := (\Phi_{t,s})_\# \mu_s,$$

which transports an initial measure at time s to a terminal measure at time t along the learned residual flow.

Stability of the reverse measure flow. To quantify the effect of initialization, consider two initial measures μ_s and ν_s, and let

$$\mu_t = (\Phi_{t,s})_\# \mu_s, \qquad \nu_t = (\Phi_{t,s})_\# \nu_s$$

be the corresponding evolved measures under the same velocity field v_θ. Assume first that $v_\theta(\cdot, t)$ is Lipschitz continuous with respect to the state variable, i.e., there exists a constant $L > 0$ such that

$$\|v_\theta(x, t) - v_\theta(y, t)\| \le L\|x - y\|, \qquad \forall x, y,\ \forall t \in [0, t^\star]. \tag{16}$$

Let X_t and Y_t be two characteristic trajectories satisfying

$$\frac{dX_t}{dt} = v_\theta(X_t, t), \qquad \frac{dY_t}{dt} = v_\theta(Y_t, t).$$

Then

$$\frac{d}{dt}(X_t - Y_t) = v_\theta(X_t, t) - v_\theta(Y_t, t).$$

By Eq. (16) and Grönwall's inequality, we obtain

$$\|X_t - Y_t\| \le e^{L(t-s)}\|X_s - Y_s\|. \tag{17}$$

Lifting Eq. (17) to the measure level through an optimal coupling yields the Wasserstein stability estimate

$$W_2(\mu_t, \nu_t) \le e^{L(t-s)} W_2(\mu_s, \nu_s). \tag{18}$$

Equation (18) shows that the reverse flow depends continuously on its initial measure. In other words, the discrepancy between terminal distributions is controlled by the discrepancy between their initializations.

Non-expansiveness and contraction. Under the standard Lipschitz assumption, Eq. (18) only guarantees that the induced flow operator is $e^{L(t-s)}$-Lipschitz in Wasserstein space. A stronger conclusion can be obtained if the velocity field additionally satisfies the one-sided Lipschitz condition

$$\langle x - y,\ v_\theta(x, t) - v_\theta(y, t)\rangle \le \lambda\|x - y\|^2, \qquad \forall x, y. \tag{19}$$

In this case,

$$\frac{1}{2}\frac{d}{dt}\|X_t - Y_t\|^2 = \langle X_t - Y_t,\ v_\theta(X_t, t) - v_\theta(Y_t, t)\rangle \le \lambda\|X_t - Y_t\|^2,$$

which implies

$$\|X_t - Y_t\| \le e^{\lambda(t-s)}\|X_s - Y_s\|. \tag{20}$$

Consequently, at the measure level,

$$W_2(\mu_t, \nu_t) \le e^{\lambda(t-s)} W_2(\mu_s, \nu_s). \tag{21}$$

Equation (21) leads to the following interpretation:
(i) if $\lambda = 0$, then

$$W_2(\mu_t, \nu_t) \le W_2(\mu_s, \nu_s),$$

and the reverse flow is non-expansive, i.e., 1-Lipschitz in Wasserstein distance;
(ii) if $\lambda < 0$, then

$$W_2(\mu_t, \nu_t) \le e^{\lambda(t-s)} W_2(\mu_s, \nu_s), \qquad e^{\lambda(t-s)} < 1,$$

and the reverse flow is contractive.

Hence, under an additional dissipativity assumption, the learned residual flow not only remains stable but also tends to reduce the discrepancy between different initial measures.

Why a structured initialization is preferable. The above stability result shows that the terminal discrepancy of the reverse flow is controlled by the discrepancy of its initialization. Let $\mu^\dagger_{t^\star}$ denote an ideal task-aware initial measure, $\mu^G_{t^\star}$ a task-agnostic Gaussian initialization, and $\mu^R_{t^\star}$ our structured residual initialization. If

$$W_2\Big(\mu^R_{t^\star}, \mu^\dagger_{t^\star}\Big) < W_2\Big(\mu^G_{t^\star}, \mu^\dagger_{t^\star}\Big), \tag{22}$$

then the stability estimate in Eq. (18) implies

$$W_2\Big(\mu^R_0, \mu^\dagger_0\Big) \le C_L W_2\Big(\mu^R_{t^\star}, \mu^\dagger_{t^\star}\Big) < C_L W_2\Big(\mu^G_{t^\star}, \mu^\dagger_{t^\star}\Big), \tag{23}$$

where $C_L = \exp(Lt^\star)$ and μ^R_0, μ^G_0, and $\mu^\dagger_0$ are the corresponding terminal measures. Therefore, Eq. (23) should be interpreted as a stability-controlled upper bound rather than an unconditional strict ordering between $W_2(\mu^R_0, \mu^\dagger_0)$ and $W_2(\mu^G_0, \mu^\dagger_0)$. It shows that, when the residual initialization is closer to the ideal task-aware initialization at $t^\star$, it admits a tighter terminal-discrepancy bound. This provides theoretical motivation for replacing task-agnostic

Gaussian initialization with a structured, observation-aware residual initialization.

Why the structured variable should be the residual. We now explain why the most natural structured variable for restoration is the residual R. Let $\hat{x}_0$ and x_0 denote the LQ and HQ images, respectively, and define

$$R = \hat{x}_0 - x_0. \tag{22}$$

Then

$$x_0 = \hat{x}_0 - R. \tag{23}$$

Therefore, conditioned on the observed degraded image $\hat{x}_0$, the target image x_0 and the residual variable R are in one-to-one correspondence. Recovering the HQ image is thus equivalent to recovering the residual. In other words, the restoration problem can be reformulated as learning a correction term rather than regenerating the full image from scratch.

Moreover, the target velocity in Residual Rectified Flow is

$$u_t = \gamma R + \epsilon - x_0. \tag{24}$$

Substituting Eq. (23) into Eq. (24) yields

$$u_t = (\gamma + 1)R + \epsilon - \hat{x}_0. \tag{25}$$

Equation (25) shows that once the observation $\hat{x}_0$ is given, the task-specific transport direction is primarily governed by the residual variable R. The degraded image $\hat{x}_0$ already provides the coarse structure and semantic layout, while ϵ only maintains the necessary stochasticity. The residual R, in contrast, explicitly captures the correction required to move from the degraded observation to the clean target. Therefore, if one aims to construct an initialization that is closer to the restoration manifold, the residual is the most direct and task-aligned choice.

Why residual-only initialization is still insufficient. The above discussion shows that the residual variable R is the most natural task-aware quantity for restoration. However, using only a deterministic residual initialization is still suboptimal. Suppose the initial state is constructed as a deterministic function of the observation, e.g.,

$$X_{t^\star}^{R\text{-only}} = g(\hat{x}_0, R).$$

Then, conditioned on $\hat{x}_0$, the induced initial measure degenerates to a Dirac mass:

$$\mu_{t^\star}^{R\text{-only}}(\cdot \mid \hat{x}_0) = \delta_{g(\hat{x}_0,R)}.$$

Since the reverse dynamics is governed by a deterministic ODE flow, its pushforward remains a Dirac measure:

$$\mu_0^{R\text{-only}}(\cdot \mid \hat{x}_0) = (\Phi_{0,t^\star})_{\#}\delta_{g(\hat{x}_0,R)} = \delta_{\Phi_{0,t^\star}(g(\hat{x}_0,R))}.$$

Therefore, a residual-only initialization cannot generate any conditional variability by itself.

This property becomes problematic when the target conditional HQ distribution is non-degenerate, i.e.,

$$\pi_0(\cdot \mid \hat{x}_0)$$

has non-zero conditional covariance. In that case, any deterministic terminal measure suffers from an irreducible distribution mismatch. For example, if $\pi_0(\cdot \mid \hat{x}_0)$ has mean m_π and covariance Σ_π, then

$$W_2^2(\delta_x, \pi_0(\cdot \mid \hat{x}_0)) = \|x - m_\pi\|^2 + \mathrm{Tr}(\Sigma_\pi),$$

Algorithm 1 Residual Rectified Flow: Main Algorithm

Training:
Require: Paired LQ-HQ samples $(\hat{x}_0, x_0) \sim D$, residual ratio γ.
$t^\star = \frac{1}{1+\gamma}$
repeat
  Sample $(\hat{x}_0^i, x_0^i) \sim D, \epsilon \sim N(0, I), t \sim \mathcal{U}(0, t^\star)$
  $R = \hat{x}_0 - x_0$
  $x_t = (1-t)x_0 + t\gamma R + t\epsilon$
  $resv = \gamma R + \epsilon - x_0$
  take gradient step on $\nabla_\theta ||resv - resv_\theta(x_t, t)||^2$
until convergence

Inference:
Require: LQ image $\hat{x}_0$, residual velocity field $resv_\theta$, residual ratio γ.
$t^\star = \frac{1}{1+\gamma}$
Sample $\epsilon \sim N(0, I)$
$x_{t^\star} = t^\star \gamma \hat{x_0} + t^\star \epsilon = \frac{\gamma}{1+\gamma}\hat{x}_0 + \frac{1}{1+\gamma}\epsilon$
Solve the ODE $\frac{dx_t}{dt} = resv_\theta(x_t, t)$ from $t = t^\star$ to 0
return x_0

which is lower bounded by $\mathrm{Tr}(\Sigma_\pi) > 0$ whenever $\Sigma_\pi \neq 0$.

In contrast, the proposed residual-based initialization preserves task structure while retaining stochastic flexibility:

$$X_{t^\star} = \frac{\gamma}{1+\gamma}\hat{x}_0 + \frac{1}{1+\gamma}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I).$$

Its conditional measure is therefore non-degenerate,

$$\mu_{t^\star}^{RN}(\cdot \mid \hat{x}_0) = \mathcal{N}\left(\frac{\gamma}{1+\gamma}\hat{x}_0, \frac{1}{(1+\gamma)^2}I\right),$$

which provides a better balance between structural alignment and distributional coverage. Hence, compared with both task-agnostic Gaussian initialization and purely deterministic residual initialization, the proposed residual-aware noisy initialization is more suitable for restoration-oriented reverse transport.

Implications for ScaleResfusion. The above analysis also explains why residual learning naturally aligns with low-rank parameter-efficient adaptation. In restoration tasks, $\hat{x}_0$ already preserves most of the image content, whereas R mainly contains restoration-specific corrections such as missing high-frequency details, denoising compensation, and deblurring offsets. As a result, the intrinsic complexity of the residual is typically lower than that of the full image signal. This suggests that the dominant task-specific transformation is concentrated in a lower-dimensional residual subspace rather than the entire image space. Consequently, parameter-efficient low-rank adaptation is sufficient to capture the principal residual transport, which provides theoretical support for adopting low-rank updates in ScaleResfusion.

In summary, the theoretical justification for residual initialization can be understood from three aspects. First, the reverse process can be formulated as a measure differential equation driven by the learned residual velocity field. Second, the induced measure

flow is stable with respect to the initial measure in Wasserstein distance, and becomes non-expansive or even contractive under an additional one-sided Lipschitz condition. Third, conditioned on the LQ observation, the residual variable R is in bijection with the target HQ image and dominates the task-relevant transport direction. Therefore, compared with task-agnostic Gaussian initialization, constructing a structured initial measure from R is better aligned with both the geometry of image restoration and the dynamics of the reverse flow.

B Algorithm

Based on the derivations from the Sec 3.3, the training and inference processes of Residual Rectified Flow can be represented as Algorithm 1. We highlight the modifications in our algorithm compared to Standard Rectified Flow in red.

Residual Rectified Flow constructs the forward path $x_t = (1-t)x_0 + t\gamma R + t\epsilon$ and learns the residual velocity $resv = \gamma R + \epsilon - x_0$ over $t \in [0, t^\star]$, where $t^\star = \frac{1}{1+\gamma}$. During inference, it starts from the acceleration point $x_{t^\star} = \frac{\gamma}{1+\gamma}\hat{x}_0 + \frac{1}{1+\gamma}\epsilon$ and solves the ODE $\frac{dx_t}{dt} = resv_\theta(x_t, t)$ from $t^\star$ to 0 to recover x_0. In this way, the reverse process can be interpreted as progressively removing the residual and noise components along the learned continuous trajectory.

As formalized in Eq. (5) and Eq. (6), and detailed in Algorithm 1, RRF preserves the Standard Rectified Flow structure and only introduces residual offsets: $x_t = x_t^{RF} + t\gamma R$ and $resv = v^{RF} + \gamma R$. Therefore, adapting a pre-trained diffusion model to RRF does not require relearning the image-noise transport, but only fitting a compact residual correction. We thus adopt **pure LoRA** [Hu et al. 2022] as the PEFT strategy, keeping the pre-trained backbone fixed while learning low-rank updates for the residual vector field.

C Experimental Setting Details

We train all models with PyTorch and the Diffusers framework following [Deutch et al. 2024; Hu et al. 2025; Liu et al. 2024b; Wang et al. 2025a; Zhang et al. 2025; Zhang and Zhang 2025], using the AdamW optimizer [Loshchilov and Hutter 2017] with default hyperparameters. For ScaleResfusion variants using SD3 (2B) and Z-Image (6B) as backbones, we use ReferenceNet to inject LQ features. For variants using FLUX2-Klein (4B, 9B) as backbones, we concatenate the LQ features with the input features along the feature dimension. This preserves the native feature extraction ability of FLUX2 for image editing. We train the SD3 and FLUX2-4B variants on 8 NVIDIA GeForce RTX 5090 GPUs, and the Z-Image and FLUX-9B variants on 8 NVIDIA RTX Pro 6000 GPUs.

We train the *w/o GAN* variant for 100K iterations with a learning rate of 5×10^{-5}, and then fine-tune it with a learning rate of 1×10^{-5} to obtain *w/ GAN*. We adopt the Dynamic Distribution Guidance (DynaDG) and Dynamic Renoise Sampling (DynaRS) following Jiang et al. [2025] for cold start during the first 10K iterations. By default, we use 4-step sampling and the residual ratio $\gamma = 1$. All testing and inference-time measurements are conducted on a single NVIDIA RTX A6000 GPU. For inference time evaluation, we test all methods at 512×512 resolution by processing 3,000 images and computing the average inference time per image after subtracting the model loading time from the total time. We employ RAM-based [Zhang et al. 2024] DAPE following OSEDiff [Wu et al. 2024a]. We use a DINOv2-based [Oquab et al. 2023] GAN discriminator identical to AddSR [Xie et al. 2024].

During training, we also observe a training–inference mismatch similar to the multi-step generator issue discussed in DMD2 [Yin et al. 2024a]. In the RRF training process, the x_0 component in the observed state x_t comes from the real HQ image. During inference, this component must instead be supplied by the generated prediction x_0^{pred}. With very few (1–4) inference steps, the x_0 component changes rapidly between adjacent sampling states, which further amplifies the input-distribution mismatch. This mismatch is less harmful for relatively deterministic restoration tasks such as low-light enhancement and deraining. However, it becomes critical for highly ill-posed problems such as super-resolution, where the generated x_0^{pred} can deviate substantially from the real x_0, leading to over-saturated and hallucinated textures. Following DMD2, we replace noisy real training states with noisy synthetic states produced by the current generator after several sampling steps.

D Comparison with Other Methods

Table 3 and Table 4 provide the full quantitative comparison on real-world benchmarks (DRealSR and RealSR) and synthetic benchmarks (DIV2K-Val and LSDIR-Val). We compare ScaleResfusion with GAN-based methods, including BSRGAN [Zhang et al. 2021], Real-ESRGAN [Wang et al. 2021b], LDL [Liang et al. 2022], and FeMaSR [Chen et al. 2022], as well as diffusion-based methods, including StableSR [Wang et al. 2024], SUPIR [Yu et al. 2024], TSD-SR [Dong et al. 2025], AddSR [Xie et al. 2024], CCSR [Sun et al. 2024], DiffBIR [Lin et al. 2024], OSEDiff [Wu et al. 2024a], PASD [Yang et al. 2024], ResShift [Yue et al. 2023], and SeeSR [Wu et al. 2024b]. These methods cover adversarial restoration models, iterative diffusion models, variants that start diffusion from noisy LQ images, and efficient one-step or few-step generators, providing a broad comparison for Real-IR. We further report ScaleResfusion variants built on SD3 (2B), FLUX2-Klein (4B), Z-Image (6B), and FLUX2-Klein (9B), with and without GAN fine-tuning. We summarize the main observations as follows:

(1) On real-world benchmarks, ScaleResfusion achieves the strongest overall fidelity and distribution alignment. On DRealSR, the FLUX2-9B *w/o GAN* variant reaches 30.17 PSNR, 0.82 SSIM, 0.25 LPIPS, 0.19 DISTS, and 109.28 FID, outperforming existing GAN-based and diffusion-based methods by a clear margin on most reference-based and distribution metrics. On RealSR, the FLUX2-9B *w/o GAN* variant obtains the best PSNR, SSIM, LPIPS, and FID, while the *w/ GAN* variant further improves DISTS. These results indicate that the proposed residual rectified-flow formulation does not merely improve perceptual appearance, but also preserves the degraded input content more faithfully than previous restoration-oriented diffusion models.

(2) On synthetic benchmarks, ScaleResfusion remains consistently competitive under both DIV2K-Val and LSDIR-Val. On DIV2K-Val, the FLUX2-9B *w/o GAN* variant achieves the best PSNR and SSIM,

Table 3. Full quantitative comparison on real-world benchmarks. Higher is better for PSNR, SSIM, MUSIQ, and MANIQA, while lower is better for LPIPS, DISTS, FID, and NIQE. Methods are grouped into GAN-based, diffusion-based, and ScaleResfusion variants. The best and second-best results of each metric are highlighted in red and blue, respectively.

Datasets	Type	Method	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑
DRealSR	GAN-based	BSRGAN	28.70	0.80	0.29	0.21	155.61	6.54	57.15	0.48
		Real-ESRGAN	28.61	0.81	0.28	0.21	147.66	6.70	54.27	0.49
		LDL	28.20	0.81	0.28	0.21	155.51	7.14	53.94	0.49
		FeMaSR	26.87	0.76	0.32	0.22	157.72	5.91	53.70	0.44
	Diffusion-based	StableSR	28.04	0.75	0.33	0.23	144.15	6.60	58.53	0.56
		SUPIR	25.09	0.65	0.42	0.28	169.48	7.39	58.79	0.55
		TSD-SR	27.77	0.76	0.30	0.21	134.98	**5.91**	**66.62**	0.59
		AddSR	26.68	0.74	0.37	0.26	164.82	7.80	65.36	0.60
		CCSR	28.24	0.78	0.32	0.23	157.30	6.81	**66.28**	0.61
		DiffBIR	25.90	0.62	0.47	0.29	180.33	6.33	66.13	0.62
		OSEDiff	27.92	0.78	0.30	0.22	135.41	6.46	64.69	0.59
		PASD	28.02	0.78	0.32	0.23	174.76	6.72	57.23	0.51
		ResShift	27.05	0.74	0.39	0.26	159.90	8.65	51.24	0.47
		SeeSR	28.07	0.77	0.32	0.23	147.37	6.41	65.09	0.61
	Ours	SD3 (w/o GAN)	28.77	0.79	0.30	0.23	149.67	6.71	63.72	0.60
		SD3 (w/ GAN)	27.86	0.76	0.32	0.22	146.93	**6.09**	64.57	0.61
		FLUX2-4B (w/o GAN)	**29.77**	**0.82**	**0.25**	0.20	118.18	6.95	62.42	0.61
		FLUX2-4B (w/ GAN)	28.26	0.78	0.29	0.21	124.03	6.21	65.16	**0.64**
		ZImage (w/o GAN)	29.35	0.80	0.28	0.22	132.31	6.78	62.52	0.59
		ZImage (w/ GAN)	28.34	0.78	0.28	0.21	127.40	6.16	65.64	**0.63**
		FLUX2-9B (w/o GAN)	**30.17**	**0.82**	**0.25**	**0.19**	**109.28**	7.22	60.54	0.59
		FLUX2-9B (w/ GAN)	28.77	0.79	0.28	**0.19**	**116.69**	6.25	62.37	0.62
RealSR	GAN-based	BSRGAN	26.38	0.77	0.27	0.21	141.24	5.64	63.28	0.54
		Real-ESRGAN	26.65	0.76	0.27	0.21	136.29	5.85	60.45	0.55
		LDL	25.28	0.76	0.28	0.21	142.74	5.99	60.92	0.55
		FeMaSR	25.06	0.74	0.29	0.23	141.01	5.77	59.05	0.49
	Diffusion-based	StableSR	24.62	0.70	0.31	0.22	128.54	5.78	65.48	0.62
		SUPIR	23.65	0.66	0.35	0.25	130.38	6.11	62.09	0.58
		TSD-SR	24.81	0.72	0.27	0.21	114.45	**5.13**	**71.19**	0.63
		AddSR	22.65	0.65	0.38	0.27	154.18	6.62	**71.41**	0.67
		CCSR	25.92	0.75	0.28	0.21	122.84	5.73	69.18	0.64
		DiffBIR	24.83	0.65	0.36	0.24	130.75	5.84	69.28	0.65
		OSEDiff	25.15	0.73	0.29	0.21	123.53	5.65	69.08	0.63
		PASD	26.04	0.74	0.28	0.21	135.48	5.71	60.03	0.56
		ResShift	25.66	0.74	0.33	0.25	128.03	8.07	56.89	0.51
		SeeSR	25.15	0.72	0.30	0.22	125.30	5.40	69.81	0.65
	Ours	SD3 (w/o GAN)	25.67	0.73	0.29	0.22	126.75	5.96	68.45	0.65
		SD3 (w/ GAN)	24.92	0.71	0.30	0.22	126.53	5.61	68.99	0.65
		FLUX2-4B (w/o GAN)	**27.05**	**0.78**	**0.24**	0.20	104.74	6.21	67.25	0.64
		FLUX2-4B (w/ GAN)	25.78	0.75	0.26	0.20	106.22	5.29	69.86	**0.68**
		ZImage (w/o GAN)	26.29	0.75	0.27	0.21	114.35	5.75	67.33	0.63
		ZImage (w/ GAN)	25.44	0.73	0.27	0.21	104.37	5.35	69.25	0.66
		FLUX2-9B (w/o GAN)	**27.30**	**0.78**	**0.23**	**0.19**	**96.38**	5.94	65.72	0.63
		FLUX2-9B (w/ GAN)	26.12	0.75	0.25	**0.18**	**98.30**	**5.11**	67.64	**0.68**

while the *w/ GAN* variant further improves perceptual and distribution metrics, obtaining the best LPIPS, DISTS, and FID. On the more challenging LSDIR-Val benchmark, the FLUX2-9B *w/o GAN* variant provides the best PSNR, SSIM, FID, and tied-best DISTS, while the *w/ GAN* variant achieves the best LPIPS and also ties for the best DISTS and MANIQA. In contrast, competing methods usually perform well only on a subset of metrics. For example, some diffusion-based methods obtain strong no-reference scores but suffer from weaker distortion fidelity, whereas GAN-based methods can preserve local structure but often show inferior distribution alignment.

(3) ScaleResfusion variants with and without GAN fine-tuning exhibit a clear perception–distortion trade-off. The *w/o GAN* variant is generally better for reconstruction fidelity, as reflected by higher PSNR/SSIM and strong LPIPS/DISTS on paired benchmarks. The *w/ GAN* variant tends to improve no-reference quality and perceptual realism, especially on MANIQA, NIQE, and MUSIQ, at the cost of a moderate decrease in pixel-level fidelity. Table 1 shows this trend clearly for FLUX2-4B: adding GAN fine-tuning improves NIQE/MUSIQ/MANIQA from 6.95/62.42/0.61 to 6.21/65.16/0.64 on DRealSR, from 6.21/67.25/0.64 to 5.29/69.86/0.68 on RealSR, and from 5.09/66.10/0.63 to 4.44/69.30/0.66 on DIV2K-Val. Meanwhile, the same fine-tuning decreases PSNR/SSIM from 29.77/0.82 to 28.26/0.78 on DRealSR, from 27.05/0.78 to 25.78/0.75 on RealSR, and from 24.90/0.64 to 23.63/0.60 on DIV2K-Val. This behavior is consistent with the known perception–distortion trade-off in image restoration

Table 4. Full quantitative comparison on synthetic benchmarks. Higher is better for PSNR, SSIM, MUSIQ, and MANIQA, while lower is better for LPIPS, DISTS, FID, and NIQE. Methods are grouped into GAN-based, diffusion-based, and ScaleResfusion variants. The best and second-best results of each metric are highlighted in red and blue, respectively.

Datasets	Type	Method	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑
DIV2K-Val	GAN-based	BSRGAN	24.58	0.63	0.35	0.23	49.55	4.75	61.68	0.50
		Real-ESRGAN	24.02	0.64	0.32	0.21	38.87	4.83	60.38	0.54
		LDL	23.83	0.63	0.33	0.22	42.28	4.86	60.04	0.53
		FeMaSR	22.45	0.59	0.34	0.22	41.97	4.87	57.94	0.48
	Diffusion-based	StableSR	23.27	0.57	0.31	0.20	24.95	4.77	65.78	0.62
		SUPIR	22.13	0.53	0.39	0.23	31.40	5.68	63.86	0.59
		TSD-SR	23.02	0.58	0.27	0.18	29.16	4.32	71.69	0.62
		AddSR	22.37	0.56	0.38	0.23	34.91	5.84	69.15	0.63
		CCSR	24.30	0.63	0.30	0.20	30.84	5.34	69.53	0.61
		DiffBIR	23.14	0.54	0.37	0.22	32.71	4.99	69.87	0.64
		OSEDiff	23.72	0.61	0.29	0.20	26.34	4.71	67.96	0.61
		PASD	24.01	0.61	0.38	0.22	37.06	4.98	63.75	0.55
		ResShift	24.59	0.62	0.31	0.21	30.81	6.92	58.90	0.53
		SeeSR	23.68	0.60	0.32	0.20	25.89	4.81	68.66	0.62
	Ours	SD3 (w/o GAN)	23.90	0.60	0.31	0.21	31.83	4.99	68.66	0.64
		SD3 (w/ GAN)	23.13	0.58	0.32	0.20	27.80	4.49	69.89	0.65
		FLUX2-4B (w/o GAN)	24.90	0.64	0.27	0.18	22.90	5.09	66.10	0.63
		FLUX2-4B (w/ GAN)	23.63	0.60	0.26	0.17	19.47	4.44	69.30	0.68
		ZImage (w/o GAN)	23.99	0.62	0.30	0.20	28.55	4.95	67.93	0.63
		ZImage (w/ GAN)	23.65	0.60	0.28	0.19	22.96	4.54	69.73	0.65
		FLUX2-9B (w/o GAN)	25.18	0.65	0.26	0.18	21.67	5.07	64.89	0.62
		FLUX2-9B (w/ GAN)	23.97	0.62	0.25	0.16	17.93	4.42	67.50	0.65
LSDIR-Val	GAN-based	BSRGAN	20.82	0.54	0.25	0.16	46.37	4.21	68.94	0.63
		Real-ESRGAN	20.58	0.55	0.24	0.15	41.28	4.18	69.52	0.64
		LDL	20.31	0.53	0.25	0.16	44.75	4.36	68.61	0.64
		FeMaSR	19.87	0.51	0.27	0.17	48.63	4.09	67.85	0.61
	Diffusion-based	StableSR	20.31	0.55	0.31	0.18	54.76	5.07	62.96	0.61
		SUPIR	20.35	0.50	0.24	0.15	43.81	4.81	71.47	0.67
		TSD-SR	19.05	0.49	0.21	0.14	45.66	3.86	74.45	0.68
		AddSR	19.20	0.45	0.34	0.20	79.80	4.99	74.20	0.70
		CCSR	20.76	0.53	0.26	0.16	56.28	4.25	72.57	0.66
		DiffBIR	20.51	0.49	0.27	0.16	58.45	4.44	73.26	0.68
		OSEDiff	20.39	0.52	0.27	0.16	59.57	4.03	72.34	0.66
		PASD	20.93	0.52	0.31	0.17	59.11	3.80	69.29	0.62
		ResShift	21.23	0.55	0.23	0.14	38.98	5.32	68.56	0.61
		SeeSR	20.69	0.52	0.25	0.15	52.06	4.10	73.27	0.68
	Ours	SD3 (w/o GAN)	20.22	0.52	0.25	0.15	63.84	4.21	74.08	0.70
		SD3 (w/ GAN)	19.93	0.51	0.26	0.16	60.61	4.12	73.75	0.70
		FLUX2-4B (w/o GAN)	21.35	0.57	0.23	0.14	41.69	4.21	70.61	0.67
		FLUX2-4B (w/ GAN)	20.64	0.55	0.21	0.14	43.56	3.82	72.59	0.69
		ZImage (w/o GAN)	20.26	0.52	0.23	0.14	54.33	4.17	73.38	0.69
		ZImage (w/ GAN)	20.41	0.53	0.24	0.15	51.57	4.07	72.92	0.68
		FLUX2-9B (w/o GAN)	21.60	0.57	0.21	0.13	34.43	4.08	69.27	0.66
		FLUX2-9B (w/ GAN)	20.75	0.55	0.20	0.13	38.34	3.82	72.29	0.70

and also explains why we report both variants: users can select the more faithful or more perceptual model depending on the target application.

(4) Stronger rectified-flow backbones further improve the robustness of ScaleResfusion. Across DRealSR, RealSR, DIV2K-Val, and LSDIR-Val, FLUX2-9B generally provides the best or second-best results among our variants. Although FLUX2-4B is not the strongest variant, it offers a practical performance-efficiency trade-off by maintaining competitive restoration quality while achieving sub-second inference. This scaling trend supports the central motivation of ScaleResfusion: by formulating restoration as residual rectified-flow adaptation, large pre-trained generative priors can be effectively reused for real-world image restoration with only task-specific residual updates.

E Caution on No-Reference Metrics

No-reference metrics are useful when ground-truth images are unavailable, but current metrics still contain biases toward certain low-level statistics, such as sharpness, contrast, and texture richness. As a result, they may assign higher scores to images with over-sharpened details, semantic drift, or local artifacts, even when these artifacts reduce restoration fidelity. Fig. 12 shows paired examples where TSD-SR obtains higher MUSIQ/MANIQA scores than ScaleResfusion, yet its outputs have lower reference fidelity and visibly stronger artifacts. This indicates that no-reference scores can

Fig. 11. No-reference metric comparison on examples without ground truth. Although TSD-SR still obtains better MUSIQ/NIQE scores than ScaleResfusion, its visual quality is clearly worse, with stronger semantic drift and local artifacts.

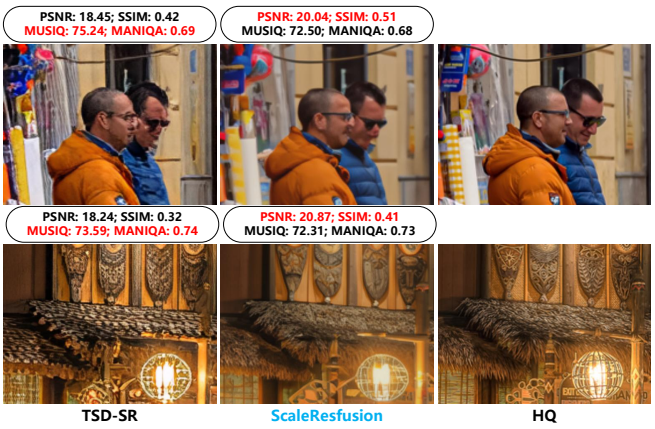


Fig. 12. Examples where no-reference metrics do not fully reflect visual quality. TSD-SR obtains higher MUSIQ/MANIQA scores, but introduces stronger artifacts and less faithful structures than ScaleResfusion.

be inconsistent with both ground-truth-based metrics and human visual judgment.

Fig. 11 further considers examples without ground-truth images, where reference-based metrics such as PSNR/SSIM cannot be computed and evaluation naturally relies more on no-reference scores. Although TSD-SR reports better no-reference scores in these cases, the visual comparison reveals facial semantic drift, unstable exposure and color appearance, and structural degradation. ScaleResfusion better preserves plausible structures and overall appearance, suggesting that the bias of no-reference metrics can become more visible when they are used in isolation. Together with the paired examples in Fig. 12, this motivates a more balanced evaluation protocol: when the gap in no-reference metrics is not substantial, consistency-oriented metrics (PSNR, SSIM, LPIPS, DISTS) and distribution-level metrics (FID) should receive more weight whenever they are available, since they better capture whether the restored image remains faithful to the input content and aligned with the target image

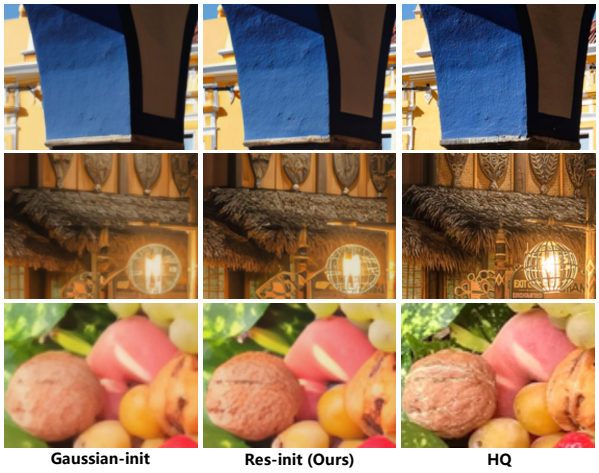


Fig. 13. Visual ablation on residual initialization. Compared with task-agnostic Gaussian initialization, residual initialization starts from a more restoration-aligned state and better preserves the observed image structure.

distribution. We thus treat no-reference scores as complementary indicators rather than the sole criterion for evaluating Real-IR quality.

F Extra Ablation Study

In this section, we provide additional ablation studies to analyze the effects of sampling steps, residual ratio, residual initialization, DMD, GAN fine-tuning, DAPE, feature extraction, and LoRA rank. We follow the same evaluation protocol as the main experiments and use the FLUX2-4B variant as the default setting unless otherwise specified.

F.1 Ablation on Res-init

The main purpose of this ablation is to verify the effectiveness of Residual Rectified Flow (RRF). To isolate the effect of the RRF

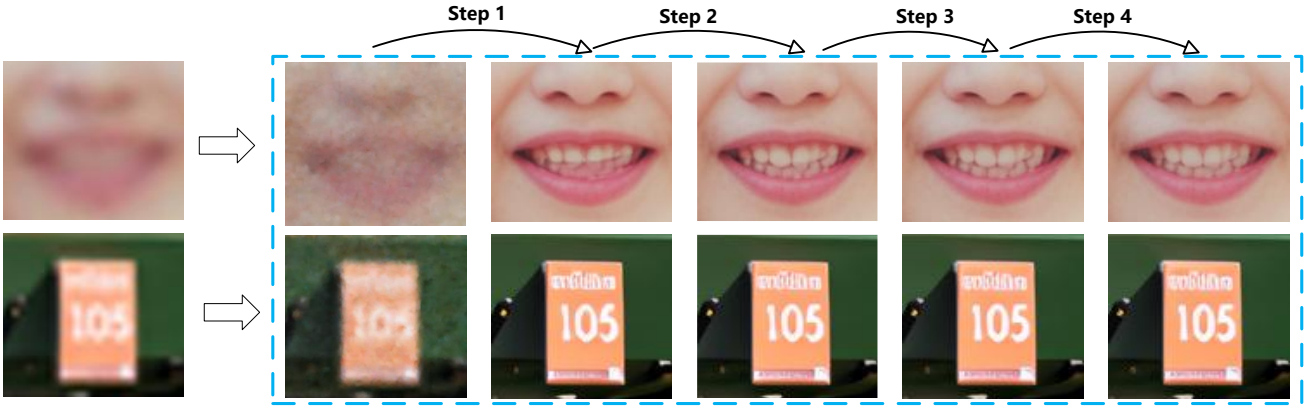


Fig. 14. Intermediate restoration results from the same multi-step model. As the number of function evaluations (NFE) increases during sampling, ScaleResfusion progressively updates the residual state, leading to clearer structures and more refined image details.

Table 5. Ablation on residual initialization strategies under SD3 and FLUX2-4B on DRealSR and LSDIR-Val.

Dataset	Backbone	Res-init	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DRealSR	SD3	Gaussian init	26.92	0.70	0.34	165.20
DRealSR	SD3	Res-init	28.77	0.79	0.30	149.67
DRealSR	FLUX2-4B	Gaussian init	26.83	0.72	0.35	137.58
DRealSR	FLUX2-4B	Res-init	29.77	0.82	0.25	118.18
LSDIR-Val	SD3	Gaussian init	17.87	0.41	0.29	79.37
LSDIR-Val	SD3	Res-init	20.22	0.52	0.25	63.84
LSDIR-Val	FLUX2-4B	Gaussian init	19.69	0.46	0.39	61.09
LSDIR-Val	FLUX2-4B	Res-init	21.35	0.57	0.23	41.69

Table 6. Ablation on the number of function evaluations (NFE). We compare 4-step, 2-step, and 1-step sampling under two backbones on DRealSR and LSDIR-Val.

Dataset	Backbone	NFE	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DRealSR	SD3	4-step	28.77	0.79	0.30	149.67
DRealSR	SD3	2-step	28.61	0.78	0.31	150.90
DRealSR	SD3	1-step	28.23	0.77	0.31	153.65
DRealSR	FLUX2-4B	4-step	29.77	0.82	0.25	118.18
DRealSR	FLUX2-4B	2-step	29.03	0.79	0.27	115.43
DRealSR	FLUX2-4B	1-step	28.93	0.78	0.28	117.56
LSDIR-Val	SD3	4-step	20.22	0.52	0.25	63.84
LSDIR-Val	SD3	2-step	19.56	0.49	0.26	65.07
LSDIR-Val	SD3	1-step	19.18	0.48	0.26	67.82
LSDIR-Val	FLUX2-4B	4-step	21.35	0.57	0.23	41.69
LSDIR-Val	FLUX2-4B	2-step	20.61	0.54	0.25	38.94
LSDIR-Val	FLUX2-4B	1-step	20.51	0.53	0.26	41.07

formulation, we train a baseline with the same backbone, data, optimizer, and knowledge-distillation based training pipeline as ScaleResfusion, but replace the RRF starting state with task-agnostic Gaussian noise. Therefore, the comparison in Table 5 and Fig. 13 directly examines whether starting from a noisy LQ-related residual state is more suitable for restoration than starting from Gaussian noise.

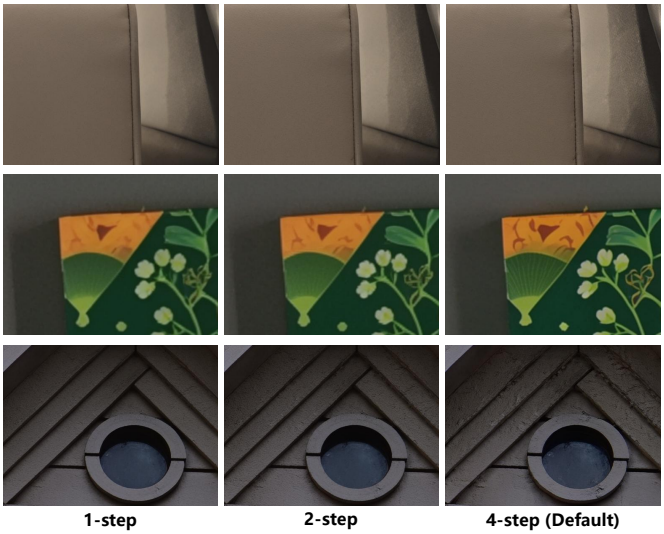


Fig. 15. Visual ablation on the number of function evaluations. Increasing the number of sampling steps generally improves restoration fidelity and detail consistency, while the proposed residual transport remains effective even with very few steps.

As shown in Table 5, residual initialization substantially improves restoration fidelity for both backbones. On DRealSR, it increases SD3 from 26.92 to 28.77 PSNR and from 0.70 to 0.79 SSIM. On FLUX2-4B, PSNR increases from 26.83 to 29.77 and SSIM increases from 0.72 to 0.82. LSDIR-Val shows the same pattern: residual initialization improves SD3 from 17.87 to 20.22 PSNR and from 0.41 to 0.52 SSIM, and improves FLUX2-4B from 19.69 to 21.35 PSNR and from 0.46 to 0.57 SSIM.

Residual initialization also improves perceptual and distribution metrics. Compared with Gaussian initialization on DRealSR, it reduces LPIPS from 0.34 to 0.30 and FID from 165.20 to 149.67 on SD3, and reduces LPIPS from 0.35 to 0.25 and FID from 137.58 to 118.18

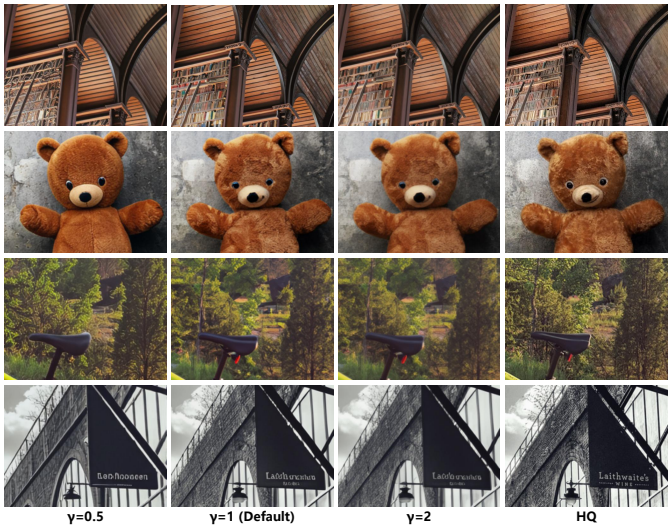


Fig. 16. Visual ablation on the residual ratio γ. The residual ratio controls the balance between deterministic LQ guidance and stochastic restoration, leading to different fidelity–realism trade-offs.

on FLUX2-4B. On LSDIR-Val, it reduces LPIPS from 0.29 to 0.25 and FID from 79.37 to 63.84 on SD3, and reduces LPIPS from 0.39 to 0.23 and FID from 61.09 to 41.69 on FLUX2-4B.

These results directly support the motivation of RRF. Starting from a noisy LQ-related state is better aligned with the restoration task, because the initial state already preserves the observed structure and the model only needs to transport residual corrections toward the HQ image. In contrast, the Gaussian-initialized baseline must synthesize both global structure and restoration details from a task-agnostic noise state, which makes few-step restoration much harder even under the same distillation pipeline.

F.2 Ablation on Number of Function Evaluations (NFE)

We study the effect of sampling steps by training separate 1-step, 2-step, and 4-step models under the same knowledge-distillation based training pipeline. This controlled setting evaluates how the target number of function evaluations affects models that are explicitly trained for different inference budgets. As shown in Table 6, more sampling steps generally improve reconstruction fidelity across both SD3 and FLUX2-4B. The 4-step setting gives the most stable distortion and perceptual similarity metrics on DRealSR and LSDIR-Val, so we use it as the default configuration. Meanwhile, the 1-step and 2-step variants remain competitive, indicating that the RRF acceleration point provides a strong initialization for fast restoration.

Beyond comparing separately trained models, Fig. 14 visualizes the intermediate predictions of the same multi-step model during sampling. Early evaluations recover the main image structure, while later evaluations progressively correct detail-level errors, reduce local inconsistencies, and refine high-frequency textures. This shows that the multi-step trajectory is not redundant: additional function evaluations allow ScaleResfusion to iteratively polish residual artifacts and improve local consistency.

Table 7. Ablation on the residual ratio γ. We evaluate three residual ratios under two backbones on DRealSR and LSDIR-Val.

Dataset	Backbone	γ	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DRealSR	SD3	2.0	29.53	0.81	0.27	139.44
DRealSR	SD3	1.0	28.77	0.79	0.30	149.67
DRealSR	SD3	0.5	28.64	0.78	0.30	148.19
DRealSR	FLUX2-4B	2.0	30.49	0.83	0.26	120.03
DRealSR	FLUX2-4B	1.0	29.77	0.82	0.25	118.18
DRealSR	FLUX2-4B	0.5	29.18	0.79	0.28	132.45
LSDIR-Val	SD3	2.0	20.48	0.52	0.22	53.61
LSDIR-Val	SD3	1.0	20.22	0.52	0.25	63.84
LSDIR-Val	SD3	0.5	19.59	0.49	0.25	62.36
LSDIR-Val	FLUX2-4B	2.0	22.07	0.58	0.24	43.54
LSDIR-Val	FLUX2-4B	1.0	21.35	0.57	0.23	41.69
LSDIR-Val	FLUX2-4B	0.5	20.76	0.54	0.26	55.96

Table 8. Ablation on DMD with the FLUX2-4B backbone. Res-init and RefNet are enabled for both variants, and GAN fine-tuning is disabled.

DMD	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓
×	21.90	0.51	0.38	0.24	51.25
✓	24.90	0.64	0.27	0.18	22.90

F.3 Ablation on Residual Ratio

We further ablate the residual ratio γ introduced in Sec. 3.3. As discussed in Sec. 3.3, γ directly determines the signal-to-noise ratio (SNR) of the RRF acceleration point: a larger γ makes the initial state closer to the LQ-related residual signal, while a smaller γ injects stronger Gaussian randomness. This parameter therefore controls how much the sampling process relies on deterministic LQ initialization versus stochastic generative refinement.

Table 7 shows that increasing γ generally improves fidelity-oriented metrics. For example, moving from $\gamma = 0.5$ to $\gamma = 2.0$ consistently improves PSNR and SSIM on both DRealSR and LSDIR-Val, suggesting that a stronger residual signal helps preserve the observed image structure. The visual comparison in Fig. 16 follows the same trend: small γ values leave more room for stochastic restoration but can introduce unstable details, while larger γ values raise the SNR of the starting state and therefore provide stronger structural guidance.

However, the largest γ is not always the best choice for perceptual and distribution-level quality. On FLUX2-4B, $\gamma = 1.0$ achieves better LPIPS and FID than $\gamma = 2.0$ on both evaluated datasets, indicating that excessive reliance on the LQ-related signal may improve pixel fidelity while limiting realistic detail synthesis and distribution alignment. We therefore use $\gamma = 1.0$ by default, as it preserves sufficient input structure while retaining enough stochasticity for realistic restoration.

F.4 Ablation on DMD Regularization

We further analyze the effect of DMD regularization using FLUX2-4B as the base model. As described in Sec. 3.4, the DMD term transfers

Fig. 17. Visual ablation on GAN fine-tuning. GAN fine-tuning enhances perceptual sharpness and local details, while occasionally reducing content consistency with the input or reference.

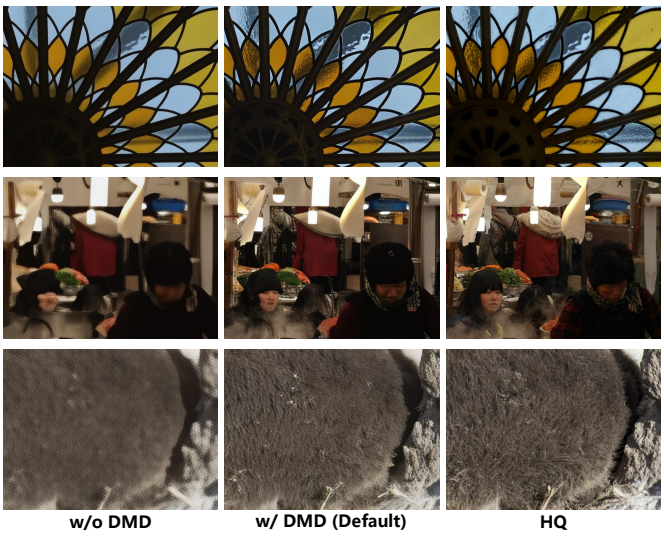


Fig. 18. Visual ablation on DMD. DMD reduces the training-inference mismatch of the few-step generator and improves both structural fidelity and perceptual detail quality.

the natural-image prior of the frozen pre-trained diffusion model to the few-step RRF generator through a teacher-student distribution-matching objective. This is especially important for ScaleResfusion because the generator is trained to produce high-quality restorations in only a few sampling steps, where small distribution errors can quickly accumulate into artifacts or structure drift. To isolate the effect of DMD, we keep residual initialization and LQ conditioning enabled, while disabling GAN fine-tuning.

Table 8 shows that DMD is crucial for stable few-step restoration. Without DMD, the FLUX2-4B variant obtains only 21.90 PSNR, 0.51 SSIM, 0.38 LPIPS, 0.24 DISTS, and 51.25 FID. Adding DMD improves PSNR to 24.90 and SSIM to 0.64, while reducing LPIPS to 0.27, DISTS to 0.18, and FID to 22.90. The visual results in Fig. 18 show the same tendency: DMD suppresses unstable textures and improves

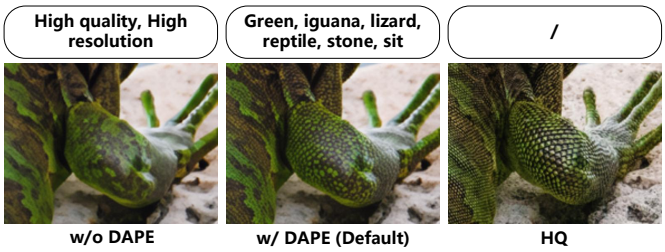


Fig. 19. Visual ablation on DAPE. Replacing DAPE with a fixed prompt simplifies the pipeline but loses image-specific semantic and degradation-aware guidance, while DAPE better preserves structures and degradation-specific details.

Table 9. Ablation on replacing DAPE with a fixed prompt using the FLUX2-4B backbone on DRealSR and LSDIR-Val.

Dataset	Variant	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DRealSR	Fixed prompt	28.39	0.78	0.28	130.98
DRealSR	DAPE	29.77	0.82	0.25	118.18
LSDIR-Val	Fixed prompt	20.83	0.55	0.23	47.95
LSDIR-Val	DAPE	21.35	0.57	0.23	41.69

structural fidelity. These results indicate that DMD regularization alleviates the mismatch between supervised training states and inference-time generated states, allowing the few-step generator to stay closer to the HQ image manifold and produce more faithful and realistic outputs.

F.5 Ablation on DAPE

We analyze whether the image-specific semantic information provided by DAPE can be replaced by a fixed restoration prompt, which would remove the extra DAPE module and simplify inference. This is a practical trade-off: the pre-trained models inherently possess strong feature extraction capabilities, so a fixed prompt may be sufficient for many cases.

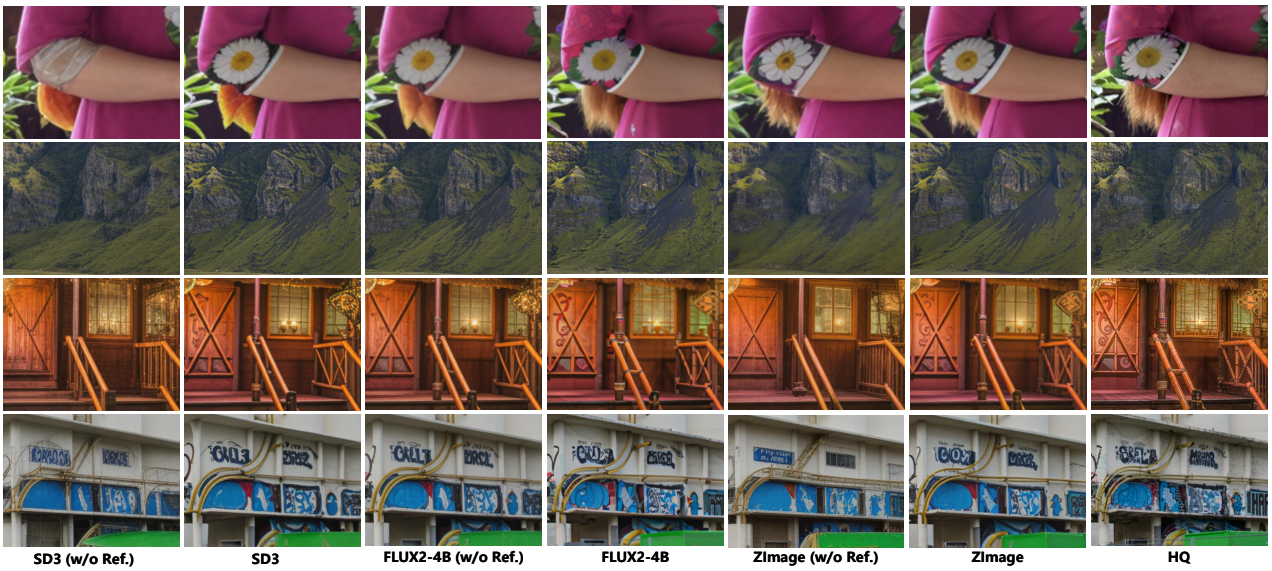


Fig. 20. Visual ablation on LQ feature conditioning. Removing LQ conditioning allows the model to synthesize more diverse details, but it also weakens input consistency and reconstruction fidelity.

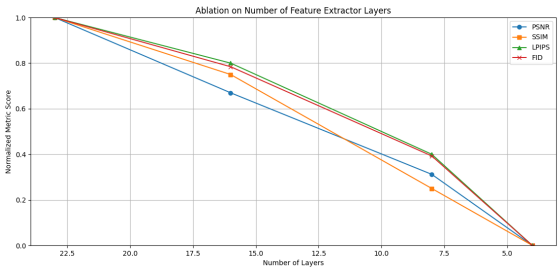


Fig. 21. Visual ablation on the number of feature extractor layers. Using more feature extraction layers provides stronger LQ conditioning and improves fine-detail restoration.

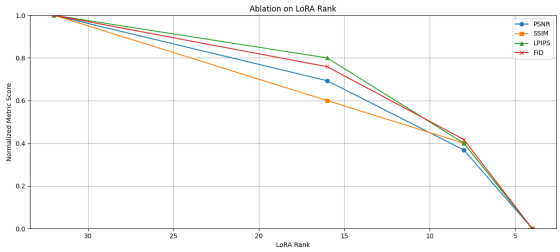


Fig. 22. Visual ablation on LoRA rank. Increasing the LoRA rank gives the residual generator more adaptation capacity and improves perceptual detail quality.

Table 10. Ablation on using the feature extractor on DRealSR and LSDIR-Val.

Dataset	Variant	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DRealSR	w/o Ref. Attention	28.93	0.79	0.28	129.86
DRealSR	w/ Ref. Attention	29.77	0.82	0.25	118.18
LSDIR-Val	w/o Ref. Attention	20.51	0.54	0.26	53.37
LSDIR-Val	w/ Ref. Attention	21.35	0.57	0.23	41.69

Table 9 and Fig. 19 show that the fixed-prompt variant is not invalid; it still produces reasonable restoration results under the same FLUX2-4B backbone. However, DAPE consistently improves PSNR, SSIM, LPIPS, and FID on DRealSR, and also improves PSNR, SSIM, and FID on LSDIR-Val. The visual comparison further suggests that image-specific textual guidance helps recover more faithful structures and degradation-specific details. Therefore, replacing DAPE with a fixed prompt is a viable lightweight option when simplicity is preferred, but the default DAPE setting offers a better quality–robustness trade-off.

F.6 Ablation on Feature Extractor

Effect of LQ Feature Conditioning. We compare ScaleResfusion with and without the feature extractor. The feature extractor provides explicit LQ feature injection as an additional conditioning signal, helping the generator align restored structures with the degraded input. Removing this conditioning gives the generative prior more freedom to synthesize diverse details, but it also weakens the constraint from the observation, which may reduce fidelity and input consistency.

Table 10 shows this trade-off quantitatively. Adding Ref. Attention improves PSNR from 28.93 to 29.77 and SSIM from 0.79 to 0.82 on DRealSR, and improves PSNR from 20.51 to 21.35 and SSIM from 0.54 to 0.57 on LSDIR-Val. The same setting also reduces LPIPS and FID on both datasets, suggesting that LQ feature conditioning improves not only distortion fidelity but also perceptual and distribution-level quality.

The visual comparison in Fig. 20 further illustrates the role of the feature extractor. Without LQ conditioning, the model can generate richer or more varied local details, but these details are less tightly

Table 11. Ablation on the number of feature extractor layers with SD3 backbone on DRealSR and LSDIR-Val.

Dataset	# Layers	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DRealSR	23	28.77	0.79	0.30	149.67
DRealSR	16	28.41	0.78	0.31	153.82
DRealSR	8	28.02	0.76	0.33	161.35
DRealSR	4	27.68	0.75	0.35	168.90
LSDIR-Val	23	20.22	0.52	0.25	63.84
LSDIR-Val	16	19.36	0.49	0.26	67.99
LSDIR-Val	8	18.97	0.47	0.28	75.52
LSDIR-Val	4	18.63	0.46	0.30	83.07

Table 12. Ablation on LoRA rank on DRealSR and LSDIR-Val.

Dataset	Rank	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
DRealSR	32	29.77	0.82	0.25	118.18
DRealSR	16	29.42	0.80	0.26	123.64
DRealSR	8	29.05	0.79	0.28	131.37
DRealSR	4	28.63	0.77	0.30	140.82
LSDIR-Val	32	21.35	0.57	0.23	41.69
LSDIR-Val	16	21.00	0.55	0.24	47.15
LSDIR-Val	8	20.63	0.54	0.26	54.88
LSDIR-Val	4	20.21	0.52	0.28	64.33

anchored to the input image and can deviate from the original structure. With feature extraction, ScaleResfusion better preserves the observed content while still benefiting from the generative prior, leading to a more reliable fidelity–diversity trade-off.

Layers of Feature Extractor. We further ablate the number of LoRA-branch feature extractor layers on SD3 to study how much LQ information should be injected into the residual generator. As shown in Table 11, using more feature extractor layers consistently improves restoration quality. Increasing the number of layers from 4 to 23 improves PSNR/SSIM and reduces LPIPS/FID on both DRealSR and LSDIR-Val, indicating that deeper feature extraction provides more complete conditioning.

This trend is consistent with the role of LQ feature injection. Shallow feature settings mainly provide coarse structural cues, which help anchor the global layout but are less effective at transferring fine textures, local degradation patterns, and high-frequency details. Deeper settings expose the residual generator to richer multi-level LQ representations, allowing it to better preserve input content while producing natural HQ details. We therefore use the full 23-layer feature extractor for SD3, as it provides the strongest and most complete conditioning signal.

F.7 Ablation on LoRA Rank

Finally, we ablate the LoRA rank, which controls the capacity of the trainable residual update while keeping the pre-trained backbone frozen. Since ScaleResfusion only learns low-rank updates for the residual vector field, the rank determines how expressive this residual adaptation subspace can be. A very small rank may be sufficient for coarse task adaptation, but it can underfit structured high-frequency corrections that are important for faithful restoration.

Table 12 shows a consistent improvement as the rank increases. On DRealSR, increasing the rank from 4 to 32 improves PSNR and SSIM while reducing LPIPS and FID. LSDIR-Val follows the same trend, with higher rank providing better fidelity and stronger distribution alignment. Notably, even rank 4 already achieves reasonable restoration quality, which supports the residual formulation in Sec. 3.3: the residual vector field differs from the standard Rectified Flow target only by a residual offset, so the adaptation does not need to relearn the full image-noise transport from scratch. This indicates that ScaleResfusion can make effective use of a compact LoRA subspace.

Rank 32 provides the best overall performance in our setting. At the same time, the stable trend across ranks supports our design choice: ScaleResfusion can adapt a large pre-trained backbone through a compact LoRA subspace, and increasing the subspace capacity further helps capture the structured residual corrections required by Real-IR.

G Theoretical Justification of Residual Rectified Flow

We justify Residual Rectified Flow (RRF) by following the marginal-preserving argument of nonlinear Rectified Flow [Liu et al. 2022]. Different from the canonical Rectified Flow, which constructs a linear interpolation between two endpoint distributions, our residual construction defines an observation-dependent stochastic path whose intermediate state admits an exact acceleration point. Therefore, RRF inherits the marginal-preserving property of nonlinear Rectified Flow, while its acceleration point comes from clean-image coefficient cancellation in the residual parameterization.

Let $X_0 \in \mathbb{R}^d$ denote the clean high-quality image and let $C = \hat{X}_0 \in \mathbb{R}^d$ denote the image-space condition obtained from the low-quality observation, such as an upsampled or preliminary restored image. Let $\epsilon \sim \mathcal{N}(0, \Sigma)$ be independent Gaussian noise. For a fixed residual strength $\gamma > 0$, define the residual endpoint

$$Y_\gamma = \gamma(C - X_0) + \epsilon. \tag{26}$$

We then construct the residual interpolation

$$X_t = (1-t)X_0 + tY_\gamma, \qquad t \in [0,1]. \tag{27}$$

Equivalently,

$$X_t = (1-t)X_0 + t\gamma(C - X_0) + t\epsilon, \tag{28}$$

and hence

$$X_t = [1 - (1+\gamma)t]X_0 + \gamma t C + t\epsilon. \tag{29}$$

The pathwise velocity is

$$\dot{X}_t = Y_\gamma - X_0 = \gamma(C - X_0) + \epsilon - X_0. \tag{30}$$

We denote $U_t := \dot{X}_t$.

Following Definition 3.1 of original Rectified Flow [Liu et al. 2022], which defines the expected velocity as $v^X(x,t) = \mathbb{E}[\dot{X}_t \mid X_t = x]$ for a pathwise differentiable process, we define the conditional expected velocity by

$$v^X(x,t,c) = \mathbb{E}[\dot{X}_t \mid X_t = x, C = c]. \tag{31}$$

In practice, v^X is approximated by a neural network $v_\theta(x,t,c)$. Following Eq. (6) of original Rectified Flow [Liu et al. 2022], which fits the nonlinear Rectified Flow velocity by matching $v(X_t,t)$ to $\dot{X}_t$, we use the least-squares objective

$$\min_\theta \int_0^1 \mathbb{E}\left[\left\|v_\theta(X_t,t,C)-\dot{X}_t\right\|_2^2\right] dt. \tag{32}$$

Here we take the weight $w_t = 1$ in Eq. (6) of original Rectified Flow [Liu et al. 2022] and add the condition C as input. Equivalently,

$$\min_\theta \int_0^1 \mathbb{E}\left[\left\|v_\theta(X_t,t,C)-\big(\gamma(C-X_0)+\epsilon-X_0\big)\right\|_2^2\right] dt. \tag{33}$$

Definition 1: Conditional residual rectified flow. For each fixed condition $C = c$, define

$$X_t^c = [1-(1+\gamma)t]X_0 + \gamma t c + t\epsilon. \tag{34}$$

Its expected velocity is

$$v^X(x,t,c) = \mathbb{E}[\dot{X}_t \mid X_t = x, C = c]. \tag{35}$$

We call $X = \{X_t : t \in [0,1]\}$ conditionally rectifiable if $v^X(\cdot,t,c)$ is locally bounded and the integral equation

$$Z_t = Z_0 + \int_0^t v^X(Z_s,s,c)\, ds \tag{36}$$

admits a unique solution for each fixed c.

Following Eq. (9) of original Rectified Flow [Liu et al. 2022], where the rectified flow is defined by the integral equation driven by v^X, the above equation is its conditional counterpart after fixing $C = c$.

Theorem 1: Conditional marginal preservation. Assume that, for each fixed $C = c$, the residual interpolation X_t is conditionally rectifiable. Let Z_t solve

$$dZ_t = v^X(Z_t,t,c)\, dt, \qquad Z_0 \mid C = c \sim X_0 \mid C = c. \tag{37}$$

Then

$$\mathrm{Law}(Z_t \mid C = c) = \mathrm{Law}(X_t \mid C = c), \qquad \forall t \in [0,1]. \tag{38}$$

Proof. Fix $C = c$. The proof is the conditional version of Theorem 3.3 in [Liu et al. 2022]. Let $\pi_t^c := \mathrm{Law}(X_t \mid C = c)$. Following Eq. (10) of original Rectified Flow [Liu et al. 2022], for any compactly supported continuously differentiable test function $h : \mathbb{R}^d \to \mathbb{R}$, the chain rule gives

$$\frac{d}{dt}\mathbb{E}[h(X_t) \mid C = c] = \mathbb{E}[\nabla h(X_t)^\top \dot{X}_t \mid C = c]. \tag{39}$$

Following the conditional expected-velocity substitution in Eq. (10) of original Rectified Flow [Liu et al. 2022], we use

$$v^X(X_t,t,c) = \mathbb{E}[\dot{X}_t \mid X_t, C = c], \tag{40}$$

and obtain

$$\mathbb{E}[\nabla h(X_t)^\top \dot{X}_t \mid C = c] = \mathbb{E}[\nabla h(X_t)^\top v^X(X_t,t,c) \mid C = c]. \tag{41}$$

This is Eq. (10) of original Rectified Flow [Liu et al. 2022] with the unconditional velocity $v^X(X_t,t)$ replaced by the conditional velocity $v^X(X_t,t,c)$. Therefore,

$$\frac{d}{dt}\mathbb{E}[h(X_t) \mid C = c] = \mathbb{E}[\nabla h(X_t)^\top v^X(X_t,t,c) \mid C = c]. \tag{42}$$

Following Eq. (11) of original Rectified Flow [Liu et al. 2022], the above weak identity is equivalently written as the continuity equation for π_t^c in the sense of distributions:

$$\dot{\pi}_t^c + \nabla \cdot \left(v^X(\cdot,t,c)\pi_t^c\right) = 0. \tag{43}$$

Following the paragraph below Eq. (11) of original Rectified Flow [Liu et al. 2022], this weak equivalence follows by multiplying the continuity equation by h and integrating by parts. Since Z_t is driven by the same velocity field $v^X(\cdot,t,c)$, its conditional law $q_t^c := \mathrm{Law}(Z_t \mid C = c)$ also satisfies

$$\dot{q}_t^c + \nabla \cdot \left(v^X(\cdot,t,c)q_t^c\right) = 0. \tag{44}$$

Moreover, $q_0^c = \pi_0^c$ by the initialization assumption. By the uniqueness of the weak solution to the continuity equation, we obtain

$$q_t^c = \pi_t^c, \qquad \forall t \in [0,1]. \tag{45}$$

Thus,

$$\mathrm{Law}(Z_t \mid C = c) = \mathrm{Law}(X_t \mid C = c). \tag{46}$$

This proves the claim.

Following Theorem 3.3 of original Rectified Flow [Liu et al. 2022], the stochastic interpolation X_t and the deterministic ODE Z_t solve the same continuity equation in Eq. (11), so their one-time marginals coincide.

Theorem 2: Exact acceleration point. For the residual interpolation

$$X_t = [1-(1+\gamma)t]X_0 + \gamma t C + t\epsilon, \tag{47}$$

there exists a unique time

$$t^\star = \frac{1}{1+\gamma} \tag{48}$$

at which the clean-image component X_0 is exactly cancelled. At this time,

$$X_{t^\star} = \frac{\gamma}{1+\gamma}C + \frac{1}{1+\gamma}\epsilon. \tag{49}$$

Proof. The coefficient of X_0 in X_t is $1-(1+\gamma)t$. Setting this coefficient to zero gives

$$1-(1+\gamma)t^\star = 0. \tag{50}$$

Therefore,

$$t^\star = \frac{1}{1+\gamma}. \tag{51}$$

Substituting $t^\star$ into the residual interpolation gives

$$X_{t^\star} = [1-(1+\gamma)t^\star]X_0 + \gamma t^\star C + t^\star\epsilon. \tag{52}$$

Since $1-(1+\gamma)t^\star = 0$, we have

$$X_{t^\star} = \gamma t^\star C + t^\star\epsilon. \tag{53}$$

Using $t^\star = 1/(1+\gamma)$, we obtain

$$X_{t^\star} = \frac{\gamma}{1+\gamma}C + \frac{1}{1+\gamma}\epsilon. \tag{54}$$

Thus, $X_{t^\star}$ depends only on the condition C and Gaussian noise ϵ, but not on the unknown clean image X_0.

If $\epsilon \sim \mathcal{N}(0,\Sigma)$, then

$$X_{t^\star} \mid C = c \sim \mathcal{N}\left(\frac{\gamma}{1+\gamma}c, \frac{1}{(1+\gamma)^2}\Sigma\right). \tag{55}$$

Hence, the reverse ODE can be initialized from

$$Z_{t^\star} = \frac{\gamma}{1+\gamma} C + \frac{1}{1+\gamma}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, \Sigma). \tag{56}$$

By Theorem 1, under the population velocity field and exact ODE integration,

$$Z_0 \mid C = c \sim X_0 \mid C = c. \tag{57}$$

Training objective of RRF.. Since

$$\dot{X}_t = \gamma(C - X_0) + \epsilon - X_0, \tag{58}$$

the population RRF objective is obtained by following Eq. (6) of original Rectified Flow [Liu et al. 2022] and substituting the residual path velocity $\dot{X}_t = \gamma(C - X_0) + \epsilon - X_0$:

$$\mathcal{L}_{\mathrm{RRF}}(\theta) = \int_0^1 \mathbb{E}\left[\|v_\theta(X_t, t, C) - (\gamma(C - X_0) + \epsilon - X_0)\|_2^2\right] dt. \tag{59}$$

Equivalently, if training is restricted to the accelerated interval $[0, t^\star]$, one may use

$$\mathcal{L}_{\mathrm{RRF}}(\theta) = \int_0^{t^\star} \mathbb{E}\left[\|v_\theta(X_t, t, C) - (\gamma(C - X_0) + \epsilon - X_0)\|_2^2\right] dt. \tag{60}$$

Following Eq. (2) of original Rectified Flow [Liu et al. 2022], where the canonical population minimizer is $v^X(x,t) = \mathbb{E}[X_1 - X_0 \mid X_t = x]$, the RRF population minimizer is

$$v^\star(x, t, c) = \mathbb{E}[\gamma(C - X_0) + \epsilon - X_0 \mid X_t = x, C = c]. \tag{61}$$

Relation to Rectified Flow. RRF preserves the linear transport form of standard Rectified Flow. Recall that canonical Rectified Flow constructs the straight interpolation

$$X_t = (1 - t)X_0 + tX_1 \tag{62}$$

and learns the velocity field associated with the path derivative $X_1 - X_0$. In RRF, we replace the standard endpoint X_1 with the residual endpoint

$$Y_\gamma = \gamma(C - X_0) + \epsilon, \tag{63}$$

which gives

$$X_t = (1 - t)X_0 + tY_\gamma. \tag{64}$$

Therefore, conditioned on C, RRF is still a linear interpolation between two endpoints, and its pathwise velocity is

$$\dot{X}_t = Y_\gamma - X_0 = \gamma(C - X_0) + \epsilon - X_0. \tag{65}$$

Following the Rectified Flow argument, the population velocity field is given by the conditional expectation

$$v^\star(x, t, c) = \mathbb{E}\left[Y_\gamma - X_0 \,\middle|\, X_t = x, C = c\right]. \tag{66}$$

At the population level, with this exact velocity field and exact ODE integration, the induced flow preserves the conditional interpolation marginals:

$$\mathrm{Law}(Z_t \mid C = c) = \mathrm{Law}(X_t \mid C = c), \qquad t \in [0, 1]. \tag{67}$$

Thus, RRF inherits the marginal-preserving property of Rectified Flow while introducing an observation-aware residual endpoint.

The residual endpoint further yields an exact acceleration point. Since

$$X_t = [1 - (1 + \gamma)t]X_0 + \gamma t C + t\epsilon, \tag{68}$$

the coefficient of the unknown clean image X_0 vanishes at

$$t^\star = \frac{1}{1+\gamma}. \tag{69}$$

Hence,

$$X_{t^\star} = \frac{\gamma}{1+\gamma} C + \frac{1}{1+\gamma}\epsilon, \tag{70}$$

which depends only on the observed condition C and Gaussian noise. This allows RRF to start sampling from a noisy LQ-related state and integrate only over the shortened interval $[0, t^\star]$.

Consequently, the RRF training objective is obtained by substituting the residual path velocity into the standard Rectified Flow objective:

$$\mathcal{L}_{\mathrm{RRF}}(\theta) = \int_0^{t^\star} \mathbb{E}\left[\|v_\theta(X_t, t, C) - (\gamma(C - X_0) + \epsilon - X_0)\|_2^2\right] dt. \tag{71}$$

In this sense, RRF keeps the linear interpolation and marginal-preserving structure of Rectified Flow, but shifts the endpoint from pure Gaussian noise to a residual-aware stochastic endpoint. This endpoint change is what enables noisy-LQ initialization and accelerated sampling.

H More Visual Results

We provide additional visual comparisons to complement the quantitative results. These examples cover real-world benchmarks, synthetic benchmarks, and challenging face restoration cases, showing the behavior of ScaleResfusion under diverse degradation patterns. Overall, ScaleResfusion better preserves input structures while recovering realistic local details, and it avoids the over-smoothed textures, semantic drift, or hallucinated artifacts that can appear in competing restoration methods.

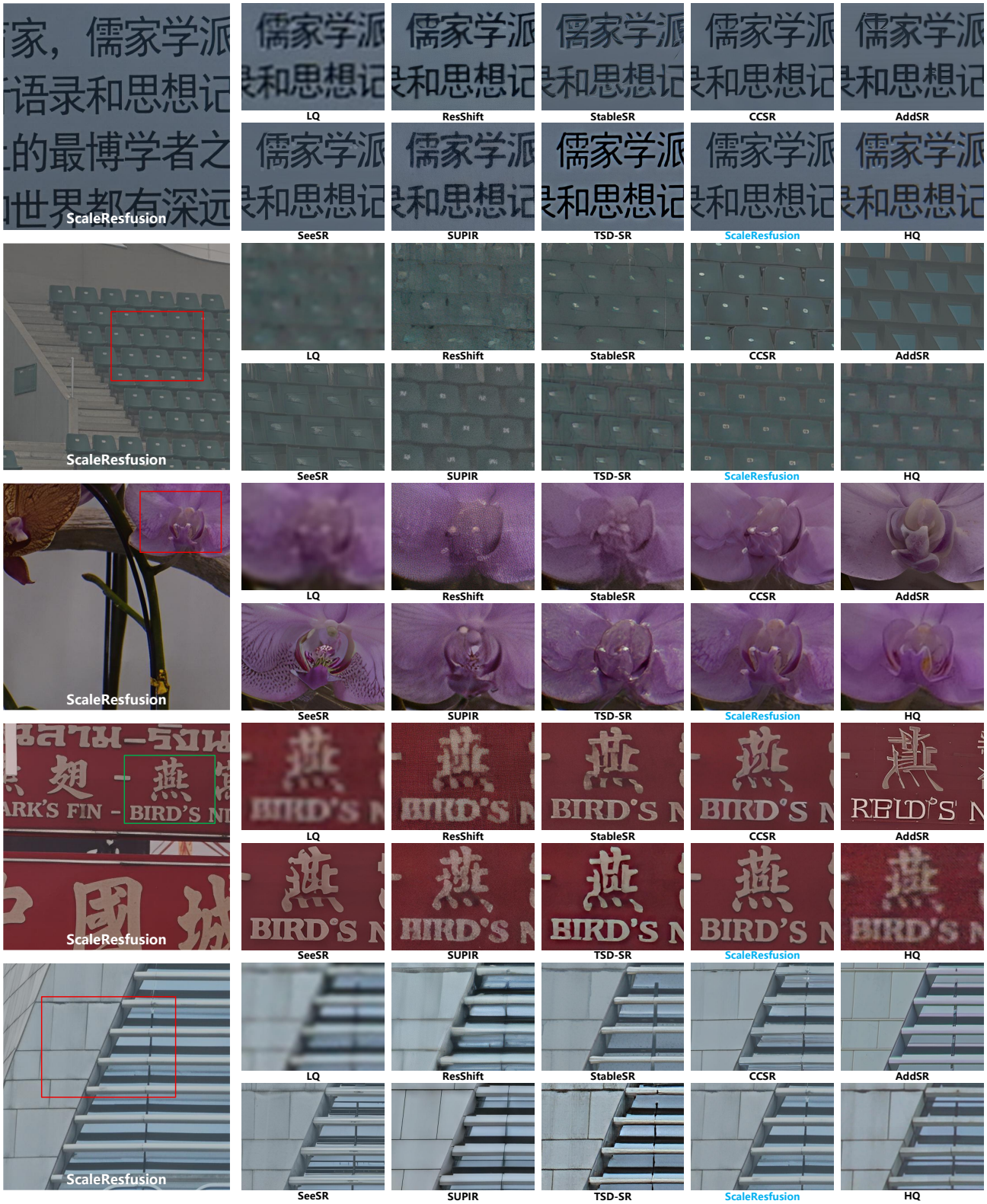


Fig. 23. More visual comparison on DRealSR with existing diffusion-based restoration methods. The left side shows the results of ScaleResfusion, while the right side compares local crops. Overall, ScaleResfusion better preserves the global structure while recovering more natural details without the over-smoothing or hallucinated patterns observed in competing methods.

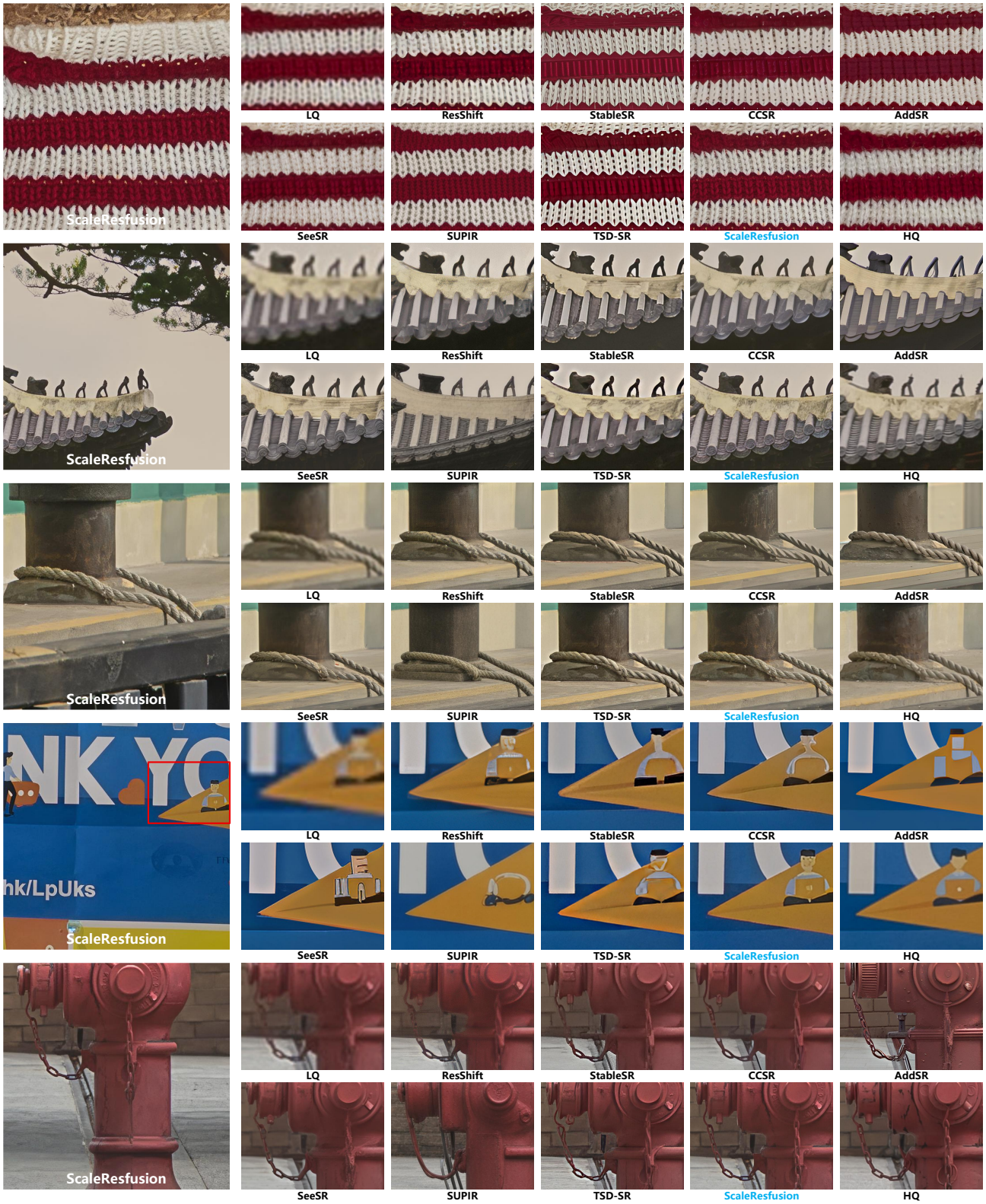


Fig. 24. More visual comparison on RealSR with existing diffusion-based restoration methods. The left side shows the results of ScaleResfusion, while the right side compares local crops. Overall, ScaleResfusion better preserves the global structure while recovering more natural details without the over-smoothing or hallucinated patterns observed in competing methods.

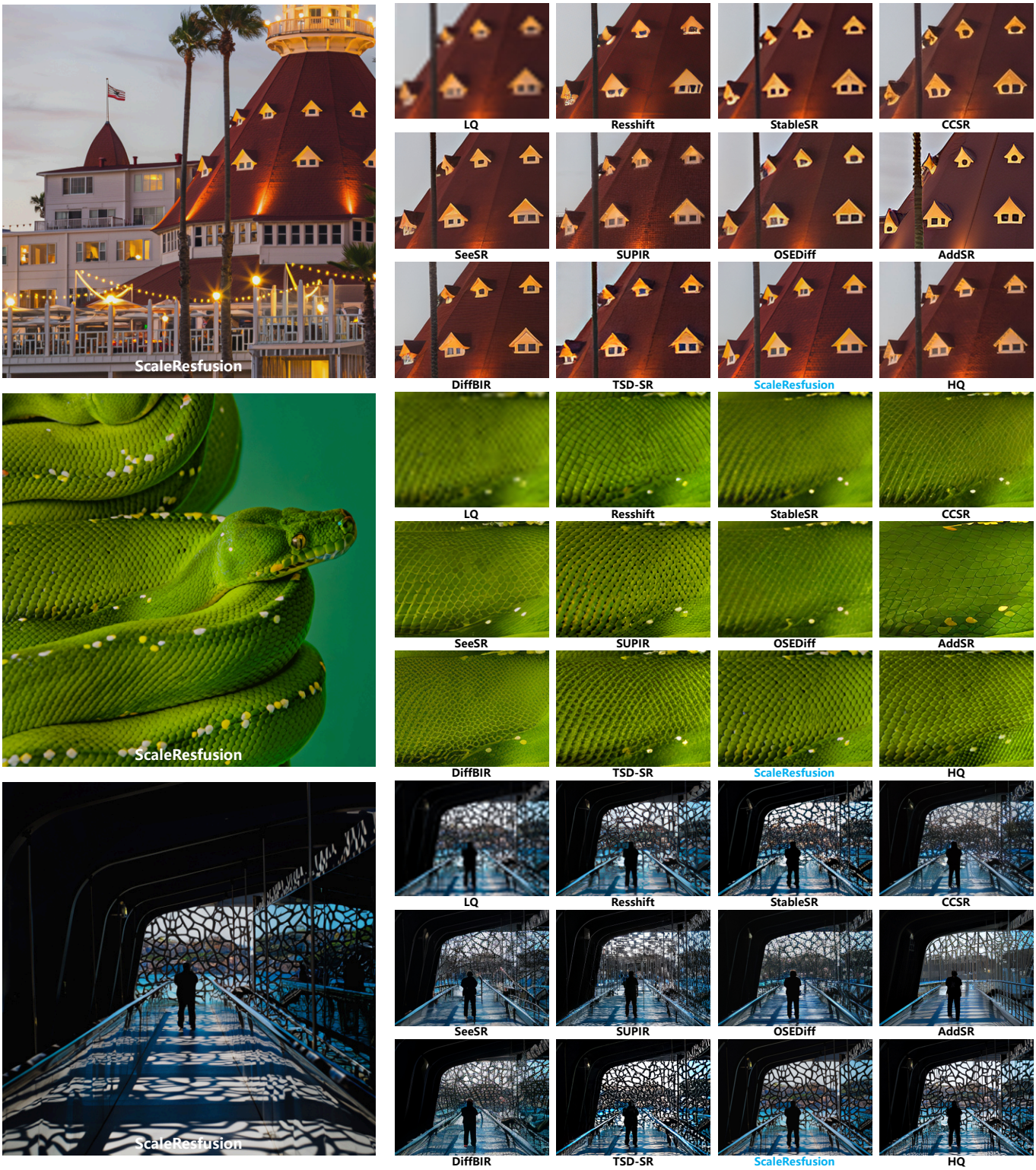


Fig. 25. More visual comparison on LSDIR-Val with existing diffusion-based restoration methods. The left side shows the results of ScaleResfusion, while the right side compares local crops. Overall, ScaleResfusion better preserves the global structure while recovering more natural details without the over-smoothing or hallucinated patterns observed in competing methods.

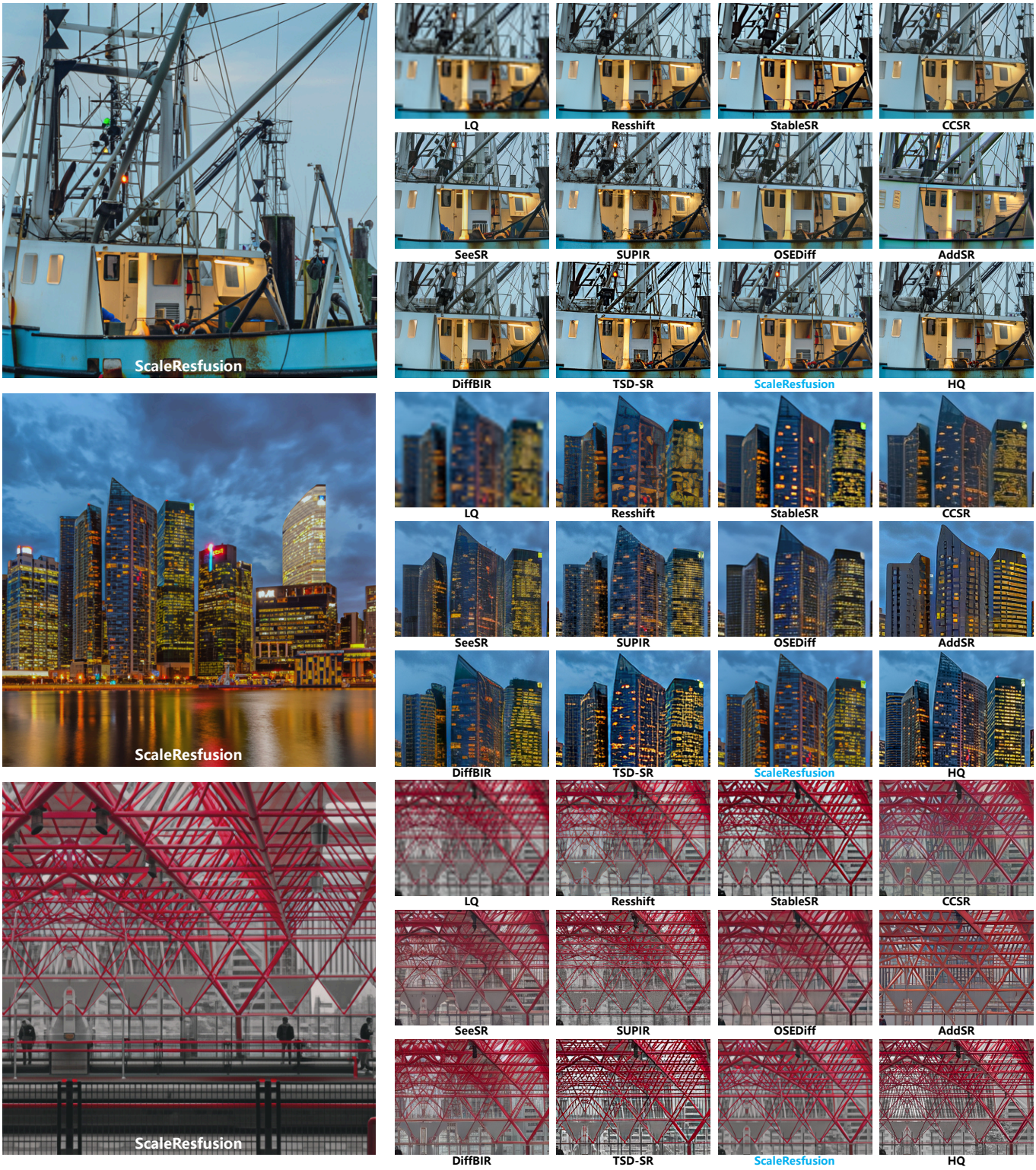


Fig. 26. More visual comparison on LSDIR-Val with existing diffusion-based restoration methods. The left side shows the results of ScaleResfusion, while the right side compares local crops. Overall, ScaleResfusion better preserves the global structure while recovering more natural details without the over-smoothing or hallucinated patterns observed in competing methods.

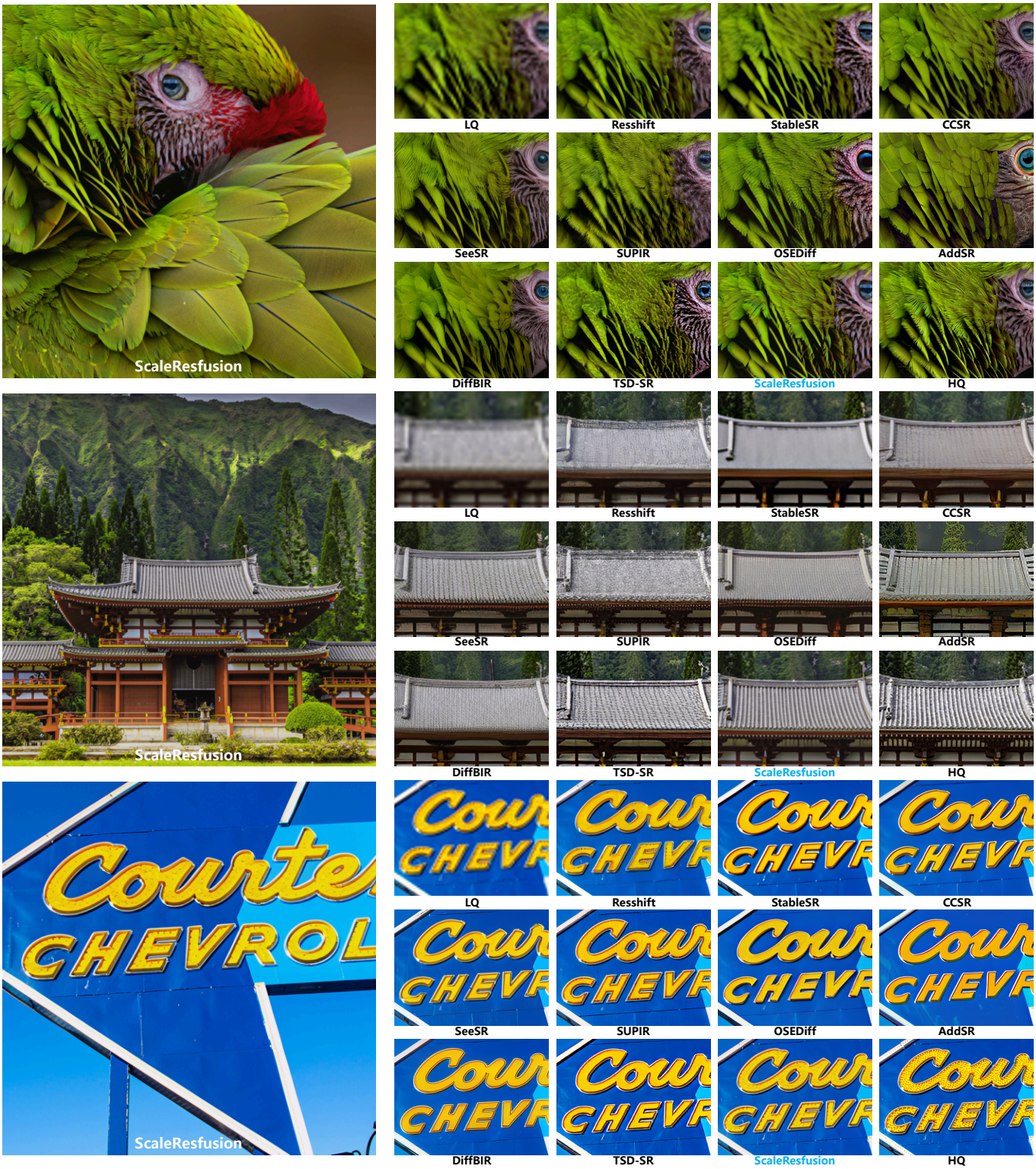


Fig. 27. More visual comparison on LSDIR-Val with existing diffusion-based restoration methods. The left side shows the results of ScaleResfusion, while the right side compares local crops. Overall, ScaleResfusion better preserves the global structure while recovering more natural details without the over-smoothing or hallucinated patterns observed in competing methods.

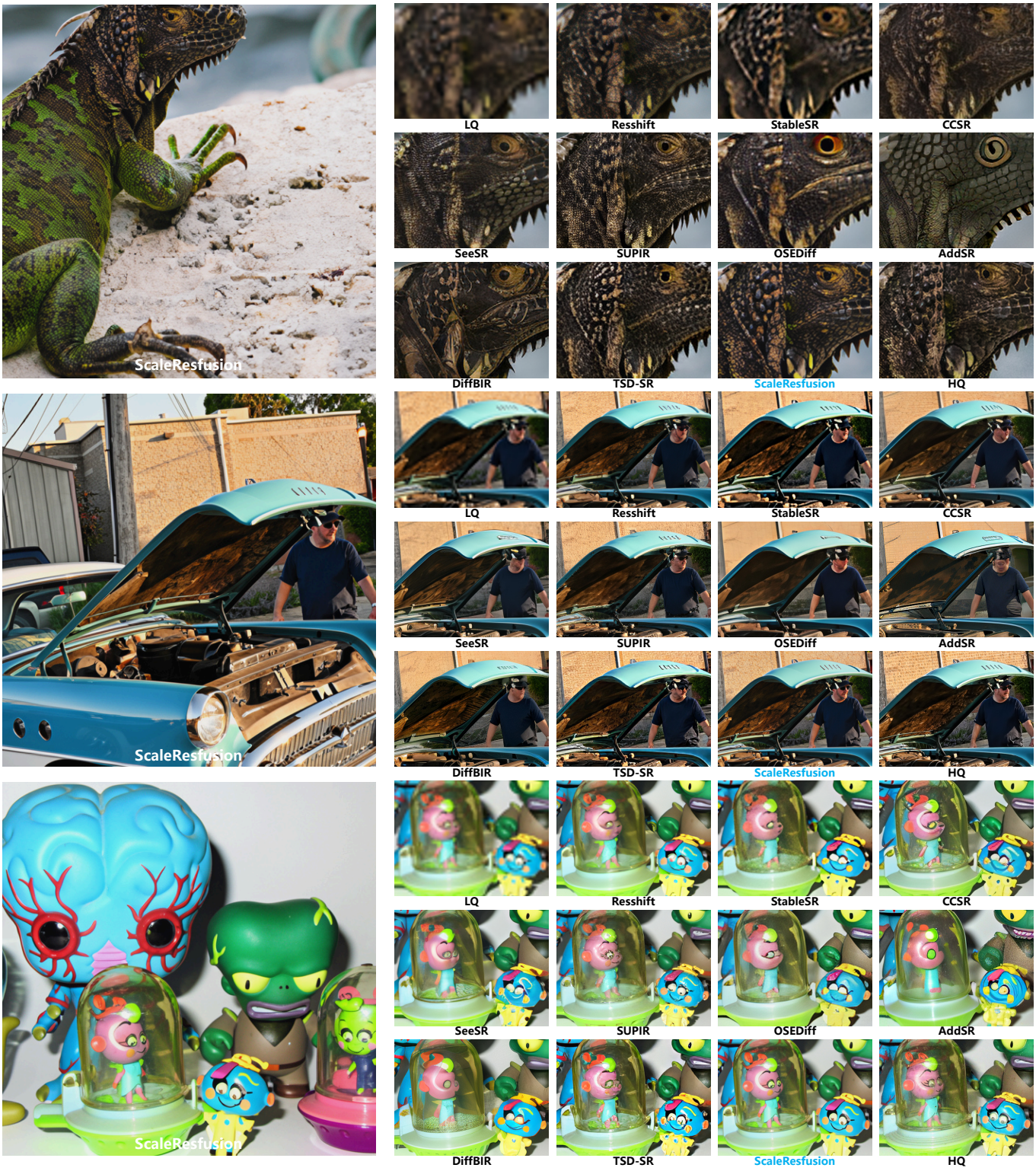


Fig. 28. More visual comparison on LSDIR-Val with existing diffusion-based restoration methods. The left side shows the results of ScaleResfusion, while the right side compares local crops. Overall, ScaleResfusion better preserves the global structure while recovering more natural details without the over-smoothing or hallucinated patterns observed in competing methods.

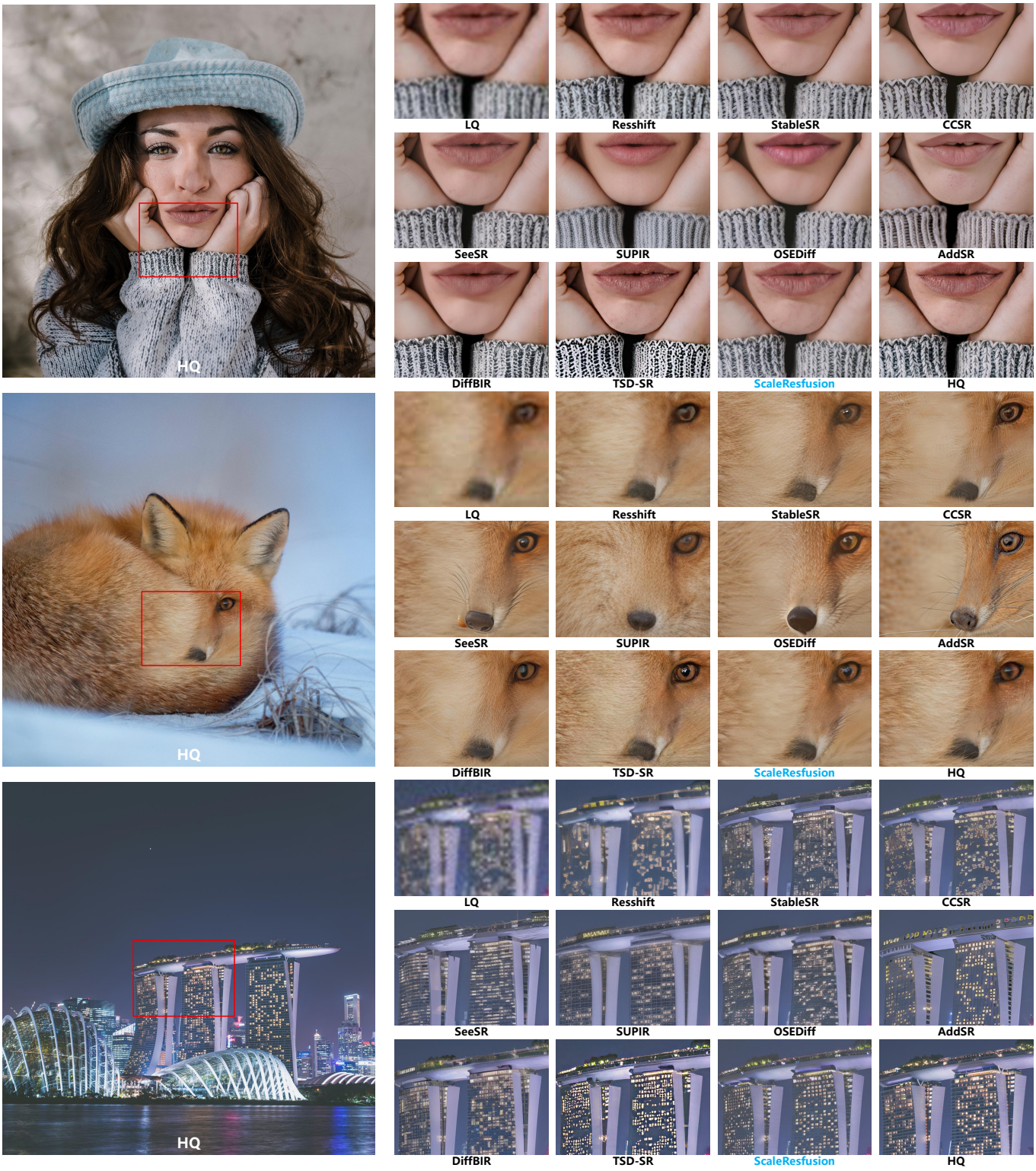


Fig. 29. More visual comparison on DIV2K-Val with existing diffusion-based restoration methods. The left side shows the HQ reference images, while the right side compares local crops. Overall, ScaleResfusion recovers more natural details while remaining faithful to the HQ references, avoiding the over-smoothing or hallucinated patterns observed in competing methods.

Fig. 30. More visual comparison on DIV2K-Val with existing diffusion-based restoration methods. The left side shows the HQ reference images, while the right side compares local crops. Overall, ScaleResfusion recovers more natural details while remaining faithful to the HQ references, avoiding the over-smoothing or hallucinated patterns observed in competing methods.

Fig. 31. More visual comparison on WebPhoto-Test [Wang et al. 2021a], an in-the-wild face restoration benchmark without paired ground truth. In this zero-shot setting, ScaleResfusion better handles out-of-distribution degradations and restores more natural facial details than competing methods.

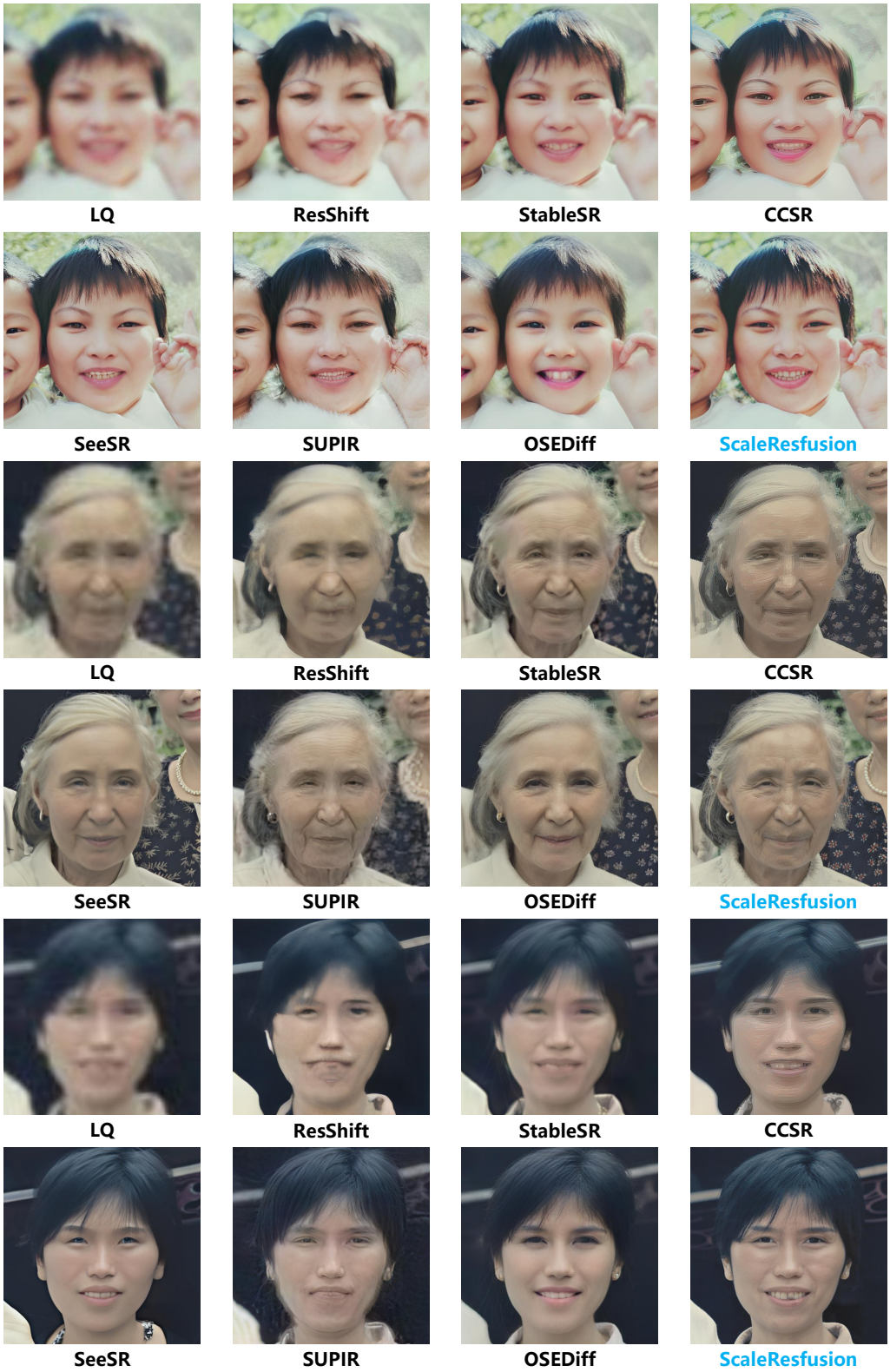


Fig. 32. More visual comparison on WebPhoto-Test [Wang et al. 2021a], an in-the-wild face restoration benchmark without paired ground truth. In this zero-shot setting, ScaleResfusion better handles out-of-distribution degradations and restores more natural facial details than competing methods.

Fig. 33. More visual comparison on WebPhoto-Test [Wang et al. 2021a], an in-the-wild face restoration benchmark without paired ground truth. In this zero-shot setting, ScaleResfusion better handles out-of-distribution degradations and restores more natural facial details than competing methods.

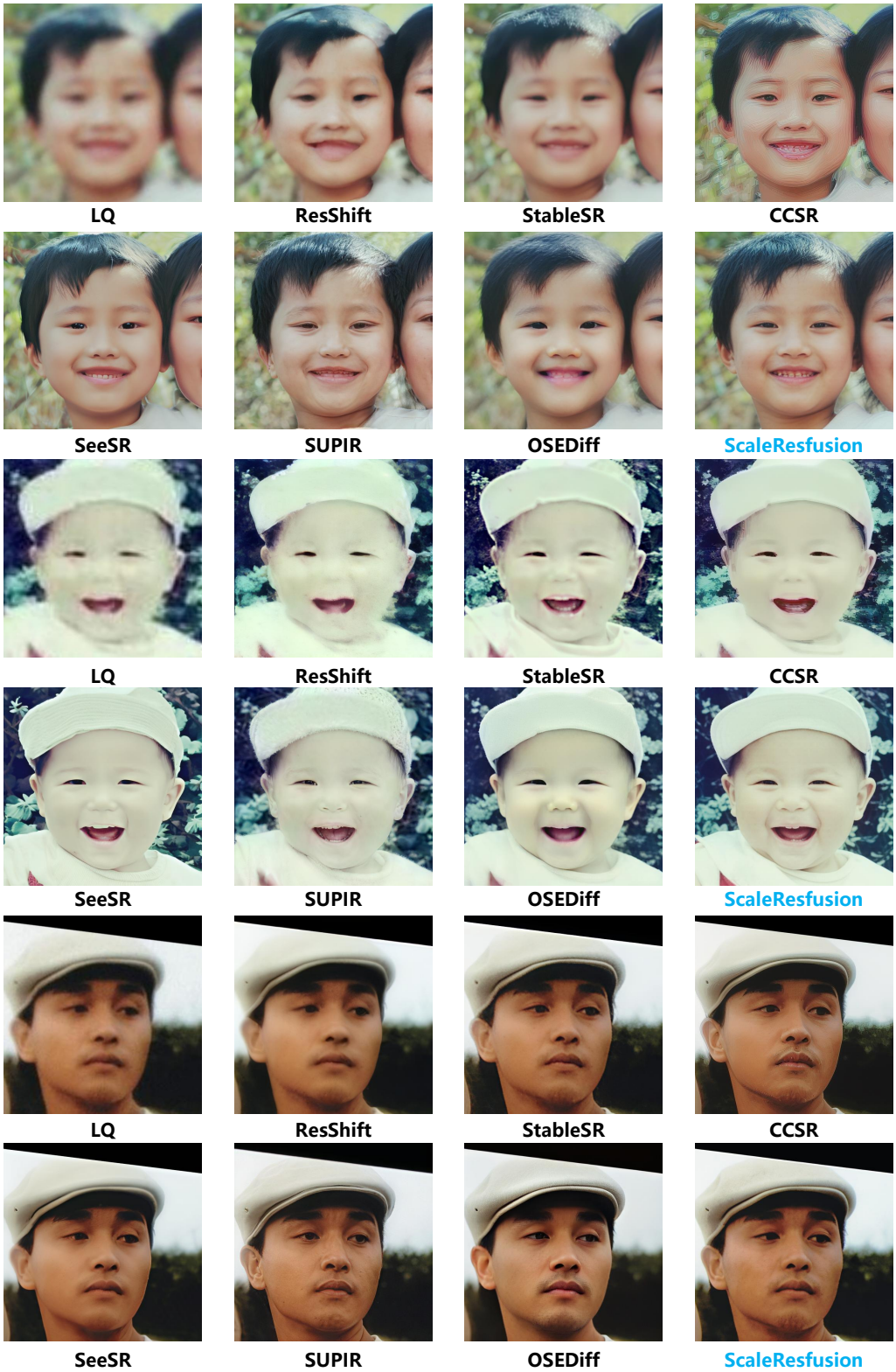


Fig. 34. More visual comparison on WebPhoto-Test [Wang et al. 2021a], an in-the-wild face restoration benchmark without paired ground truth. In this zero-shot setting, ScaleResfusion better handles out-of-distribution degradations and restores more natural facial details than competing methods.