RETHINKING FIDELITY–PERCEPTION TRADE-OFF IN LATENT-FREE SUPER-RESOLUTION VIA PIXEL FLOW

Anonymous authors
Paper under double-blind review

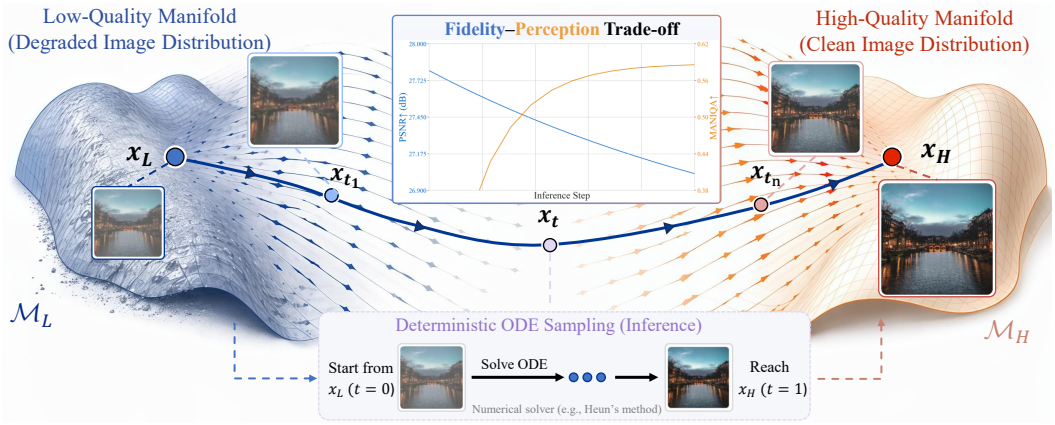

Figure 1: **Conceptual illustration of PixelSR.** PixelSR formulates real-world SR as a latent-free pixel-flow transformation from the low-quality manifold $\mathcal{M}_L$ to the high-quality manifold $\mathcal{M}_H$. Starting from the degraded input $\mathbf{x}_L$ at $t = 0$, the model solves a deterministic ODE to progressively move along an image-space restoration trajectory toward the clean image $\mathbf{x}_H$ at $t = 1$. By varying steps along this trajectory, PixelSR provides a simple mechanism to navigate the fidelity–perception trade-off, where fewer steps favor fidelity preservation and more steps improve perception.

ABSTRACT

Real-world image super-resolution (SR) aims to reconstruct high-resolution images from low-resolution observations. Recent models using diffusion models and flow matching have shown strong perceptual restoration ability, yet we observe two key limitations that hinder their robustness and fidelity: 1) their performance is sensitive to different noise and sampling stochasticity; 2) per-step noise prediction may fail under composite degradations and accumulate residual errors, causing identity, texture, and structural drift. To address these issues, we rethink the fidelity–perception trade-off from a latent-free perspective and propose PixelSR, a pixel-flow SR framework that directly models the restoration trajectory in image space. Instead of predicting per-step noise in a compressed latent space, PixelSR predicts clean high-resolution images and derives the corresponding pixel-space flow for iterative refinement, thereby anchoring the generation process to the high-quality image manifold while preserving low-level structures from the input. This clean-data-centric formulation reduces error accumulation and provides a more stable path toward perceptually realistic yet faithful reconstruction. Interestingly, our experiments reveal that varying the number of sampling steps provides a simple and effective mechanism to navigate the fidelity–perception trade-off. Extensive experiments on synthetic and real-world SR benchmarks demonstrate that PixelSR achieves a better fidelity–perception balance than strong diffusion-based baselines.

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

1 INTRODUCTION

Recent diffusion-based Ho et al. (2020) and flow-matching-based Lipman et al. (2022) approaches have shown strong perceptual restoration ability for real-world image super-resolution (SR) Wu et al. (2024a); Wang et al. (2024a); Chen et al. (2025). By leveraging powerful generative priors and iterative sampling processes, these methods can synthesize visually realistic high-frequency details from severely degraded low-resolution (LR) inputs. However, we observe two key limitations that hinder their robustness and fidelity in practice. First, their performance is often sensitive to noise, as shown in Fig. 2. Small changes in the sampling process may lead to noticeable variations in texture, identity, and structural details, making the restoration behavior unstable, as shown in Fig. 2. Second, many existing diffusion/flow-based SR pipelines rely on per-step noise prediction or noise-related targets, such as velocity, rather than directly reconstructing clean high-resolution (HR) images. Since these prediction targets are not themselves clean images on the high-quality image manifold, inaccurate estimation at intermediate steps can accumulate along the iterative trajectory, especially under complex real-world degradations.

In practical real-world SR scenarios, LR images are rarely degraded by a single, well-defined corruption process Zhou et al. (2025b); Wang et al. (2021). Instead, they often suffer from unknown and composite degradations, including blur, sensor noise, compression artifacts, demosaicing errors, and other complex distortions that deviate from predefined degradation assumptions. Under such conditions, purely predicting per-step noise may fail to remove all corruption at each step, and the remaining residual errors can be propagated or amplified during subsequent iterations. This can cause identity drift, hallucinated textures, and structural inconsistency, ultimately weakening the balance between distortion fidelity and perceptual realism. These observations motivate us to rethink the fidelity–perception trade-off in real-world SR from a latent-free and clean-data-centric perspective.

To this end, we propose **PixelSR**, a pixel-flow framework for latent-free real-world image super-resolution. As illustrated in Fig. 1, PixelSR directly models the restoration trajectory in image space, without relying on compressed latent-space generation. Instead of taking stochastic noise as the primary prediction target, PixelSR predicts clean HR images and derives the corresponding pixel-space flow for iterative refinement. This formulation anchors the restoration process to the high-quality image manifold while preserving low-level structures from the LR input. By allocating model capacity to clean-image prediction rather than off-manifold noise estimation, PixelSR reduces error accumulation along the sampling trajectory and provides a more stable path toward faithful and perceptually realistic reconstruction.

An interesting property of PixelSR is that the number of sampling steps naturally provides a simple mechanism to navigate the fidelity-perception trade-off, as shown in Fig. 1. With fewer sampling steps, the model tends to better preserve pixel-level fidelity and identity consistency. With more sampling steps, the model can progressively introduce richer perceptual details and more realistic textures. This controllability allows PixelSR to adjust the balance between fidelity and perception without changing the model architecture or requiring retraining.

Extensive experiments on synthetic and real-world SR benchmarks demonstrate that PixelSR achieves a better fidelity–perception balance than strong diffusion-based baselines. In particular, PixelSR improves identity preservation, structural consistency, and perceptual quality under complex and unknown degradations. These results suggest that latent-free pixel-flow modeling provides an effective alternative to conventional noise-centric generative SR pipelines.

Our contributions are summarized as follows:

- We revisit the fidelity–perception trade-off in real-world SR from a latent-free pixel-flow perspective, highlighting the instability and error-accumulation issues of existing noise-centric generative pipelines using diffusion models or flow matching.
- We propose PixelSR, a clean-data-centric pixel-flow framework that directly predicts clean HR images and derives pixel-space flow for iterative refinement.
- We reveal that the number of sampling steps provides a simple and effective mechanism to control the fidelity–perception trade-off.
- Extensive experiments show that PixelSR achieves strong perceptual quality while better preserving identity, structure, and low-level fidelity compared with diffusion-based baselines.

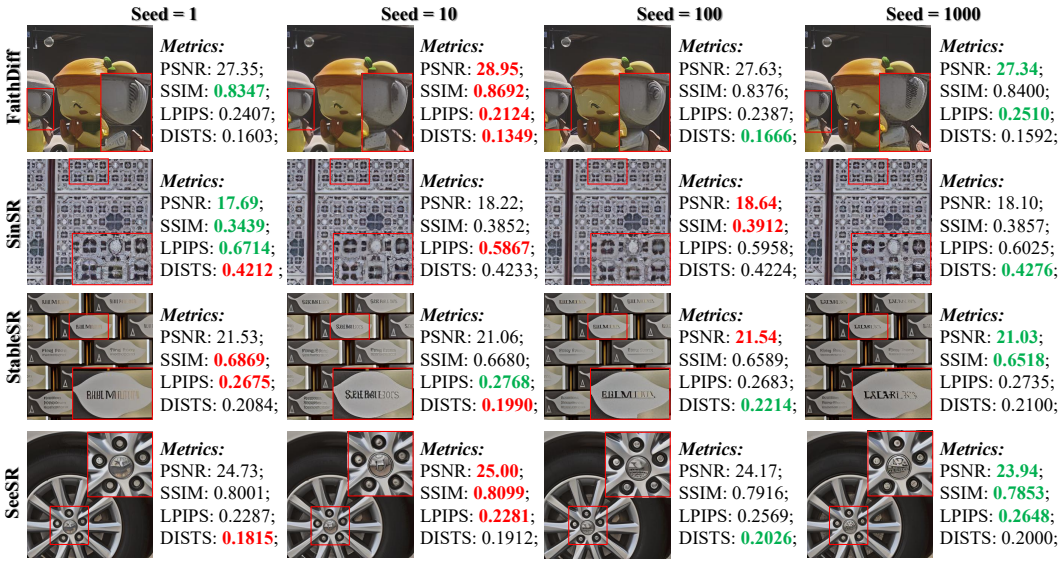

Figure 2: The impact of random sampling strategy on restoration results. The **red** markers indicate the optimal values, while the **green** markers indicate the worst values for each metric.

2 Related Works

Diffusion model-based SR. Compared with previous approaches (Dong et al., 2014; Lim et al., 2017; Zhang et al., 2018b; Liang et al., 2021; Chen et al., 2023), diffusion models have recently become a powerful paradigm for real-world image SR, as their iterative generative process can synthesize realistic high-frequency details from severely degraded LR inputs. As shown in Fig. 3 early conditional diffusion methods, such as SR3 Saharia et al. (2022b) and Palette Saharia et al. (2022a), formulate SR as a conditional denoising process. Recent SR methods further exploit large-scale pre-trained diffusion priors. StableSR Wang et al. (2024a) adapts latent diffusion models for blind SR with a time-aware encoder and progressive aggregation sampling, while DiffBIR Lin et al. (2024a) decomposes blind restoration into degradation removal and generative content regeneration. PASD Yang et al. (2024) introduces pixel-aware guidance to improve local structure preservation, SeeSR Wu et al. (2024b) employs degradation-aware semantic prompts to enhance semantic fidelity, and SUPIR Yu et al. (2024) scales up generative restoration with multimodal prompts and high-quality data. ResShift Yue et al. (2023) improves the efficiency of diffusion SR by constructing a residual-shifting Markov chain from LR to HR images.

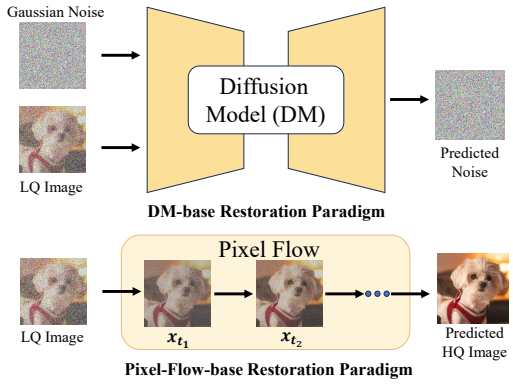

Figure 3: Differences between DM-based restoration paradigms and our PixelSR.

Despite their strong perceptual quality, they are still largely built upon multi-step denoising, latent-space generation, or noise-related prediction targets. Their outputs can therefore be sensitive to noise, as shown in Fig. 2. The fidelity-perception trade-off remains difficult to control in noise-centric diffusion SR. Moreover, under complex real-world degradations, residual errors may not be fully removed at each denoising step and can accumulate during sampling, causing hallucinated details, identity drift, and structural inconsistency.

Flow matching-based SR. Flow matching and rectified flow have recently emerged as efficient alternatives to diffusion models by learning a vector field that transports samples from a source distribution to a target distribution Lipman et al. (2022); Liu et al. (2023). Compared with diffusion models, flow-based formulations can provide straighter sampling trajectories and improved sampling efficiency. In SR, recent works have explored rectified-flow or flow-matching formulations to bridge

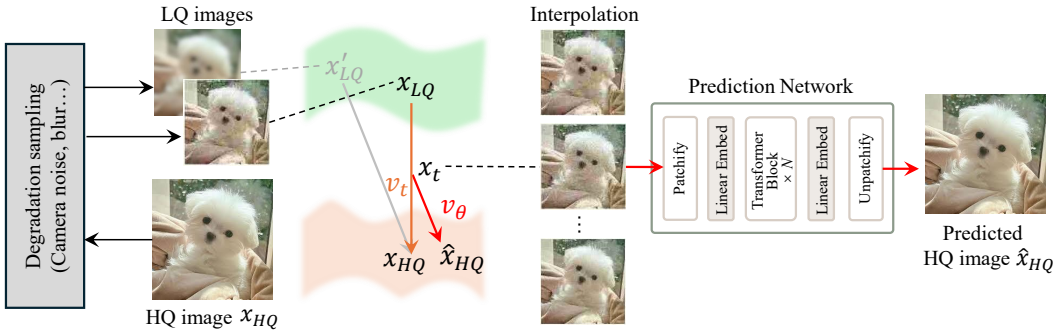

Figure 4: Overview of our proposed PixelSR. PixelSR constructs a latent-free pixel-space trajectory between the degraded LQ image $\mathbf{x}_{LQ}$ and the clean HQ image $\mathbf{x}_{HQ}$. An intermediate state $\mathbf{x}_t$ is sampled through interpolation and fed into the prediction network, which directly estimates the clean HQ image $\hat{\mathbf{x}}_{HQ}$. The corresponding pixel-flow velocity $\mathbf{v}_\theta$ is then derived from $\hat{\mathbf{x}}_{HQ}$ and supervised to match the ground-truth restoration direction $\mathbf{v}_t$, enabling stable and faithful LQ-to-HQ restoration.

LR and HR image distributions. FlowSR Xu et al. (2025) formulates SR as a rectified flow between LR and HR images and employs consistency learning to distill a high-quality multi-step sampling process into faster inference. Other flow-based SR methods, such as FluxSR Li et al. (2025), improve efficiency while preserving the fidelity–realism trade-off. In addition, rectified flow has been used to model realistic degradation processes for real-world SR, where flow trajectories are learned to synthesize realistic LR-HR pairs from unpaired data Zhou et al. (2025a).

Although flow matching-based methods reduce the sampling burden compared with conventional diffusion models, many of them still inherit a noise-to-data transport formulation, start from noisy or noise-augmented states, or depend on teacher trajectories learned from stochastic generative processes. As a result, the restoration path can still be affected by noise initialization, trajectory mismatch, and accumulated prediction errors. This is especially problematic for real-world SR, where preserving identity, geometry, and low-level structures is as important as generating realistic textures. In contrast, our PixelSR directly predicts clean HR images and derives the corresponding pixel-space flow for iterative refinement. This latent-free, clean-data-centric formulation anchors the sampling trajectory to the high-quality image manifold and provides a simple way to navigate the fidelity-perception trade-off by varying the number of sampling steps.

3 Proposed Method

3.1 Trajectory Reformulation

Following the rectified flow framework Esser et al. (2024), we reformulate real-world IR as a deterministic trajectory transformation from the LQ image manifold $\mathcal{M}_L$ to the HQ image manifold $\mathcal{M}_H$ as shown in Algorithm 1. Given an HQ target image $\mathbf{x}_H$, we synthesize an LQ image $\mathbf{x}_L$ by `random_degradation` using the pipelines from Wang et al. (2021), the interpolated state at time $t \in [0, 1]$ is defined as:

$$\mathbf{x}_t = t\mathbf{x}_H + (1-t)\mathbf{x}_L. \tag{1}$$

Under this formulation, $\mathbf{x}_t{=}\mathbf{x}_L$ when $t{=}0$, and $\mathbf{x}_t{=}\mathbf{x}_H$ when $t{=}1$, such that the restoration process corresponds to traversing the trajectory from $t{=}1$ to $t{=}0$. Unlike conventional diffusion formulations that interpolate between the target image and Gaussian noise, i.e., $\mathbf{x}_t = t\mathbf{x}_H + (1-t)\boldsymbol{\varepsilon}$, we replace the stochastic noise term $\boldsymbol{\varepsilon}$ with the structured degraded image $\mathbf{x}_L$. This modification constrains the entire restoration trajectory to remain on the image manifold, thereby avoiding the instability introduced by random noise initialization and providing a more reliable restoration foundation. The resulting velocity field is obtained by differentiating the trajectory with respect to time:

$$\mathbf{v}_t = \frac{d\mathbf{x}_t}{dt} = \mathbf{x}_H - \mathbf{x}_L. \tag{2}$$

This constant velocity field explicitly describes the degradation direction from HQ images to LQ images. Consequently, the objective of PixelSR is to learn reverse velocity field for progressive IR.

3.2 Velocity Prediction and Training

To model the reverse restoration process, we train a neural network f_θ to directly predict the clean HQ image $\mathbf{x}_H$ from any intermediate state $\mathbf{x}_t$, conditioned on both the degradation time step t and the input LQ image $\mathbf{x}_L$, as shown in Fig. 4. This design follows the $\mathbf{x}$-prediction paradigm, ensuring that the network output consistently resides on the natural image manifold. The predicted velocity field is therefore defined as:

$$\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{x}_L) = \frac{f_\theta(\mathbf{x}_t, t, \mathbf{x}_L) - \mathbf{x}_t}{1-t} = \frac{\mathbf{x}_\theta - \mathbf{x}_t}{1-t}, \quad (3)$$

where $\mathbf{x}_\theta$ denotes the predicted HQ image. To supervise velocity prediction accuracy, we calculate the discrepancy between the predicted velocity and the ground-truth velocity from Eq. (2) using the flow-matching loss:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\mathbf{x}_H,\mathbf{x}_L}\left[\left|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{x}_L) - \mathbf{v}\right|_2^2\right]. \quad (4)$$

Algorithm 1 Training stage

Input: JiT network $\texttt{net}(\mathbf{x}_t, t)$
HQ images $\mathbf{x}_H$
$t \leftarrow \texttt{sample_t()}$
$\mathcal{D} \leftarrow \texttt{random_degradation()}$
$\mathbf{x}_L \leftarrow \mathcal{D}(\mathbf{x}_H)$
$\mathbf{x}_t \leftarrow (1-t)\cdot\mathbf{x}_H + t\cdot\mathbf{x}_L$
$\mathbf{v} \leftarrow (\mathbf{x}_H - \mathbf{x}_t)/(1-t)$
$\mathbf{x}_\theta \leftarrow \texttt{net}(\mathbf{x}_t, t)$
$\mathbf{v}_\theta \leftarrow (\mathbf{x}_\theta - \mathbf{x}_t)/(1-t)$
$\mathcal{L} \leftarrow \mathcal{L}_{\text{FM}} + \lambda_1\mathcal{L}_{\text{LPIPS}} + \lambda_2\mathcal{L}_{\text{DINO}}$

To further balance restoration fidelity with perceptual quality, we introduce a dual-stream complementary perceptual supervision strategy. Specifically, LPIPS Zhang et al. (2018a) is employed to capture low-level perceptual discrepancies in texture, color consistency, local contrast, and fine detail. Using a frozen pre-trained VGG network Simonyan & Zisserman (2014), the LPIPS loss is formulated as:

$$\mathcal{L}_{\text{LPIPS}} = \sum_l \left|w_l \odot \left(f^l_{\text{VGG}}(\mathbf{x}_\theta) - f^l_{\text{VGG}}(\mathbf{x}_H)\right)\right|_2^2, \quad (5)$$

where $\odot$ denotes the Hadamard product, and w_l represents the learned channel-wise weighting at the l-th feature VGG layer f^l_{VGG}. In parallel, to enhance semantic consistency and high-level structural restoration, we incorporate a DINO-based perceptual loss Oquab et al. (2023):

$$\mathcal{L}_{\text{DINO}} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\left(1 - \cos\left(f^p_{\text{DINO}}(\mathbf{x}_\theta), f^p_{\text{DINO}}(\mathbf{x}_H)\right)\right), \quad (6)$$

where $f_{\text{DINO}}(\cdot)$ denotes patch-level feature extraction using a frozen DINOv2-B model Oquab et al. (2023), and $\mathcal{P}$ represents the set of extracted patches. This loss emphasizes global semantic alignment, structural coherence, and scene-level consistency. The final training objective integrates flow supervision with multi-scale perceptual constraints:

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda_1\mathcal{L}_{\text{LPIPS}} + \lambda_2\mathcal{L}_{\text{DINO}}, \quad (7)$$

where λ_1 and λ_2 are balancing coefficients that regulate the relative contributions of low-level and high-level perceptual supervision. This unified objective enables PixelSR to jointly optimize deterministic flow learning, perceptual realism, and semantic restoration quality.

3.3 Deterministic Sampling for Stable Restoration

During inference, PixelSR performs restoration through a fully deterministic sampling process as shown in Algorithm 2. Starting from the degraded input image $\mathbf{x}_L$, the restoration trajectory is generated by solving the learned reverse ordinary differential equation (ODE) backward from $t = 0$ to $t = 1$:

$$\frac{d\mathbf{x}_t}{dt} = \mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{x}_L), \mathbf{x}_H = \mathbf{x}_t|_{t=0}. \quad (8)$$

This ODE can be efficiently solved using numerical integration techniques such as Heun's method Rapp (2017). Through this process, the degraded image is progressively refined along the learned manifold-constrained trajectory, continuously enhancing structural details, texture fidelity, and perceptual quality. Unlike conventional DM-based models, PixelSR eliminates sampling randomness entirely, thereby avoiding restoration instability and output inconsistency across repeated runs. As a result, PixelSR provides stable and reliable restoration results for any given LQ input. This deterministic manifold-guided restoration process aligns naturally with the core objective of real-world super-resolution and image restoration tasks, enabling more faithful recovery of HQ visual content.

Algorithm 2 Inference stage

Input: Low Quality Images $\mathbf{x}_L$
Inference Steps s
$\mathbf{x}_0 \leftarrow \mathbf{x}_L$
for i in $[0, s-1]$:
$t \leftarrow i/s$
$t_{\text{next}} \leftarrow (i+1)/s$
$\mathbf{x}_\theta \leftarrow \texttt{net}(\mathbf{x}_t, t)$
$\mathbf{v}_\theta \leftarrow (\mathbf{x}_\theta - \mathbf{x}_t)/(1-t)$
$\mathbf{x}_t \leftarrow \mathbf{x}_t + (t_{\text{next}} - t)\,\mathbf{v}_\theta$
Output: $\mathbf{x}_1$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Training Datasets. Unlike prior studies Lin et al. (2024b); Wang et al. (2024a); Yang et al. (2024), which commonly integrate multiple training datasets to develop real-world IR models, we adopt the training configurations of SeeSR Wu et al. (2024b) and OSEDiff Wu et al. (2024a) to establish a more streamlined evaluation protocol and ensure fairer performance comparisons. Specifically, all restoration models are trained exclusively on the LSDIR dataset Li et al. (2023), which contains 84,991 high-quality real-world images. To faithfully emulate real-world degradation processes, we employ the two-stage stochastic mixed degradation pipeline proposed in Real-ESRGAN Wang et al. (2021) to generate LQ and HQ image pairs. For test set selection, to comprehensively assess the generalization ability and robustness of the PixelSR model across both synthetic and real-world restoration scenarios, we utilize the three benchmark datasets adopted by StableSR Wang et al. (2024a): DIV2K-Val Agustsson & Timofte (2017), RealSR Cai et al. (2019), and DRealSR Wei et al. (2020). DIV2K-Val Agustsson & Timofte (2017) consists of 100 synthetic images at a resolution of 512×512, whereas RealSR Cai et al. (2019) and DRealSR Wei et al. (2020). each provide 100 real-world LQ-HQ image pairs with resolutions of 128×128 and 512×512, respectively. Notably, for DIV2K-Val Wei et al. (2020), the Real-ESRGAN degradation simulation pipeline Wang et al. (2021) is further applied to perform center cropping and degradation synthesis, thereby producing the corresponding degraded inputs and ground-truth reference images.

Compared Methods. We evaluate PixelSR against a diverse set of state-of-the-art (SOTA) diffusion model (DM) based real-world IR approaches, including StableSR Wang et al. (2024a), DiffBIR Lin et al. (2024b), SeeSR Wu et al. (2024b), PASD Yang et al. (2024), ResShift Yue et al. (2024), SinSR Wang et al. (2024b), OSEDiff Wu et al. (2024a), FaithDiff Chen et al. (2025), XPSR Qu et al. (2024) and PiSA-SR Sun et al. (2025). Among all, StableSR, DiffBIR, SeeSR, PASD, ResShift, FaithDiff, and XPSR operate as multi-step restoration frameworks, whereas SinSR is a single-step model distilled from ResShift, and OSEDiff represents an efficient single-step DM based IR method.

Evaluation Metrics. To ensure a comprehensive quantitative evaluation of restoration performance, we employ both full-reference and no-reference image quality assessment metrics. Specifically, PSNR and SSIM Wang et al. (2004) are adopted to measure pixel-level fidelity between restored and reference images based on the Y channel in the YCbCr color space. LPIPS Zhang et al. (2018a) and DISTS Ding et al. (2020) are utilized as perceptual full-reference metrics to assess perceptual similarity. FID Heusel et al. (2017) is further employed to evaluate the distributional discrepancy between restored images and ground-truth images. In addition, no-reference quality metrics, including NIQE Mittal et al. (2012), MANIQA Yang et al. (2022), MUSIQ Ke et al. (2021), and CLIPIQA Wang et al. (2023), are incorporated to provide a broader assessment of perceptual image quality.

Implementation Details. PixelSR is trained using the AdamW optimizer Loshchilov & Hutter (2017) with an initial learning rate of 5×10^{-5} and a warm-up period of 5 epochs. The training process is conducted on a server equipped with 8 NVIDIA H20 GPUs using a batch size of 16. To enhance training stability, an exponential moving average (EMA) strategy is employed to smooth parameter updates over time and mitigate training fluctuations. For image prediction backbone, we adopt the JiT-H/32 architecture Li & He (2025). The model is initialized with pre-trained weights obtained from ImageNet-22K Deng et al. (2009) to leverage rich and generalizable visual representations, followed by fine-tuning on the LSDIR dataset. Regarding the loss formulation, the weighting coefficients λ_1 and λ_2 are empirically set to 0.1 and 0.01, respectively.

4.2 PERFORMANCE EXPERIMENTS

The quantitative performance of all compared image restoration methods is presented in Table 1. Based on the results, several key observations can be drawn: 1) The proposed PixelSR framework consistently achieves superior performance across multiple evaluation metrics on all three benchmark datasets. In particular, PixelSR demonstrates substantial advantages in full-reference perceptual metrics, including PSNR, SSIM, LPIPS, and DISTS, indicating its enhanced capability to preserve structural consistency while more faithfully reconstructing fine-grained textures and visual details. These improvements suggest that, relative to existing DM-based restoration approaches, PixelSR effectively mitigates the adverse effects of stochastic noise perturbation and random sampling, leading to more accurate restoration outcomes; 2) The performance gains observed in texture and detail

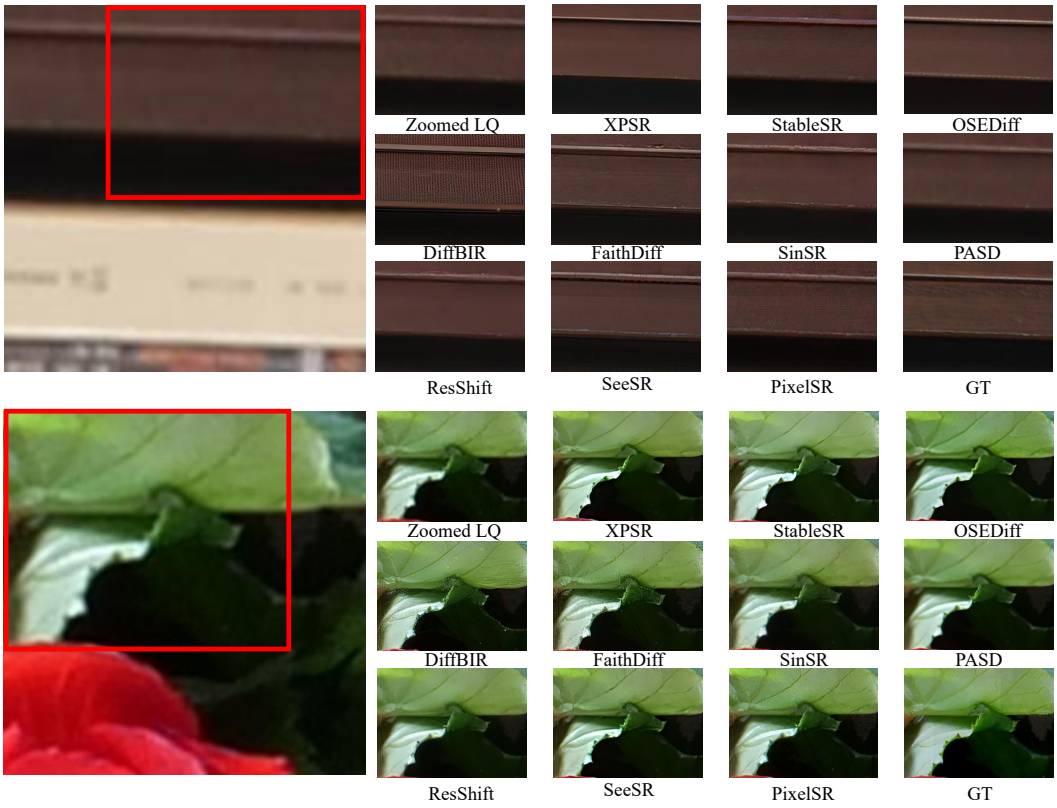

Figure 5: Results of different image restoration methods. Please zoom in for a better view.

fidelity are consistently maintained across both synthetic and real-world datasets, further validating the generalization ability, robustness, and effectiveness of PixelSR under diverse restoration scenarios; 3) From the perspective of restoration conditions, several competitive DM-based methods, such as SeeSR, OSEDiff, FaithDiff, and XPSR, incorporate additional textual prompts as auxiliary guidance to enhance restoration quality. In contrast, PixelSR achieves highly competitive and often superior performance by relying solely on pixel-space flow modelling for LQ image restoration, without requiring supplementary multimodal conditional inputs. This highlights the efficiency and strength of the proposed framework in leveraging intrinsic image information alone for HQ restoration.

To illustrate more clearly the differences in restoration performance between PixelSR and other algorithms, we have compared the results of various algorithms, as shown in Figure 1. By examining Figure 1, we can draw the following conclusions: 1) Existing diffusion-based restoration models often suffer from 'over-optimisation' in terms of detail handling, thereby overlooking some of the authentic textures and details inherent in the original image, and are at risk of introducing colour bias. In contrast, PixelSR eliminates the introduction of noise and focuses on directly learning the restoration path from low-quality (LQ) to high-quality (HQ) images, resulting in restoration results that more closely align with the original image and exhibit more authentic textures and details.

4.3 Experiment on Sampling Steps

To systematically examine the influence of sampling step count on the restoration performance of PixelSR, we conduct a series of experiments by varying the number of sampling steps during the inference stage. The corresponding restoration results across the three benchmark datasets under different sampling configurations are summarized in Table 2. Based on these results, several important observations can be made: 1) Under the single-step sampling setting, PixelSR achieves the highest PSNR and SSIM scores across all three datasets, indicating that single-step restoration primarily emphasizes pixel-level fidelity and structural consistency with the reference images; 2) As the number of sampling steps increases, perceptual quality metrics such as LPIPS, DISTS, and several no-reference evaluation metrics exhibit noticeable improvements. This trend suggests that multi-step sampling enables the model to generate richer textures and finer visual details, thereby enhancing the

Table 1: Quantitative comparison with state-of-the-art methods on both synthetic and real-world benchmarks. 's' denotes the number of diffusion reverse steps in the method. The best and second best results of each metric are highlighted in **red** and **blue**, respectively.

Datasets	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑
DIV2K-Val	StableSR-s200	23.26	0.5726	0.3113	0.2048	**24.44**	4.7581	65.92	0.6192	0.6771
	DiffBIR-s50	23.64	0.5647	0.3524	0.2128	30.72	4.7042	65.81	0.6210	0.6704
	SeeSR-s50	23.68	0.6043	0.3194	0.1968	25.90	4.8102	68.67	0.6240	**0.6936**
	PASD-s20	23.14	0.5505	0.3571	0.2207	29.20	**4.3617**	68.95	0.6483	0.6788
	ResShift-s15	**24.65**	**0.6181**	0.3349	0.2213	36.11	6.8212	61.09	0.5454	0.6071
	SinSR-s1	24.41	0.6018	0.3240	0.2066	35.57	6.0159	62.82	0.5386	0.6471
	OSEDiff-s1	23.72	0.6108	0.2941	0.1976	26.32	4.7097	67.97	0.6148	0.6683
	FaithDiff-s20	22.40	0.5844	**0.2564**	**0.1645**	32.98	**3.9147**	72.27	**0.7082**	0.6597
	XPSR-s50	21.32	0.5506	0.3136	0.1911	35.65	4.9449	**72.97**	0.6775	**0.7082**
	PiSA-SR-1s	23.87	0.6058	0.2823	0.1934	**25.09**	4.5563	69.67	0.6375	0.6928
	PixelSR-s50 (Ours)	**24.44**	**0.6494**	**0.2485**	**0.1632**	30.84	4.3880	**73.04**	**0.6873**	0.6674
DrealSR	StableSR-s200	28.03	0.7536	0.3284	0.2269	148.98	6.5239	58.51	0.5601	0.6356
	DiffBIR-s50	26.71	0.6571	0.4557	0.2748	166.79	6.3124	61.07	0.5930	0.6395
	SeeSR-s50	28.17	0.7691	0.3189	0.2315	147.39	6.3967	64.93	0.6042	0.6804
	PASD-s20	27.36	0.7073	0.3760	0.2531	156.13	**5.5474**	64.87	0.6169	0.6808
	ResShift-s15	**28.46**	0.7673	0.4006	0.2656	172.26	8.1249	50.60	0.4586	0.5342
	SinSR-s1	28.36	0.7515	0.3665	0.2485	170.57	6.9907	55.33	0.4884	0.6383
	OSEDiff-s1	27.92	**0.7835**	**0.2968**	**0.2165**	**135.30**	6.4902	64.65	0.5899	**0.6963**
	FaithDiff-s20	27.40	0.7144	0.3511	0.2389	152.56	6.1242	66.30	**0.6388**	0.6291
	XPSR-s50	26.47	0.7276	0.3758	0.2621	164.16	7.1603	67.39	**0.6245**	**0.7427**
	PiSA-SR-1s	28.32	**0.7804**	**0.2960**	0.2169	**130.48**	6.1803	**66.11**	0.6161	0.6918
	PixelSR-s50 (Ours)	**28.73**	0.7776	0.3188	**0.2104**	160.54	**5.8456**	**64.99**	0.5891	0.6283
RealSR	StableSR-s200	24.70	0.7085	0.3018	0.2288	128.51	5.9122	65.78	0.6221	0.6178
	DiffBIR-s50	24.75	0.6567	0.3636	0.2312	128.99	5.5346	64.98	0.6246	0.6463
	SeeSR-s50	25.18	0.7216	0.3009	0.2223	125.55	5.4081	69.77	0.6442	0.6612
	PASD-s20	25.21	0.6798	0.3380	0.2260	124.29	5.4137	68.75	0.6487	0.6620
	ResShift-s15	**26.31**	**0.7421**	0.3460	0.2498	141.71	7.2635	58.43	0.5285	0.5444
	SinSR-s1	26.28	0.7347	0.3188	0.2353	135.93	6.2872	60.80	0.5385	0.6122
	OSEDiff-s1	25.15	0.7341	0.2921	0.2128	**123.49**	5.6476	69.09	0.6326	0.6693
	FaithDiff-s20	25.31	0.7083	0.2879	0.2105	**108.67**	**5.3367**	68.46	**0.6703**	0.6190
	XPSR-s50	23.72	0.6754	0.3573	0.2557	143.61	6.0293	**71.31**	**0.6638**	**0.7295**
	PiSA-SR-1s	25.50	0.7418	**0.2672**	**0.2044**	124.18	5.5054	**70.14**	0.6551	**0.6928**
	PixelSR-s50 (Ours)	**26.39**	**0.7448**	**0.2653**	**0.1980**	130.68	**5.1550**	67.98	0.5928	0.5647

perceptual realism and overall visual quality of the restored outputs; 3) Across all datasets, PixelSR demonstrates the existence of a critical sampling step threshold beyond which restoration performance either reaches its optimal point or gradually stabilizes. This finding highlights the trade-off between pixel-wise accuracy and perceptual quality, while also providing practical guidance for selecting an appropriate sampling configuration to balance restoration efficiency and output quality.

4.4 Ablation Experiment

We conduct ablation experiments on the loss function used, and the results are shown in Table 3. The experimental results from Table 3 lead to the following observations: 1) The complete loss configuration achieves the best performance across all full-reference image quality metrics on the three benchmark datasets, demonstrating that PixelSR effectively balances restoration fidelity and perceptual quality under the joint supervision of the three complementary loss functions; 2) In the absence of velocity field prediction constraints, PixelSR attains strong performance on no-reference quality assessment metrics; however, it also exhibits a tendency for the restored outputs to deviate from the target image distribution, indicating reduced restoration faithfulness; 3) Regarding the respective contributions of the LPIPS and DINO losses, the LPIPS loss provides more pronounced improvements in perceptual quality metrics such as MUSIQ, MANIQA, and CLIPIQA, whereas the DINO loss yields superior performance on several other evaluation metrics. These complementary findings further validate the effectiveness and necessity of incorporating dual-stream perceptual supervision within the proposed framework.

5 Broader Impacts and Limitations

This paper introduces PixelSR for the first time, which directly predicts clean images via pixel flow, providing a novel SR paradigm for real-world image restoration that balances both fidelity and perceptual quality. In the pixel space, we define the SR task directly as image prediction, thereby enhancing the model's restoration performance and ensuring deterministic outputs. Furthermore, the concept of pixel flow can be transferred to other related tasks such as image restoration, low-light

Table 2: Quantitative results of PixelSR at different sampling steps. The best and second best results of each metric are highlighted in **red** and **blue**, respectively.

Datasets	Sampling Step	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑
DIV2K-Val	1	**25.31**	**0.6850**	0.3356	0.2364	33.11	5.8978	48.16	0.3635	0.3151
	10	24.03	0.6422	**0.2494**	0.1810	**30.83**	4.4977	70.18	0.6320	0.5686
	25	23.41	0.6253	0.2538	0.1776	**29.31**	**4.4377**	72.09	0.6529	0.5913
	50	**24.44**	**0.6494**	**0.2485**	**0.1632**	30.84	**4.3880**	**73.04**	**0.6873**	**0.6674**
	100	23.14	0.6186	0.2577	**0.1772**	32.75	4.4476	**72.39**	**0.6560**	**0.5938**
DrealSR	1	**30.14**	**0.8398**	**0.2873**	**0.2026**	**153.97**	8.6616	46.04	0.3449	0.3336
	10	28.75	**0.7931**	**0.2921**	0.2075	**149.40**	**5.4929**	**65.90**	**0.6084**	**0.6448**
	25	**28.96**	0.7781	0.3002	0.2175	158.70	5.8449	64.98	0.5849	0.6234
	50	28.73	0.7776	0.3188	0.2104	160.54	5.8456	64.99	**0.5891**	0.6283
	100	28.60	0.7735	0.3125	**0.2017**	161.12	**5.7508**	**65.56**	0.5810	**0.6308**
RealSR	1	**27.60**	**0.7913**	0.2727	0.2138	137.30	7.3024	53.15	0.3970	0.3296
	10	**26.73**	**0.7611**	**0.2557**	**0.1981**	**131.05**	5.4910	65.89	0.5724	0.5358
	25	26.21	0.7483	0.2662	0.2001	131.53	5.2690	67.87	0.5911	**0.5728**
	50	26.39	0.7448	**0.2653**	**0.1980**	**130.68**	**5.1550**	**67.98**	**0.5928**	0.5647
	100	26.01	0.7432	0.2706	0.2012	132.33	**5.2634**	**68.14**	**0.5945**	**0.5756**

Table 3: Ablation experiment results on the Real-ISR testset with ground truth. The best and second best results of each metric are highlighted in **red** and **blue**, respectively.

Datasets	Loss Function	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑
DIV2K-Val	$w/o\ \mathcal{L}_{\text{FM}}$	21.48	0.5708	0.2925	0.1770	44.15	**4.0976**	72.39	0.6752	**0.6491**
	$w/o\ \mathcal{L}_{\text{LPIPS}}$	**23.06**	**0.6101**	**0.2577**	0.1739	36.44	4.4238	**72.59**	0.6665	0.5992
	$w/o\ \mathcal{L}_{\text{DINO}}$	22.63	0.5984	0.2581	**0.1706**	**35.67**	4.5318	**73.04**	**0.6824**	0.6094
	All	**24.44**	**0.6494**	**0.2485**	**0.1632**	**30.84**	**4.3880**	**73.04**	**0.6873**	**0.6674**
DrealSR	$w/o\ \mathcal{L}_{\text{FM}}$	25.27	0.6553	0.4216	0.2713	185.03	**5.1316**	**66.95**	**0.6181**	**0.6392**
	$w/o\ \mathcal{L}_{\text{LPIPS}}$	**27.29**	**0.7258**	**0.3564**	**0.2338**	172.16	**5.4339**	65.93	0.5989	0.6032
	$w/o\ \mathcal{L}_{\text{DINO}}$	26.60	0.7097	0.3754	0.2494	**169.64**	5.8772	**66.89**	**0.6102**	**0.6372**
	All	**28.73**	**0.7776**	**0.3188**	**0.2104**	**160.54**	5.8456	64.99	0.5891	0.6283
RealSR	$w/o\ \mathcal{L}_{\text{FM}}$	24.18	0.6723	0.3240	0.2353	140.90	**4.8337**	**69.37**	**0.6238**	**0.6153**
	$w/o\ \mathcal{L}_{\text{LPIPS}}$	**25.61**	**0.7251**	**0.2837**	**0.2034**	**131.74**	**5.0496**	68.65	0.6132	0.5829
	$w/o\ \mathcal{L}_{\text{DINO}}$	24.78	0.7037	0.2978	0.2152	135.05	5.3888	**69.85**	**0.6254**	**0.6020**
	All	**26.39**	**0.7448**	**0.2653**	**0.1980**	**130.68**	5.1550	67.98	0.5928	0.5647

enhancement Ciubotariu et al. (2026), and vintage photo restoration Umamaheshwari et al. (2026). In particular, this method also provides valuable insights for the design of video SR algorithms Du et al. (2025); Kong et al. (2025).

Although the proposed method demonstrates strong restoration performance across extensive experiments, several avenues remain for further improvement: 1) Under single-step restoration settings, PixelSR achieves high restoration fidelity but exhibits comparatively limited perceptual quality. To further enhance its single-step restoration capability, incorporating textual guidance or semantic priors may provide additional contextual information, thereby improving the perceptual realism and visual quality of the restored outputs; 2) The current implementation adopts the JiT-H/32 Li & He (2025) backbone, which involves substantial model complexity and computational cost due to its large parameter scale. Consequently, future work should explore the development of more lightweight prediction architectures or efficient restoration strategies to better balance restoration performance with inference speed and computational efficiency.

6 Conclusion

In this paper, we revisited the fidelity-perception trade-off in real-world image SR from a latent-free perspective. We identified two key limitations of existing diffusion- and flow matching-based SR methods: their sensitivity to noise and sampling stochasticity, and the accumulation of residual errors caused by per-step noise prediction under complex degradations. To address these issues, we proposed PixelSR, a clean-data-centric pixel-flow framework that directly predicts clean high-resolution images and derives pixel-space flow for iterative refinement in image space. By anchoring the restoration trajectory to the high-quality image manifold, PixelSR reduces error accumulation and better preserves identity, structure, and low-level fidelity. Moreover, we showed that varying the number of sampling steps provides a simple and effective mechanism to navigate the fidelity–perception trade-off. Extensive experiments on synthetic and real-world SR benchmarks demonstrate that PixelSR achieves a better balance between faithful reconstruction and perceptual realism than strong diffusion-based baselines.

AI USE STATEMENT

Large Language Models (LLMs) are utilized to assist in the writing and polishing of this manuscript. Specifically, we employ an LLM to refine the language, improve readability, and ensure clarity across various sections of the paper. The model assists with tasks such as sentence rephrasing, grammar checking, enhancing the overall flow of the text, and maintaining consistency in terminology and style throughout the manuscript.

It is important to note that the LLM is not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses are developed and conducted solely by the authors. The contributions of the LLM are strictly limited to improving the linguistic quality of the paper, with no involvement in the scientific content, data collection, or data interpretation.

The authors take full responsibility for the entire content of the manuscript, including any text that is generated or polished by the LLM. We ensure that all LLM-assisted text adheres to ethical guidelines, maintains academic integrity, and does not contribute to plagiarism or any form of scientific misconduct. Furthermore, we confirm that the use of LLMs in this work is transparently disclosed and complies with the ethical standards of our research community.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model, 2019.

Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28188–28197, 2025.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22367–22377, 2023.

George Ciubotariu, Abdur Rehman, Fayaz Ali Dharejo, Rizwan Ali Naqvi, Marcos V Conde, Radu Timofte, Zhi Jin, Hongjun Wu, Wenjian Zhang, Chang Ye, et al. Low light image enhancement challenge at ntire 2026. *arXiv preprint arXiv:2604.17669*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *TPAMI*, 2020.

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

Shian Du, Menghan Xia, Chang Liu, Xintao Wang, Jing Wang, Pengfei Wan, Di Zhang, and Xiangyang Ji. Patchvsr: Breaking video diffusion resolution limits with patch-wise video super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17799–17809, 2025.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 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. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021.

Zhe Kong, Le Li, Yong Zhang, Feng Gao, Shaoshu Yang, Tao Wang, Kaihao Zhang, Zhuoliang Kang, Xiaoming Wei, Guanying Chen, et al. Dam-vsr: Disentanglement of appearance and motion for video super-resolution. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Jianze Li, Jiezhang Cao, Yong Guo, Wenbo Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation. In *International Conference on Machine Learning*, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *CVPRW*, 2023.

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

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144, 2017.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European Conference on Computer Vision*, 2024a.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Ben Fei, Bo Dai, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. In *ECCV*, 2024b.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

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

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE SPL*, 20(3):209–212, 2012.

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

Yunpeng Qu, Kun Yuan, Kai Zhao, Qizhi Xie, Jinhua Hao, Ming Sun, and Chao Zhou. Xpsr: Cross-modal priors for diffusion-based image super-resolution. In *European Conference on Computer Vision*, pp. 285–303. Springer, 2024.

Bastian E Rapp. Numerical methods for solving differential equations. *Microfluidics: Modelling, Mechanics and Mathematics, Micro and Nano Technologies*, pp. 549–593, 2017.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022a.

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

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. 2025.

KM Umamaheshwari, Bhuvan Beera, Himal Pandey, and Sowjanya Dandapani. Deep latent space translation for restoration and enhancement of vintage photographs. In *Intelligent and Sustainable Systems*, pp. 360–363. CRC Press, 2026.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 2024a.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021.

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

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *European conference on computer vision*, pp. 101–117. Springer, 2020.

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

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024b.

Jiaqi Xu, Wenbo Li, Haoze Sun, Fan Li, Zhixin Wang, Long Peng, Jingjing Ren, Haoran Yang, Xiaowei Hu, Renjing Pei, et al. Fast image super-resolution via consistency rectified flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, 2022.

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

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. 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*, 2024.

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

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018a.

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

Hongyang Zhou, Xiaobin Zhu, Liuling Chen, Junyi He, Jingyan Qin, Xu-Cheng Yin, and xiaoxing Zhang. Unsupervised real-world super-resolution via rectified flow degradation modeling. *arXiv preprint arXiv:2508.07214*, 2025a.

Yingjie Zhou, Jiezhang Cao, Zicheng Zhang, Farong Wen, Yanwei Jiang, Jun Jia, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Q-agent: quality-driven chain-of-thought image restoration agent through robust multimodal large language model. *arXiv preprint arXiv:2504.07148*, 2025b.