FreqOrtho-SR: Frequency-Guided Orthogonal Expert Learning for Real-World Image Super-Resolution

Minh Son Hoang*, Dinh Phu Tran*, Quyen Nguyen Duc, Dam Hoang Phuong, and Daeyoung Kim**

School of Computing, KAIST, Republic of Korea
{sonhm2910,phutx2000,nguyenducquyen2311,hoangphuong1211,kimd}@kaist.ac.kr

Abstract. Diffusion prior-based methods have shown impressive results in real-world image super-resolution (ISR), yet two key challenges persist: balancing pixel-level fidelity with semantic quality, and adapting to diverse degradations. Existing dual-branch approaches freeze the pixel module during semantic training, but the semantic branch can still expand capacity within the pixel subspace, precluding genuine perceptual improvement. Moreover, using a single static adapter cannot generalize across heterogeneous real-world corruptions. To address both issues, we propose FreqOrtho-SR, which comprises: **Freq**uency-guided Mixture of LoRA Experts (FreqMoE), it routes inputs to specialized experts via a non-parametric FFT-based degradation-feature extractor that encodes frequency-domain signatures, enabling stable and interpretable specialization across corruption types; and **Ortho**gonal Gradient Projection (OGP), which reframes the dual-objective optimization as a subspace-constrained problem: by extracting the pixel-fidelity subspace via SVD on combined expert weight deltas and projecting semantic gradients onto its null space, OGP guarantees orthogonality between the two objectives, enabling genuinely complementary learning without mutual interference. Experiments show that FreqOrtho-SR achieves competitive overall performance and a strong fidelity-perception trade-off across multiple benchmarks with efficient single-step inference. The source code of our method can be found at `sonhm3029/FreqOrtho-SR`.

Keywords: Image Super-Resolution · Frequency-guided Mixture of Experts · Orthogonal Learning · Diffusion Models

1 Introduction

Image super-resolution (ISR) [53, 62] aims to produce high-resolution (HR) images from degraded low-resolution (LR) inputs. A main challenge in ISR is the

* Equal contribution.
** Corresponding author.

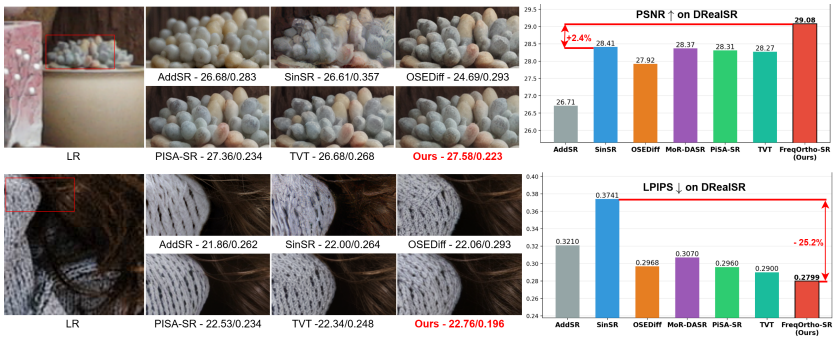

Fig. 1: Visual and quantitative comparison on DRealSR. Values below each patch denote PSNR↑/LPIPS↓, with best in red. Our FreqOrtho-SR achieves the best fidelity and perceptual quality among one-step diffusion methods.

inherent trade-off between perception and distortion [2]. While pixel-level regression methods using $\mathcal{L}_1$ or $\mathcal{L}_2$ losses achieve high fidelity metrics (*e.g.*, PSNR, SSIM), their results are perceptually unsatisfying and overly smooth [30]. Differently, generative methods create visually appealing textures but also introduce artifacts that compromise content fidelity [33]. The challenge becomes even more complex in real-world ISR, which must handle diverse, unknown degradations [59, 76].

Early methods [9, 12, 26, 28, 54, 79] employed deep neural networks trained with pixel-level losses, but struggled with real-world degradations, producing outputs that lacked perceptual realism. GAN-based approaches [30,33,34,59,60] significantly enhanced visual realism by aligning outputs with natural image distributions. However, their limited generative capacity and training instability often lead to unnatural artifacts. Recent text-to-image diffusion models, particularly Stable Diffusion (SD), have transformed real-world ISR [35,51,58,67,69–71] by providing powerful generative prior knowledge from vast image datasets and yielding more realistic SR outputs than GAN-based methods. However, the iterative sampling process inherent to dynamic models typically requires multiple sequential steps, imposing significant latency that hinders practical deployment.

Recent efforts [19,51,61,66,70] address these limitations via one-step diffusion models, achieved either by distilling multi-step diffusion models or fine-tuning pre-trained diffusion models with Low-Rank Adaptation (LoRA). OSEDiff [66] improves the process by inputting LR latent features and uses Variational Score Distillation to condense multi-step diffusion into a single step via LoRA fine-tuning. TVT [70] addresses fine-structure preservation by transferring 8× downsampled VAE of SD to a variant that supports 4× downsampling. PiSA-SR [51] pushes the field forward by using a dual-LoRA framework that decouples pixel-level regression from semantic enhancement, better improving both objectives.

However, these methods still face two main challenges. First, many monolithic architectures apply a uniform balance of fidelity and perceptual quality across different degradation types, failing to adapt to complex scenarios where conflicting restoration needs exist. Second, jointly optimizing pixel-level fidelity

and semantic-level enhancement tends to entangle the two objectives, making it difficult to balance faithful structure preservation and perceptual detail synthesis. PiSA-SR [51] addresses this issue by decoupling them into two LoRA modules and freezing the pixel branch during semantic training. However, we argue that this passive decoupling is still insufficient because it does not explicitly control the geometry of semantic updates. Even with the pixel weights frozen, the semantic branch can receive update components along the pixel-fidelity subspace and learn features already covered by the pixel branch, leading to redundant subspace overlap. This redundancy hampers the model's ability to fully utilize its capacity for genuine perceptual enhancement.

To mitigate these limitations, we propose FreqOrtho-SR, a novel framework for real-world SR with two key innovations: 1) *Frequency-Guided Mixture of Experts* (FreqMoE) employs multiple LoRA experts tailored to specific degradation patterns. Using a lightweight Fast Fourier Transform (FFT) based [8] degradation feature extractor to guide a gating network that routes inputs to appropriate experts, enabling stable and interpretable specialization. 2) *Orthogonal Gradient Projection* (OGP) to mitigate pixel-subspace gradient interference in multi-objective optimization. Inspired by continual learning techniques, we view fidelity preservation and perceptual enhancement as sequential tasks requiring orthogonality. By projecting away semantic-gradient components that lie in the pixel-level subspace, OGP enables the semantic LoRA to learn in an independent subspace, resulting in significantly improved perceptual quality with minimal fidelity trade-off.

Our main contributions can be summarized as follows:

1. We propose FreqOrtho-SR, a novel method for degradation-adaptive optimization that introduces a frequency-aware gating mechanism based on explicit FFT-based degradation signatures. This enables stable and interpretable expert specialization, where different experts operate at distinct fidelity-perceptual levels tailored to specific degradation characteristics.
2. We introduce orthogonal gradient projection from continual learning to real-world ISR task, reducing overlap between pixel and semantic subspaces and enabling the semantic LoRA to learn more effectively in an independent subspace.
3. Our extensive experiments show that FreqOrtho-SR achieves competitive overall performance, with best or second-best results on several key image quality metrics across multiple datasets.

2 Related Work

2.1 Diffusion Model-based Super-Resolution

Latent Diffusion Models (LDMs) [47] extend DDPMs [21] by performing denoising in a compressed latent space, yielding the Stable Diffusion (SD) model whose generative priors are now widely used for real-world SR.

Multi-step methods employ these priors at 20–50 denoising steps. StableSR [58] injects LR features into a frozen SD model via spatial feature transform layers; DiffBIR [35] first removes degradations and then enhances details with ControlNet [77]; PASD [69] adds pixel-aware cross-attention for structure guidance; and SeeSR [67] steers generation with degradation-aware text prompts. Despite strong visual quality, iterative sampling remains a practical bottleneck. One-step alternatives address this cost via consistency distillation (SinSR [61]), residual-shifting Markov chains (ResShift [74]), Variational Score Distillation with LoRA [22] tuning (OSEDiff [66]), adversarial diffusion distillation (AddSR [52]), target score distillation (TSD-SR [13]), diffusion-inversion noise prediction (InvSR [73]), VAE transfer from 8× to 4× downsampling (TVT [70]), dual-LoRA decoupling of pixel regression and semantic enhancement (PiSA-SR [51]), and CLIP-guided mixture-of-ranks routing (MoR-DASR [19]). Unlike MoR-DASR, which routes via a frozen CLIP encoder with predefined prompt pairs, our FreqMoE employs non-parametric FFT-based degradation signatures for interpretable routing without external pretrained models.

However, all these methods face a fundamental contradiction between pixel-level fidelity and semantic-level enhancement. Monolithic architectures (*e.g.*, TVT [70], PiSA-SR [51], OSEDiff [66]) lack the flexibility to adapt to diverse real-world degradations, while dual-branch designs (*e.g.*, PiSA-SR [51]) can suffer from subspace overlap between pixel- and semantic-level objectives. FreqOrtho-SR addresses both issues via frequency-guided expert specialization and orthogonal constraint theory, transforming the dual-branch architecture from a rigid structure into a dynamic, mathematically-constrained framework.

2.2 Frequency Analysis in Image Processing

Frequency domain analysis has long been a fundamental aspect of image processing [7,31,43,46,54,55]. Our insight is that different types of degradation exhibit unique frequency patterns. For example, blur suppresses high frequencies [29,42], Gaussian noise elevates all frequencies uniformly [4,17], JPEG compression introduces artifacts in 8×8 blocks that are visible in the frequency spectrum [39], and downsampling creates distinct aliasing patterns [3]. These observations have been utilized for various tasks such as blur detection [16,56], noise estimation [37,45], and quality assessment [41, 50]. Recent methods have incorporated frequency information through spectral losses [23] and frequency-domain adversarial training [14]. However, frequency features have not been systematically applied to routing in mixture of experts (MoE) architectures, which remains a significant challenge for effective routing. Traditional gating based on spatial features may suffer from training instability and load imbalance [15,49], leading some experts to dominate routing while others remain underutilized. Additionally, learned spatial representations may struggle to capture underlying degradation patterns without explicit structural priors [44, 80]. We address this gap by integrating explicit frequency-based degradation features into the gating mechanism.

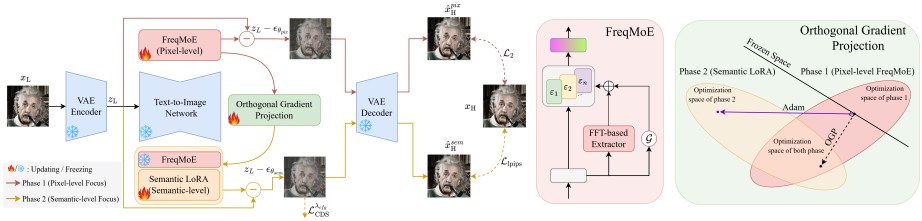

Fig. 2: Overall architecture of our FreqOrtho-SR. The FreqMoE and Semantic LoRA modules are optimized for pixel-level and semantic-level enhancements, respectively. FreqMoE utilizes the power of Mixture of LoRA Experts to adaptively handle diverse degradation types via FFT-based gating, while Orthogonal Gradient Projection ensures the semantic LoRA learns in a subspace orthogonal to the pixel-level representations, reducing subspace interference between fidelity and perceptual objectives.

2.3 Continual Learning and Gradient Projection

Continual learning aims to acquire sequential tasks without forgetting previously acquired knowledge [18,40]. One prominent method is Gradient Projection Memory (GPM) [48], which preserves task-specific knowledge by constraining gradient updates via subspace projection. After learning task A, GPM extracts important parameter directions using Singular Value Decomposition (SVD) [48] and projects the gradients of subsequent tasks into the null space of this subspace to prevent interference [6,75]. Related methods include Orthogonal Weights Modification [75] and Gradient Episodic Memory [38], both of which constrain gradient updates to preserve prior knowledge. While these methods are widely used in classification [6, 48] and reinforcement learning [65], they have not yet been explored in ISR. Therefore, we propose orthogonal projection to real-world ISR, treating pixel-level fidelity and semantic-level enhancement as sequential learning tasks. By extracting pixel-level LoRA subspaces via SVD and projecting semantic gradients into the null space, we remove pixel-subspace gradient components that would otherwise drive redundant semantic updates. This enables semantic LoRA to learn in an independent subspace, thereby improving perceptual quality while maintaining competitive fidelity.

3 Methodology

This section first formulates the SD-based SR as a residual learning model. It then introduces our proposed FreqOrtho-SR, a novel framework designed for true subspace disentanglement and adaptive restoration in real-world ISR, aiming to achieve a better balance between fidelity and perceptual quality. In this study, we denote the low-resolution (LR) and high-resolution (HR) images as x_L and x_H. Their corresponding latent codes are represented as $z_L = \mathcal{E}(x_L)$ and $z_H = \mathcal{E}(x_H)$. We can approximate that $x_L \approx \mathcal{D}(z_L)$ and $x_H \approx \mathcal{D}(z_H)$, where $\mathcal{E}$ and $\mathcal{D}$ are the encoder and decoder of a pre-trained Variational Autoencoder (VAE).

3.1 Model Formulation

Multi-step DM-based SR methods [35,58,67] perform T-step denoising to transform Gaussian noise z_T into the HR latent z_H, conditioned on x_L. This iterative process is computationally expensive and causes instability from random noise sampling. In this work, we adopt a one-step approach that starts directly from z_L and leverages residual learning scheme: $z_H = z_L - \epsilon_\theta(z_L)$, where ϵ_θ is the SD UNet parameterized by θ, and the noise schedule coefficients are absorbed into ϵ_θ following [51, 66]. This forces the network to learn the high-frequency difference between z_L and z_H, accelerating training convergence.

Following the dual-LoRA paradigm [51], we introduce two sets of LoRA [22] modules upon the frozen SD weights θ_{sd}: a *pixel-level* LoRA $\Delta\theta_{pix}$ trained with $\mathcal{L}_2$ loss for degradation removal, and a *semantic-level* LoRA $\Delta\theta_{sem}$ trained with $\mathcal{L}_2$, $\mathcal{L}_{lpips}$ [78], and $\mathcal{L}_{csd}$ [51, 72] losses for detail and semantic enhancement. The pixel-level LoRA is optimized first, then the semantic-level LoRA is trained while $\Delta\theta_{pix}$ remains frozen as follows:

$$z_H^{pix} = z_L - \epsilon_{\theta_{pix}}(z_L), \quad \theta_{pix} = \{\theta_{sd}, \Delta\theta_{pix}\}, \tag{1}$$

$$z_H^{sem} = z_L - \epsilon_{\theta_{full}}(z_L), \quad \theta_{full} = \{\theta_{sd}, \Delta\theta_{pix}, \Delta\theta_{sem}\}. \tag{2}$$

The decoded outputs $\hat{x}_H^{pix} = \mathcal{D}(z_H^{pix})$ and $\hat{x}_H^{sem} = \mathcal{D}(z_H^{sem})$ are used for loss computation in their respective phases.

Our overall framework is illustrated in Fig. 2, which extends dual-LoRA paradigm in two ways. First, we replace the single pixel-level LoRA $\Delta\theta_{pix}$ with a robust Frequency-guided Mixture of LoRA Experts (Sec. 3.2), enabling degradation-adaptive restoration rather than applying a static uniform approach. Second, beyond simply freezing $\Delta\theta_{pix}$ during semantic training, we additionally employ Orthogonal Gradient Projection (Sec. 3.3) to constrain semantic gradients to the null space of the pixel-level subspace, providing a mathematically stronger guarantee against fidelity degradation.

3.2 Frequency-guided Mixture of LoRA Experts

Recent DM-based methods [51, 66] employ a single LoRA adapter for pixel-level restoration, applying a uniform strategy regardless of degradation type. Since real-world images exhibit diverse, mixed degradations requiring distinct frequency-domain restoration behaviors (Fig. 3a) [59, 67], a single LoRA lacks the capacity to specialize across such heterogeneous patterns. To address this challenge, we replace the monolithic pixel-level LoRA $\Delta\theta_{pix}$ with a Frequency-guided Mixture of LoRA Experts (FreqMoE; Fig. 3b). Specifically, each target layer in the UNet now contains N parallel LoRA expert pairs $\{(\mathbf{A}_i, \mathbf{B}_i)\}_{i=1}^{N}$, where $\mathbf{A}_i \in \mathbb{R}^{r \times d_{in}}$ and $\mathbf{B}_i \in \mathbb{R}^{d_{out} \times r}$ are the low-rank down- and up-projection matrices with rank r, respectively. A gating network $\mathcal{G}$ routes each input to the top-k most suitable experts using both local spatial features from the layer input and global frequency-domain degradation features extracted from x_L. The

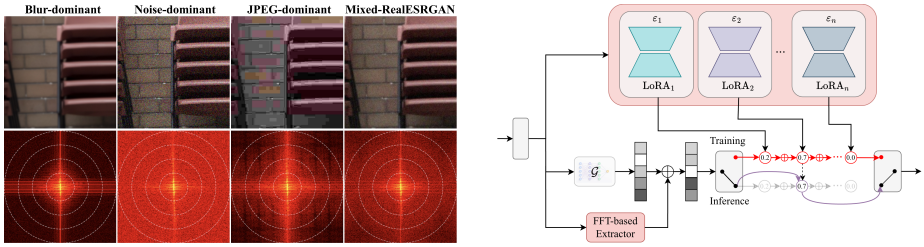

Fig. 3: (a) FFT magnitude spectra (log scale) of LR images under different degradations. Blur suppresses high-frequency energy (dim outer ring), noise raises it uniformly (bright outer ring), and JPEG introduces periodic block artifacts (grid pattern). These signatures enable our non-parametric extractor to reliably identify degradation types. (b) Structure of the FreqMoE module which routes inputs to specialized LoRA experts based on frequency-domain degradation features.

FreqMoE output for a given layer is represented as follows:

$$\Delta \mathbf{h}_{pix} = \sum_{i \in \text{Top-}k(\boldsymbol{\ell})} g_i(\mathbf{x}, \mathbf{f}) \cdot \mathbf{B}_i \mathbf{A}_i \mathbf{x}, \tag{3}$$

where $\mathbf{x}$ is the layer input, $\mathbf{f} \in \mathbb{R}^{d_f}$ is the degradation feature vector extracted from x_L, $\boldsymbol{\ell}$ are the gating logits, and $g_i(\mathbf{x}, \mathbf{f})$ is the routing weight for expert i among the selected top-k experts. This sparse routing reduces computational cost while enabling complementary experts to collaborate when beneficial. Ablations on the number of experts, top-k routing, and expert routing behavior are provided in Sec. 11 of the **supplementary materials**.

Degradation feature extraction. We design a lightweight, non-parametric feature extractor that captures degradation signatures from the LR input x_L in the frequency domain. Real-world LR images are typically corrupted by a composition of multiple degradation types. Following the second-order degradation model of RealESRGAN [59], which cascades blur, resize, noise, and JPEG compression in two sequential rounds, each corruption leaves a characteristic fingerprint in the frequency spectrum (Fig. 3a): blur suppresses high-frequency content [29], noise elevates energy uniformly across frequencies [4], JPEG compression introduces periodic artifacts at 8×8 block boundaries [39], and resize operations produce aliasing patterns visible in the radial energy distribution [3].

Given an input image $x_L \in \mathbb{R}^{C \times H \times W}$, we first convert it to a grayscale image x_g and compute its centered 2D FFT magnitude spectrum $\mathbf{M} = |\mathcal{F}(x_g)|$, where $\mathcal{F}$ is the shift-centered 2D FFT magnitude operator. The spectrum is divided into K concentric radial bands $\{R_k\}_{k=1}^{K}$ from the center (low frequency) to the edges (high frequency), where each band spans the radial range $[(k-1)\rho/K,\ k\rho/K)$ and ρ is the maximum radius. The normalized energy for each band is:

$$\bar{M}_k = \frac{1}{|R_k|} \sum_{(u,v) \in R_k} \mathbf{M}(u, v), \quad e_k = \frac{\bar{M}_k}{\sum_{k'=1}^{K} \bar{M}_{k'}}, \tag{4}$$

where $|R_k|$ is the pixel count of band R_k. This area-normalized mean prevents larger outer bands from dominating, yielding a probability distribution $\mathbf{e} = [e_1, \ldots, e_K]$. We further compute three scalar degradation indicators:

1. *Blur score*: $s_{blur} = e_1/(e_K + \epsilon)$, measuring the ratio of low- to high-frequency energy. A high score indicates blur-typical low-frequency dominance.
2. *JPEG blocking score*: $s_{jpeg} = \bar{g}_{block}/(\bar{g}_{all} + \epsilon)$, where $\bar{g}_{block}$ is the mean absolute gradient at 8×8 block boundaries and $\bar{g}_{all}$ is the overall mean gradient. Elevated boundary gradients signal JPEG compression artifacts.
3. *Noise score*: $s_{noise} = \text{mean}(|\mathcal{L} * x_L|)$, where $\mathcal{L}$ is the 3×3 Laplacian kernel applied via depthwise convolution. A high Laplacian response indicates the presence of high-frequency noise.

The final feature vector $\mathbf{f} = [\mathbf{e};\, s_{blur};\, s_{jpeg};\, s_{noise}] \in \mathbb{R}^{d_f}$ $(d_f = K+3)$ concatenates the band energies with the three scalar scores. While the three scalar scores target the dominant degradation types, the radial band energies $\mathbf{e}$ implicitly capture other corruptions such as resize-induced aliasing, which manifests as characteristic shifts in the radial energy distribution. Together, $\mathbf{f}$ provides a compact yet comprehensive degradation descriptor. This extractor is non-parametric and computed with gradients disabled, achieving negligible overhead. The effect of the number of frequency bands K is analyzed in Sec. 11 of the **supplementary materials**.

Frequency-modulated gating network. The gating network $\mathcal{G}$ combines spatial information from the layer activations with the global degradation features $\mathbf{f}$ to produce routing weights. Given the layer input $\mathbf{x} \in \mathbb{R}^{d_{in}}$ and degradation features $\mathbf{f} \in \mathbb{R}^{d_f}$, the gating logits are computed as:

$$\boldsymbol{\ell} = \mathbf{W}_g \mathbf{x} + \phi(\mathbf{f}), \tag{5}$$

where $\mathbf{W}_g \in \mathbb{R}^{N \times d_{in}}$ is the base gate projection and $\phi : \mathbb{R}^{d_f} \to \mathbb{R}^N$ is a small MLP consisting of two linear layers with a ReLU activation: $\phi(\mathbf{f}) = \mathbf{W}_2\, \text{ReLU}(\mathbf{W}_1 \mathbf{f})$, with $\mathbf{W}_1 \in \mathbb{R}^{2d_f \times d_f}$ and $\mathbf{W}_2 \in \mathbb{R}^{N \times 2d_f}$.

This additive design allows $\phi(\mathbf{f})$ to act as a degradation-dependent bias that shifts the base routing logits. Since $\mathbf{f}$ is computed once per image and shared across all spatial positions and layers, the frequency bias provides a consistent global routing signal. Experts are not pre-assigned to specific degradation types; instead, they emerge through end-to-end training guided by the frequency features. For instance, experts may specialize in high-frequency recovery (activated for blurred inputs) or denoising (activated for noisy inputs), while mixed-degradation images activate multiple experts via top-k routing.

For routing, we select the top-k experts per token. The gating weights are obtained by applying softmax over the selected logits:

$$g_i(\mathbf{x}, \mathbf{f}) = \begin{cases} \frac{\exp(\ell_i)}{\sum_{j \in \text{Top-}k} \exp(\ell_j)}, & \text{if } i \in \text{Top-}k(\boldsymbol{\ell}) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Top-$k(\boldsymbol{\ell})$ returns the indices of k largest logits. This sparse routing lowers computation over dense mixtures while still enabling complementary expert collaboration when useful. For stable training, we initialize $\mathbf{W}_g$ with small random values (σ=0.01) and set the output layer of ϕ to zero, so the gating starts from near-uniform routing and gradually learns degradation-specific specialization.

To address expert collapse in MoE frameworks [11, 15, 36], we employ the Load Balancing Loss (LBL) [15], which encourages uniform expert utilization by linking the frequency of expert selection to the routing weights each expert receives. Let f_i be the fraction of tokens directed to expert i, P_i denote the average routing probability for expert i, and N be the total number of experts. The LBL is defined as: $\mathcal{L}_{lbl} = \alpha \cdot N \cdot \sum_{i=1}^{N} f_i \cdot P_i$.

3.3 Orthogonal Gradient Projection

Problem formulation. A core challenge in decoupled real-world ISR is ensuring that distinct adapters capture truly independent features. We have identified a fundamental phenomenon in subspace interference: when a semantic-level adapter is trained on top of an existing fidelity-level module, the optimization process often collapses into the subspace already covered by the fidelity weights.

Formally, consider the joint inference state $\theta_{full} = \{\theta_{sd}, \Delta\theta_{pix}, \Delta\theta_{sem}\}$. While $\Delta\theta_{pix}$ is optimized for structural alignment, the subsequent optimization of $\Delta\theta_{sem}$ lacks a mechanism to prevent it from re-learning structural features that are already encoded in $\Delta\theta_{pix}$. If the update directions of the semantic LoRA are not constrained to be orthogonal to the fidelity subspace, the model suffers from representational redundancy. This redundancy wastes the model's limited rank capacity on repetitive structural information instead of capturing novel perceptual textures. This prevents the model from reaching the optimal or near-optimal boundary of the perception-fidelity trade-off, irrespective of the parameter budget. To address this issue, we propose explicitly isolating these objectives by projecting semantic updates into the null space of the fidelity manifold.

Subspace extraction via SVD. After the first phase, we extract the pixel-level subspace for each layer using Singular Value Decomposition (SVD) [48]. For a layer with weight W, the pixel LoRA introduces a change $\Delta W_{pix} = B_{pix}A_{pix} \in \mathbb{R}^{d_{out} \times d_{in}}$. In a MoE framework with N experts, we concatenate all expert weight changes to capture the union of their subspaces:

$$\Delta W_{all} = [\Delta W_1 \mid \Delta W_2 \mid \cdots \mid \Delta W_N] \in \mathbb{R}^{d_{out} \times (N \cdot d_{in})}$$

We then perform SVD on $\Delta W_{all} = U\Sigma V^T$ and retain the top $\tilde{k}$ singular vectors that account for 95% of the energy:

$$\tilde{k} = \arg\min_{\tilde{k}} \left\{ \frac{\sum_{i=1}^{\tilde{k}} \sigma_i^2}{\sum_j \sigma_j^2} \geq 0.95 \right\}$$

The truncated left singular matrix $U_{\tilde{k}} \in \mathbb{R}^{d_{out} \times \tilde{k}}$ defines the essential pixel subspace preserved for Phase 2. This SVD is performed only once offline between

training phases, not per iteration, and operates on relatively small matrices. On our hardware, the entire SVD extraction across all layers completes in under 5 seconds, introducing negligible overhead relative to the total training time. Ablations on the SVD energy threshold are provided in Sec. 11 of the **supplementary materials**.

Gradient projection during semantic LoRA training. During the second phase, we register backward hooks on all semantic LoRA B weight matrices. When gradients are computed, the hook projects them into the null space of $U_{\tilde{k}}$:

$$G_{ortho} = G - U_{\tilde{k}}(U_{\tilde{k}}^T G), \tag{7}$$

where $G \in \mathbb{R}^{d_{out} \times r}$ is the original gradient of the LoRA B weight. Since $U_{\tilde{k}}$ has orthonormal columns ($U_{\tilde{k}}^T U_{\tilde{k}} = I$), it follows that $U_{\tilde{k}}^T G_{ortho} = 0$, ensuring that the semantic gradient has zero component along the pixel-level subspace directions and removing pixel-subspace gradient interference. As a result, the semantic LoRA is guided to learn in a subspace independent of the pixel-level representations, enabling more effective perceptual enhancement. The projection is computationally efficient with negligible overhead. The projection target choice is further justified and ablated in Sec. 11 of the **supplementary materials**.

3.4 Training Strategy

The loss design in both phases builds upon the combination of $\mathcal{L}_2$, $\mathcal{L}_{lpips}$, and $\mathcal{L}_{csd}$ losses, which has been shown effective for balancing pixel-level fidelity and semantic enhancement in diffusion-based SR [51,66]. We further use an intermediate SVD extraction step to enable orthogonal gradient projection.

Phase 1 - FreqMoE Training. In the first phase, we train the FreqMoE module using the $\mathcal{L}_2$ loss to measure pixel-level fidelity between the predicted HQ image and ground truth, combined with the load balancing loss $\mathcal{L}_{lbl}$ (Sec. 3.2) to encourage balanced expert utilization. All FreqMoE parameters $\Delta\theta_{pix}$ are updated in this phase while the pre-trained SD parameters (θ_{sd}) remain frozen.

SVD-based subspace extraction. Between the two training phases, we extract the pixel-level subspace from the learned FreqMoE weights via SVD as described in Sec. 3.3. The resulting orthogonal bases $U_{\tilde{k}}$ are stored and used to construct the projection operators for Phase 2.

Phase 2 - Semantic LoRA Training with OGP. In this phase, we train the semantic-level LoRA $\{\Delta\theta_{sem}\}$ while keeping both the pre-trained SD θ_{sd} and the pixel-level FreqMoE ($\Delta\theta_{pix}$) frozen. The training objective combines three complementary losses: $\mathcal{L}_2$ loss to maintain pixel-level consistency, $\mathcal{L}_{lpips}$ loss to align high-level features with a pre-trained VGG for perceptual quality, and $\mathcal{L}_{csd}$ loss to leverage semantic priors from the pre-trained SD for detail enhancement. Detailed loss configurations are provided in Sec. 6 of the **supplementary materials**. Critically, during backpropagation, we employ the orthogonal gradient projection (Sec. 3.3) to all gradients flowing into the semantic LoRA B matrices,

ensuring the semantic LoRA learns in a subspace orthogonal to the pixel-level representations.

Inference. In the default setting, both the FreqMoE and semantic LoRA are active, and the SR output is produced in a single forward pass via $z_H = z_L - \epsilon_{\theta_{full}}(z_L)$. Our dual-LoRA design also supports an adjustable mode by introducing pixel-level and semantic-level guidance scales λ_{pix} and λ_{sem}:

$$\epsilon_\theta(z_L) = \lambda_{pix}\,\epsilon_{\theta_{pix}}(z_L) + \lambda_{sem}\big(\epsilon_{\theta_{full}}(z_L) - \epsilon_{\theta_{pix}}(z_L)\big), \tag{8}$$

where $\epsilon_{\theta_{pix}}(z_L)$ is the output with only the FreqMoE active, and $\epsilon_{\theta_{full}}(z_L) - \epsilon_{\theta_{pix}}(z_L)$ isolates the semantic-level contribution. Setting λ_{pix}=λ_{sem}=1 recovers the default single-pass mode. This adjustable mode requires two forward passes but enables controllable fidelity-perception trade-off without re-training.

4 Experiments

4.1 Experimental Settings

Training settings. We train FreqOrtho-SR using the SD 2.1-base [47] for the $\times$4 SR task. A pixel-level FreqMoE and semantic LoRA modules are applied to the weights of all convolutional and MLP layers, both initialized using a Gaussian distribution with a rank of 4. The FreqMoE uses N=4 experts with top-k=2 routing, and the frequency band number is set to K=4 (d_f=7). Following recent methods [19, 51, 66, 70], we use LSDIR [32] and the first 10K images from FFHQ [24] dataset as training data. We generate paired training data using RealESRGAN's degradation pipeline [59]. The batch size and patch size are set to 16 and 512×512. We use the AdamW optimizer [27] with a learning rate of 5e-5. We train Phase 1 and Phase 2 for 4K and 26K iterations, respectively.

Compared methods. We compare FreqOrtho-SR with several recent one-step DM-based methods: AddSR [52], SinSR [61], OSEDiff [66], MoR-DASR [19], PiSA-SR [51], and TVT [70]. Additional comparisons with GAN-based methods (BSRGAN [76], RealESRGAN [59], LDL [33]) and multi-step DM-based methods (DiffBIR [35], PASD [69], SeeSR [67]) are provided in Sec. 7 and Sec. 8 of the **supplementary materials**, respectively. All comparative results are obtained from officially released codes.

Test datasets and metrics. Following prior work [19, 51], we test on both synthetic and real-world data. Synthetic test samples are obtained by applying RealESRGAN's degradation pipeline [59] to DIV2K dataset [1]. Real-world data are center-cropped from the RealSR [5] and DRealSR [64] datasets. To measure fidelity, we use PSNR and SSIM [63], computed on the Y channel in YCbCr color space. We also evaluate perceptual quality using LPIPS [78] and DISTS [10], both computed in RGB space. FID [20] is used to evaluate the distribution similarity between GT and SR images. For image quality assessment without reference GT, we use NIQE [41], CLIPIQA [57], MUSIQ [25], and MANIQA [68].

Table 1: Quantitative comparison among the state-of-the-art one-step DM-based SR methods on synthetic and real-world test datasets. The best and the second-best results are highlighted in red and blue, respectively. "-" indicates metrics not reported in the original paper due to unavailable code.

Datasets	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
RealSR	AddSR	23.12	0.6550	0.3090	0.2703	154.14	6.66	0.5520	67.14	0.4880
	SinSR	26.30	0.7354	0.3212	0.2346	137.05	6.31	0.6204	60.41	0.5389
	OSEDiff	25.15	0.7341	0.2921	0.2128	123.50	5.65	0.6693	69.09	0.6339
	MoR-DASR	25.32	0.7280	0.2910	-	-	-	0.6910	69.78	0.512
	PiSA-SR	25.50	0.7417	0.2672	0.2044	124.09	5.50	0.6702	70.15	0.6560
	TVT	25.81	0.7596	0.2587	0.2061	109.93	5.92	0.6882	69.89	0.6228
	FreqOrtho-SR	26.27	0.7481	0.2537	0.1951	108.91	5.32	0.6630	69.39	0.6586
DRealSR	AddSR	26.71	0.7380	0.3210	0.2631	164.79	7.72	0.5930	62.13	0.4580
	SinSR	28.41	0.7495	0.3741	0.2488	177.05	7.02	0.6367	55.34	0.4898
	OSEDiff	27.92	0.7835	0.2968	0.2165	135.29	6.49	0.6963	64.65	0.5899
	MoR-DASR	28.37	0.7760	0.3070	-	-	-	0.7170	65.94	0.5090
	PiSA-SR	28.31	0.7804	0.2960	0.2169	130.61	6.20	0.6970	66.11	0.6156
	TVT	28.27	0.7899	0.2900	0.2205	134.19	7.02	0.7220	65.56	0.5783
	FreqOrtho-SR	29.08	0.7911	0.2799	0.2110	130.85	6.29	0.6879	65.89	0.6182
DIV2K	AddSR	23.26	0.5900	0.3620	0.2344	35.03	5.82	0.5730	63.39	0.4050
	SinSR	24.43	0.6012	0.3262	0.2066	35.45	6.02	0.6499	62.80	0.5395
	OSEDiff	23.72	0.6108	0.2941	0.1976	26.32	4.71	0.6683	67.97	0.6148
	MoR-DASR	24.01	0.6060	0.2890	-	-	-	0.6810	68.09	0.4750
	PiSA-SR	23.87	0.6058	0.2823	0.1934	25.07	4.55	0.6927	69.68	0.6400
	TVT	24.23	0.6292	0.2773	0.1860	24.78	5.60	0.6986	68.67	0.6037
	FreqOrtho-SR	24.35	0.6189	0.2748	0.1889	23.54	4.63	0.6702	68.67	0.6355

4.2 Comparisons with State-of-the-Art Methods

Quantitative results. Table 1 compares the performance of our FreqOrtho-SR with SOTA one-step DM-based SR methods across both synthetic and real-world test datasets. From this comparison, we can make the following observations:

- First, our method demonstrates clear advantages over competing methods in full-reference fidelity metrics, such as PSNR and SSIM, as well as in perceptual quality metrics including LPIPS and DISTS. Notably, these advantages are especially evident on the two real-world datasets, DRealSR and RealSR.
- Second, FreqOrtho-SR consistently performs well on non-reference metrics like NIQE and MANIQA, which are widely recognized for evaluating real-world ISR. Although PiSA-SR surpasses us on one non-reference metric (MUSIQ), our performance on this metric remains comparable. This is satisfactory given that we excel on most of the metrics listed in Table 1.

Notably, FreqOrtho-SR manages to achieve competitive results on both fidelity and perceptual quality metrics, which is a significant challenge in real-world ISR. This is achieved through our proposed components: Frequency-guided Mixture of LoRA Experts and orthogonal gradient projection, in a two-phase training setup that mimics a sequential task setup. Generally, there is no single method that consistently improves performance at all metrics; results vary substantially

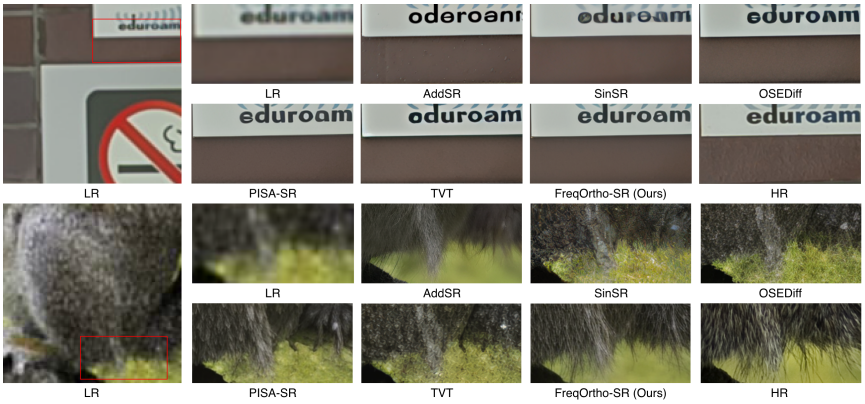

Fig. 4: Visual comparisons of one-step DM-based SR methods. Zoom in for best view.

across metrics. In contrast, FreqOrtho-SR achieves the best or second-best in several cases, indicating substantial improvements in the real-world ISR field.

Qualitative results. Fig. 4 presents visual comparisons of FreqOrtho-SR with SOTA one-step DM-based SR methods. AddSR [52] and SinSR [61] produce over-smoothed textures, failing to recover fine-grained details. OSEDiff [66] generates more consistent outputs but with limited semantic details. While PiSA-SR [51] and TVT [70] can restore richer textures, they occasionally introduce visual artifacts or structural distortions. In contrast, FreqOrtho-SR reconstructs more accurate structures (*e.g.*, the sharp text in the first example) and produces more natural, realistic details (*e.g.*, the fine textures in the second example), benefiting from frequency-guided expert learning that effectively disentangles pixel-level and semantic-level enhancements. As described in Sec. 3, our dual-LoRA design naturally supports an adjustable inference mode (Eq. (8)), enabling controllable fidelity-perception trade-off via pixel-semantic guidance scales. Detailed adjustable SR experiments are provided in Sec. 9 of the **supplementary materials**.

Complexity analysis. Table 2 reports run time, model size, and representative metrics of recent SOTA methods on RealSR dataset. FreqOrtho-SR has same parameter count as PiSA-SR (1.30 B); lower than OSEDiff (1.77 B) and TVT (1.72 B). For FLOPs/peak memory, PiSA-SR, TVT, and FreqOrtho-SR require 2.23T/3.00 GB, 2.97T/7.31 GB, and 2.25T/3.03 GB, respectively. Thus, our activated compute and memory remain close to PiSA-SR, while being much lower

Table 2: Complexity and performance of one-step DM-based methods. Time on 128×128 inputs (A100). LPIPS/FID on RealSR. †: no code.

Method	Time (s)	Params (B)	LPIPS↓	FID↓
SinSR	0.13	0.18	0.3121	137.05
OSEDiff	0.12	1.77	0.2921	123.50
MoR-DASR†	-	-	0.2910	-
PiSA-SR	0.09	1.30	0.2672	124.09
TVT	0.28	1.72	0.2587	109.93
Ours	0.30	1.30	0.2537	108.91

Table 3: OOD evaluation on RealLR200 [67] (real-world, no GT). The image shows an in-the-wild LR input (left) and our output (right).

Method	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
PiSA-SR	4.00	0.6869	70.5058	0.6271
TVT	4.73	0.6909	70.5275	0.5954
Ours	3.90	0.6912	70.6040	0.6321

than TVT in memory. The additional inference time stems from FreqMoE routing: unlike a single LoRA that can merge into the base weights for free, MoE structure requires computing top-k experts per token along with the lightweight FFT-based gating, yielding a higher run time. Despite this overhead, FreqOrtho-SR has comparable inference time compared to the current SOTA model TVT [70] while achieving the best LPIPS and FID among all compared methods, yielding a favorable quality-efficiency trade-off. We note that MoE inference overhead can potentially be reduced via expert pruning or structured sparsity that we leave for future work.

OOD real-world evaluation. To further evaluate robustness to unknown real-world degradations, we test on RealLR200 [67], a no-reference real LR benchmark without ground-truth HR images. As shown in Table 3, FreqOrtho-SR achieves the best no-reference scores among the strongest one-step baselines, supporting its generalization beyond the RealESRGAN degradation pipeline.

4.3 Ablation Study

We conduct ablation studies on the RealSR dataset to validate all proposed components of FreqOrtho-SR (Table 4) by progressively removing each component. For a fair comparison, we run all experiments on the same standard settings. Additional branch-design ablations, including applying FreqMoE to both the pixel and semantic branches, are reported in Sec. 11 of the **supplementary materials**.

FreqMoE with spatial-only gating (i.e., without frequency features) in the second row already improves fidelity (PSNR +0.87 dB, SSIM +0.015), showing the benefit of multi-expert capacity. Adding frequency-guided routing (first row) further improves LPIPS and FID, confirming that FFT-based degradation signatures provide a more discriminative routing signal than spatial features alone. OGP (third row) also independently reduces FID, even without multi-expert routing. The complete FreqOrtho-SR (last row), which combines FreqMoE and OGP, achieves the best DISTS, FID, NIQE, and MANIQA scores among all variants, with a slight PSNR trade-off as OGP constrains the semantic LoRA to an independent subspace. The significant FID improvement confirms better output distribution, validating that OGP effectively mitigates subspace-level redundancy.

Table 4: Ablations on the RealSR dataset. We validate the effectiveness of each proposed component. The best results are highlighted in red.

MoE	Freq	OGP	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	MANIQA↑
√	√	×	26.3097	0.7520	0.2517	0.1968	111.45	5.4812	0.6510
√	×	×	26.3718	0.7571	0.2584	0.2017	118.35	5.678	0.6560
×	×	√	26.1992	0.7432	0.2651	0.2021	113.39	5.4390	0.6557
√	√	√	26.27	0.7481	0.2537	0.1951	108.91	5.3200	0.6586

Subspace orthogonality analysis. To directly verify that OGP enforces subspace independence and explain the perceptual gains when adding OGP to FreqMoE in Table 4 (row 1 vs. last row), we measure the per-layer projection energy ratio $\frac{\|U_k^{(l)\top} \Delta W_{sem}^{(l)}\|_F^2}{\|\Delta W_{sem}^{(l)}\|_F^2}$, which quantifies how much of the semantic LoRA weights lie within the pixel-level subspace (lower is better). As shown in Fig. 5, the overlap without OGP is highly non-uniform: the output layer reaches 87% and encoder attention layers up to 32%, meaning nearly all semantic capacity at those layers redundantly duplicates pixel-level features. OGP suppresses the overlap to near-zero across all 258 layers, confirming that the semantic LoRA learns in a truly independent subspace, thereby improving perceptual metrics DISTS, FID, NIQE, and MANIQA simultaneously.

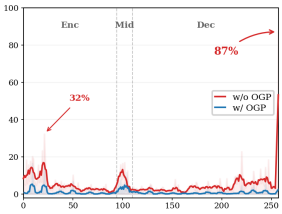

Fig. 5: Per-layer projection energy ratio onto the pixel-level subspace.

5 Conclusion

We propose FreqOrtho-SR, a one-step diffusion framework that introduces two novel principles for real-world image super-resolution. FreqMoE introduces frequency-domain degradation signatures as a principled routing signal for expert specialization, yielding stable and interpretable gating without additional learned feature extractors. OGP bridges continual learning with multi-objective SR optimization, providing a provable orthogonality guarantee that moves beyond simple parameter freezing. Together, these contributions demonstrate a strong fidelity-perception balance across RealSR, DRealSR, and DIV2K, with competitive or best performance on key metrics among one-step diffusion methods.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00573160); the Technology Innovation Program (RS-2025-02222776, Development and Demonstration of AI and Lightweight Technology-Based Automated E-Waste Sorting and Retrieval System) funded by the Ministry of Trade, Industry and Resources (MOTIR, Korea); the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2026-RS-2023-00259703); and the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT).

This work was also supported by Hyundai Motor Chung Mong-Koo Global Scholarship to Dinh Phu Tran (co-first author, equal contribution).

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
2. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6228–6237 (2018)
3. Blu, T., Thévenaz, P., Unser, M.: Linear interpolation revitalized. IEEE Transactions on Image Processing **13**(5), 710–719 (2004)
4. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale modeling & simulation **4**(2), 490–530 (2005)
5. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: 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)
6. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P., Torr, P., Ranzato, M.: Continual learning with tiny episodic memories. In: Workshop on Multi-Task and Lifelong Reinforcement Learning (2019)
7. Chen, K., Li, L., Liu, H., Li, Y., Tang, C., Chen, J.: Swinfsr: Stereo image super-resolution using swinir and frequency domain knowledge. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1764–1774 (2023)
8. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex fourier series. Mathematics of computation **19**(90), 297–301 (1965)
9. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11065–11074 (2019)
10. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. IEEE transactions on pattern analysis and machine intelligence **44**(5), 2567–2581 (2020)
11. Do, G., Le, H., Tran, T.: Simsmoe: Toward efficient training mixture of experts via solving representational collapse. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 2012–2025 (2025)
12. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence **38**(2), 295–307 (2015)
13. Dong, L., Fan, Q., Guo, Y., Wang, Z., Shan, Q., Li, J., Liu, J., Liao, Y., Cheng, S., Pei, S.: TSD-SR: One-step diffusion with target score distillation for real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
14. Durall, R., Keuper, M., Keuper, J.: Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7890–7899 (2020)
15. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)
16. Ferzli, R., Karam, L.J.: A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). IEEE transactions on image processing **18**(4), 717–728 (2009)

17. Foi, A., Katkovnik, V., Egiazarian, K.: Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images. IEEE transactions on image processing **16**(5), 1395–1411 (2007)
18. French, R.M.: Catastrophic forgetting in connectionist networks. Trends in cognitive sciences **3**(4), 128–135 (1999)
19. He, X., Tu, Z., Cheng, K., Zhu, M., Hu, J., Wang, N., Gao, X.: Mixture of ranks with degradation-aware routing for one-step real-world image super-resolution. arXiv preprint arXiv:2511.16024 (2025)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
22. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
23. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13919–13929 (2021)
24. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
25. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021)
26. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1646–1654 (2016)
27. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
28. Kong, X., Zhao, H., Qiao, Y., Dong, C.: Classsr: A general framework to accelerate super-resolution networks by data characteristic. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12016–12025 (2021)
29. Kundur, D., Hatzinakos, D.: Blind image deconvolution. IEEE signal processing magazine **13**(3), 43–64 (1996)
30. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017)
31. Li, X., Zhang, Y., Yuan, J., Lu, H., Zhu, Y.: Discrete cosin transformer: Image modeling from frequency domain. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5468–5478 (2023)
32. Li, Y., Zhang, K., Liang, J., Cao, J., Liu, C., Gong, R., Zhang, Y., Tang, H., Liu, Y., Demandolx, D., 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 (2023)
33. Liang, J., Zeng, H., Zhang, L.: 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)

34. Liang, J., Zeng, H., Zhang, L.: Efficient and degradation-adaptive network for real-world image super-resolution. In: European Conference on Computer Vision. pp. 574–591. Springer (2022)
35. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: Diffbir: Toward blind image restoration with generative diffusion prior. In: European conference on computer vision. pp. 430–448. Springer (2024)
36. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
37. Liu, X., Tanaka, M., Okutomi, M.: Single-image noise level estimation for blind denoising. IEEE transactions on image processing **22**(12), 5226–5237 (2013)
38. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. Advances in neural information processing systems **30** (2017)
39. Luo, W., Huang, J., Qiu, G.: Jpeg error analysis and its applications to digital image forensics. IEEE Transactions on Information Forensics and Security **5**(3), 480–491 (2010)
40. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989)
41. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters **20**(3), 209–212 (2012)
42. Narvekar, N.D., Karam, L.J.: A no-reference image blur metric based on the cumulative probability of blur detection (cpbd). IEEE Transactions on Image Processing **20**(9), 2678–2683 (2011)
43. Pratt, W.K.: Digital image processing: PIKS Scientific inside, vol. 4. Wiley Online Library (2007)
44. Puigcerver, J., Riquelme, C., Mustafa, B., Houlsby, N.: From sparse to soft mixtures of experts. arXiv preprint arXiv:2308.00951 (2023)
45. Pyatykh, S., Hesser, J., Zheng, L.: Image noise level estimation by principal component analysis. IEEE transactions on image processing **22**(2), 687–699 (2012)
46. Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J.: Global filter networks for image classification. Advances in neural information processing systems **34**, 980–993 (2021)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
48. Saha, G., Garg, I., Roy, K.: Gradient projection memory for continual learning. arXiv preprint arXiv:2103.09762 (2021)
49. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
50. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on image processing **15**(2), 430–444 (2006)
51. Sun, L., Wu, R., Ma, Z., Liu, S., Yi, Q., Zhang, L.: Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2333–2343 (2025)
52. Tai, Y., Xie, R., Zhao, C., Zhang, K., Zhang, Z., Zhou, J., Yang, J.: Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. Pattern Recognition p. 113012 (2026)
53. Tran, D.P., Do, T., Wazir, S., Kim, S., Kim, S.K., Kim, D.: Sat: Selective aggregation transformer for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4982–4992 (2026)

54. Tran, D.P., Hung, D.D., Kim, D.: Channel-partitioned windowed attention and frequency learning for single image super-resolution. arXiv preprint arXiv:2407.16232 (2024)
55. Tran, D.P., Hung, D.D., Kim, D.: Vsrm: A robust mamba-based framework for video super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14711–14721 (2025)
56. Vu, C.T., Phan, T.D., Chandler, D.M.: S_3: a spectral and spatial measure of local perceived sharpness in natural images. IEEE transactions on image processing **21**(3), 934–945 (2011)
57. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2555–2563 (2023)
58. Wang, J., Yue, Z., Zhou, S., Chan, K.C., Loy, C.C.: Exploiting diffusion prior for real-world image super-resolution. International Journal of Computer Vision **132**(12), 5929–5949 (2024)
59. Wang, X., Xie, L., Dong, C., Shan, Y.: 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)
60. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops. pp. 0–0 (2018)
61. Wang, Y., Yang, W., Chen, X., Wang, Y., Guo, L., Chau, L.P., Liu, Z., Qiao, Y., Kot, A.C., Wen, B.: 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 (2024)
62. Wang, Z., Chen, J., Hoi, S.C.: Deep learning for image super-resolution: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3365–3387 (2020)
63. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
64. Wei, P., Xie, Z., Lu, H., Zhan, Z., Ye, Q., Zuo, W., Lin, L.: Component divide-and-conquer for real-world image super-resolution. In: European conference on computer vision. pp. 101–117. Springer (2020)
65. Wołczyk, M., Zając, M., Pascanu, R., Kuciński, Ł., Miłoś, P.: Continual world: A robotic benchmark for continual reinforcement learning. Advances in Neural Information Processing Systems **34**, 28496–28510 (2021)
66. Wu, R., Sun, L., Ma, Z., Zhang, L.: One-step effective diffusion network for real-world image super-resolution. Advances in Neural Information Processing Systems **37**, 92529–92553 (2024)
67. Wu, R., Yang, T., Sun, L., Zhang, Z., Li, S., Zhang, L.: 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 (2024)
68. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: 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)
69. Yang, T., Wu, R., Ren, P., Xie, X., Zhang, L.: Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In: European conference on computer vision. pp. 74–91. Springer (2024)

70. Yi, Q., Li, S., Wu, R., Sun, L., Wu, Y., Zhang, L.: Fine-structure preserved real-world image super-resolution via transfer vae training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12415–12426 (2025)
71. Yu, F., Gu, J., Li, Z., Hu, J., Kong, X., Wang, X., He, J., Qiao, Y., Dong, C.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 25669–25680 (2024)
72. Yu, X., Guo, Y.C., Li, Y., Liang, D., Zhang, S.H., Qi, X.: Text-to-3d with classifier score distillation. arXiv preprint arXiv:2310.19415 (2023)
73. Yue, Z., Liao, K., Loy, C.C.: Arbitrary-steps image super-resolution via diffusion inversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
74. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
75. Zeng, G., Chen, Y., Cui, B., Yu, S.: Continual learning of context-dependent processing in neural networks. Nature Machine Intelligence **1**(8), 364–372 (2019)
76. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: 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)
77. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
78. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
79. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV). pp. 286–301 (2018)
80. Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A.M., Le, Q.V., Laudon, J., et al.: Mixture-of-experts with expert choice routing. Advances in Neural Information Processing Systems **35**, 7103–7114 (2022)