SPECTRAL-AWARE ADAPTATION FOR FIDELITY-PRESERVING DIFFUSION SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based methods have recently advanced real-world image super-resolution (SR) by leveraging powerful generative priors, typically adapting frozen backbones via low-rank updates such as LoRA. Existing approaches assign LoRA rank uniformly across layers, implicitly assuming homogeneous structural importance. We show that diffusion UNets instead exhibit pronounced intrinsic rank heterogeneity: effective rank is lowest at the input and output extremities of the network and rises toward the bottleneck. Motivated by this observation, we propose **SAFA** (**S**pectral-**A**ware **F**idelity **A**daptation), a parameter-efficient adaptation framework that estimates the effective rank of every candidate module via singular value decomposition, keeps the most spectrally expressive ones, and distributes a fixed trainable-parameter budget among them in proportion to effective rank, without adding any inference-time cost. Compared to uniform allocation on the same backbone, SAFA concentrates capacity in structurally expressive layers and produces $2.27\times$ stronger total update energy on RealSR while training $6.1\times$ fewer LoRA parameters; against prior diffusion-based SR methods it improves fidelity metrics (PSNR, SSIM) and reference-based perceptual similarity (LPIPS, DISTS, FID) on DIV2K, RealSR, and DRealSR. Qualitatively, SAFA holds structure together in regions that are easy to hallucinate – text, repetitive patterns and fine textures – where prior diffusion-based SR methods introduce geometrically inconsistent detail. We find that this fidelity-oriented allocation trades off some no-reference perceptual scores relative to methods tuned for aesthetic realism, and we report this trade-off explicitly rather than claiming uniform improvement. We further verify that the heterogeneity is not checkpoint-specific: the same macro-pattern holds across SD1.5, SD2.1 and SDXL, where decoder layers average $1.8\times$ the effective rank of encoder layers, although the precise spatial profile is architecture-dependent. Our results suggest that pre-adaptation spectral structure is a lightweight, gradient-free signal for guiding parameter-efficient adaptation, and that it transfers across diffusion UNet backbones.

1 INTRODUCTION

Single image super-resolution (SR) reconstructs a high-quality image from a degraded observation and is fundamentally ill-posed: the missing high-frequency content is ambiguous. Pixel-wise regression losses preserve fidelity but over-smooth, while perceptual and adversarial objectives improve realism at the cost of instability and hallucinated artifacts (Ledig et al., 2017; Wang et al., 2018). Large-scale diffusion models (Ho et al., 2020; Rombach et al., 2022) have since emerged as strong generative priors for SR (Wu et al., 2024b; Lin et al., 2024; Wang et al., 2024a), yielding richer textures and better robustness than GAN-based methods. Most of them, however, adapt the pretrained backbone with uniformly assigned low-rank updates (e.g., LoRA), implicitly assuming that every layer contributes equally to adaptation, which wastes capacity under a fixed parameter budget. Even PiSA-SR (Sun et al., 2025), which decouples pixel-level and perceptual objectives into dual LoRA branches for inference-time control, still assigns rank uniformly across the selected layers.

In practice, diffusion UNets are structurally heterogeneous: representational roles and spectral characteristics differ markedly across encoder, bottleneck, and decoder stages. Giving every layer the same rank therefore spends budget on layers that barely need it, while the richest-structured layers get no more capacity than the rest, which surfaces as the characteristic artifacts shown in Fig. 1:

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.



HR SinSR OSEDiff PiSA-SR SAFA

Figure 1: Typical restoration artifacts of diffusion-based Real-SR methods, with the ground-truth HR crop shown first for reference. SinSR (Wang et al., 2024b) produces over-smoothed textures, OSEDiff (Wu et al., 2024a) partially loses structural consistency, and PiSA-SR (Sun et al., 2025) introduces excessive high-frequency details. In contrast, SAFA reconstructs sharper and more coherent textures while preserving structural fidelity.

over-smoothed textures, structural inconsistencies, and hallucinated high-frequency detail. This motivates distributing adaptation capacity according to layer structure rather than uniformly across the backbone.

In this work, we introduce **SAFA** (**S**pectral-**A**ware **F**idelity **A**daptation), an architecture-aware low-rank adaptation framework for diffusion-based SR. We estimate each layer's effective rank via singular value decomposition, defined as the minimum number of leading singular components capturing a fixed proportion of the nuclear norm, keep the top $K\%$ of candidate modules by this measure, and allocate LoRA rank among them in proportion to it under a fixed trainable-parameter budget. Unlike prior approaches that assign identical rank across modules, this concentrates adaptation in structurally expressive layers and avoids redundant updates elsewhere. Because the learned low-rank updates are merged into the backbone weights after training, the deployed model has exactly the architecture and inference cost of the frozen backbone. Training follows a stable two-stage schedule, optimizing reconstruction fidelity first and introducing perceptual refinement second, with the allocation held fixed. On synthetic and real-world benchmarks, SAFA improves both distortion metrics (PSNR, SSIM) and reference-based perceptual quality (LPIPS, DISTS, FID) over prior diffusion-based SR methods, including dual-branch approaches (Sun et al., 2025), at the same single diffusion step and with $6.1\times$ fewer trainable LoRA parameters (1.32M against 8.11M for PiSA-SR), and structural analysis shows that it redistributes adaptation energy into stronger, more concentrated updates. Our contributions are summarized as follows:

- We analyze the spectral characteristics of diffusion UNet layers and reveal significant intrinsic rank heterogeneity across the network, indicating that uniform low-rank adaptation is inefficient for real-world super-resolution.
- We propose **SAFA**, an effective rank allocation framework that estimates module-wise effective rank via singular value decomposition, adapts only the top $K\%$ of modules, and sizes their adapters in proportion to that measure under a fixed parameter budget.
- We show that intrinsic rank-guided allocation redistributes adaptation energy more effectively than uniform updates, producing stronger and more structured parameter updates and leading to consistent improvements in both reconstruction fidelity and reference-based perceptual quality without increasing inference cost.

2 Related Work

Image Super-Resolution. Early deep SR methods targeted fidelity under synthetic bicubic degradation, using residual (Lim et al., 2017; Zhang et al., 2018b;c; Ledig et al., 2017) and attention-based (Liang et al., 2021; Chen et al., 2022; Zhang et al., 2022; Chen et al., 2023b;a; Korkmaz & Tekalp, 2024) architectures to improve PSNR and SSIM. Such settings, however, do not capture real-world degradations. Later work either collected real paired data (Cai et al., 2019; Wei et al., 2020) or synthesized realistic degradations (Zhang et al., 2021; Sun et al., 2024): BSRGAN (Zhang et al., 2021) randomly composes degradation operators, and RealESRGAN (Wang et al., 2021) applies high-order degradation processes. Perceptual and adversarial objectives (Ledig et al., 2017; Wang et al., 2018) pushed realism beyond pixel-wise regression, but introduce artifacts and instability, motivating frequency-domain regularization (Fritsche et al., 2019; Korkmaz et al., 2024) and loss balancing with structural priors (Liang et al., 2022; Blau & Michaeli, 2018; Sun et al., 2024). CNN- and GAN-based models nonetheless exploit large-scale semantic priors only to a limited extent, which motivated the shift to diffusion models.

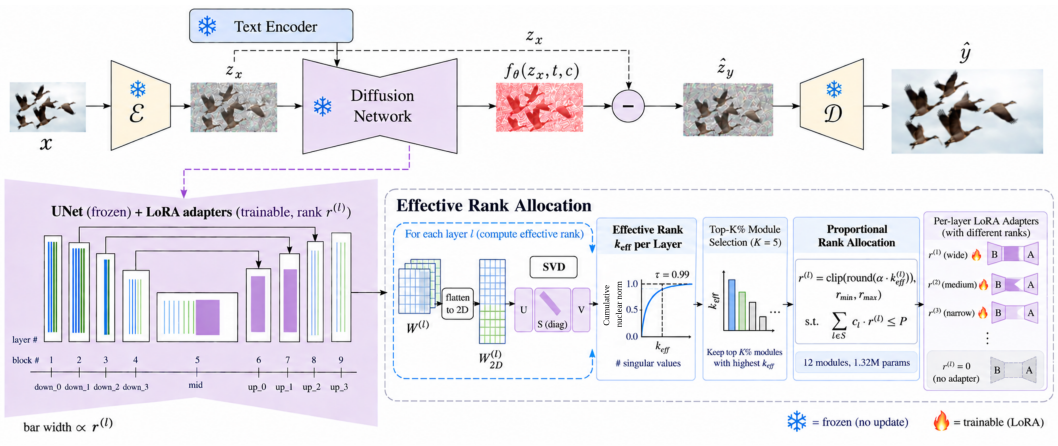


Figure 2: Overview of SAFA. We first perform SVD on the pretrained UNet weights to obtain the effective rank of every candidate module at a 99% nuclear-norm energy threshold ($\tau = 0.99$). The top K% of modules by effective rank are then selected, and a fixed trainable-parameter budget is split among them in proportion to those estimates, so spectrally rich modules receive wider LoRA adapters while the rest stay frozen.

Diffusion-Based Image Super-Resolution. Diffusion models (DMs) (Ho et al., 2020; Rombach et al., 2022) have been adopted for SR for their strong generative priors. Early methods conditioned the reverse process directly (Kawar et al., 2022; Yue et al., 2023); later ones build on Stable Diffusion (SD) pretrained on text-to-image data (Wang et al., 2024a; Lin et al., 2024; Yang et al., 2024; Wu et al., 2024b;a) and differ mainly in conditioning and architecture. StableSR (Wang et al., 2024a) and DiffBIR (Lin et al., 2024) add dedicated encoders, PASD (Yang et al., 2024) fuses pixel and semantic features via cross-attention, and SeeSR (Wu et al., 2024b) uses degradation-aware prompts. For finer detail, TVT (Yi et al., 2025) rescales the SD VAE from $8\times$ to $4\times$ downsampling in alignment with the pretrained UNet, while DiT4SR (Duan et al., 2025) injects low-resolution features into the attention of a diffusion Transformer. One-step and distilled models such as SinSR (Wang et al., 2024b), OSEDiff (Wu et al., 2024a), and AdcSR (Chen et al., 2025) compress multi-step sampling to cut inference cost. Across these designs, pixel-level fidelity and semantic enhancement remain entangled in shared parameters, yielding the perception–distortion trade-off (Blau & Michaeli, 2018).

Parameter-Efficient Adaptation for Diffusion SR. Parameter-efficient fine-tuning (PEFT) (Hou et al., 2023; Dettmers et al., 2024; Xie et al., 2023), in particular LoRA (Hu et al., 2022), adapts large pretrained models while keeping the backbone frozen, and is widely used in vision and language (Li et al., 2023a; Lian et al., 2022; Jia et al., 2022). AdaLoRA (Zhang et al., 2023) makes the allocation non-uniform, resizing adapters during training from gradient-based importance scores. In diffusion SR, PiSA-SR (Sun et al., 2025) decouples pixel-level fidelity from semantic enhancement with a dual-branch LoRA, but still spreads its adapters uniformly over a hand-picked set of layers. SAFA shares AdaLoRA's non-uniform ranks but not its machinery: the allocation is read off the pretrained spectrum *before* training and then frozen, so it needs no gradients, no importance statistics and no budget re-estimation during optimization.

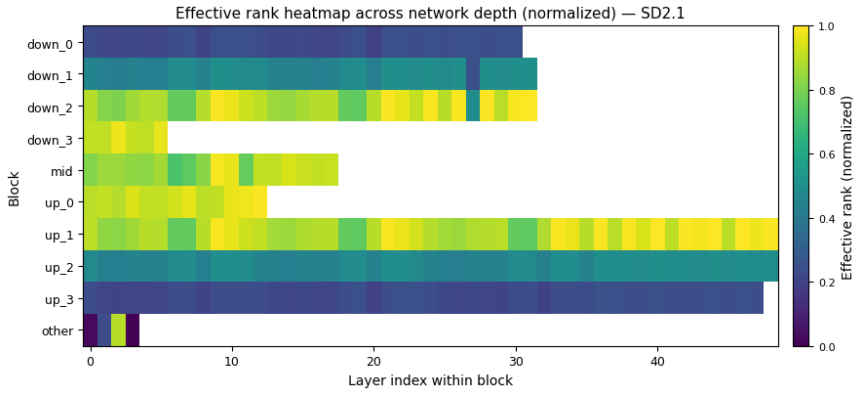
3 Method

3.1 Diffusion-Based SR Formulation

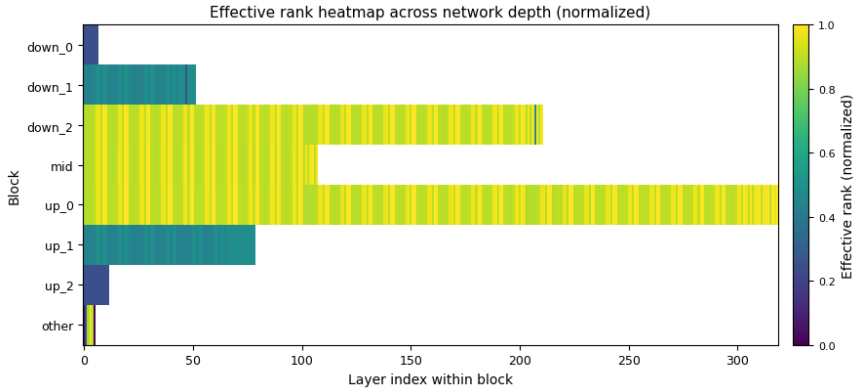
Let x be a low-quality (LQ) image and y its high-quality (HQ) counterpart. We adopt a latent diffusion model (LDM) (Rombach et al., 2022) as the backbone: a frozen VAE encoder $\mathcal{E}$ gives $z_x = \mathcal{E}(x)$, the diffusion UNet f_θ predicts a residual conditioned on timestep t and optional context c, and a decoder $\mathcal{D}$ restores the image,

$$\hat{z}_y = z_x - f_\theta(z_x, t, c), \qquad \hat{y} = \mathcal{D}(\hat{z}_y). \tag{1}$$

For parameter-efficient adaptation, LoRA (Hu et al., 2022) replaces a backbone weight $W \in \mathbb{R}^{d_1 \times d_2}$ by $W' = W + BA$, with learnable $B \in \mathbb{R}^{d_1 \times r}$, $A \in \mathbb{R}^{r \times d_2}$ and $r \ll \min(d_1, d_2)$. Existing diffusion



(a) SD2.1-base



(b) SDXL

Figure 3: Min–max normalized effective rank across UNet depth, computed at the same 99% nuclear-norm energy threshold ($\tau = 0.99$) used throughout. Rows are UNet blocks and columns are layers within a block; `other` collects the layers that belong to no resolution block (e.g., the input and output convolutions and the time embedding). Both backbones are lowest at the input and output extremities and peak around the bottleneck, but SD2.1-base transitions gradually and asymmetrically while SDXL saturates sharply and near-symmetrically.

SR methods assign a uniform rank r across a fixed set of layers; we instead let the backbone's own spectral structure set a separate rank $r^{(l)}$ per layer.

3.2 SAFA

Figure 2 illustrates SAFA on top of the pipeline of Sec. 3.1: the frozen encoder $\mathcal{E}$ maps the LQ image x to z_x, the UNet f_θ predicts the residual conditioned on z_x, the timestep and optional text features, and the frozen decoder $\mathcal{D}$ turns $\hat{z}_y$ into the reconstruction $\hat{y}$.

The base model stays frozen and is adapted with lightweight LoRA modules. Rather than giving every module in a hand-picked set the same adapter, SAFA sizes each adapter from that layer's intrinsic complexity, measured as the effective rank of its weight matrix $\mathbf{W}$: the number of singular components needed to capture a predefined fraction of the nuclear-norm energy (99% here). As illustrated in Fig. 2, this means (1) flattening the layer weights to a two-dimensional matrix, (2) computing the SVD and estimating the effective rank, and (3) keeping the top $K\%$ of modules and splitting a fixed trainable-parameter budget among them in proportion to their effective ranks.

3.3 Effective Rank Allocation

Standard LoRA adaptation places adapters of one uniform rank over a fixed, hand-picked set of layers, implicitly assuming they all require similar capacity. Diffusion UNets are not uniform: layers

differ substantially in intrinsic rank, singular value spectrum, and representational complexity. To test whether this heterogeneity is specific to SD2.1 or a broader property of diffusion UNets, we repeat the effective-rank analysis on SDXL and compare the normalized spatial distributions in Fig. 3. Both backbones share a macro-pattern: normalized effective rank is lowest at the input and output extremities (`down_0` and the final decoder blocks) and rises toward the bottleneck, where the mid block and its neighbours reach near-maximal rank. The transition differs, however. SDXL (Fig. 3b) is sharp and near-symmetric, with `down_2`, `mid`, and `up_0` saturating while `down_0` and `up_2` stay sharply low, reflecting its shallower three-stage resolution hierarchy. SD2.1 (Fig. 3a) is gradual and asymmetric: rank climbs steadily across `down_0`–`down_2`, stays elevated through `mid` and `up_1`, then decays gradually over `up_2`–`up_3` rather than dropping abruptly at the decoder's end, its finer four-stage hierarchy affording more blocks for the transition. SD1.5 follows the same pattern; averaged over all three backbones, decoder layers carry $1.8\times$ the effective rank of encoder layers. Pronounced rank heterogeneity is thus shared across the Stable Diffusion UNet family rather than being an artifact of one checkpoint, while its precise spatial profile is architecture-dependent; we have not tested non-UNet backbones such as diffusion Transformers. A uniform rank therefore wastes budget on layers that are already effectively low-rank, and the allocation signal should be read off the pretrained weights themselves rather than fixed by a positional heuristic such as "always favor decoder layers"; Sec. 4.4 shows it also matters which weight statistic is used.

We therefore let effective rank decide both *which* parts of the backbone to adapt and *how much* capacity each receives. Allocation operates at the granularity of UNet *modules* (a module being an adaptable submodule such as a transformer block, which contains several weight matrices). Let $\mathcal{C}$ be the set of candidate modules and $k_{\text{eff}}^{(l)}$ the effective rank of module l, obtained by applying Algorithm 1 to its weight matrices (Sec. 3.4). We first keep the top $K\%$ by effective rank,

$$\mathcal{S} = \Big\{ l \in \mathcal{C} : \pi(l) \le \big\lceil \tfrac{K}{100} |\mathcal{C}| \big\rceil \Big\}, \tag{2}$$

where $\pi(l)$ is the position of l when $\mathcal{C}$ is sorted by k_{eff} in descending order. Within $\mathcal{S}$ we then set ranks proportional to effective rank, subject to a global budget P on trainable parameters,

$$r^{(l)} = \text{clip}\big(\text{round}(\alpha\, k_{\text{eff}}^{(l)}),\ r_{\min},\ r_{\max}\big), \qquad \text{s.t.} \quad \sum_{l \in \mathcal{S}} c_l\, r^{(l)} \le P, \tag{3}$$

where c_l is the per-rank parameter cost of module l and α is the largest scaling factor for which the budget holds; any leftover budget goes to the largest fractional remainders. Modules outside $\mathcal{S}$ stay frozen. Inside a selected module, every weight matrix $W \in \mathbb{R}^{d_1 \times d_2}$ receives its own update $W' = W + BA$ at that module's rank, with $B \in \mathbb{R}^{d_1 \times r^{(l)}}$ and $A \in \mathbb{R}^{r^{(l)} \times d_2}$. With $K = 5$ this selects 12 UNet modules and 1.32M trainable parameters, against 8.11M when the full candidate set is adapted uniformly. Note that k_{eff} enters only through the ratio in $\alpha k_{\text{eff}}^{(l)}$ and is never used as a rank directly. We use $\tau = 0.99$ and $K = 5$ throughout; Sec. 4.4 ablates K and Appendix D ablates τ.

3.4 Effective Rank Estimation

Let $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ be a weight matrix (a linear layer, or a convolution kernel flattened to two dimensions), with SVD $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, full rank $\rho = \min(d_1, d_2)$ and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_\rho)$, $\sigma_1 \ge \cdots \ge \sigma_\rho \ge 0$. Let $p_i = \sigma_i / \|\mathbf{W}\|_*$ be the share of the nuclear norm $\|\mathbf{W}\|_* = \sum_j \sigma_j$ carried by the i-th component, so $\sum_i p_i = 1$, and let $P_k = \sum_{i \le k} p_i$ be the cumulative energy. Given a threshold $\tau \in (0, 1]$ (we use $\tau = 0.99$), the **effective rank** is the smallest number of leading components capturing a fraction τ of that energy, clamped to the valid range:

$$k_{\text{eff}}(\mathbf{W}; \tau) = \max\Big(1,\ \min\big(\min\{\, k : P_k \ge \tau \,\},\ \rho\big)\Big). \tag{4}$$

A small k_{eff} relative to ρ means the spectrum is concentrated in few directions and the matrix is effectively low-rank; a large one means the singular values are spread out, which we read as the layer being more expressive and thus a better place to spend adaptation capacity. Algorithm 1 states the per-layer estimator, including the reshape applied to 4D convolution kernels.

Algorithm 1 Effective rank via SVD (per-layer)

Require: Weight tensor $\mathbf{W}$ (2D or 4D); energy threshold $\tau \in (0, 1]$ (e.g., $\tau = 0.99$)
Ensure: Effective rank $k_{\text{eff}} \in \{1, \ldots, \min(\texttt{shape}(\mathbf{W}_{\text{2d}}))\}$

$d_1 \leftarrow \texttt{shape}(\mathbf{W})[0]$
if $\mathbf{W}$ is 4D (convolution) **then**
  $\mathbf{W}_{\text{2d}} \leftarrow \texttt{reshape}(\mathbf{W},\ (d_1,\ -1))$ ▷ $(c_{\text{out}},\ c_{\text{in}} \cdot h \cdot w)$
else if $\mathbf{W}$ is 2D **then**
  $\mathbf{W}_{\text{2d}} \leftarrow \mathbf{W}$
else
  return $\min(\texttt{shape}(\mathbf{W}))$
end if
$\rho \leftarrow \min(\texttt{shape}(\mathbf{W}_{\text{2d}}))$
$\mathbf{U}, \boldsymbol{\sigma}, \mathbf{V}^\top \leftarrow \texttt{SVD}(\mathbf{W}_{\text{2d}},\ \texttt{full_matrices} = \texttt{False})$ ▷ $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_\rho), \sigma_1 \geq \cdots \geq \sigma_\rho$
$s \leftarrow \sum_{i=1}^{\rho} \sigma_i$
$p_i \leftarrow \sigma_i / s$ for $i = 1, \ldots, \rho$ ▷ Normalized singular values (energy distribution)
$P_k \leftarrow \texttt{cumulative_sum}(p_1, \ldots, p_\rho)$ ▷ $P_k = \sum_{i=1}^{k} p_i$
$k_{\text{eff}} \leftarrow 1 + \left|\{k : P_k < \tau\}\right|$ ▷ Smallest k such that $P_k \geq \tau$
$k_{\text{eff}} \leftarrow \texttt{clamp}(k_{\text{eff}},\ 1,\ \rho)$
return k_{eff}

3.5 Training Strategy

We adopt the two-stage schedule of Sun et al. (2025). The first stage optimizes the reconstruction loss $\mathcal{L}_{\text{pix}} = \|y - \hat{y}\|_2^2$ alone; the second adds perceptual supervision,

$$\mathcal{L}_{\text{rec}} = \lambda_{pix}\, \mathcal{L}_{\text{pix}} + \lambda_{perc}\, \mathcal{L}_{\text{perc}}, \tag{5}$$

where $\mathcal{L}_{\text{perc}}$ is the LPIPS perceptual loss and $\lambda_{pix}, \lambda_{perc}$ are fixed weights. Fidelity is therefore established first and texture refined second. The rank allocation $\{r^{(l)}\}$ is computed once before training, from the pretrained weights alone, and held fixed throughout both stages.

4 Experiments

4.1 Experimental Setup

Datasets. We train on LSDIR (Li et al., 2023b), generating LQ–HQ pairs with the Real-ESRGAN degradation pipeline (Wang et al., 2021) as is standard for real-world SR. For synthetic evaluation we crop 3,000 512×512 images from the DIV2K validation set (Agustsson & Timofte, 2017) and degrade them to 128×128 with the same pipeline. For real-world evaluation we use RealSR (Cai et al., 2019) and DRealSR (Wei et al., 2020), center-cropped into aligned 128×128 LQ and 512×512 HQ pairs.

Evaluation Metrics. PSNR and SSIM (Wang et al., 2004) measure pixel-level fidelity; LPIPS (Zhang et al., 2018a) and DISTS (Ding et al., 2020) measure deep-feature perceptual similarity; FID (Heusel et al., 2017) measures distributional similarity to the ground-truth set. We also report the no-reference metrics CLIPIQA (Wang et al., 2023), NIQE (Zhang et al., 2015), MUSIQ (Ke et al., 2021), and MANIQA (Yang et al., 2022).

Implementation Details. We build on Stable Diffusion 2.1-base (Stability AI, 2022), keeping the pretrained VAE and UNet frozen and training only the LoRA modules, whose per-layer ranks $r^{(l)}$ follow the allocation of Sec. 3.3. Training runs on two NVIDIA H100 GPUs, batch size 8, 200K iterations, AdamW (Kingma, 2014) at learning rate 5×10^{-5}, with prompts from RAM (Zhang et al., 2024) and $\lambda_{\text{pix}} = \lambda_{\text{perc}} = 0.5$.

4.2 Comparisons with the State-of-the-Arts

Quantitative Comparisons Table 1 compares SAFA with existing diffusion-based real-world SR methods on synthetic and real-world benchmarks. SAFA attains the best PSNR and SSIM on all three datasets, indicating that restorations stay closely aligned with the ground-truth content, together with

Table 1: Quantitative comparison of diffusion-based SR methods on the synthetic DIV2K benchmark and the real-world RealSR and DRealSR benchmarks. The suffix -sN denotes the number of diffusion steps N. ↑ indicates higher is better and ↓ indicates lower is better. The best and the second-best results are highlighted in red and blue, respectively.

Method	DIV2K								
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
ResShift-s15	24.69	0.6175	0.3374	0.2215	36.01	6.82	0.6089	60.92	0.5450
StableSR-s200	23.31	0.5728	0.3129	0.2138	24.67	4.76	0.6682	65.63	0.6188
DiffBIR-s50	23.67	0.5653	0.3541	0.2129	30.93	4.71	0.6652	65.66	0.6204
PASD-s20	23.14	0.5489	0.3607	0.2219	29.32	4.40	0.6711	68.83	0.6484
SeeSR-s50	23.71	0.6045	0.3207	0.1967	25.83	4.82	0.6857	68.49	0.6239
SinSR-s1	24.43	0.6012	0.3262	0.2066	35.45	6.02	0.6499	62.80	0.5395
OSEDiff-s1	23.72	0.6108	0.2941	0.1976	26.32	4.71	0.6683	67.97	0.6148
PiSA-SR-s1	23.87	0.6058	0.2823	0.1934	25.07	4.55	0.6927	69.68	0.6400
AdcSR-s1	23.74	0.6017	0.2853	0.1899	25.52	4.36	0.6763	67.99	0.6098
TVT-s1	24.23	0.6292	0.2773	0.1860	24.78	5.59	0.6986	68.67	0.6060
SAFA-s1	24.91	0.6395	0.2710	0.1871	23.19	4.84	0.6003	65.11	0.6057
	RealSR								
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
ResShift-s15	26.31	0.7411	0.3489	0.2498	142.81	7.27	0.5450	58.10	0.5305
StableSR-s200	24.69	0.7052	0.3091	0.2167	127.20	5.76	0.6195	65.42	0.6211
DiffBIR-s50	24.88	0.6673	0.3567	0.2290	124.56	5.63	0.6412	64.66	0.6231
PASD-s20	25.22	0.6809	0.3392	0.2259	123.08	5.18	0.6502	68.74	0.6461
SeeSR-s50	25.33	0.7273	0.2985	0.2213	125.66	5.38	0.6594	69.37	0.6439
SinSR-s1	26.30	0.7354	0.3212	0.2346	137.05	6.31	0.6204	60.41	0.5389
OSEDiff-s1	25.15	0.7341	0.2921	0.2128	123.50	5.65	0.6693	69.09	0.6339
PiSA-SR-s1	25.50	0.7417	0.2672	0.2044	124.09	5.50	0.6702	70.15	0.6560
AdcSR-s1	25.47	0.7301	0.2885	0.2129	118.43	5.35	0.6729	69.89	0.6360
TVT-s1	25.81	0.7596	0.2597	0.2061	109.89	5.92	0.6882	69.89	0.6232
SAFA-s1	26.73	0.7676	0.2445	0.1932	109.30	5.99	0.5960	65.69	0.6240
	DRealSR								
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
ResShift-s15	28.45	0.7632	0.4073	0.2700	175.92	8.28	0.5259	49.86	0.4573
StableSR-s200	28.04	0.7460	0.3354	0.2287	147.03	6.51	0.6171	58.50	0.5602
DiffBIR-s50	26.84	0.6660	0.4446	0.2706	167.38	6.02	0.6292	60.68	0.5902
PASD-s20	27.48	0.7051	0.3854	0.2535	157.36	5.57	0.6714	64.55	0.6130
SeeSR-s50	28.26	0.7698	0.3197	0.2306	149.86	6.52	0.6672	64.84	0.6026
SinSR-s1	28.41	0.7495	0.3741	0.2488	177.05	7.02	0.6367	55.34	0.4898
OSEDiff-s1	27.92	0.7835	0.2968	0.2165	135.29	6.49	0.6963	64.65	0.5899
PiSA-SR-s1	28.31	0.7804	0.2960	0.2169	130.61	6.20	0.6970	66.11	0.6156
AdcSR-s1	28.09	0.7726	0.3046	0.2200	134.05	6.45	0.7049	66.26	0.5913
TVT-s1	28.27	0.7899	0.2899	0.2205	134.29	7.03	0.7218	65.56	0.5776
SAFA-s1	29.60	0.8143	0.2627	0.2003	128.24	6.76	0.6149	60.73	0.5821

the best LPIPS throughout and the best DISTS on RealSR and DRealSR (second on DIV2K, behind TVT) – notable given that many baselines rely on multi-step sampling to refine detail. It also attains the best FID on all three benchmarks, so the outputs remain close to natural-image statistics while preserving structure. SAFA does not lead on the no-reference metrics (NIQE, CLIPIQA, MUSIQ, MANIQA), where methods leaning on semantic priors or tuned toward aesthetic realism score higher; it sits mid-pack and instead keeps a balance across fidelity- and perception-oriented measures. This matters in real-world SR, where excessive perceptual enhancement introduces hallucinated structures that deviate from the input content. Overall, SAFA achieves the strongest reconstruction accuracy among diffusion-based Real-SR methods while maintaining single-step inference.

Qualitative Results Figure 4 shows qualitative comparisons on real-world images. SeeSR and SinSR struggle to recover reliable fine structure, producing over-smoothed textures or irregular high-frequency patterns. OSEDiff is visually sharper, but its details drift from the underlying structure, particularly around repetitive patterns. PiSA-SR raises perceptual sharpness, yet its textures can appear over-enhanced or geometrically distorted in complex regions. AdcSR and TVT preserve coarse structure but blur thin edges and repeated textures such as roof tiles or letter strokes. SAFA instead restores fine detail while staying structurally consistent with the input: around letters, repetitive roof patterns, and subtle surface textures it remains geometrically stable, with textures that are sharper yet faithful to the low-quality observation. Appendix E repeats the comparison against GAN-based

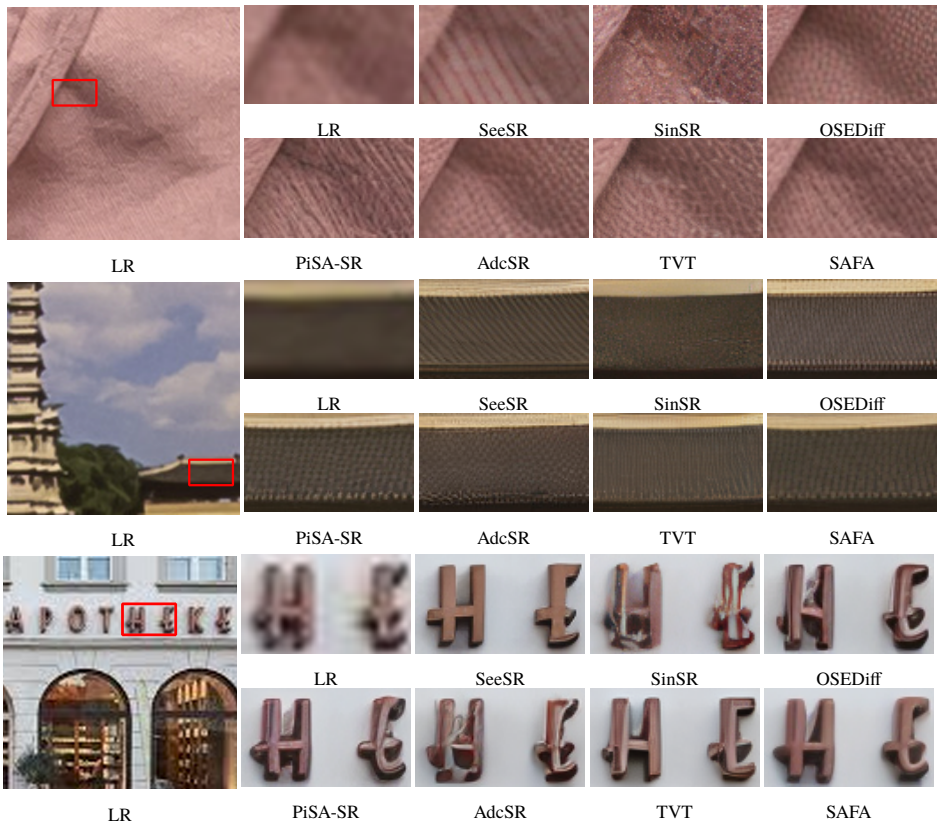


Figure 4: Qualitative comparison with state-of-the-art diffusion-based SR methods on RealSR and real-world photographs. In each example the leftmost panel is the full LQ input with the crop region boxed; the remaining panels zoom into that region for the LQ input, SeeSR, SinSR, OSEDiff, PiSA-SR, AdcSR, TVT, and SAFA. The examples are chosen to stress structure that is easy to hallucinate: a repetitive fabric weave, roof tiles, and letter strokes. SAFA keeps these patterns geometrically regular, whereas the baselines either over-smooth them or synthesize detail that drifts from the input.

methods, where RealESRGAN (Wang et al., 2021) and LDL (Liang et al., 2022) sharpen textures but break the geometry of grid patterns and letter strokes, and SAFA leads on every metric except NIQE. Further visual comparisons are given in Appendix F.

4.3 Adaptation Energy Redistribution

To understand how different allocation strategies modify the backbone, we merge the trained LoRA into the base UNet and measure the per-layer change $\|\Delta W_l\|_F = \|W'_l - W_l\|_F$. Table 2 compares uniform LoRA allocation with our intrinsic-rank-guided allocation on RealSR. We take PiSA-SR as the uniform counterpart, since it adapts the same SD2.1 backbone with a uniform rank; its scores are the published ones rather than a re-training under our schedule. The two are not budget-matched, and the gap runs in our favour: SAFA trains 1.32M LoRA parameters against 8.11M. With that smaller budget it still improves PSNR and SSIM by 4.82% and 3.49% and cuts LPIPS by 8.50% and DISTS by 5.48%. Structurally it also moves the backbone further, raising the total adaptation energy $\sum_l \|\Delta W_l\|_F^2$ by 126.5% (2.27×) and concentrating it slightly more, with the top 20% of layers carrying 86.96% of the energy against 84.70%. A smaller trainable set therefore does not imply a smaller change to the backbone. Because this particular comparison is not controlled, we do not read causality from it; the controlled evidence that the *signal* matters is the matched-budget comparison in Sec. 4.4.

Fig. 5 makes this concrete: the distribution is head-heavy in every group, with a minority of layers far above the uniform mean – the largest in the decoder and the bottleneck – and a long tail below

Table 2: **Uniform vs. effective rank allocation**, on RealSR. The uniform row is PiSA-SR, whose scores are reproduced from Table 1. Params counts trainable LoRA parameters, so its Gain is a reduction; total adaptation energy is $\sum_l \|\Delta W_l\|_F^2$ and Top-20% is the share of it held by the top 20% of layers.

Method	Params↓	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	Total Energy	Top-20%
Uniform allocation (PiSA-SR)	8.11M	25.50	0.7417	0.2672	0.2044	49,325.81	84.70%
SAFA (Ours)	1.32M	26.73	0.7676	0.2445	0.1932	111,725.16	86.96%
Gain(%)	83.72	4.82	3.49	8.50	5.48	126.50	2.67

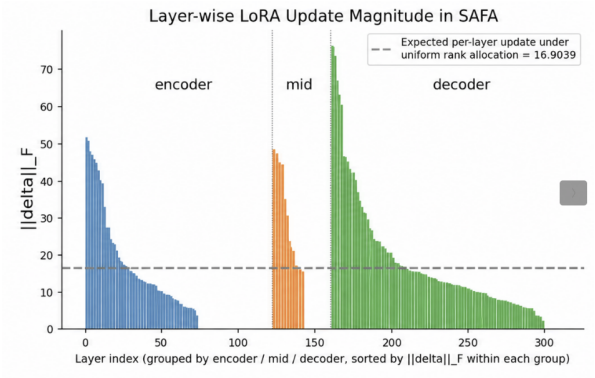


Figure 5: Layer-wise LoRA update magnitude $\|\Delta W_l\|_F$ after training with SAFA, grouped into encoder, bottleneck and decoder and sorted within each group. The dashed line marks the mean per-layer update magnitude under uniform rank allocation.

it. SAFA also trains at 40.76 GB and 1.24 s/iteration, against 43.87 GB / 1.60 s for PiSA-SR and 56.28 GB / 2.26 s for OSEDiff (Appendix C).

4.4 Ablation Study

Table 3 varies K and swaps the allocation signal for the Frobenius norm. Sparse selection matters most: every $K \leq 20$ beats uniform adaptation of the full candidate set on a fraction of its parameters, and accuracy peaks at moderate K. At a matched 1.32M budget, effective rank gives a clearly better PSNR than magnitude (26.73 vs 26.56) at the same LPIPS; the two correlate ($\rho = 0.79$ encoder, 0.65 decoder) but only weakly elsewhere (0.21), so they are not interchangeable. Frobenius at $K = 10$ edges ahead of SAFA, but spends $1.8\times$ the parameters. The threshold τ matters little (Appendix D).

Table 3: **Selection ratio and allocation signal**, on RealSR. "Uniform" adapts every candidate module at one rank and is the PiSA-SR row of Table 2; the Frobenius rows replace k_{eff} with the Frobenius norm as the allocation signal.

Allocation signal	Top-K%	Params↓	PSNR↑ / LPIPS↓
Uniform (PiSA-SR)	100	8.11M	25.50 / 0.267
Frobenius norm	2.5	0.72M	25.82 / 0.255
Frobenius norm	5	1.32M	26.56 / 0.244
Frobenius norm	10	2.32M	26.75 / 0.240
Frobenius norm	20	4.10M	26.47 / 0.248
Effective rank (ours)	5	1.32M	26.73 / 0.245

5 Conclusion

We presented SAFA, which adapts only the diffusion-UNet modules an SVD-based effective-rank measure marks as expressive and sizes their adapters in proportion to it. Because intrinsic complexity varies sharply across the network, this outperforms a uniform rank over a hand-picked set while training $6.1\times$ fewer parameters at no extra inference cost, making pre-adaptation spectral structure a lightweight signal that transfers across backbones.

Ethics Statement

This work studies parameter-efficient adaptation of pretrained diffusion models for image super-resolution. All training and evaluation data come from publicly available benchmarks released for research use (LSDIR, DIV2K, RealSR, DRealSR), together with photographs taken by the authors; no human subjects were involved and no personally identifying data was collected or released. Because our method inherits the pretrained Stable Diffusion prior, it also inherits whatever biases that prior carries, and like any generative restoration model it can synthesize detail that is not present in the low-quality input. SAFA is explicitly fidelity-oriented and reduces such hallucination relative to perception-tuned baselines, but its outputs should still not be treated as evidence in forensic, medical, or legal settings. We declare no conflicts of interest or external sponsorship influencing this work.

Reproducibility Statement

The allocation procedure is specified in full: the effective-rank estimator is defined in Sec. 3.4 and given as pseudocode in Algorithm 1, and the proportional rank-allocation rule is given in Sec. 3.3. Sec. 4.1 lists the backbone, datasets and degradation pipeline, optimizer, learning rate, batch size, iteration count, hardware, and loss weights needed to rerun training, and the threshold τ used throughout is ablated in Table 4. The appendix reports the block-wise rank statistics, the layer-wise update magnitudes, and the training memory and time measurements behind our efficiency claims. Source code and the trained LoRA weights will be released publicly upon publication; we do not include a code archive with this submission in order to preserve anonymity.

AI Use Statement

We used generative AI tools as writing and formatting assistants: polishing and condensing author-written prose. We did not use generative AI tools for research ideation, for deriving or verifying the method, for generating experimental results, or for retrieving or summarizing related work; all citations were selected and checked by the authors. All AI-assisted text was reviewed and edited by the authors, and we take full responsibility for the final content of this work, including all claims, numbers, and figures.

References

E Agustsson and R. Timofte. NTIRE 2017 Challenge on single image super-resolution: Dataset and study. In *IEEE/CVF Conf. Comp. Vis. and Patt. Recog. (CVPR) Workshops*, 2017.

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

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3086–3095, 2019.

Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression for real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28208–28220, 2025.

Xiangyu Chen, X. Wang, J. Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pp. 22367–22377, June 2023a.

Zheng Chen, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration. In *NeurIPS*, 2022.

Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, pp. 12312–12321, October 2023b.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

K. Ding, K. Ma, S. Wang, and E. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Trans. Patt Anal. Mach. Intel.*, 44:2567–2581, 2020.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S Ren, Chunle Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18948–18958, 2025.

Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 3599–3608. IEEE, 2019.

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

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

Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727. Springer, 2022.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

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

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Cansu Korkmaz and A. Murat Tekalp. Training transformer models by wavelet losses improves quantitative and visual performance in single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2024.

Cansu Korkmaz, A Murat Tekalp, and Zafer Dogan. Training generative image super-resolution models by wavelet-domain losses enables better control of artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5926–5936, 2024.

C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *IEEE/CVF Conf. on Comp. Vision and Patt. Recog. (CVPR)*, pp. 105–114, 2017.

Wenbo Li, Xin Lu, Shengju Qian, and Jiangbo Lu. On efficient transformer-based image pre-training for low-level vision. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 1089–1097, 2023a.

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

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *Advances in Neural Information Processing Systems*, 35: 109–123, 2022.

Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5657–5666, 2022.

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

B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE/CVF CVPR Workshops*, 2017.

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

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*, pp. 10684–10695, 2022.

Stability AI. Stable diffusion, 2022. URL `https://stability.ai/stable-diffusion`.

Lingchen Sun, Jie Liang, Shuaizheng Liu, Hongwei Yong, and Lei Zhang. Perception-distortion balanced super-resolution: A multi-objective optimization perspective. *IEEE Transactions on Image Processing*, 33:4444–4458, 2024.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.

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

X. Wang, Ke Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy. ESRGAN: enhanced super-resolution generative adversarial networks. In *European Conf. on Comp. Vision (ECCV) Workshops*, 09 2018.

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

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

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

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pp. 101–117. Springer, 2020.

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

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

Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Diffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4230–4239, 2023.

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

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *The European Conference on Computer Vision (ECCV) 2024*, 2024.

Qiaosi Yi, Shuai Li, Rongyuan Wu, Lingchen Sun, Yuhui Wu, and Lei Zhang. Fine-structure preserved real-world image super-resolution via transfer vae training. *Proceedings of the IEEE/CVF international conference on computer vision*, 2025.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36: 13294–13307, 2023.

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

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

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023.

R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conf. Comp. Vision and Patt. Recog. (CVPR)*, pp. 586–595, 2018a.

Xindong Zhang, Hui Zeng, Shi Guo, and L. Zhang. Efficient long-range attention network for image super-resolution. In *Euro. Conf. Comp. Vision (ECCV)*, 2022.

Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *IEEE/CVF ECCV*, 2018b.

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

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2472–2481, 2018c.