

One Small Step in Latent, One Giant Leap for Pixels: Fast Latent Upscale Adapter for Your Diffusion Models

Aleksandr Razin*

razin.ad@edu.spbstu.ru

Kazantsev Danyl*

311495@niuitmo.ru

Ilya Makarov

iamakarov@hse.ru

Abstract

Diffusion models struggle to scale beyond their training resolutions, as direct high-resolution sampling is slow and costly, while post-hoc image super-resolution (ISR) introduces artifacts and additional latency by operating after decoding. We present the Latent Upscaler Adapter (LUA), a lightweight module that performs super-resolution directly on the generator’s latent code before the final VAE decoding step. LUA integrates as a drop-in component, requiring no modifications to the base model or additional diffusion stages, and enables high-resolution synthesis through a single feed-forward pass in latent space. A shared Swin-style backbone with scale-specific pixel-shuffle heads supports $\times 2$ and $\times 4$ factors and remains compatible with image-space SR baselines, achieving comparable perceptual quality with nearly 3 \times lower decoding and upscaling time (adding only +0.42 s for 1024px generation from 512px, compared to 1.87 s for pixel-space SR using the same SwinIR architecture). Furthermore, LUA shows strong generalization across the latent spaces of different VAEs, making it easy to deploy without retraining from scratch for each new decoder. Extensive experiments demonstrate that LUA closely matches the fidelity of native high-resolution generation while offering a practical and efficient path to scalable, high-fidelity image synthesis in modern diffusion pipelines.

1. Introduction

Diffusion models have transformed image synthesis, progressing from pixel-space formulations [16, 33] to *Latent Diffusion Models* (LDMs), which shift computation into compact latent representations [31]. This latent formulation underpins contemporary systems for image generation, editing, and translation [2, 11]. Despite these advances, current models are effectively constrained by the spatial resolutions seen during training (typically 512^2 or 1024^2).

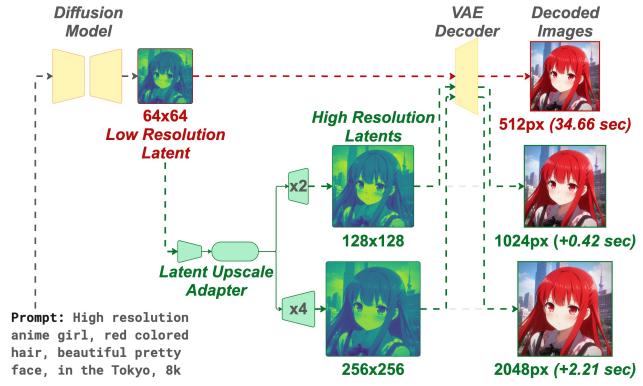


Figure 1. Our proposed lightweight *Latent Upscaler Adapter* (LUA) integrates into diffusion pipelines *without* retraining the generator/decoder and *without* an extra diffusion stage. The example uses a FLUX [2] generator: it produces a 64×64 latent for a 512 px image (red dashed path decodes directly). Our path (green dashed) upsamples the same latent to 128×128 ($\times 2$) or 256×256 ($\times 4$) and decodes once to 1024 px or 2048 px, adding only +0.42 s (1K) and +2.21 s (2K) on an NVIDIA L40S GPU. LUA outperforms multi-stage high-resolution pipelines while avoiding their extra diffusion passes, and achieves efficiency competitive with image-space SR at comparable perceptual quality, all via a single final decode.

Naïvely sampling beyond these scales often yields repetition, geometric distortions, and texture breakdown [14, 40]. While retraining or high-resolution fine-tuning can reduce such artifacts [23], these remedies demand substantial compute and data.

A pragmatic alternative is to generate at the native resolution and subsequently upsample. Two main paradigms implement this strategy. *Pixel-space super-resolution* (SR) applies an external SR model to the decoded image. This approach is simple but reconstructs fine structure purely from pixels, which encourages oversmoothing, semantic drift, and a computational cost that grows quadratically with output size [6, 38]. *Latent-space upsampling* instead enlarges the latent representation prior to decoding, reducing inference cost and better preserving semantics. Recent reference-based methods, such as DemoFusion-style

* Equal Contribution.

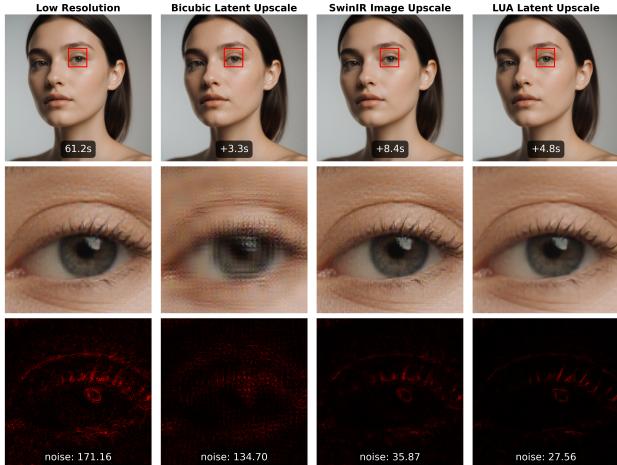


Figure 2. Upscaling FLUX outputs [2] from $1024^2 \rightarrow 2048^2$. Columns: (1) base decode, (2) bicubic *latent*, (3) SwinIR *image-space* SR, (4) LUA *latent-space* SR. Top: runtime overhead vs. (1). Middle ($8 \times$ crops): bicubic blurs/aliases; SwinIR sharpens but adds noise/textural drift; LUA preserves eyelashes and skin with stable edges. Bottom: Laplacian-variance maps (darker = less noise) with means—LUA attains the lowest residual noise and the smallest overhead via single-decode latent upscaling.

pipelines [10] and LSRNA [17], first generate a low-resolution reference, upsample it, and then run a second diffusion stage guided by the upsampled latent. Although effective, these pipelines require multi-stage inference, auxiliary noise or guidance branches, and tight coupling to specific VAEs, which increases latency and limits generality across model families.

In practice, the *upsampling step* is the principal bottleneck. Bicubic resizing or naïve latent interpolation departs from the manifold of valid latents, producing unnatural textures after decoding (Fig. 2, column 2). Conversely, pixel-space SR applied after decoding can improve fidelity but incurs higher latency and may introduce noise (Fig. 2, column 3). The core challenge is to increase latent resolution while preserving manifold geometry and high-frequency latent details that decode into realistic images—*without* invoking an additional diffusion process.

We address this challenge with the **Latent Upscaler Adapter (LUA)**, a lightweight module inserted between the generator and VAE decoder. Given a latent $z \in \mathbb{R}^{h \times w \times C}$, LUA predicts an upscaled latent $\hat{z} \in \mathbb{R}^{\alpha h \times \alpha w \times C}$ for $\alpha \in \{2, 4\}$, which is decoded once to produce the final image. Because VAE decoders typically expand spatial dimensions by stride $s=8$, a $\times 2$ latent upscaling yields a $16 \times$ increase in pixel count with no extra denoising. LUA employs a shared SwinIR-style backbone [24] with lightweight, scale-specific pixel-shuffle heads and is trained directly in the latent space with complementary latent- and pixel-domain objectives to

preserve high-frequency consistency and enable photorealistic decoding (Fig. 2, column 4).

Our study focuses on three questions:

1. Can a simple latent adapter deliver higher-quality images at lower cost than native high-resolution synthesis or pixel-space SR?
2. Can a single adapter handle multiple upscaling factors within one unified framework?
3. Can an adapter trained for one model’s VAE transfer to other generators with minimal fine-tuning?

We evaluate LUA across backbones and resolutions using FID/KID/CLIP and local-detail metrics, with ablations on architecture, objectives, and scale handling. Preserving latent microstructure via a single upscaling stage narrows the gap to native high-res synthesis while reducing complexity and latency relative to multi-stage diffusion and pixel-space SR.

Our contributions are threefold:

- We train *only* a latent upscaler adapter with a multi-stage curriculum and show that latent-space super-resolution attains quality comparable to modern high-resolution diffusion pipelines while being more efficient and less noisy than pixel-space SR.
- We design a single model that supports multiple scale factors ($\times 2, \times 4$) via a shared backbone with jointly trained, scale-specific heads—avoiding retraining from scratch.
- We demonstrate cross-VAE generalization: the same backbone operates across SD3 [11], SDXL [28], and FLUX [2] by changing only the first layer to match input channels with minimal fine-tuning.

2. Related Work

This section reviews three lines of work relevant to high-resolution image synthesis: (i) efficient diffusion-based generation at large scales, (ii) super-resolution in pixel and latent spaces, and (iii) multi-scale (discrete vs. continuous) super-resolution. For each, we outline representative methods and the limitations that motivate our approach.

Efficient high-resolution generation with diffusion models. Diffusion models in compressed latent spaces [31] enable controllable synthesis, yet sampling beyond the training scale (512^2 – 1024^2) often yields repetition, distortions, or texture loss [14, 40]. Direct high-resolution training has been demonstrated in large systems such as SDXL [28], SD3 [11], and related frameworks [23], but it demands massive datasets and compute. Inference-time strategies avoid full retraining: tiling/blending in MultiDiffusion [1] preserves locality but risks seams; receptive-field expansions via adaptive/dilated or sparse convolutions [12] and step-reduction by distillation/consistency methods [27, 34] improve speed but can reduce fidelity at extreme resolu-

tions; progressive pipelines such as HiDiffusion and ScaleCrafter [14, 40] iteratively upsample and refine; reference-based approaches like DemoFusion [10] synthesize a low-resolution latent, upsample it, and re-diffuse. These families rely on additional diffusion passes with method-specific schedules and step counts, increasing latency and coupling execution to particular resolution settings.

Super-resolution in image and latent spaces. Pixel-space SR advanced from CNN regressors (SRCNN, EDSR) [9, 25] to adversarial/perceptual (SRGAN, ESRGAN) [20, 37] and transformer models (SwinIR, HAT) [7, 24]; diffusion-based SR (SR3, SRDiff, SeeSR, StableSR, DiffBIR, SUPIR) [22, 26, 32, 36, 38, 39] further boosts perceptual fidelity. However, all denoise at the target resolution, incurring quadratic compute/memory with image size and risking semantic drift in fine textures. Latent-space SR reduces cost by upsampling before decoding, but naïve latent interpolation (e.g., bicubic/linear) departs from the generative manifold, and learned mappings such as LSRNA or the latent guidance used in DemoFusion [10, 17] still depend on a subsequent diffusion stage, yielding multi-stage, latency-heavy pipelines.

Discrete vs. continuous multi-scale SR. Discrete-factor SR (e.g., $\times 2$, $\times 4$) commonly trains separate networks or employs a shared backbone with scale-specific heads (MDSR, SwinIR) [24, 25], which is effective but resource intensive to train and store for multiple scales. Arbitrary-scale methods (LIIF, LTE, CiaoSR) [4, 8, 21] predict continuous coordinates from learned features, yet often underperform direct upscaling on fine textures where high-frequency structures dominate.

In contrast, we target a single-pass *latent* upscaler that avoids extra diffusion stages, side-steps the quadratic cost and drift of pixel-space SR, and replaces naïve latent resizing with a dedicated training curriculum. The design supports multiple scales via a shared backbone with lightweight, scale-specific heads rather than separate per-scale models or weaker arbitrary-scale decoders. This aims to deliver high-resolution fidelity with substantially lower latency and practical deployment characteristics.

3. Proposed Method

We target high-resolution synthesis without an extra diffusion stage or retraining the generator/VAE by inserting a single-pass *Latent Upscaler Adapter* (LUA) between the generator and the frozen decoder: LUA enlarges the latent, then one decode produces the final image with near-base-decode overhead. In Sec. 3.1 we formalize the upscaling operator U_α , establish its computational efficiency relative to image-space SR and multi-stage diffusion, and

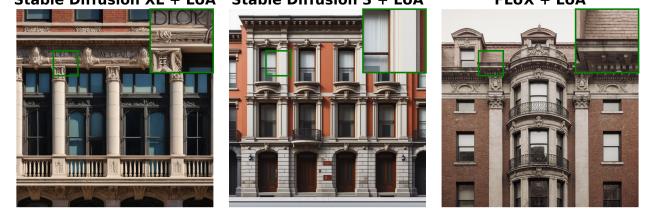


Figure 3. Cross-model $2\times$ latent upscaling with a single adapter. For SDXL [28], SD3 [11], and FLUX [2], a 128×128 latent is upscaled to 256×256 by the same LUA and decoded once by each model’s native VAE to yield 2048^2 images. SD3 and FLUX share $C=16$ latents; SDXL ($C=4$) is supported by changing only the first convolution. Insets show artifact-free detail preservation; green boxes mark $\times 8$ zooms.

show cross-model generalization (FLUX, SD3, SDXL) via changing only the first convolution layer and minimal fine-tuning. We then describe the multi-scale *architecture*—a shared backbone with scale-specific heads for $\times 2$ and $\times 4$ (Sec. 3.2)—and the *multi-stage training* strategy that preserves latent microstructure and stabilizes decoded appearance (Sec. 3.3).

3.1. Latent Upscaling

Formulation. Given text condition c and noise ϵ , a pre-trained generator G produces a latent $z \in \mathbb{R}^{h \times w \times C}$:

$$z = G(c, \epsilon). \quad (1)$$

A frozen VAE decoder D with spatial stride s (typically $s=8$) maps z to an RGB image $x \in \mathbb{R}^{(sh) \times (sw) \times 3}$:

$$x = D(z). \quad (2)$$

We introduce a deterministic latent upscaler U_α with scale factor $\alpha \in \{2, 4\}$ that maps $z \in \mathbb{R}^{h \times w \times C}$ to $\hat{z} \in \mathbb{R}^{\alpha h \times \alpha w \times C}$:

$$\hat{z} = U_\alpha(z). \quad (3)$$

A single decode yields the high-resolution image:

$$\hat{x} = D(\hat{z}). \quad (4)$$

Here, h, w are the latent spatial dimensions, C is the latent channel width, s is the decoder stride, c is the conditioning (e.g., text embedding), and ϵ is the generator noise. All generative stochasticity resides in G ; U_α is a feed-forward operator trained to remain on the latent manifold and to preserve the fine-scale statistics required for photorealistic decoding.

Computational efficiency. Pixel-space SR operates on $(sh) \times (sw)$ positions; LUA operates on $h \times w$ positions and still decodes *once*. The cost ratio scales as

$$\frac{O((sh)(sw))}{O(hw)} \approx s^2, \quad (5)$$

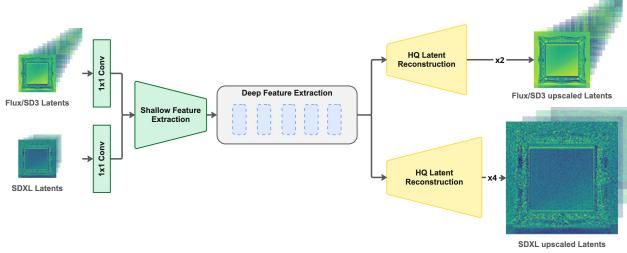


Figure 4. Architecture of the Latent Upscaler Adapter (LUA). A SwinIR-style backbone [24] is shared across scales; a 1×1 input conv adapts the VAE latent width ($C=16$ for FLUX/SD3; $C=4$ for SDXL). Scale-specific pixel-shuffle heads output $\times 2$ or $\times 4$ latents. At inference, the path selects the input adapter, runs the shared backbone, and activates the requested head. The schematic shows FLUX/SD3 $\times 2$ and SDXL $\times 4$.

so for typical $s=8$ our upscaler touches about 1/64 as many spatial elements as image-space SR. In addition, unlike progressive or reference-based pipelines, LUA does not add a second diffusion pass or any full-resolution refinement stage. Empirically, this yields lower wall-clock overhead and memory traffic while achieving comparable high-resolution fidelity (Sec. 4).

Cross-model generalization. LUA acts on latents, not backbone internals, and therefore does not require training from scratch for each VAE. The same model generalizes across FLUX, SD3, and SDXL by changing only the *first convolution layer* to match the input channel count and fine-tuning the adapter on a small set of latents from the target model. The backbone and scale heads remain unchanged. Architectural details are given in Sec. 3.2; the fine-tuning protocol and data sizes are in Sec. 4.

3.2. Architecture

We adopt a Swin Transformer restoration backbone in the spirit of SwinIR [24], which has proven effective for super-resolution in RGB space, and consistent with latent-domain adaptations such as LSRNA [17]. Windowed self-attention provides long-range context while preserving locality, matching the spatial-statistical structure of VAE latents. Given an input latent $z \in \mathbb{R}^{h \times w \times C}$, the backbone $\phi(\cdot)$ extracts features through an encoder-decoder with residual connections. Upscaling is realized with explicit SR heads—shallow convolutions followed by pixel-shuffle—that directly predict an enlarged latent; compared with implicit coordinate decoders (e.g., LIIF [8]), this better preserves high-frequency latent microstructure that decodes into sharper textures.

To support multiple scale factors without duplicating capacity, a single backbone ϕ is shared across scale-specific heads for $\times 2$ and $\times 4$ (denoted $U_{\times 2}$ and $U_{\times 4}$). Joint training

with balanced sampling encourages the backbone to learn scale-agnostic representations while each head specializes to its factor’s aliasing and artifact profile. At inference, the backbone runs once and the head corresponding to the requested scale $\alpha \in \{2, 4\}$ is applied:

$$\hat{x} = D(U_{\times \alpha}(\phi(z))), \quad \alpha \in \{2, 4\}, \quad (6)$$

where D is the frozen VAE decoder and z is the input latent with channel width C . The overall module layout is depicted in Fig. 4.

To operate across VAEs with different latent channel widths (e.g., SD3/FLUX with $C=16$ and SDXL with $C=4$), we change only the first convolution layer to match the input channels; the shared backbone and the scale heads are reused unchanged. A brief fine-tuning on a small set of latents from the target model aligns statistics and enables cross-VAE transfer without training from scratch. The fine-tuning protocol and data sizes are detailed in Sec. 4, and cross-model results are shown in Fig. 3.

3.3. Multi-stage Training Strategy

Single-domain objectives are insufficient for latent SR: optimizing *only* in latent space preserves coarse structure but yields decoded images with residual grid/blur and spurious high-frequency noise, while optimizing *only* in pixel space is unstable—gradients backpropagated through the frozen VAE decoder interact with unnormalized latents and fail to converge. To address both issues, we adopt a progressive three-stage curriculum that first secures latent structural and spectral alignment, then couples latent fidelity to decoded appearance with high-frequency emphasis, and finally refines edges in pixel space without re-diffusion. Throughout this section, z and \hat{z} denote input and upscaled latents, z_{HR} the reference HR latent (from the frozen VAE encoder), and $x=D(z)$, $\hat{x}=D(\hat{z})$, $x_{\text{HR}}=D(z_{\text{HR}})$ their decoded images via the frozen decoder D . Superscripts z and x indicate losses computed in latent and pixel domains, respectively.

Stage I — Latent-domain structural alignment. This stage learns a stable mapping $\hat{z} = U_{\alpha}(z)$ that matches high-resolution latent structure and spectra while avoiding over-smoothing. The objective is

$$\mathcal{L}_{\text{SI}} = \alpha_1 \mathcal{L}_{\text{L1}}^{\hat{z}} + \beta_1 \mathcal{L}_{\text{FFT}}^{\hat{z}}, \quad (7)$$

where $\alpha_1, \beta_1 \geq 0$ balance reconstruction and spectral alignment in latent space, with

$$\mathcal{L}_{\text{L1}}^{\hat{z}} = \|\hat{z} - z_{\text{HR}}\|_1, \quad (8)$$

$$\mathcal{L}_{\text{FFT}}^{\hat{z}} = \|\mathcal{F}(\hat{z}) - \mathcal{F}(z_{\text{HR}})\|_1. \quad (9)$$

Here $\hat{z} \in \mathbb{R}^{\alpha h \times \alpha w \times C}$ and $\mathcal{F}(\cdot)$ is the channel-wise 2D FFT magnitude [13]. $\mathcal{L}_{\text{L1}}^{\hat{z}}$ enforces element-wise correspondence; $\mathcal{L}_{\text{FFT}}^{\hat{z}}$ aligns high-frequency latent statistics to preserve microstructure [18].

Stage II — Joint latent–pixel consistency. This stage links latent fidelity to decoded appearance by adding image-domain constraints while retaining the Stage I latent terms:

$$\mathcal{L}_{\text{SII}} = \alpha_2 \mathcal{L}_{\text{L1}}^z + \beta_2 \mathcal{L}_{\text{FFT}}^z + \gamma_2 \mathcal{L}_{\text{DS}}^x + \delta_2 \mathcal{L}_{\text{HF}}^x, \quad (10)$$

with $\alpha_2, \beta_2, \gamma_2, \delta_2 \geq 0$ and

$$\mathcal{L}_{\text{DS}}^x = |\downarrow_d(\hat{x}) - \downarrow_d(x_{\text{HR}})|_1, \quad (11)$$

$$\mathcal{L}_{\text{HF}}^x = \left| (\hat{x} - \mathcal{G}_\sigma(\hat{x})) - (x_{\text{HR}} - \mathcal{G}_\sigma(x_{\text{HR}})) \right|_1. \quad (12)$$

Here $\downarrow_d(\cdot)$ denotes bicubic downsampling [29] with $d=2$ for $\times 2$ and $d=4$ for $\times 4$, and $\mathcal{G}_\sigma(\cdot)$ is a Gaussian blur with $\sigma=1.0$ (applied channel-wise). $\mathcal{L}_{\text{DS}}^x$ enforces coarse appearance consistency at a common reduced scale; $\mathcal{L}_{\text{HF}}^x$ emphasizes edges and textures by matching high-frequency residuals.

Stage III — Edge-aware image refinement. The final stage sharpens edges and suppresses residual ringing/grids in pixel space, without any additional denoising:

$$\mathcal{L}_{\text{SIII}} = \alpha_3 \mathcal{L}_{\text{L1}}^x + \beta_3 \mathcal{L}_{\text{FFT}}^x + \gamma_3 \mathcal{L}_{\text{EAGLE}}^x, \quad (13)$$

with $\alpha_3, \beta_3, \gamma_3 \geq 0$ and

$$\mathcal{L}_{\text{L1}}^x = |\hat{x} - x_{\text{HR}}|_1, \quad (14)$$

$$\mathcal{L}_{\text{FFT}}^x = |\mathcal{F}(\hat{x}) - \mathcal{F}(x_{\text{HR}})|_1, \quad (15)$$

where $\mathcal{L}_{\text{EAGLE}}^x$ is an edge-aware gradient localization loss that enforces crisp boundaries and reduces staircase artifacts [35].

Weight selection. Weights were set via grid search: ℓ_1 and FFT terms (latent and image domains) receive the largest coefficients as primary reconstruction and texture-sensitive objectives, while downsampling and edge-aware terms use smaller auxiliary regularizers. Please see the appendix for detailed grid-search settings and per-stage loss weights.

Our multi-stage latent–pixel curriculum adapts LUA to the latent domain and the frozen decoder. Stage I aligns the upscaled latent with high-resolution structure and spectra (Fig. 5, col. 2). Stage II introduces joint latent–image supervision (downsampled and high-frequency terms) to normalize latents before decoding and suppress noise (col. 3). Stage III trains in pixel space only, selectively removing redundant noise and grid artifacts while preserving necessary high-frequency detail (col. 4). The curriculum enables coherent cross-domain training and yields single-decode synthesis without an extra diffusion stage; per-stage loss weights are provided in the appendix.

4. Experiments

We evaluate LUA on OpenImages with a unified train/val setup, comparing to representative high-resolution baselines and reporting fidelity (FID/KID and patch variants),

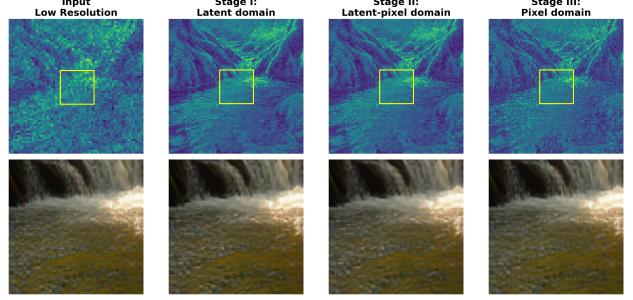


Figure 5. Effect of the three-stage curriculum on latent reconstruction and decoded appearance (FLUX backbone). The 2×4 grid shows *top*: latent feature maps (channel 10, min–max normalized); *bottom*: corresponding $8 \times$ zoomed decodes. Columns: (1) original low-resolution latent (128^2) and decode; (2–4) LUA upscaled latents to 256^2 after Stage I–III with their decodes. Yellow boxes mark the zoomed region. From (2) to (4), decodes become less noisy and more structured; Stage III concentrates high-frequency energy around details, indicating that controlled latent noise aids faithful VAE decoding.

text–image alignment (CLIP), and wall-clock latency. We also study cross-model and multi-scale generalization and ablate the curriculum and architecture. Further implementation details and extended results are provided in the appendix.

4.1. Experiment Settings

Training data. Following LSRNA [17], we use OpenImages [19]; photos with both sides ≥ 1440 px are tiled into non-overlapping 512×512 crops. HR/LR pairs are made by bicubic downsampling at the target scale ($\times 2$, $\times 4$). Crops are encoded with the FLUX VAE (stride $s=8$, $C=16$) to produce latent: for $\times 2$, $16 \times 32 \times 32 \rightarrow 16 \times 64 \times 64$; for $\times 4$, $16 \times 16 \times 16 \rightarrow 16 \times 64 \times 64$. The final set contains 3.8M pairs.

Multi-stage training. We train the three stages (Sec. 3.3) with Adam (lr 2×10^{-4} , weight decay 0), EMA 0.999, grad-clip 0.4, and MultiStepLR (milestones 62.5k/93.75k/112.5k, $\gamma=0.5$); each stage runs 125k steps. Hyperparameters follow SwinIR [24]. We keep the same initial lr across stages; Stage III uses a short warm-up. Effective batch sizes: 2,048 (Stage I) and 32 (Stages II–III). Losses match Sec. 3.3.

Scalability: multi-scale and cross-VAE. A single Swin backbone is shared by the $\times 2/\times 4$ heads ($U_{\times 2}$, $U_{\times 4}$) and trained jointly with balanced sampling (Sec. 3, Fig. 4). For cross-VAE transfer, we build 500k LR/HR latent pairs for SDXL ($C=4$) and SD3 ($C=16$), replace only the first convolution to match channels, and perform a brief Stage III-style fine-tune, reusing the backbone and heads without retraining from scratch.

Table 1. OpenImages validation. Metrics follow Sec. 4.2 (FID/pFID, KID/pKID, CLIP) and runtime (median s). H100, batch size 1. For **1K** (1024^2), SDXL samples at 512^2 then upscales to 1024^2 via LUA (latent) or SwinIR (pixel). For **2K/4K**, images are sampled at 1024^2 then upscaled to $2048^2/4096^2$. LUA achieves the lowest latency at all resolutions and the strongest fidelity at 2K/4K; best marked in **bold**.

Resolution	Method	FID \downarrow	pFID \downarrow	KID \downarrow	pKID \downarrow	CLIP \uparrow	Time (s)
1024×1024	HiDiffusion	232.55	230.39	0.0211	0.0288	0.695	1.54
	DemoFusion	195.82	193.99	0.0153	0.0229	0.725	2.04
	LSRNa–DemoFusion	194.55	192.73	0.0151	0.0228	0.734	3.09
	SDXL (Direct)	194.53	192.71	0.0151	0.0225	0.731	1.61
	SDXL + SwinIR	210.40	204.23	0.0313	0.0411	0.694	2.47
	SDXL + LUA (ours)	209.80	191.75	0.0330	0.0426	0.738	1.42
2048×2048	HiDiffusion	200.72	114.30	0.0030	0.0090	0.738	4.97
	DemoFusion	184.79	177.67	0.0030	0.0100	0.750	28.99
	LSRNa–DemoFusion	181.24	98.09	0.0019	0.0066	0.762	20.77
	SDXL (Direct)	202.87	116.57	0.0030	0.0086	0.741	7.23
	SDXL + SwinIR	183.16	100.09	0.0020	0.0077	0.757	6.29
	SDXL + LUA (ours)	180.80	97.90	0.0018	0.0065	0.764	3.52
4096×4096	HiDiffusion	233.65	95.95	0.0158	0.0214	0.698	122.62
	DemoFusion	185.36	177.89	0.0043	0.0113	0.749	225.77
	LSRNa–DemoFusion	177.95	62.07	0.0023	0.0071	0.757	91.64
	SDXL (Direct)	280.42	101.89	0.0396	0.0175	0.663	148.71
	SDXL + SwinIR	183.15	65.71	0.0018	0.0103	0.756	7.29
	SDXL + LUA (ours)	176.90	61.80	0.0015	0.0152	0.759	6.87

4.2. Evaluation Protocol

Validation data. We evaluate on 1,000 held-out high-resolution OpenImages photos (disjoint from training). Prompts are obtained via captioning and shared across methods; see the appendix for details.

Metrics. We report FID [15], KID [3], and CLIP [30]. Images are synthesized at the target resolution ($1024^2/2048^2/4096^2$). To better capture fine detail, we also report patch metrics (pFID/pKID) on random 1024×1024 crops with an identical crop policy across methods, following LSRNA and Anyres-GAN [5]. Additional metrics and computation details are provided in the appendix.

Baselines. We compare ScaleCrafter [14], HiDiffusion [40] (progressive), DemoFusion [10], LSRNA–DemoFusion [17] (reference-based re-diffusion), SDXL (Direct), and SDXL+SwinIR [24] (pixel-space SR). Samplers, steps, guidance, and prompts are held constant where applicable; see the appendix for configuration details.

Our method. *SDXL+LUA*: generate at 1024^2 , upscale in latent space, decode once. We report SDXL-anchored comparisons in Table 1 and summarize cross-model/multi-scale results (FLUX/SD3/SDXL at $\times 2/\times 4$) in Table 2.

Runtime measurement. Per-image wall-clock time is the median over N runs after 20 warm-ups on a single GPU with AMP and batch size 1. Composite pipelines include generation, upscaling, and decode.

4.3. Quantitative Results

Table 1 summarizes fidelity–efficiency trade-offs across resolutions. At 1024^2 , SDXL+LUA attains the lowest latency (1.42 s) but lags strongest single-stage baselines on fidelity (FID 209.80 vs. 194.53 for SDXL (Direct) and 194.55 for LSRNA–DemoFusion). We attribute this gap to low-resolution input latent (64×64 for 512 px, stride 8), which constrains recoverable detail in $\times 2$ setting; notably, our patch fidelity is competitive (pFID 191.75, best in row), indicating preserved local structure despite lower scores.

Beyond 1K, LUA consistently improves both quality and speed. At 2048^2 , SDXL+LUA delivers the best fidelity among single-decode pipelines while remaining the fastest: FID 180.80, pFID 97.90, KID 0.0018, CLIP 0.764 in 3.52 s, outperforming SDXL+SwinIR (183.16 / 100.09 / 0.0020 / 0.757 in 6.29 s) and substantially undercutting multi-stage re-diffusion (LSRNA–DemoFusion: FID 181.24 in 20.77 s; DemoFusion: 184.79 in 28.99 s). At 4096^2 , SDXL+LUA again sets the best single-pass fidelity (FID 176.90; pFID 61.80) with the lowest runtime (6.87 s), surpassing SDXL+SwinIR (183.15; 65.71; 7.29 s) and avoiding the quality collapse of direct high-res sampling (SDXL (Direct) FID 280.42). LSRNA–DemoFusion attains slightly lower pKID (0.0071) but is an order of magnitude slower (91.64 s), underscoring LUA’s favorable accuracy–latency frontier. These results highlight the efficacy of LUA for high-resolution synthesis.

Table 2 further shows that a single backbone generalizes across models and scales with minimal adaptation: at

Table 2. Cross-model, multi-scale results for LUA. Metrics, crop protocol, runtimes, and hardware match Table 1. We evaluate $\times 2/\times 4$ latent upscaling ($1024^2 \rightarrow 2048^2/4096^2$) on FLUX, SDXL, and SD3; best results are in **bold**.

Scale	Diffusion Model	FID \downarrow	pFID \downarrow	KID \downarrow	pKID \downarrow	CLIP \uparrow	Time (s)
$\times 2$	FLUX + LUA	180.99	100.40	0.0020	0.0079	0.773	29.829
	SDXL + LUA	183.15	101.18	0.0020	0.0087	0.753	3.52
	SD3 + LUA	184.58	103.94	0.0022	0.0083	0.768	20.292
$\times 4$	FLUX + LUA	181.06	62.30	0.0018	0.0085	0.772	31.908
	SDXL + LUA	182.42	71.92	0.0015	0.0152	0.754	6.87
	SD3 + LUA	183.34	67.25	0.0016	0.0095	0.769	21.843

$\times 2$, FLUX+LUA reaches FID 180.99 and CLIP 0.773; at $\times 4$, SDXL+LUA yields KID 0.0015 while FLUX+LUA attains the strongest pFID (62.30). These results confirm robust transfer across SDXL, SD3, and FLUX and consistent gains at both $\times 2$ and $\times 4$ without retraining the generator or adding diffusion stages. These findings highlight the adaptability of LUA across models and scales.

4.4. Qualitative Results

Figure 6 presents side-by-side comparisons at 2048^2 (top two rows) and 4096^2 (bottom two), starting from the same 1024^2 SDXL bases (identical seeds/prompts). Direct high-resolution sampling (*SDXL Direct*) exhibits canonical large-scale failures: duplicated structures and geometry drift in the crab legs and sand granules (row 1), brittle clumps in dog fur and snow particles (row 2), and warped tail-lights/reflections in the street scene (row 3). *HiDiffusion* shows similar breakdowns at 4K (row 3), indicating that training-free escalation struggles to maintain global layout at extreme scales. In contrast, SR-from-base approaches avoid these hallucinations by inheriting the cleaner 1024^2 generation.

Among SR methods, pixel-space SwinIR sharpens but introduces ringing/haloing and plastic textures: overshot specular ridges on the crab shell (row 1), halos at fur boundaries (row 2), glare around car edges (row 3), and granular noise on petals (row 4). DemoFusion+LSRINA restores rich texture but requires a second diffusion stage with much higher latency. SDXL+LUA (ours) preserves edge continuity and microstructure with fewer artifacts: crisp eyelashes and shell ridges without halos (row 1), distinct yet stable fur strands (row 2), sharp panel seams and reflections without brittleness (row 3), and coherent high-frequency petal detail (row 4). Per-image runtimes overlaid above each panel align with Table 1: LUA attains comparable or better visual quality at the lowest latency via single-decode latent upscaling. Additional examples are provided in the appendix.

4.5. Ablation Studies

Multi-stage training effectiveness. We assess the curriculum by removing individual stages and comparing against a latent ℓ_1 -only baseline. Table 3 shows that the

Table 3. Ablation on the multi-stage latent-pixel curriculum for $\times 2$ ($512 \rightarrow 1024$) and $\times 4$ ($256 \rightarrow 1024$). Best results are in **bold**.

Configuration	PSNR \uparrow	LPIPS \downarrow	PSNR \uparrow	LPIPS \downarrow
Latent ℓ_1	28.53	0.198	26.16	0.236
I+II (w/o III)	28.96	0.172	26.67	0.213
I+III (w/o II)	31.05	0.150	27.10	0.198
II+III (w/o I)	31.60	0.145	27.40	0.192
Full (I+II+III)	32.54	0.138	27.94	0.184

Table 4. Ablation of upsampling strategies for $\times 2$ ($512 \rightarrow 1024$) and $\times 4$ ($256 \rightarrow 1024$). Best results are in **bold**.

Variant	PSNR \uparrow	LPIPS \downarrow	PSNR \uparrow	LPIPS \downarrow
LIIF	29.10	0.210	26.10	0.235
Separated- $\times 2$	31.92	0.150	—	—
Separated- $\times 4$	—	—	27.71	0.189
Joint Multi-Head	32.54	0.138	27.94	0.184

full configuration (I+II+III) yields the best fidelity at both scales. At $\times 2$, the full model attains PSNR 32.54 / LPIPS 0.138 versus 28.96 / 0.172 for I+II (w/o III). At $\times 4$, II+III (w/o I) reaches 27.40 / 0.192 compared to 27.94 / 0.184 for the full model. Removing Stage III reduces edge sharpness and perceptual quality; removing Stage II weakens latent-pixel consistency and increases artifacts. A pixel-only variant fails to converge with a frozen VAE. Please, follow appendix for training settings details.

Multi-scale super-resolution adaptation. We compare three design choices per scale factor: (i) an implicit, coordinate-based upsampler (LIIF), (ii) per-scale specialist models trained only for a single factor ($\times 2$ or $\times 4$), and (iii) a joint multi-head model that shares a backbone and uses scale-specific heads. Table 4 reports within-scale results. The joint multi-head design attains the best PSNR/LPIPS at both $\times 2$ and $\times 4$, while also consolidating capacity in a single backbone, reducing storage and training overhead (see appendix for resource analysis and experiments settings).

5. Discussion

Our approach inherits limitations from its adapter nature: errors or biases in the generator’s latent are faithfully upscaled, so artifacts in the base sample can per-

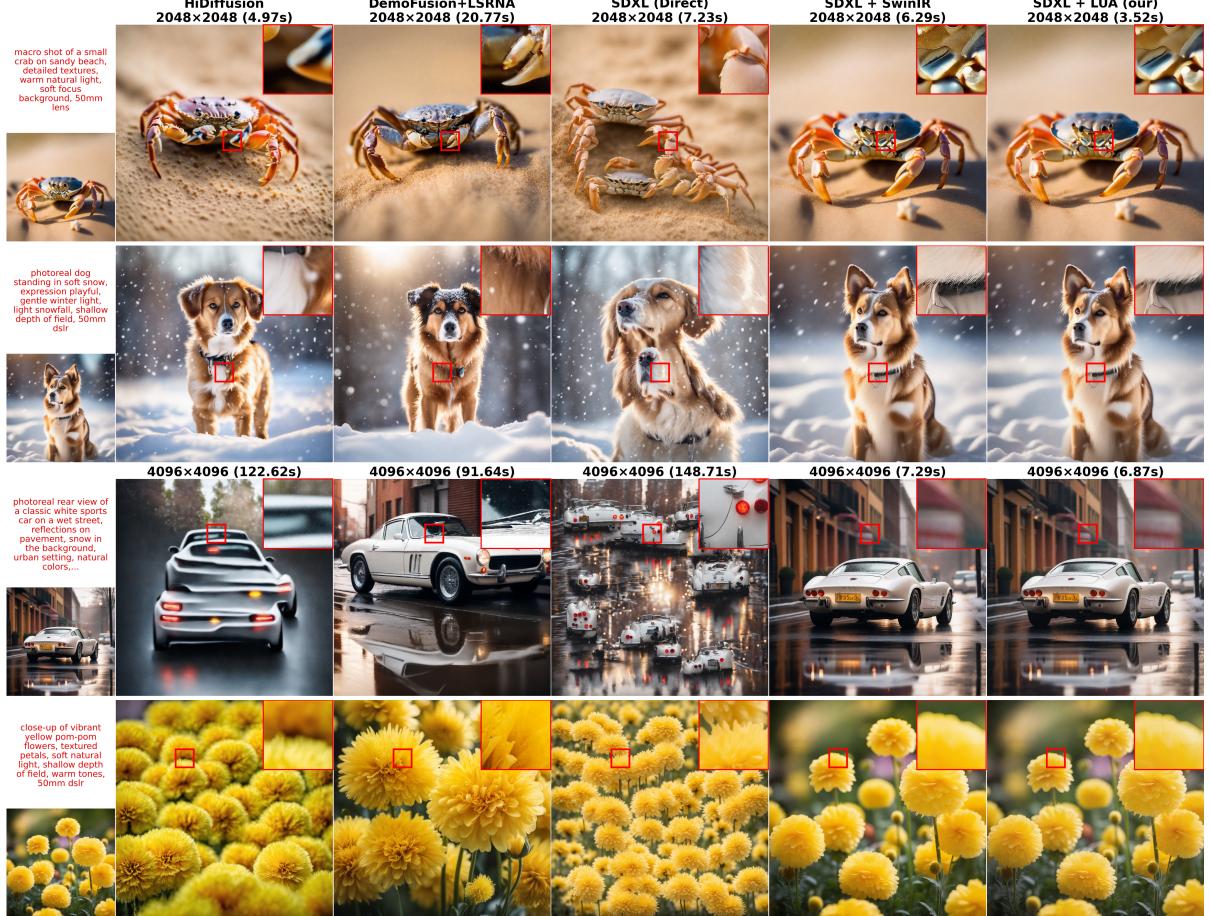


Figure 6. Qualitative comparison at 2048^2 and 4096^2 starting from the same 1024^2 SDXL base generations. Each row uses identical seeds and prompts (GPT-generated captions from OpenImages validation). Red boxes indicate $12\times$ magnified crops; titles report per-image runtime. For visual clarity we show the DemoFusion+LSRNA variant in place of plain DemoFusion. The column with *SDXL+LUA (ours)* achieves the lowest latency and produces clean, stable textures without the high-resolution artifacts (e.g., repetition, gridding) seen in direct high-res sampling, and without the sharpening noise typical of pixel-space SR.

sist at higher resolution. A promising direction is joint refinement-and-upscaling directly in latent space before decoding, e.g., lightweight consistency modules to suppress artifacts while preserving semantics, with uncertainty-aware gating to invoke refinement only when needed. Beyond text-to-image, the same mechanism can serve image-to-image tasks requiring high-resolution outputs—such as depth-to-image or semantic maps, where preserving structure during upscaling is critical. Finally, extending the adapter to video with temporal consistency (e.g., recurrent latent refinement or temporal priors) remains essential for practical high-resolution synthesis in dynamic settings.

6. Conclusion

We introduced LUA, a single-pass latent upscaler inserted between a pretrained generator and a frozen VAE decoder, and demonstrated that latent-space upscaling is train-

able via a multi-stage latent-pixel curriculum with scale-specific $\times 2 / \times 4$ heads. On OpenImages, at 2048^2 and 4096^2 , SDXL+LUA achieves state-of-the-art single-decode fidelity (FID 180.80 / 176.90; pFID 97.90 / 61.80) while remaining the fastest; at 2048^2 it runs in 3.52 s versus 7.23 s for SDXL (Direct), outperforms pixel-space SR (SDXL+SwinIR), and approaches multi-stage re-diffusion quality at a fraction of the runtime. At 1024^2 , performance trails the strongest baselines due to the 64×64 input latent constraint, though patch-level fidelity remains competitive. A single backbone transfers across SDXL, SD3, and FLUX with only a first-layer channel change and brief fine-tuning, demonstrating robust cross-VAE and multi-scale generalization without modifying the generator or adding diffusion stages. Overall, these results establish single-decode latent upscaling as a practical alternative to multi-stage high-resolution pipelines.

References

- [1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation.(2023). *URL https://arxiv.org/abs/2302.08113*, 2023. 2
- [2] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 1, 2, 3
- [3] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021. 6
- [4] Jiezhang Cao, Qin Wang, Yongqin Xian, Yawei Li, Bingbing Ni, Zhiming Pi, Kai Zhang, Yulun Zhang, Radu Timofte, and Luc Van Gool. Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution, 2023. 3
- [5] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis, 2022. 6
- [6] Weimin Chen, Yuqing Ma, Xianglong Liu, and Yi Yuan. Hierarchical generative adversarial networks for single image super-resolution. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 355–364, 2021. 1
- [7] Xiangyu Chen, Xintao Wang, Wenlong Zhang, Xiangtao Kong, Yu Qiao, Jiantao Zhou, and Chao Dong. Hat: Hybrid attention transformer for image restoration, 2025. 3
- [8] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function, 2021. 3, 4
- [9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015. 3
- [10] Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no \$. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6159–6168, 2024. 2, 3, 6
- [11] 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. 1, 2, 3
- [12] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *European conference on computer vision*, pages 396–414. Springer, 2022. 2
- [13] Dario Fuoli, Luc Van Gool, and Radu Timofte. Fourier space losses for efficient perceptual image super-resolution, 2021. 4
- [14] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 2, 3, 6
- [15] 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, 2018. 6
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1
- [17] Jinho Jeong, Sangmin Han, Jinwoo Kim, and Seon Joo Kim. Latent space super-resolution for higher-resolution image generation with diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2355–2365, 2025. 2, 3, 4, 5, 6
- [18] Xingyu Jiang, Xiuhui Zhang, Ning Gao, and Yue Deng. When fast fourier transform meets transformer for image restoration. In *European Conference on Computer Vision*, pages 381–402. Springer, 2024. 4
- [19] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 5
- [20] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network, 2017. 3
- [21] Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function, 2022. 3
- [22] Haoying Li, Yifan Yang, Meng Chang, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models, 2021. 3
- [23] Hao Li, Yang Zou, Ying Wang, Orchid Majumder, Yusheng Xie, R Manmatha, Ashwin Swaminathan, Zhuowen Tu, Stefano Ermon, and Stefano Soatto. On the scalability of diffusion-based text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9400–9409, 2024. 1, 2
- [24] 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*, pages 1833–1844, 2021. 2, 3, 4, 5, 6
- [25] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution, 2017. 3
- [26] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior, 2024. 3

- [27] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14297–14306, 2023. [2](#)
- [28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. [2, 3](#)
- [29] Mohammad Saeed Rad, Thomas Yu, Claudiu Musat, Hazim Kemal Ekenel, Behzad Bozorgtabar, and Jean-Philippe Thiran. Benefiting from bicubically down-sampled images for learning real-world image super-resolution, 2020. [5](#)
- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. [6](#)
- [31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. [1, 2](#)
- [32] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021. [3](#)
- [33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. [1](#)
- [34] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023. [2](#)
- [35] Yipeng Sun, Yixing Huang, Zeyu Yang, Linda-Sophie Schneider, Mareike Thies, Mingxuan Gu, Siyuan Mei, Siming Bayer, Frank G Zöllner, and Andreas Maier. Eagle: an edge-aware gradient localization enhanced loss for ct image reconstruction. *Journal of Medical Imaging*, 12(1):014001–014001, 2025. [5](#)
- [36] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution, 2024. [3](#)
- [37] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaou Tang. Esrgan: Enhanced super-resolution generative adversarial networks, 2018. [3](#)
- [38] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution, 2024. [1, 3](#)
- [39] 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, 2024. [3](#)
- [40] Shen Zhang, Zhaowei Chen, Zhenyu Zhao, Yuhao Chen, Yao Tang, and Jiajun Liang. Hidiffusion: Unlocking higher-resolution creativity and efficiency in pretrained diffusion models. In *European Conference on Computer Vision*, pages 145–161. Springer, 2024. [1, 2, 3, 6](#)