SINGLE IMAGE SUPER-RESOLUTION VIA BIVARIATE À TROUS WAVELET DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

The effectiveness of super-resolution (SR) models hinges on their ability to recover high-frequency structure without introducing artifacts. Diffusion-based approaches have recently advanced the state of the art in SR. However, most diffusion-based SR pipelines operate purely in the spatial domain, which may yield high-frequency details that are not well supported by the underlying low-resolution evidence. On the other hand, unlike supervised SR models that may inject dataset-specific textures, single-image SR relies primarily on internal image statistics and can therefore be less prone to dataset-driven hallucinations; nevertheless, ambiguity in the LR observation can still lead to inconsistent high-frequency details. To tackle this problem, we introduce BATDiff, an unsupervised Bivariate À trous Wavelet Diffusion model designed to provide structured cross-scale guidance during the generative process. BATDiff employs an à Trous wavelet transform that constructs an undecimated multiscale representation, progressively revealing high-frequency details while preserving the full spatial resolution. As the core inference mechanism, BATDiff includes a bivariate cross-scale module that models parent–child dependencies between adjacent scales. It improves high-frequency coherence and reduces mismatch artifacts in diffusion-based SR. Experiments on standard benchmarks demonstrate that BATDiff produces sharper and more structurally consistent reconstructions than existing diffusion and non-diffusion baselines, achieving improvements in fidelity and perceptual quality.

1 INTRODUCTION

Single-image super-resolution (SISR) addresses the inverse problem of recovering a high-resolution (HR) image from a single low-resolution (LR) observation, where substantial high-frequency information is irreversibly lost during image acquisition (Wang et al., 2021). Despite the remarkable progress achieved by deep learning-based approaches, the faithful reconstruction of fine-scale structures remains a persistent challenge, particularly in real-world scenarios characterized by complex degradations such as unknown blur kernels, noise contamination, and non-ideal downsampling processes (Zhang et al., 2021; Liang et al., 2022; Lepcha et al., 2023). Existing methods often struggle to balance distortion minimization with perceptual fidelity, frequently resulting in either oversmoothed outputs or visually sharp yet structurally inconsistent artifacts (Blau & Michaeli, 2018). Consequently, developing SR frameworks capable of restoring high-frequency details while preserving global coherence continues to be a central objective in the field.

Recent advances in generative models have transformed SR from deterministic regression into a probabilistic reconstruction problem (Zhai et al., 2023). Generative approaches learn rich priors over natural image statistics and use them to model the conditional distribution of HR images given LR observations. This formulation supports the recovery of fine-scale structures that remain ill-constrained under LR inputs. Among these approaches, diffusion probabilistic models have gained increasing attention due to their stable training dynamics and strong performance in image generation and restoration tasks (Dhariwal & Nichol, 2021; Song et al., 2020; Karras et al., 2022). Diffusion-based methods cast SR within a conditional denoising framework (Li et al., 2025b), where successive refinement steps progressively approximate the target distribution and facilitate structured high-frequency reconstruction.

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.

Despite their strong empirical performance, prevailing diffusion-based SISR formulations typically perform reconstruction at a single effective resolution scale, whether defined in pixel space (Zhang et al., 2024) or a latent/feature space (Wang et al., 2024a). From a multiscale perspective, an image can be represented through wavelet decompositions as a collection of subband coefficients at progressively finer frequency levels defined on a common spatial grid (Moser et al., 2024). Within this representation, classical statistical analyses have demonstrated that high-frequency coefficients are conditionally dependent on their corresponding parent coefficients at the adjacent coarser level (Portilla & Simoncelli, 2000; Achim & Kuruoglu, 2005; Sendur & Selesnick, 2002; Şendur & Selesnick, 2002). These parent–child relationships capture intrinsic cross-scale structure in natural images. However, such cross-scale statistical dependencies are not explicitly modeled in most existing diffusion-based SISR formulations. Consequently, high-frequency details may be generated without structural alignment to the underlying coarse information inferred from the LR observation. This limitation may become particularly evident in challenging reconstruction scenarios, where stable texture generation requires consistent interaction between low- and high-frequency components (Lin et al., 2024).

In this paper, we propose BATDiff, an unsupervised (no external paired LR–HR supervision) diffusion-based framework for SISR that incorporates à trous wavelet decomposition to construct a shift-invariant multiscale representation (Yu & Koltun, 2015). Owing to its undecimated formulation, the à trous transform preserves full spatial resolution at each decomposition level, maintaining precise spatial alignment across frequency subbands. Building upon this aligned multiscale structure, BATDiff introduces a bivariate cross-scale reconstruction strategy within the reverse diffusion process. At each reverse step, the reconstruction at a given frequency band is conditioned not only on its current noisy estimate but also on the estimate from the adjacent coarser band at the same diffusion timestep. This explicit parent–child conditioning embeds structured cross-scale statistical dependencies directly into diffusion inference and promotes consistency between coarse structural components and generated fine details.

Our contributions can be summarized as follows:

- We introduce a bivariate cross-scale conditioning mechanism for SISR that captures statistical dependencies across scales during reverse diffusion.
- We instantiate this mechanism in BATDiff using à trous wavelet transform, which provides an undecimated, spatially aligned multiscale representation that enables stable cross-scale conditioning during reconstruction.
- We develop an internal-learning formulation for BATDiff that requires no paired LR–HR training data and enforces consistency between the reconstructed image and the observed LR input.

We further show in Appendix C.5 that the same bivariate cross-scale framework extends directly to image denoising.

Experiments show that BATDiff outperforms state-of-the-art SISR methods across diverse SR settings. Under the $\times 4$ bicubic protocol, BATDiff achieves the best PSNR on three of four benchmarks, including 29.34 dB on DIV2K and 30.12 dB on Set14, and obtains the highest SSIM across all evaluated benchmarks. Under the more challenging blind and Real-ISR protocols, BATDiff achieves the best DISTS and FID on DIV2K, while also obtaining the best DISTS, FID, and NIQE on DRealSR, highlighting the benefit of cross-scale conditioning for structurally consistent and perceptually faithful reconstruction.

2 Related Work

Early deep-learning approaches formulated super-resolution (SR) as deterministic regression. Methods such as SRCNN and EDSR (Dong et al., 2014; Lim et al., 2017), together with later architectures including RDN and RCAN (Zhang et al., 2018b;a), were trained using pixel-wise reconstruction losses. While effective at reducing distortion, these models often produce oversmoothed textures due to their limited ability to synthesise realistic high-frequency details.

To improve perceptual quality, subsequent work introduced stronger generative priors. GAN-based methods (Ledig et al., 2017; Wang et al., 2018) enhance perceptual realism but may suffer from

training instability and hallucinated textures. More recently, transformer-based architectures such as SwinIR (Liang et al., 2021), SwinFIR (Zhang et al., 2022), and SRFormer (Zhou et al., 2023), together with degradation-aware pipelines like BSRGAN (Zhang et al., 2021), have further improved robustness under complex real-world degradations (Mei et al., 2020; Chen et al., 2021a).

Recent advances in generative modelling have also shifted attention toward diffusion-based formulations. Methods including SRDiff (Li et al., 2022), IDM (Gao et al., 2023), ACDMSR (Niu et al., 2024), ResShift (Yue et al., 2023), StableSR (Wang et al., 2024a), and related DDPM/DDIM-inspired frameworks demonstrate that iterative denoising can provide a stable pathway for generative reconstruction (Noble et al., 2026; He et al., 2026; Li et al., 2025c; Zhu et al., 2025).

In parallel, multiscale representations offer a principled framework for modelling hierarchical image structures. Wavelet and frequency-domain decompositions separate coarse and fine components while preserving spatial localisation, enabling access to cross-resolution statistics. Recent SR approaches operating in wavelet or frequency domains (Moser et al., 2024; Wang et al., 2024c;b) suggest that frequency-aware processing can regularise texture synthesis and facilitate coarse-to-fine refinement.

Despite their strong generative capacity, most learning-based SR frameworks, including diffusion-based models, rely on large-scale external datasets with paired LR–HR supervision. This reliance makes multi-scale approaches difficult to apply in practice and ties performance closely to the availability and representativeness of training data. In contrast, SISR departs from this paradigm by exploiting internal image statistics within the input image itself.

Single-Image Super-Resolution. SISR adopts an image-adaptive learning model in which the reconstruction prior is inferred directly from internal statistics of the observed image. Rather than relying on externally learned mappings, this paradigm exploits the recurrence of local patterns and self-similarity inherent in natural images.

Early formulations instantiated this idea through explicit image-specific regression models. ZSSR (Shocher et al., 2018) trained a lightweight predictor on patches extracted from the test image itself, demonstrating that internal examples can compensate for unknown degradations. Complementary approaches replaced explicit regression with implicit priors induced by network structure and test-time optimization, as in Deep Image Prior (Ulyanov et al., 2018). Extending this concept to generative modeling, SinGAN (Shaham et al., 2019) learned a multiscale internal generator by matching patch distributions across resolutions.

More recently, diffusion-based models have been trained directly on a single image and sampled in a coarse-to-fine manner (Kulikov et al., 2023; Nikankin et al., 2023), indicating that probabilistic generative priors can also be realized within an internal learning setting. Nevertheless, the LR observation remains inherently ambiguous. While internal learning can capture image-specific statistics, it does not explicitly encode structured dependencies across frequency levels, which are crucial for stabilizing high-frequency generation.

3 PRELIMINARIES

Motivated by the multiscale and cross-scale considerations, we formulate the reconstruction problem within a structured probabilistic framework.

Degradation model. We observe a low-resolution (LR) image $y \in \mathbb{R}^{h\times w}$ generated from a high-resolution (HR) image $x \in \mathbb{R}^{H\times W}$, where $H = rh$ and $W = rw$ for an upsampling factor r, via

$$y = \mathcal{D}(x) + n, \tag{1}$$

where $\mathcal{D}(\cdot)$ denotes a degradation operator (e.g., blur, downsampling) and n is additive noise. Since x is unobserved in our formulation, the LR observation is enforced through an explicit LR-consistency constraint during inference. In practice, this update uses a differentiable instantiation of $\mathcal{D}$ consistent with the assumed degradation model in the experimental setting.

DDPM recap. A denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) defines a forward noising process

$$q(x_t \mid x_0) = \mathcal{N}\big(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)I\big)\,, \tag{2}$$

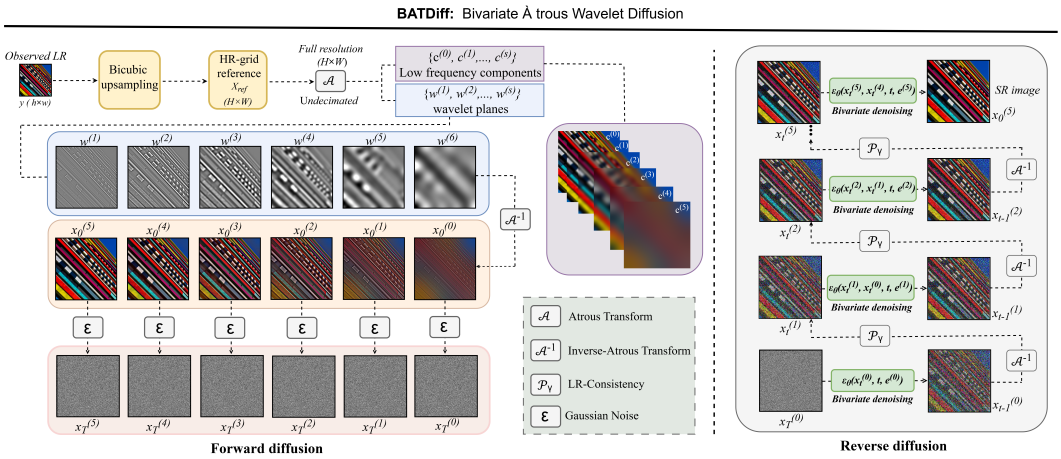

Figure 1: **Overview of BATDiff.** Given an observed LR image y, we form an HR-grid reference $x_{\text{ref}} = \mathcal{U}(y)$ via bicubic upsampling and build an undecimated à trous pyramid $\mathcal{A}$ to obtain multi-scale wavelet planes $\{w^{(s)}\}$ (S=6). We construct progressively refined clean targets $\{x_0^{(s)}\}$ (Equation 6) and train a shared DDPM by forward noising at each scale. At inference, reverse diffusion proceeds from coarse to fine with *bivariate* denoising conditioned on the time-aligned parent state $x_t^{(s-1)}$, followed by an LR-consistency correction $\mathcal{P}_y$ after each step.

with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, where $\beta_t \in (0, 1)$ denotes the noise variance at diffusion timestep t. The reverse process is parameterized as Gaussian transitions

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I\big)\,, \tag{3}$$

where $\mu_\theta(\cdot)$ is computed from a noise-prediction network ϵ_θ using the standard DDPM parameterization.

À trous multiscale representation. Let $x_{\text{ref}} \in \mathbb{R}^{H \times W}$ denote an HR-grid reference defined on the target HR. An *undecimated* à trous transform produces a sequence of smooth components $\{c^{(s)}\}_{s=0}^{S}$ and detail planes $\{w^{(s)}\}_{s=1}^{S}$ on the HR grid:

$$c^{(0)} = x_{\text{ref}}, \qquad c^{(s)} = c^{(s-1)} * k^{(s)}, \qquad w^{(s)} = c^{(s-1)} - c^{(s)}, \tag{4}$$

where $k^{(s)}$ denotes the dilated low-pass filter associated with the standard à trous cubic B-spline (B3-spline) scaling function, implemented by inserting zeros between smoothing kernel coefficients across scales. The reconstruction identity is:

$$x_{\text{ref}} = c^{(S)} + \sum_{s=1}^{S} w^{(s)}. \tag{5}$$

Within this specified setting, the proposed reconstruction model can be derived as shown in the following section.

4 Proposed Method

Figure 1 summarizes BATDiff. Given an LR observation y, we first construct an HR-grid reference $x_{\text{ref}} = \mathcal{U}(y)$ using an upsampling operator $\mathcal{U}$. An à trous hierarchy (Section 3) is then built from x_{ref} to define scale-indexed clean targets $\{x_0^{(s)}\}_{s=0}^{S}$ that progressively reveal high-frequency content while preserving spatial alignment. We train a shared DDPM noise-predictor on these targets by synthetically noising them with the DDPM forward process. During inference, given an LR observation y (Equation 1), we sample sequentially from coarse to fine scales using the bivariate reverse transitions in Section 4.2, while enforcing LR-consistency via Equation 12 after each reverse step.

4.1 DIFFUSION VARIABLES

From à trous decomposition (Equation 4–Equation 5), we form the partial reconstructions

$$x_0^{(0)} = c^{(S)}, \qquad x_0^{(s)} = c^{(S)} + \sum_{j=1}^{s} w^{(j)}, \quad s = 1, \ldots, S, \tag{6}$$

which define a coarse-to-fine refinement path on a common spatial grid. For each scale s, we run the standard DDPM forward process (Equation 2) on $x_0^{(s)}$ to obtain noisy states $\{x_t^{(s)}\}_{t=1}^{T}$. The noise schedule $\{\beta_t\}$ and the forward dynamics are shared across scales; scale enters only through the clean target $x_0^{(s)}$.

4.2 BIVARIATE REVERSE DIFFUSION

Cross-scale structure. A scale-separated construction would model each hierarchy level independently via $p_\theta(x_{t-1}^{(s)} \mid x_t^{(s)})$, which leaves the fine-scale trajectory weakly constrained by the evolving coarse structure. Parent–child interscale dependencies are well documented in wavelet-domain restoration and have been exploited via bivariate estimators that condition a coefficient on its parent to improve stability and suppress artifacts (Achim & Kuruoglu, 2005). BATDiff instead imposes explicit parent-child coupling across adjacent à trous scales by modeling, for $s = 1, \ldots, S$,

$$p_\theta\Big(x_{t-1}^{(s)} \mid x_t^{(s)}, x_t^{(s-1)}\Big), \tag{7}$$

with the coarsest scale as the standard univariate special case $p_\theta(x_{t-1}^{(0)} \mid x_t^{(0)})$. The conditioning variable $x_t^{(s-1)}$ provides a coarse-scale reference at the same diffusion timestep as $x_t^{(s)}$, helping reduce cross-scale drift during sampling. Algorithm 1 summarizes the sampling procedure. We parameterize all reverse steps using a single noise-prediction network ϵ_θ shared across scales. The scale index is provided through a learned embedding $\mathbf{e}(s)$, allowing the same network to represent scale-dependent denoising behaviour. Concretely,

$$\hat{\epsilon}_t^{(0)} = \epsilon_\theta\Big(x_t^{(0)}, t, \mathbf{e}(0)\Big), \tag{8}$$

$$\hat{\epsilon}_t^{(s)} = \epsilon_\theta\Big(x_t^{(s)}, x_t^{(s-1)}, t, \mathbf{e}(s)\Big), \qquad s = 1, \ldots, S. \tag{9}$$

This bivariate input is the defining structural element of BATDiff; the reverse diffusion parameterization remains the standard DDPM Gaussian form (Equation 3).

Reverse update Given $\hat{\epsilon}_t^{(s)}$, we compute the reverse mean using the standard DDPM noise-parameterization

$$\mu_\theta^{(s)} \;=\; \frac{1}{\sqrt{\alpha_t}}\left(x_t^{(s)} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\hat{\epsilon}_t^{(s)}\right), \tag{10}$$

and sample $x_{t-1}^{(s)} \sim \mathcal{N}(\mu_\theta^{(s)}, \sigma_t^2 I)$. The parent $x_t^{(s-1)}$ influences the update only through $\hat{\epsilon}_t^{(s)}$ in Equation 9.

LR-consistency. We enforce fidelity to the LR observation y by penalizing violations of the degradation model in Equation 1. Concretely, we use the data-consistency objective

$$\mathcal{L}_{\mathrm{lr}}(x) = \|\mathcal{D}(x) - y\|_2^2, \tag{11}$$

where $\mathcal{D}$ is implemented as a differentiable degradation operator consistent with the assumed evaluation setting (e.g., blur+downsampling). After each reverse diffusion step at scale s and timestep t (i.e., after obtaining the intermediate iterate $x_{t-1}^{(s)}$), we apply a lightweight correction

$$x_{t-1}^{(s)} \leftarrow x_{t-1}^{(s)} - \eta_t \nabla_{x_{t-1}^{(s)}} \Big\|\mathcal{D}\Big(x_{t-1}^{(s)}\Big) - y\Big\|_2^2, \tag{12}$$

with step size η_t kept constant across diffusion timesteps. Note that all $x_t^{(s)}$ live on the HR grid (undecimated à trous), hence $\mathcal{D}$ is applied consistently at every scale. When $\mathcal{D}$ admits a closed-form projection, Equation 12 can be replaced by the corresponding projection operator. This design keeps the LR observation as an explicit inference-time constraint, while the diffusion prior supplies the missing high-frequency content.

Algorithm 1: BATDiff inference

Input: LR y, scales S, steps T, noise schedule $\{\beta_t\}$, denoiser ϵ_θ, degradation operator $\mathcal{D}$, step sizes $\{\eta_t\}$
Output: HR estimate $\hat{x}$

Sample $x_T^{(0)} \sim \mathcal{N}(0, I)$
for $s = 0$ **to** S **do**
  if $s > 0$ **then**
    $x_T^{(s)} \leftarrow \hat{x}_0^{(s-1)}$
  for $t = T, T-1, \ldots, 1$ **do**
    if $s = 0$ **then**
      $\hat{\epsilon}_t^{(0)} \leftarrow \epsilon_\theta(x_t^{(0)}, t, \mathbf{e}(0))$
    else
      $\hat{\epsilon}_t^{(s)} \leftarrow \epsilon_\theta(x_t^{(s)}, x_t^{(s-1)}, t, \mathbf{e}(s))$
    $\mu_\theta^{(s)} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(x_t^{(s)} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\hat{\epsilon}_t^{(s)}\right)$
    Sample $x_{t-1}^{(s)} \sim \mathcal{N}(\mu_\theta^{(s)}, \sigma_t^2 I)$
    $x_{t-1}^{(s)} \leftarrow x_{t-1}^{(s)} - \eta_t \nabla_x \|\mathcal{D}(x_{t-1}^{(s)}) - y\|_2^2$
  $\hat{x}_0^{(s)} \leftarrow x_0^{(s)}$
return $\hat{x} \leftarrow \hat{x}_0^{(S)}$

4.3 Training objective

We optimize the standard DDPM noise-prediction objective, averaged over wavelet scales, diffusion timesteps, and Gaussian noise:

$$\mathcal{L}(\theta) = \mathbb{E}_{\substack{s\sim\mathrm{Unif}\{0,\ldots,S\} \\ t\sim\mathrm{Unif}\{1,\ldots,T\} \\ \epsilon\sim\mathcal{N}(0,I)}} \left[\left\| \epsilon - \hat{\epsilon}_t^{(s)} \right\|_2^2 \right]. \tag{13}$$

For a given scale s, the clean target $x_0^{(s)}$ is constructed deterministically from the reference image x_{ref} via Equation 6. A noisy sample is then generated using the forward diffusion process (Equation 2),

$$x_t^{(s)} = \sqrt{\bar{\alpha}_t}\, x_0^{(s)} + \sqrt{1-\bar{\alpha}_t}\, \epsilon. \tag{14}$$

The predicted noise $\hat{\epsilon}_t^{(s)}$ is given by Equation 8 for $s = 0$ and Equation 9 for $s \geq 1$. No paired LR–HR training set is used; the LR observation y is used only at inference time via the LR-consistency updates (Equation 12).

5 Experiments

Evaluation benchmarks and protocol. We evaluate BATDiff on six standard SR datasets, namely DIV2K (Timofte et al., 2017), Set5 (Bevilacqua et al., 2012), Set14 (Zeyde et al., 2012), Urban100 (Huang et al., 2015), DRealSR (Wei et al., 2020), and Manga109 (Matsui et al., 2017). Bicubic interpolation is used as the upsampling operator $\mathcal{U}$. Following standard SR evaluation practice, PSNR, SSIM, LPIPS, DISTS, FID, NIQE, MANIQA, MUSIQ and CLIPIQA are computed. Since BATDiff operates in a single image/internal learning regime, no external paired LR–HR training set is used. Instead, for each test image, scale-specific training samples are constructed from the image itself.

Compared methods. We compare BATDiff with representative methods from several SR families, GAN-based, transformers and diffusion-based approaches. Specifically, the comparison includes ZSSR (Shocher et al., 2018), ESRGAN (Wang et al., 2018), EDSR (Lim et al., 2017), BSRGAN (Zhang et al., 2021), SwinIR (Liang et al., 2021), SwinFIR (Zhang et al., 2022), SRFormer (Zhou et al., 2023), SRDiff (Li et al., 2022), DiWa (Moser et al., 2024), FDDiff (Wang et al., 2024b), ResShift (Yue et al., 2023), StableSR (Wang et al., 2024a), SinSR (Wang et al., 2024c), PASD (Yang et al., 2024) SeeSR (Wu et al., 2024), DiffBIR (Lin et al., 2024), and HAT (Chen et al., 2023b).

Implementation details. The model is optimized using Adam with an initial learning rate of 10^{-3} and a training batch size of 16. For each test image, we perform image specific optimization for 120k steps using a multi step learning rate schedule. The diffusion process is defined over 100 timesteps, and the final reconstruction is obtained using 100 reverse diffusion steps. Also, to construct the multiscale representation, we use a 6-level à trous pyramid. Unless otherwise stated, the LR-consistency

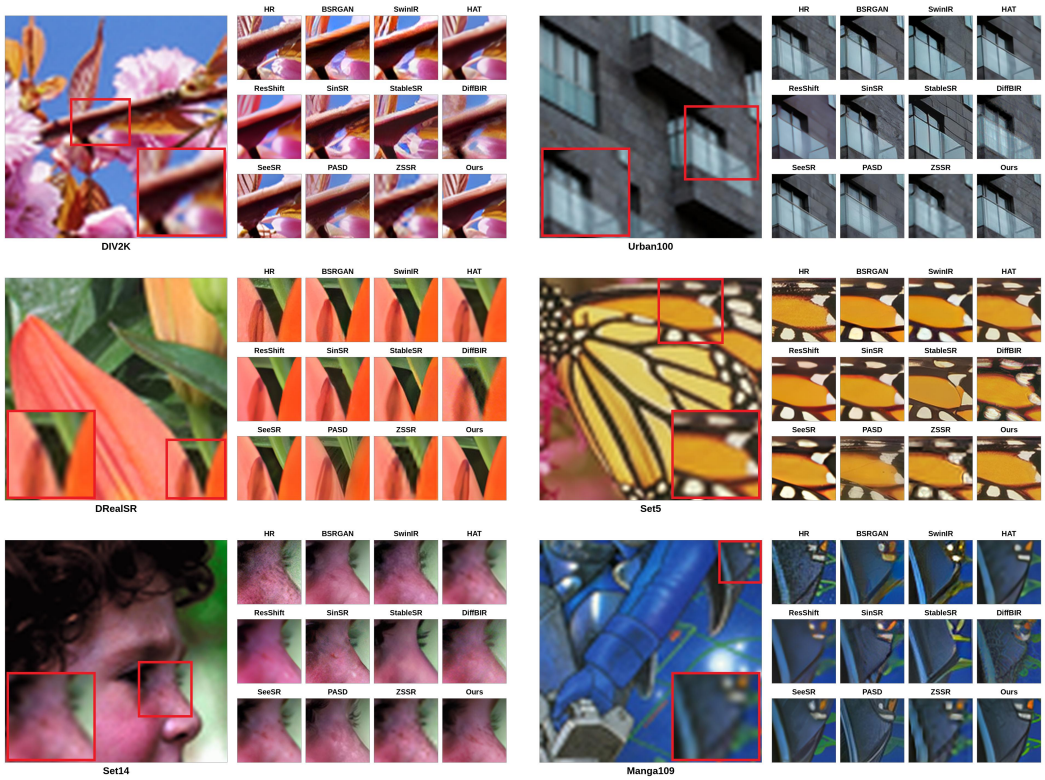

Figure 2: Qualitative comparison on standard SR benchmarks ($\times 4$).

Table 1: Quantitative comparison under the bicubic degradation protocol ($\times 4$). The Regime row marks unsupervised (U) and supervised (S) methods. The best and second best results of each metric are highlighted in **red** and blue, respectively.

DIV2K	EDSR	SRDiff	DiWa	FDDiff	SinSR	**Ours**
Regime	S	S	S	S	S	U
PSNR↑	28.98	27.41	28.09	27.53	23.12	**29.34**
SSIM↑	0.8300	0.7900	0.7800	0.8500	0.7700	**0.8948**
Set5	SwinIR	SwinFIR	SinSR	HAT	ZSSR	**Ours**
Regime	S	S	S	S	U	U
PSNR↑	32.72	**33.08**	23.91	33.04	31.13	32.89
SSIM↑	0.9021	0.9048	0.8717	0.9056	0.8796	**0.9063**
Set14	SwinIR	SRFormer	SinSR	HAT	ZSSR	**Ours**
Regime	S	S	S	S	U	U
PSNR↑	28.94	29.01	29.47	29.13	28.01	**30.12**
SSIM↑	0.7914	0.7919	0.7360	0.7973	0.7651	**0.8134**
Urban100	SwinIR	SRFormer	SinSR	HAT	ZSSR	**Ours**
Regime	S	S	S	S	U	U
PSNR↑	27.07	27.85	23.84	27.97	24.69	**28.53**
SSIM↑	0.8164	0.8338	0.5964	0.8368	0.6732	**0.8502**

weight ω and correction step size d are fixed across all datasets. All experiments are conducted on an NVIDIA B200 GPU, and the corresponding runtime characteristics are summarized in Table 3. Full implementation details, including the training/inference protocol and the shared denoiser design, are provided in Appendix A.

5.1 Quantitative Comparison

Table 1 reports quantitative comparisons at the $\times 4$ upsampling scale, where the LR input follows the standard bicubic SISR setting. BATDiff belongs to the unsupervised single image setting, since it

Table 2: Comparison under the $\times 4$ blind and Real-ISR protocol on DIV2K and DRealSR. The Regime row marks unsupervised (U) and supervised (S) methods. Best/second-best per metric are **red**/blue.

Dataset	Metric	BSRGAN	ResShift	PASD	SeeSR	DiffBIR	BATDiff
	Regime	S	S	S	S	S	U
DIV2K	PSNR↑	**21.87**	21.75	20.77	21.19	20.94	21.28
	SSIM↑	**0.5539**	0.5422	0.4958	0.5386	0.4938	0.5524
	LPIPS↓	0.4136	0.4284	0.4410	**0.3843**	0.4270	0.4175
	DISTS↓	0.2737	0.2606	0.2538	0.2257	0.2471	**0.2205**
	FID↓	64.28	55.77	40.77	31.93	40.42	**30.46**
	NIQE↓	4.7615	6.9731	4.8328	4.9275	4.7211	**4.6885**
	MANIQA↑	0.4834	0.5232	0.6049	0.6198	**0.6205**	0.6153
	MUSIQ↑	59.11	58.23	66.85	**68.33**	65.23	66.59
	CLIPIQA↑	0.5183	0.5948	0.6799	**0.6946**	0.6664	0.6876
DRealSR	PSNR↑	**28.75**	28.46	27.00	28.17	26.76	27.55
	SSIM↑	**0.8031**	0.7673	0.7084	0.7691	0.6576	0.7823
	LPIPS↓	**0.2883**	0.4006	0.3931	0.3189	0.4599	0.3118
	DISTS↓	0.2142	0.2656	0.2515	0.2315	0.2749	**0.2138**
	FID↓	155.63	172.26	159.24	147.39	166.79	**145.12**
	NIQE↓	6.5192	8.1249	**5.8595**	6.3967	6.2935	6.1326
	MANIQA↑	0.4878	0.4586	0.5850	**0.6042**	0.5923	0.5879
	MUSIQ↑	57.14	50.60	64.81	**64.93**	61.19	62.35
	CLIPIQA↑	0.4915	0.5342	0.6773	**0.6804**	0.6346	0.6583

Table 3: Runtime profile of BATDiff.

Component	Configuration / runtime
Image specific optimization	120k steps, ∼9 minutes / image
Reverse diffusion sampling	100 steps, ∼2 seconds / image

Table 4: PSNR↑ comparison with a scale of $\times 8$.

Dataset	Supervised						Unsupervised
	MetaSR	LIIF	LTE	CLIT	CiaoSR	HIIF	Ours×8
Set5	27.02	27.36	27.35	27.62	27.45	27.56	**28.74**
Set14	25.09	25.34	25.42	25.55	25.42	25.59	**27.12**
Urban100	22.75	23.14	23.17	23.33	23.34	23.36	**23.47**

adapts to the given input image and does not rely on external paired LR–HR training data. Therefore, ZSSR (Shocher et al., 2018) is the most directly comparable baseline where available. Under the bicubic setting, BATDiff achieves the best performance across all four benchmarks, including 29.34 dB PSNR on DIV2K and 30.12 dB PSNR on Set14. In Table 2, BATDiff still remains competitive under different assumptions (blind and Real-ISR). Table 4 demonstrates consistent performance at the more challenging $\times 8$ scale factor across MetaSR (Hu et al., 2019), LIIF (Chen et al., 2021b), LTE (Lee & Jin, 2022), CLIT (Chen et al., 2023a), CiaoSR (Cao et al., 2023) and HIIF (Jiang et al., 2025). In addition, since reconstruction is performed on an HR-grid reference rather than scale specific upsampling layers, the proposed framework can naturally support non-integer magnification factors without architectural modification. Additional evaluation at $\times 6$ and non-integer scale factors are provided in Appendix C.

5.2 Qualitative Comparison

Representative visual comparisons are shown in Figure 2. BATDiff tends to reconstruct cleaner edges, sharper contours, and more plausible fine structures than other approaches. Improvements are especially evident in regions containing repeated textures, thin line patterns, and high contrast boundaries. Additional qualitative comparisons, including challenging Urban100 scenes and representative failure cases, are provided in Appendix C.3.

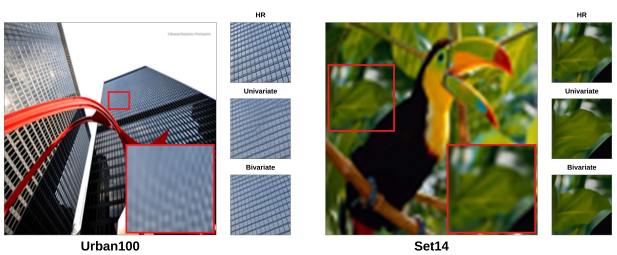

Figure 3: Effect of bivariate sampling($\times 4$).

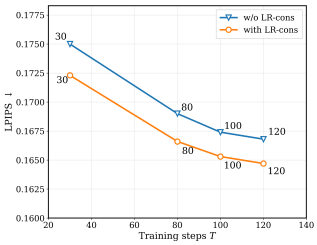

Figure 4: Effect of LR.

5.3 Ablation Study

We analyze the contribution of each component on Urban100 under the $\times 4$ bicubic setting, chosen for its strong geometric regularity and dense high-frequency textures.

Effect of the parent signal. Table 5 studies different choices for the parent representation used in the bivariate reverse process. Removing parent conditioning entirely reduces the model to an univariate variant and yields the weakest results. Replacing the time-aligned parent with the parent state at timestep $t-1$, $x_{t-1}^{(s-1)}$, yields only a modest improvement, showing that cross-scale context is useful but sensitive to temporal inconsistency. Using the final coarse estimate $x_0^{(s-1)}$ performs better, as it provides a more stable guide. However, the best performance is achieved when the denoiser is conditioned on the time-aligned parent state $x_t^{(s-1)}$. This comparison clarifies the gain depends specifically on preserving temporal alignment across adjacent scales during sampling. Figure 3 compares the univariate baseline with the proposed bivariate model on representative examples from Urban100 and Set14. In both cases, the bivariate formulation reconstructs better boundaries and visually cleaner local textures.

Effect of core components. Table 6 shows that each component contributes cumulatively, with LR-consistency, the à trous decomposition, and the bivariate reverse diffusion together yielding the largest gains in PSNR, SSIM, and LPIPS. Figure 4 further confirms that LR-consistency stabilizes the sampling trajectory, reducing perceptual error across training steps.

Sensitivity to training steps and hyperparameters. Table 7 analyzes sensitivity to the training steps T and the hyperparameters ω and d. Here, ω controls the posterior noise level during reverse diffusion, while d denotes the detail-gain coefficient in the à trous branch which controls the strength of wavelet-detail injection during multiscale reconstruction. Increasing the training budget from 30 to 120 steps consistently improves all metrics, with the best performance obtained at $T = 120$. We further observe that the best operating point is obtained at $\omega = 0.3$ and $d = 0.8$. Increasing ω beyond this value degrades both fidelity and perceptual quality, suggesting that excessively strong correction perturbs the generative trajectory. Similarly, moving d away from the selected setting leads to a gradual reduction in performance. The selected setting, $\omega = 0.3$ and $d = 0.8$, is used unchanged across other datasets.

Table 5: Ablation of parent choice.

Parent in Equation 7	PSNR↑	SSIM↑	LPIPS↓
None (univariate)	27.46	0.8377	0.1723
Misaligned $x_{t-1}^{(s-1)}$	27.92	0.8426	0.1704
Coarse final $x_0^{(s-1)}$	28.14	0.8450	0.1668
Time-aligned $x_t^{(s-1)}$	**28.53**	**0.8502**	**0.1647**

Table 6: Ablation of core components.

LR-cons.	À trous	Bivariate	PSNR↑	SSIM↑	LPIPS↓
✓	✗	✗	26.93	0.8240	0.1784
✓	✓	✗	27.46	0.8377	0.1723
✓	✓	✓	**28.53**	**0.8502**	**0.1647**

Table 7: Sensitivity to training steps and hyperparameters.

	T sweep				ω sweep				d sweep			
T	30	80	100	120	120	120	120	120	120	120	120	120
ω	0.3	0.3	0.3	0.3	0.3	0.5	1.0	2.0	0.3	0.3	0.3	0.3
d	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.5	0.8	1.0	1.5
PSNR↑	28.05	28.42	28.49	**28.53**	**28.53**	28.32	28.38	28.12	28.50	**28.53**	28.47	28.22
SSIM↑	0.8440	0.8485	0.8494	**0.8502**	**0.8502**	0.8491	0.8488	0.8450	0.8495	**0.8502**	0.8427	0.8416
LPIPS↓	0.1759	0.1697	0.1663	**0.1647**	**0.1647**	0.1654	0.1710	0.1769	0.1685	**0.1647**	0.1732	0.1761

6 Conclusion

This paper proposed BATDiff, an unsupervised diffusion based framework for single-image super-resolution (SISR) that addresses a structural limitation in most existing diffusion based SR frameworks: high-frequency details are often reconstructed without an explicit mechanism for preserving structural coherence. BATDiff addresses this limitation through a progressive multiscale reconstruction strategy, where fine details are recovered step by step under the guidance of aligned coarser scale information. In parallel, the reconstruction is constrained throughout the inference to remain consistent with the observed low resolution input, rather than relying solely on the generative prior. The experimental results on standard benchmarks showed that this formulation yields strong performance and improved structural consistency relative to the representative baselines.

AI USE DISCLOSURE

We used LLMs to aid and polish the writing of this manuscript and to assist with retrieval and discovery of related work. All central ideas, including the algorithmic design of BATDiff, the theoretical formulation, the implementation of experiments, and the analysis of results, were conceived and conducted entirely by the human authors.

REPRODUCIBILITY STATEMENT

To facilitate the reproducibility of our work, we provide detailed descriptions of each module of BATDiff in the main text, including the à trous multiscale representation (Section 3), the bivariate reverse diffusion formulation (Section 4.2), and the complete inference procedure summarized in Algorithm 1. Comprehensive implementation details, including the optimizer, learning rate schedule, number of diffusion timesteps, number of à trous levels, and hardware configuration, are specified in Appendix A and summarized in Table 8. The LR-consistency weight ω and correction step size d, together with an extensive sensitivity analysis across training steps and hyperparameters, are reported in Section 5.3 and further examined in Appendix B, including the effect of the number of wavelet levels (Table 9), the LR-consistency step size (Table 10), and the number of reverse diffusion steps (Table 12).

All evaluation benchmarks (DIV2K, Set5, Set14, Urban100, DRealSR, and Manga109) are publicly available, and the evaluation protocol, including all reported metrics (PSNR, SSIM, LPIPS, DISTS, FID, NIQE, MANIQA, MUSIQ, and CLIPIQA), follows standard practice in the super-resolution literature. Since BATDiff operates under a single-image internal learning paradigm, no external paired LR–HR training set is required, which further simplifies reproduction: for any test image, the model can be retrained directly from the image itself following the procedure detailed in Appendix A. Additional results supporting the generality of our approach, including its extension to Gaussian denoising, are provided in Appendix C.5. We believe that the detailed descriptions provided throughout the main text and appendix will facilitate the reproduction of our results.

ETHICS STATEMENT

This work presents a methodological contribution to single-image super-resolution and does not involve human subjects, personally identifiable information, or sensitive attributes. All experiments are conducted on publicly available, widely used super-resolution benchmarks (DIV2K, Set5, Set14, Urban100, DRealSR, and Manga109) and a standard denoising benchmark (Kodak24), none of which contain personal identifying information. To the best of our knowledge, this study does not raise ethical concerns, as it introduces no foreseeable risks of misuse and no apparent biases toward particular demographic groups. The intended applications of this work, such as improving image quality in resource-constrained or supervision-limited settings, are aligned with the broader goals of responsible and beneficial machine learning research.

REFERENCES

Alin Achim and Ercan E. Kuruoglu. Image denoising using bivariate α-stable distributions in the complex wavelet domain. *IEEE Sign. Process. Letters*, 12(1):17–20, 2005.

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, 2012.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, pp. 6228–6237, 2018.

Jiezhang Cao, Qilong Wang, Yibo Xian, Yuchao Li, Bingbing Ni, Zhijie Pi, and Luc Van Gool. Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution. In *CVPR*, pp. 1796–1807, 2023.

Hanting Chen, Yunhe Wang, Tao Guo, Chang Xu, Yong Deng, Zhen Liu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, pp. 12299–12310, 2021a.

Hsuan-Wei Chen, Yu-Sheng Xu, Ming-Fang Hong, Yi-Min Tsai, Hsin-Kai Kuo, and Chia-Yu Lee. Cascaded local implicit transformer for arbitrary-scale super-resolution. In *CVPR*, pp. 18257–18267, 2023a.

Junyang Chen, Jinshan Pan, and Jiangxin Dong. FaithDiff: Unleashing diffusion priors for faithful image super-resolution. In *CVPR*, pp. 28188–28197, 2025.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, pp. 22367–22377, 2023b.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *CVPR*, pp. 8628–8638, 2021b.

L. Şendur and I. W. Selesnick. Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. *IEEE Transactions on Signal Processing*, 50(11):2744–2756, 2002.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, volume 34, pp. 8780–8794, 2021.

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

Rich Franzen. Kodak lossless true color image suite. `http://r0k.us/graphics/kodak/`, 1999.

Shunyu Gao, Xiaodong Liu, Bo Zeng, Sheng Xu, Yi Li, Xiang Luo, and Bin Zhang. Implicit diffusion models for continuous super-resolution. In *CVPR*, pp. 10021–10030, 2023.

Yuanbiao Gou, Haiyu Zhao, Boyun Li, Xinyan Xiao, and Xi Peng. Test-time degradation adaptation for open-set image restoration. In *ICML*, volume 235 of *PMLR*, pp. 16167–16177, 2024.

Hao He, Xinyu Zhan, Yifan Bai, Rui Lan, Lei Sun, and Xiaoqiang Chu. Texts-diff: Texts-aware diffusion model for real-world text image super-resolution, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Xintao Hu, Haoran Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. Meta-sr: A magnification-arbitrary network for super-resolution, 2019.

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pp. 5197–5206, 2015.

Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2Neighbor: Self-supervised denoising from single noisy images. In *CVPR*, pp. 14781–14790, 2021.

Yifan Jiang, Ho Man Kwan, Tian Peng, Guang Gao, Fan Zhang, Xiaoyu Zhu, and David Bull. Hiif: Hierarchical encoding based implicit image function for continuous super-resolution. In *CVPR*, pp. 2289–2299, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, volume 35, pp. 26565–26577, 2022.

Vladimir Kulikov, Shahar Yadin, Matan Kleiner, and Tomer Michaeli. SinDDM: A single image denoising diffusion model. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17920–17930. PMLR, 2023.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zhaowen Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pp. 4681–4690, 2017.

Jaejun Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *CVPR*, pp. 1929–1938, 2022.

D. C. Lepcha, B. Goyal, A. Dogra, and V. Goyal. Image super-resolution: A comprehensive review, recent trends, challenges and applications. *PR*, 91:230–260, 2023.

Haoyang Li, Yifan Yang, Mingchang Chang, Shijian Chen, Haoran Feng, Zhiqiang Xu, Chunping Tu, and Yanning Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022.

Huaqiu Li, Yong Wang, Tongwen Huang, Hailang Huang, Haoqian Wang, and Xiangxiang Chu. LD-RPS: Zero-shot unified image restoration via latent diffusion recurrent posterior sampling. In *ICCV*, pp. 13684–13694, 2025a.

J. Li, H. Wang, Y. Li, and H. Zhang. A comprehensive review of image restoration research based on diffusion models. *Mathematics*, 13(13):2079, 2025b.

Jianze Li, Junjie Cao, Yifan Guo, Wei Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation, 2025c.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pp. 1833–1844, 2021.

Jingyun Liang, Hong Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *ECCV*, pp. 574–591, 2022.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 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 *ECCV*, pp. 430–448, 2024.

Youssef Mansour and Reinhard Heckel. Zero-Shot Noise2Noise: Efficient image denoising without any data. In *CVPR*, pp. 14018–14027, 2023.

Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using Manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017.

Yiqun Mei, Yuchen Fan, Yuqian Zhou, Li Huang, Thomas S. Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *CVPR*, pp. 5690–5699, 2020.

Bernhard B. Moser, Stanislav Frolov, Federico Raue, Sebastian Palacio, and Andreas Dengel. Waving goodbye to low-res: A diffusion-wavelet approach for image super-resolution. pp. 1–8, 2024.

Yaniv Nikankin, Niv Haim, and Michal Irani. SinFusion: Training diffusion models on a single image or video. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26199–26214. PMLR, 2023.

Aobo Niu, Trung Xuan Pham, Kai Zhang, Jian Sun, Yi Zhu, Qiong Yan, and Yanning Zhang. Acdmsr: Accelerated conditional diffusion models for single image super-resolution. 70(2):492–504, 2024.

Maxence Noble, Gonzalo I. Quintana, Benjamin Aubin, and Clément Chadebec. Fast, faithful and photorealistic diffusion-based image super-resolution with enhanced flow map models, 2026.

Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *IJCV*, 40(1):49–71, 2000.

Levent Sendur and Ivan W. Selesnick. Bivariate shrinkage with local variance estimation. *IEEE Sign. Process. Letters*, 9(12):438–441, 2002.

Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. SinGAN: Learning a generative model from a single natural image. In *ICCV*, pp. 4570–4580, 2019.

Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *CVPR*, pp. 3118–3126, 2018.

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

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 114–125, 2017.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *CVPR*, pp. 9446–9454, 2018.

Jianyi Wang, Zongsheng Yue, Shiyu Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 132(12):5929–5949, 2024a.

Xiaoyu Wang, Lin Chai, and Jie Chen. Frequency-domain refinement with multiscale diffusion for super-resolution, 2024b.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCV*, 2018.

Yu Wang, Wenming Yang, Xiang Chen, Yan Wang, Lin Guo, Lap-Pui Chau, and Bihan Wen. Sinsr: Diffusion-based image super-resolution in a single step. In *CVPR*, pp. 25796–25805, 2024c.

Z. Wang, J. Chen, and S. C. Hoi. Deep learning for image super-resolution: A survey. *IEEE TPAMI*, 43(10):3365–3387, 2021.

Zejin Wang, Jiazheng Liu, Guoqing Li, and Hua Han. Blind2Unblind: Self-supervised image denoising with visible blind spots. In *CVPR*, pp. 2027–2036, 2022.

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 *ECCV*, pp. 101–117, 2020.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. In *CVPR*, pp. 25456–25467, 2024.

Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *ECCV*, pp. 74–91, 2024.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions, 2015.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, volume 36, pp. 13294–13307, 2023.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, volume 6920 of *Lecture Notes in Computer Science*, pp. 711–730. Springer, 2012.

L. Zhai, Y. Wang, S. Cui, and Y. Zhou. A comprehensive review of deep learning-based real-world image restoration. *IEEE Access*, 11:21049–21067, 2023.

Di Zhang, Fei Huang, Shuo Liu, Xiaojie Wang, and Zhi Jin. Swinfir: Revisiting swinir with fast fourier convolution and improved training for image super-resolution, 2022.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, pp. 4791–4800, 2021.

Xinyi Zhang, Qiqi Bao, Qingmin Liao, Lu Tian, Zicheng Liu, Zhongdao Wang, and Emad Barsoum. Taming diffusion prior for image super-resolution with domain shift sdes. In *NeurIPS*, volume 37, pp. 42765–42797, 2024.

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

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, pp. 2472–2481, 2018b.

Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, and Wei-Shi Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In *CVPR*, pp. 25445–25455, 2024.

Yifan Zhou, Zhen Li, Chenlin Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *ICCV*, pp. 12780–12791, 2023.

Chao Zhu, Yang Liu, Shuang Huang, and Fang Wang. Taming a diffusion model to revitalize remote sensing image super-resolution. *Remote Sensing*, 17(8):1348, 2025.

Outline of the Appendix

This appendix provides supplementary materials to support the main content of the paper. Appendix A provides additional implementation details, including the training and inference configuration and the design of the shared denoiser. Appendix B presents additional ablation studies on the number of wavelet levels, the LR-consistency step size, the shared versus scale-specific denoiser design, and the number of reverse diffusion steps. Appendix C reports additional quantitative and qualitative results, including comparison under the FaithDiff degradation protocol (Appendix C.1), evaluation at multiple scale factors, additional qualitative results on Urban100 (Appendix C.3), failure cases, and an extension of BATDiff to image denoising (Appendix C.5).

A Implementation Details

For completeness and reproducibility, this section provides additional implementation details. In particular, we describe the internal learning setting, the construction of the multiscale training data, the optimization protocol, and the inference configuration used throughout the experimental evaluation.

A.1 Training Setup

BATDiff is trained under a single-image internal learning paradigm. For each test image, the model is optimized directly on samples constructed from the image itself, rather than relying on an external paired LR–HR training set. An HR-grid reference image is first obtained via bicubic upsampling of the observed LR input, after which an undecimated à trous decomposition (Yu & Koltun, 2015) is applied in order to generate the scale-wise representations used during training.

This multiscale construction allows the model to learn reconstruction dynamics across progressively finer frequency bands while preserving spatial correspondence between scales. The complete set of implementation choices adopted across all experiments is summarized in Table 8.

A.2 Inference Details

At inference time, reconstruction proceeds sequentially from coarse to fine wavelet scales. At each scale, reverse diffusion refines the current estimate by sampling from the posterior distribution conditioned on the representation at the immediately coarser scale, ensuring that fine-scale details remain consistent with the previously reconstructed structure.

Following each reverse diffusion step, the intermediate reconstruction is further corrected with the LR-consistency update described in the main paper. This update enforces consistency with the

Table 8: Implementation details used in all experiments.

Parameter	Value
Optimizer	Adam
Initial learning rate	1×10^{-3}
Batch size	16
Training iterations	120k
Diffusion timesteps	100
Number of à trous levels	6
Upsampling initialization	Bicubic
GPU	NVIDIA B200

observed LR input while allowing the diffusion prior to recover high-frequency structures that are not explicitly present in the degraded observation. The influence of several inference-related design choices is further examined in Tables 10 and 12.

A.3 Shared Denoiser

BATDiff employs a single denoising network that is shared across all wavelet scales. Rather than training independent denoisers for different scales, the model conditions the shared network on the current scale index through a learned scale embedding, enabling the same denoiser to adapt its behaviour according to the frequency level being reconstructed.

This design substantially improves parameter efficiency and avoids the need to optimize multiple models in a scale-adaptive fashion. A direct comparison between the shared denoiser design and a scale-specific alternative is reported in Table 11.

B Ablation Studies

This section presents additional ablation experiments that complement the analyses reported in the main paper. Unless otherwise specified, all results are reported on Urban100 (Huang et al., 2015) under the $\times 4$ super-resolution setting.

B.1 Effect of the Number of Wavelet Levels

We first examine the influence of the number of à trous decomposition levels used in the multiscale hierarchy. As shown in Table 9, increasing the number of scales improves reconstruction performance up to six levels, beyond which the gain becomes marginal. We observe similar behavior for larger magnification factors ($\times 6$ and $\times 8$), and therefore adopt six wavelet levels as the default configuration throughout the experiments.

Table 9: Effect of the number of à trous levels on Urban100 ($\times 4$).

Scales	PSNR↑	SSIM↑	LPIPS↓
4	27.92	0.8421	0.1712
5	28.21	0.8465	0.1680
6	**28.53**	**0.8502**	**0.1647**
7	28.49	0.8496	0.1651

These results suggest that shallow pyramids do not expose sufficient cross-scale structure, whereas deeper hierarchies offer only limited additional benefit beyond the selected configuration.

B.2 Effect of LR-Consistency Strength

We further investigate the influence of the LR-consistency correction strength. Recall that, after each reverse diffusion step, the intermediate estimate is updated according to

$$x_{t-1}^{(s)} \leftarrow x_{t-1}^{(s)} - \eta \nabla_{x_{t-1}^{(s)}} \left\| \mathcal{D}\left(x_{t-1}^{(s)}\right) - y \right\|_2^2, \tag{15}$$

where η controls the strength of the correction toward consistency with the observed LR input. The quantitative results reported in Table 10 show that a moderate correction strength provides the most favorable balance between fidelity to the LR observation and generative flexibility.

Table 10: Effect of LR-consistency step size on Urban100 ($\times 4$).

η	PSNR↑	SSIM↑	LPIPS↓
0.1	28.12	0.8467	0.1689
0.3	**28.53**	**0.8502**	**0.1647**
0.5	28.28	0.8489	0.1668

When the correction is too weak, the reconstruction remains insufficiently constrained by the observed LR input. Conversely, an overly strong correction perturbs the generative trajectory and leads to a modest degradation in reconstruction quality.

B.3 Shared vs. Scale-Specific Denoisers

We compare the proposed shared denoiser formulation with a variant that trains an independent denoising network for each scale. As shown in Table 11, the shared design achieves slightly better reconstruction performance while being substantially more parameter efficient.

Table 11: Shared vs. scale-specific denoisers on Urban100 ($\times 4$).

Model	PSNR↑	SSIM↑	LPIPS↓
Separate network per scale	28.41	0.8485	0.1669
Shared network (ours)	**28.53**	**0.8502**	**0.1647**

This result indicates that exposing a single denoiser to multiple scales is beneficial, likely because it promotes parameter sharing across structurally related reconstruction stages.

B.4 Effect of Reverse Diffusion Steps

This experiment examines the trade-off between the number of reverse diffusion steps and reconstruction quality. The corresponding evaluation results are summarized in Table 12. As the number of reverse diffusion steps increases, reconstruction quality gradually improves. The best performance is achieved with 100 steps, suggesting that this setting provides a good balance between reconstruction accuracy and computational efficiency, while further increasing the number of steps yields only marginal gains.

Table 12: Effect of reverse diffusion steps on Urban100 ($\times 4$).

Reverse steps	PSNR↑	SSIM↑	LPIPS↓
50	28.11	0.8459	0.1698
75	28.36	0.8487	0.1672
100	**28.53**	**0.8502**	**0.1647**

C Quantitative and Qualitative Results

C.1 Comparison under the FaithDiff D-Level Degradation Protocol

Table 13 provides an additional comparison with FaithDiff (Chen et al., 2025) on DIV2K (Timofte et al., 2017) under the $\times 4$ Level-I setting of the FaithDiff D-level degradation protocol, which corresponds to a mild synthetic real-world degradation regime. Compared with FaithDiff, BATDiff obtains a slightly lower PSNR (-0.08 dB) but improves SSIM and LPIPS, indicating stronger structural and perceptual fidelity to the ground truth. On no-reference perceptual metrics, FaithDiff performs better on MUSIQ, CLIPIQA+ and MANIQA, whereas BATDiff achieves a lower NIQE, and the gaps are small in all cases.

Table 13: Comparison with FaithDiff on DIV2K under the $\times 4$ Level-I setting of the FaithDiff D-level degradation protocol. **Red** indicates the better result per metric.

Method	Fidelity			Perceptual			
	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	CLIPIQA+↑	NIQE↓	MANIQA↑
FaithDiff	**24.29**	0.6668	0.2187	**66.53**	**0.5432**	3.7755	**0.5945**
BATDiff	24.21	**0.6815**	**0.2093**	65.88	0.5237	**3.7532**	0.5924

C.2 Evaluation at Multiple Scale Factors

To further assess the flexibility of BATDiff under different super-resolution factors, we evaluate it in the $\times 6$ setting. Table 14 reports PSNR comparisons on Set5 (Bevilacqua et al., 2012), Set14 (Zeyde et al., 2012), and Urban100 (Huang et al., 2015) against representative arbitrary-scale SR approaches, including MetaSR (Hu et al., 2019), LIIF (Chen et al., 2021b), LTE (Lee & Jin, 2022), CLIT (Chen et al., 2023a), CiaoSR (Cao et al., 2023), and HIIF (Jiang et al., 2025).

Table 14: Quantitative comparison for $\times 6$ super-resolution (PSNR). **Red** and blue indicate the best and second-best results, respectively.

Dataset	MetaSR	LIIF	LTE	CLIT	CiaoSR	HIIF	Ours$\times 6$
Set5	28.74	28.93	28.93	28.95	29.10	29.17	**29.86**
Set14	26.51	26.64	26.71	26.83	26.79	26.86	**28.54**
Urban100	23.99	24.20	24.28	24.44	24.58	24.59	**24.66**

To study the behaviour of BATDiff under non-integer scaling ratios, we also consider magnification factors of $\times 3.15$ and $\times 5.20$ (Table 15). Since reconstruction is performed on an HR-grid reference rather than through scale-specific upsampling layers, these factors require no architectural modification. The results suggest that the multiscale diffusion formulation provides stable reconstruction at intermediate scale factors, while the cross-scale conditioning helps maintain structural consistency when recovering high-frequency details at larger magnifications.

Table 15: Performance of BATDiff for continuous super-resolution scales (PSNR).

Scale	Dataset	PSNR↑
$\times 3.15$	Set5	33.26
	Set14	30.41
	Urban100	28.79
$\times 5.20$	Set5	30.35
	Set14	28.96
	Urban100	25.11

C.3 Additional Results on Urban100

This section provides additional qualitative comparisons on challenging urban scenes containing repeated windows, thin structural boundaries and facade patterns. Such regions are particularly informative for assessing whether a super-resolution method can recover high-frequency detail while preserving geometric consistency.

Figure 5 presents three representative Urban100 (Huang et al., 2015) examples under the $\times 8$ setting, comparing several baseline methods (BSRGAN (Zhang et al., 2021), SwinIR (Liang et al., 2021), StableSR (Wang et al., 2024a) and SeeSR (Wu et al., 2024)), our BATDiff reconstruction, and the HR reference. The selected crops highlight structures that are typically difficult to reconstruct faithfully, including narrow window frames, repetitive brick layouts, dense line patterns, and sharp building contours.

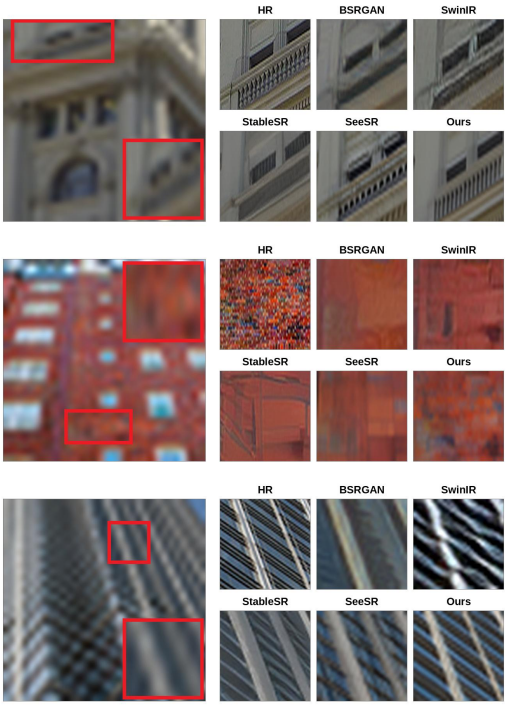

Figure 5: Qualitative comparison on Urban100 under the $\times 8$ super-resolution setting.

C.4 Failure Cases

Although BATDiff generally improves structural reconstruction and texture fidelity, certain challenging cases remain difficult to recover reliably. Figure 6 illustrates two representative examples under the $\times 8$ setting. Nevertheless, even in these challenging scenarios, the proposed method preserves the global geometry and major structural components of the scene without introducing strong hallucinated artifacts.



Figure 6: Representative failure cases under the $\times 8$ setting.

C.5 Beyond Super-Resolution: BATDiff for Image Denoising

Why BATDiff transfers to other restoration tasks. None of the three components of BATDiff is specific to super-resolution. (i) The prior is learned by internal learning from the observed image itself, so no task-specific training set is required. (ii) The à trous hierarchy and the bivariate reverse transition (Equation 7) model the cross-scale statistics of natural images, which do not depend on the degradation. (iii) The degradation enters the framework *only* through the reference operator $\mathcal{U}$ used to build x_{ref} and the operator $\mathcal{D}$ used in the consistency correction (Equation 12). Adapting BATDiff to a new inverse problem therefore amounts to redefining $(\mathcal{U}, \mathcal{D})$, while the network, the training objective (Equation 13) and the sampling procedure (Algorithm 1) remain unchanged. This transfer is also natural from a historical perspective: parent–child bivariate models were originally introduced for wavelet-domain denoising, and BATDiff embeds the same cross-scale dependency into a generative diffusion prior. Table 16 summarizes the instantiations used in this work.

Table 16: Instantiating BATDiff for gaussian denoising . Only the reference operator $\mathcal{U}$ and the consistency operator $\mathcal{D}$ change; the network, training objective and sampler are identical.

Task	Observation model	Reference $x_{\text{ref}} = \mathcal{U}(y)$
Super-resolution	$y = \mathcal{D}(x) + n$	bicubic upsampling of y
Gaussian denoising	$y = x + n,\ n \sim \mathcal{N}(0, \sigma^2 I)$	y

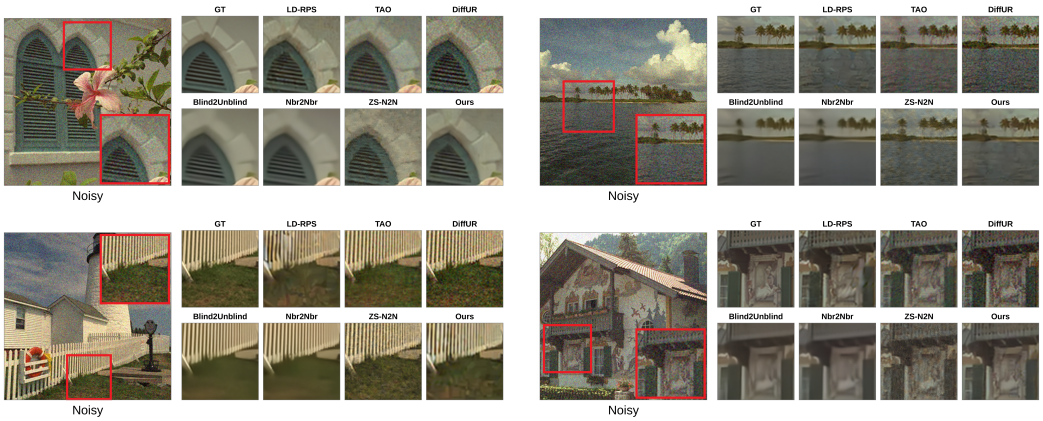

Figure 7: Qualitative comparison for Gaussian denoising on Kodak24 ($\sigma = 30$).

For denoising, $\mathcal{D}$ is the identity, so the correction in Equation 12 pulls the estimate towards the noisy observation. We therefore use a smaller step size η than in SR, so that the correction preserves the low-frequency content of y without re-injecting noise into the fine scales. All other hyperparameters follow Table 8.

Gaussian denoising. We evaluate on Kodak24 (Franzen, 1999) with additive white Gaussian noise of standard deviation $\sigma = 30$. We compare against a supervised unified restoration model (DiffUIR (Zheng et al., 2024)), task-specific self-supervised denoisers (Blind2Unblind (Wang et al., 2022), Neighbor2Neighbor (Huang et al., 2021), ZS-N2N (Mansour & Heckel, 2023)) and diffusion posterior-sampling methods (TAO (Gou et al., 2024), LD-RPS (Li et al., 2025a)). As shown in Table 17, BATDiff achieves the best PSNR and SSIM, improving PSNR by 0.65 dB over ZS-N2N, the strongest competing method that, like ours, is both dataset-free and unsupervised. BATDiff obtains the second-best LPIPS, only 0.005 behind Neighbor2Neighbor, which requires an external training set of noisy images. Figure 7 shows that BATDiff removes noise in flat regions while retaining fine textures and edges that are over-smoothed by the other methods.

Table 17: Gaussian denoising on Kodak24 ($\sigma = 30$). D: dataset-free (no external training set); U: unsupervised (no paired clean/noisy data). The best and second-best results are highlighted in **red** and blue, respectively.

Method	D	U	PSNR↑	SSIM↑	LPIPS↓
Supervised unified restoration					
DiffUIR	✗	✗	22.86	0.789	0.219
Task-specific self-supervised					
Blind2Unblind	✗	✓	29.35	0.836	0.141
Neighbor2Neighbor	✗	✓	29.09	0.821	**0.119**
ZS-N2N	✓	✓	30.14	0.862	0.132
Diffusion posterior sampling					
TAO	✓	✓	27.72	0.815	0.179
LD-RPS	✓	✓	28.64	0.841	0.175
Single-image internal learning					
BATDiff (ours)	✓	✓	**30.79**	**0.865**	0.124