Mixture of Ranks with Degradation-Aware Routing for One-Step Real-World Image Super-Resolution

Xiao He[1], Zhijun Tu[2], Kun Cheng[1], Mingrui Zhu[1], Jie Hu[2], Nannan Wang[1]*, Xinbo Gao[1]

[1]State Key Laboratory of Integrated Services Networks, Xidian University
[2]Huawei Noah's Ark Lab
xiaohe366@gmail.com, zhijun.tu@huawei.com, nnwang@xidian.edu.cn

Abstract

The demonstrated success of sparsely-gated Mixture-of-Experts (MoE) architectures, exemplified by models such as DeepSeek and Grok, has motivated researchers to investigate their adaptation to diverse domains. In real-world image super-resolution (Real-ISR), existing approaches mainly rely on fine-tuning pre-trained diffusion models through Low-Rank Adaptation (LoRA) module to reconstruct high-resolution (HR) images. However, these dense Real-ISR models are limited in their ability to adaptively capture the heterogeneous characteristics of complex real-world degraded samples or enable knowledge sharing between inputs under equivalent computational budgets. To address this, we investigate the integration of sparse MoE into Real-ISR and propose a Mixture-of-Ranks (MoR) architecture for single-step image super-resolution. We introduce a fine-grained expert partitioning strategy that treats each rank in LoRA as an independent expert. This design enables flexible knowledge recombination while isolating fixed-position ranks as shared experts to preserve common-sense features and minimize routing redundancy. Furthermore, we develop a degradation estimation module leveraging CLIP embeddings and predefined positive-negative text pairs to compute relative degradation scores, dynamically guiding expert activation. To better accommodate varying sample complexities, we incorporate zero-expert slots and propose a degradation-aware load-balancing loss, which dynamically adjusts the number of active experts based on degradation severity, ensuring optimal computational resource allocation. Comprehensive experiments validate our framework's effectiveness and state-of-the-art performance.

Introduction

In image super-resolution (SR) tasks (Dong et al. 2014; Zhang, Zhang, and Bovik 2015; Wang et al. 2018; Liang et al. 2021), models process low-resolution (LR) inputs to reconstruct high-resolution (HR) outputs with enhanced high-fidelity details. Traditional SR methodologies (Kim, Lee, and Lee 2016; Bao et al. 2022; Chen et al. 2023) typically generate LR images via bicubic downsampling of HR counterparts. However, real-world image degradation processes exhibit inherent complexity, encompassing multifaceted distortions such as blurring, sensor noise, and other ill-defined artifacts. This discrepancy between synthetic and real-world degradation has positioned real-world image super-resolution (Real-ISR) (Wang et al. 2021; Zhang et al. 2021) as a critically challenging problem, driving substantial research efforts in recent years.

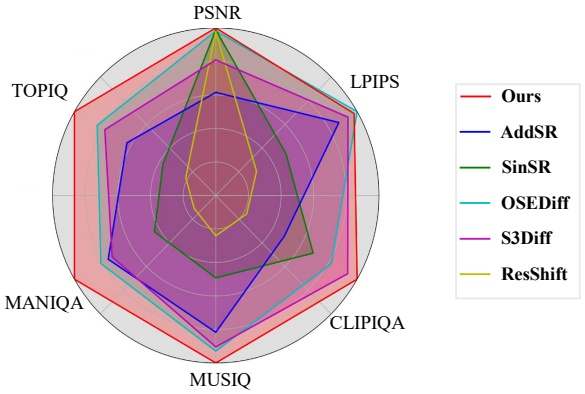

Figure 1: Performance Comparison. Compared to other Real-ISR methods, MoR-DASR achieves superior performance with just a single diffusion step.

Generative adversarial networks (GANs) (Goodfellow et al. 2020) and diffusion models (DMs) (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) currently constitute the predominant architectures for Real-ISR tasks, with diffusion models garnering particular research interest due to their superior generative priors. However, the iterative sampling process inherent to DMs—typically requiring multiple sequential steps—imposes prohibitive computational latency, hindering practical deployment. Recent efforts (Wang et al. 2023a; Wu et al. 2024a; He et al. 2024) mitigate this limitation through single-step SR frameworks, achieved either by distilling multi-step diffusion models or fine-tuning pre-trained diffusion models via Low-Rank Adaptation (LoRA) modules (Hu et al. 2022). While these advancements have accelerated progress in Real-ISR, they remain suboptimal in fully leveraging model capabilities under constrained computational budgets.

Recently, the remarkable success of Sparse Mixture-of-Experts (MoE) (Shazeer et al. 2017) architectures in

*Corresponding author.

transformer-based large language models (LLMs), exemplified by DeepSeek (Liu et al. 2024) and Grok, has reinvigorated advancements in deep learning. The MoE framework is grounded in a conceptually elegant principle: decomposing models into specialized subnetworks ("experts") optimized for distinct data patterns or tasks. By dynamically activating only task-relevant experts per input, this paradigm retains access to extensive domain expertise while preserving computational efficiency. Inspired by this success, we explore the application of sparse MoE architectures to Real-ISR tasks, where input samples exhibit substantial heterogeneity in both degradation types (e.g., blur and noise) and severity levels.

In this paper, we integrate MoE with LoRA to propose a novel mixture-of-ranks (MoR) architecture for one-step Real-ISR tasks. Instead of simply dividing each LoRA module into experts, we treat each rank within the LoRA decomposition as an independent expert, allowing the model to capture fine-grained differences and relationships within the data more effectively. This design also facilitates more flexible decomposition and recombination of expert knowledge, thereby enhancing the model's learning capacity. Additionally, we isolate fixed-position ranks as shared experts to capture common features and reduce routing complexity. To ensure degradation-aware expert activation, we design a degradation estimation module that computes relative degradation scores using CLIP's cross-modal alignment capabilities. This module calculates cosine similarity between low-resolution (LR) images and predefined multi-dimensional positive/negative text prompts, generating adaptive routing weights. Additionally, we introduce zero-expert slots and a degradation-aware load-balancing loss, which dynamically scales the number of active experts based on input degradation severity: severely degraded samples engage more experts for enhanced reconstruction, while simpler cases utilize fewer. This adaptive scaling ensures robust restoration quality across diverse input conditions. As demonstrated in Figure 1, our framework achieves single-step generation of high-resolution images with enhanced fidelity. Furthermore, when compared to the multi-step Real-ISR method SeeSR, MoR-DASR achieves a $40\times$ speedup in inference time while maintaining comparable reconstruction quality. Our main contribution can be summarized as follows:

- We explore the integration of sparse MoE architectures into Real-ISR tasks, introducing a novel MoE architecture for single-step Real-ISR tasks that achieves high-fidelity reconstruction while maintaining resource efficiency.
- We propose a Mixture-of-Ranks (MoR) architecture that designates each LoRA rank as an independent expert, enabling dynamic knowledge recombination, while isolating fixed-position ranks as shared experts to capture common knowledge and mitigate redundancy in routed experts.
- We design a degradation estimation module to dynamically guide expert activation. This module leverages the cross-modal alignment capabilities to derive relative degradation scores of inputs. These scores are then integrated into the expert routing module to activate the relevant experts.
- Furthermore, we introduce zero experts and design a degradation-aware load-balancing loss, which ensures dynamic allocation of computational resources based on the severity of degradation in each sample.

Related Works

Mixture of Experts

The Mixture of Experts (MoE) paradigm, initially introduced in (Jacobs et al. 1991; Jordan and Jacobs 1994), has undergone significant refinement through subsequent research (Collobert, Bengio, and Bengio 2001; Aljundi, Chakravarty, and Tuytelaars 2017). In MoE architecture, routers dynamically select specialized parameter subsets to process input tokens, with outputs aggregated to form the final output. This sparse activation paradigm remains prevalent in contemporary architectures and has proven instrumental in scaling large language models (LLMs). DeepSeek-MoE (Liu et al. 2024) further advanced this sparse activation paradigm through two innovations: (1) Fine Expert Partitioning: Dividing the model into a larger number of specialized sub-networks. (2) Shared Expert Isolation: Designating a subset of experts to capture domain-invariant knowledge, reducing routing redundancy. These refinements have injected vitality into the development of LLMs. Even some works have extended sparse MoE to visual tasks—for instance, DiT-MoE (Fei et al. 2024) scales diffusion models to 16B parameters via 2 shared and 8 routed experts, advancing image generation. Concurrently, multi-task learning studies (Dou et al. 2023; Liu et al. 2023; Zhao et al. 2025) integrate LoRA with MoE architectures, enhancing downstream task performance without proportional computational overhead.

Real-world Image Super-Resolution

Early Real-ISR approaches (Ledig et al. 2017; Tu et al. 2023, 2024) leveraged generative adversarial networks (GANs), combining adversarial and perceptual losses (Zhang et al. 2018; Ding et al. 2020) to ensure reconstruction fidelity and quality. However, the inherent complexity of real-world degradation processes causes existing methods to struggle with satisfactory restoration of severely degraded samples. Recently, diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Ding et al. 2025) has gradually emerged as a successor to GANs (Goodfellow et al. 2020) in various downstream tasks. StableSR (Wang et al. 2024b) introduces an auxiliary encoder to project low-resolution (LR) features into the latent space of a pre-trained text-to-image diffusion model, fine-tuning the architecture for Real-ISR tasks. Subsequent studies (Yang et al. 2023; Wu et al. 2024b; Yue, Wang, and Loy 2024) systematically investigated methods for injecting LR image information into diffusion models. Despite the improved perceptual quality, these approaches suffer from computational inefficiency, often requiring 50–100 iterative sampling steps during inference. To address this, recent work (He et al. 2024; Wang et al. 2023a; Wu et al. 2024a; Xie et al. 2024; Zhang et al. 2024; Wang et al. 2024a) achieves single-step super-resolution via knowledge Distillation or fine-tuning

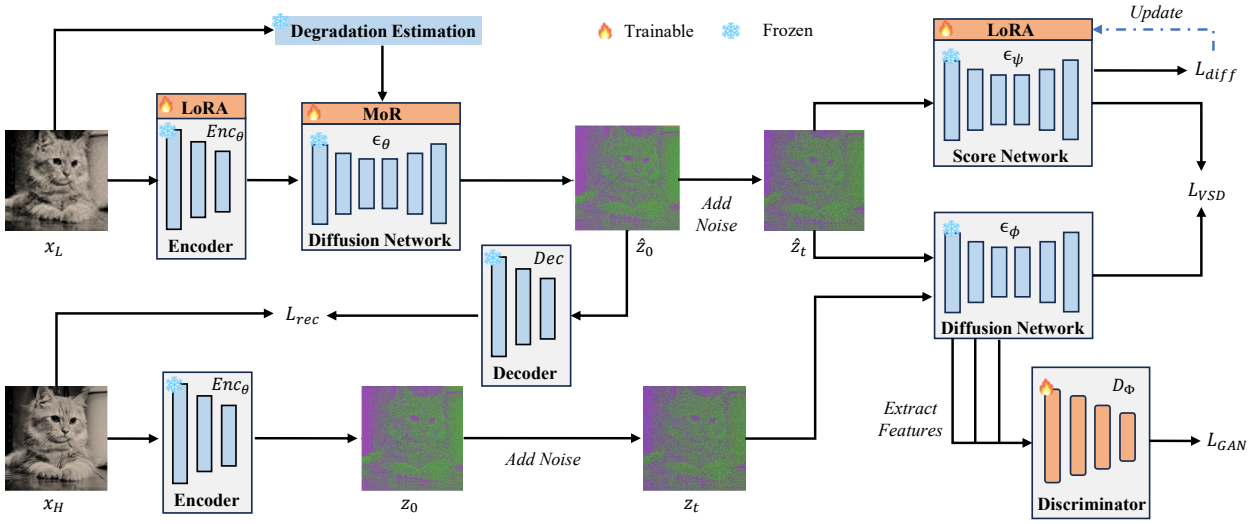

Figure 2: The training framework of MoR-DASR. The LR image is passed through a trainable encoder Enc_θ, a diffusion network with a MoR module ϵ_θ and a frozen decoder Dec to obtain the desired HR image. The training procedure alternates between two phases: 1. Optimizing the variational score network ϵ_ψ through diffusion loss $\mathcal{L}_{diff}$ to fit the distribution of the generated samples. 2. Finetune the diffusion model ϵ_θ and encoder Enc_θ to generate high-quality samples through reconstruction loss $\mathcal{L}_{rec}$, variational score distillation loss $\mathcal{L}_{VSD}$, and GAN loss $\mathcal{L}_{GAN}$.

pre-trained diffusion models. However, the significant input heterogeneity inherent to Real-ISR—spanning diverse degradation types and severity levels—limits the effectiveness of these computationally intensive methods, as they fail to fully utilize model capacity under constrained computational budgets.

Methodology

Preliminaries

Problem Modeling. Real-ISR task aims to reconstruct HR images $\hat{x}_H$ form LR input x_L. Given a pre-trained text-to-image diffusion model, existing methods fine-tune the model to adapt to Real-ISR tasks under paired data supervision. The diffusion model receives a noisy version z_t of the latent representation z_0 encoded by the HR image, with the condition of LR images or features extracted from it. The model is then optimized to accurately predict the noise in the latent code at each time step, which can be represented as:

$$\mathcal{L} = \mathbb{E}_{z_0,t,x_L,\epsilon}\left[\epsilon - \epsilon_\theta\left(z_t, t, x_L\right)\right], \tag{1}$$

where $z_t = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\epsilon$, $\epsilon \in \mathcal{N}(0, I)$. During inference, the model takes gaussian noise z_T as input and iteratively transforms it to the clean latent codes $\hat{z}_0$. To accelerate the sampling process, existing ISR methods employ distillation to generate the clean latent code $\hat{z}_0$ in a single step. The computation of $\hat{z}_0$ is formulated as:

$$\hat{z}_0 = \frac{z_T - \sqrt{1-\alpha_T}\epsilon_\theta(z_T, x_L, T)}{\sqrt{\alpha_T}}. \tag{2}$$

While these methods accelerate diffusion model inference, the random noise inherent to its input may degrade fidelity in Real-ISR tasks. To mitigate this, OSEDiff directly fine-tunes the pre-trained diffusion model to learn the LR to HR mapping, formulated as:

$$\hat{z}_0 = \frac{z_L - \sqrt{1-\alpha_T}\epsilon_\theta(z_L, c, T)}{\sqrt{\alpha_T}}, \tag{3}$$

where c is the prompt obtained by applying text prompt extractor DAPE to LR input. We follow this paradigm to fine-tune the pre-trained diffusion model to adapt Real-ISR tasks.

Variational Score Distillation. Variational Score Distillation (VSD) leverages the priors of a pretrained text-to-image diffusion model to optimize generative models, ensuring that generated images align semantically with the input text prompts.

Within the VSD framework, the generator's output $\hat{z}_0$ is re-noised and fed into both a pre-trained diffusion model ϵ_ϕ and an online-trained variational score network ϵ_ψ. The generator is optimized by minimizing the discrepancy between the two models' predictions. As formalized in (Wang et al. 2023b), this process is expressed as:

$$\nabla_\theta \mathcal{L}_{VSD} = \mathbb{E}_{t,\epsilon}\left[\omega(t)(\epsilon_\phi(\hat{z}_t, t, c) - \epsilon_\psi(\hat{z}_t, t, c))\frac{\partial \hat{z}_t}{\partial \theta}\right], \tag{4}$$

where $\hat{z}_t = \sqrt{\alpha_t}\hat{z}_0 + \sqrt{1-\alpha_t}\epsilon$, ω_t is a weighting function.

Overview of MoR-DASR

The framework of MoR-DASR is illustrated in Figure 2. Given an LR input, the degradation estimation module first computes its degradation score. This score is fed into the dynamic routing mechanism of the Mixture-of-Ranks (MoR) module, which activates corresponding experts (ranks) to reconstruct the HR image. During fine-tuning of the diffusion model, we employ reconstruction losses (e.g., L2 and

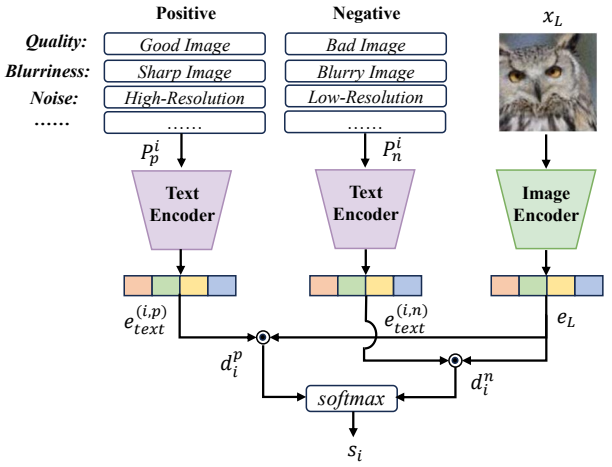

Figure 3: The framework of degradation estimation module.

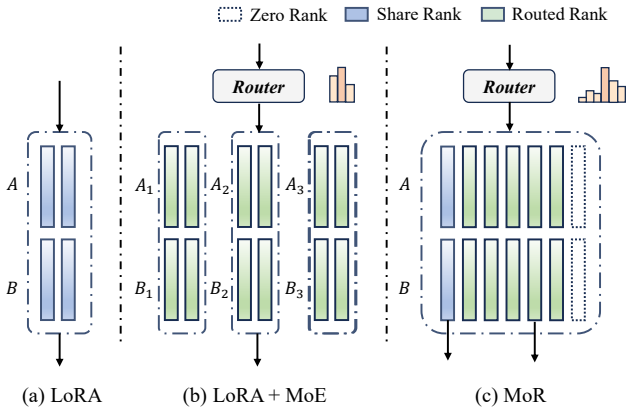

Figure 4: Comparison of LoRA, LoRA MoE and MoR. In MoR, each rank is treated as an expert. A subset of these ranks is designated as shared experts to process all samples, while the remaining ranks function as routed experts that are selectively activated to process specific samples.

LPIPS) to enforce visual similarity between the output and the ground truth. Additionally, we align the output's data distribution with high-quality image distributions using priors from the pre-trained diffusion model, implemented via Variational Score Distillation and GAN losses.

Degradation Estimation Module

We integrates dynamic computation—inherent to MoE—into super-resolution tasks to enhance model performance while maintaining computational efficiency. A critical challenge lies in determining how to categorize input data for the RealISR task, specifically identifying optimal criteria to guide the expert routing mechanism. A naive approach involves routed experts based on LR input features alone. However, semantic-driven data partitioning proves suboptimal for Real-ISR, as empirical studies (Liang, Zeng, and Zhang 2022b; Zhang et al. 2024) emphasize the critical role of degradation characteristics. Through empirical analysis, we observe that reconstruction quality is inversely correlated with degradation severity: images with mild degradation achieve superior reconstruction, whereas severe degradation images present greater reconstruction challenges. This finding motivates our proposal of degradation intensity as a critical criterion for dynamic computation in expert-based super-resolution frameworks.

Based on the above analysis, we utilize the cross-modality ability of CLIP model and proposed a novel degradation estimation module. We first defined a set of positive and negative prompt pairs from multiple perspectives for evaluating image quality, which includes the overall quality of the image, as well as the degree of blur, whether it contains noise, and so on. Then, for each input image, we obtain its embedding e_L through CLIP image encoder:

$$e_L = f_{img}(x_L). \tag{5}$$

Then, we calculate the embedding vectors corresponding to the positive and negative prompts on each evaluation dimension:

$$e_{text}^{(i,p)} = f_{text}(P_p^i), \quad e_{text}^{(i,n)} = f_{text}(P_n^i), \tag{6}$$

where $e_{text}^{(i,p)}$ and $e_{text}^{(i,n)}$ denote the embeddings of the positive (P_p^i) and negative prompts (P_n^i) of the i-th evaluation dimension, respectively. We then combine image embedding with the distance between positive and negative text embeddings to obtain the degradation score:

$$d^{(i,p)} = \frac{e_L \odot e_{text}^{(i,p)}}{||e_L|| \cdot ||e_{text}^{(i,p)}||}, \quad d^{(i,n)} = \frac{e_L \odot e_{text}^{(i,n)}}{||e_L|| \cdot ||e_{text}^{(i,n)}||}, \tag{7}$$

$$s_i = \frac{\exp(d^{(i,n)})}{\exp(d^{(i,p)}) + \exp(d^{(i,n)})}, \tag{8}$$

where $d^{(i,p)}$ and $d^{(i,n)}$ denote the cosine distance between the LR image and positive and negative prompts on the i-th evaluation dimension,respectively. s_i represents the final degradation score on the i-th evaluation dimension.

MoR with Degradation-Aware Router

Existing Parameter-Efficient Fine-Tuning (PEFT) methods treat each LoRA module as a standalone expert, assigning individual experts to specific tasks (as shown in Figure 4). However, this static partitioning of the parameter space restricts the dynamic fusion and decomposition of expert knowledge, as entire modules are activated or deactivated holistically—a coarse-grained strategy that results in suboptimal utilization of learned features. In contrast, we treat each rank of the LoRA as an independent expert, enabling fine-grained expert activation and allowing knowledge to be flexibly decomposed and recombined.

The architecture of MoR is illustrated in Figure 4. Building upon a pre-trained text-to-image diffusion model, we integrate trainable low-rank matrices $A \in R^{d \times r}$ and $B \in R^{r \times d}$ to fine-tune the model parameters, where d denotes the pre-trained weights' channel dimension, r denotes the

rank of the LoRA module. We treat each rank as an independent expert, partitioned into two categories: shared experts and routed experts. Routed experts are activated via gating mechanisms and top-k selection strategies, whereas shared experts remain persistently active to capture common knowledge. Specifically, the degradation score s of input LR images, computed by the degradation estimation module, is fed into a gating function to generate logits. Experts corresponding to the top-k logits are subsequently activated. This process is formalized as follows:

$$g_i(s) = TopK(Softmax(sW_g)), \tag{9}$$

where W_g is a learnable gating matrix, $g_i(s)$ denotes the logits of the selected experts i. Following the selection of k experts, the forward pass of MoR is formulated as:

$$MoR(z_t) = W_0 x + \sum_{i=1}^{k} g_i(s) B_i A_i x + \sum_{j=1}^{m} B_j A_j x, \tag{10}$$

where W_0 represents the parameter matrix of the backbone model, m is the number of shared experts.

Building on the aforementioned framework, we introduce zero experts to enable dynamic rank adaptation, addressing two key challenges: (1) input samples with varying degradation severity necessitate distinct computational budgets, and (2) the optimal LoRA rank may differ across model layers. For mildly degraded samples, the network activates zero experts—effectively bypassing unnecessary computations—to prevent over-restoration. Conversely, severely degraded inputs trigger the activation of more real experts, allocating greater computational resources to achieve high-fidelity reconstruction. This adaptive mechanism ensures efficient resource distribution, balancing restoration quality and computational cost across diverse degradation scenarios.

Loss Functions

To enable single-step high-quality image reconstruction, we employ a composite loss function comprising reconstruction loss and feature-distribution matching loss. The reconstruction loss combines L2 loss and LPIPS loss to enforce structural and perceptual fidelity. The feature-distribution matching loss integrates two components: (1) VSD loss, which aligns the model's output distribution with the high-quality prior embedded in the pre-trained text-to-image diffusion model, and (2) GAN loss, leveraging the diffusion model's feature extraction capability to align outputs with real data distributions under varying noise perturbations through adversarial training. Furthermore, we introduce a degradation-aware load-balancing loss to mitigate suboptimal resource allocation and improve overall computational efficiency. Thus, the full optimization objective can be expressed as follows:

$$\mathcal{L}_\theta = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{VSD} + \lambda_2 \mathcal{L}_{GAN} + \mathcal{L}_{balance}. \tag{11}$$

GAN loss. Following (He et al. 2024), we leverage a pre-trained diffusion model to extract features from both synthetic and real data under varying noise perturbations. These features are then processed by multi-scale tiny discriminator heads D_Φ to distinguish between the two domains. The adversarial training objective is formalized as:

$$\mathcal{L}_{adv}^{\epsilon_\theta} = -\mathbb{E}_{\hat{z}_0}\left[D_\Phi\left(\hat{z}_t, t\right)\right], \tag{12}$$

where $\hat{z}_t$ denotes the latent representation derived by re-adding noise to the generator's initial output.

Degradation-aware load-balancing loss. In MoE's training process, load-balancing loss plays a critical role. It penalizes imbalanced expert selection and encourages equitable utilization of experts. It prevents the model from over-relying on specific experts, thereby mitigating suboptimal resource allocation and improving overall computational efficiency, which can be formalized as:

$$\mathcal{L}_{balance} = \alpha N \sum_{i=1}^{N} f_i \, \mathcal{P}_i, \tag{13}$$

where N is the number of routed experts, α is a weighting factor and f_i is the fraction of samples dispatched to expert i:

$$f_i = \frac{1}{b} \sum_{x \in \mathcal{B}} 1\left\{\text{argmax } p(x) = i\right\}, \tag{14}$$

where $\mathcal{B}$ is the sample batch, b is the batch size, $p(x)$ denotes the probability value of the input after gating calculation and $\mathcal{P}_i$ is the fraction of router probability allocated for expert i:

$$\mathcal{P}_i = \frac{1}{b} \sum_{x \in \mathcal{B}} p_i(x). \tag{15}$$

However, the formulation treats zero and real experts homogeneously for load balancing, assigning zero experts a static activation probability of approximately k/N, independent of input degradation severity. This violates our core objective: adaptive expert allocation, where simple samples activate more zero experts (minimizing computation) and complex samples prioritize real experts (maximizing reconstruction quality). To resolve this, we propose a degradation-aware load-balancing loss by revising Eq. 13 to incorporate degradation severity into the balancing criterion, ensuring expert activation aligns dynamically with input complexity.

$$\mathcal{L}_{balance} = N \sum_{i=1}^{N} \alpha_i \, f_i \, p_i, \tag{16}$$

where

$$\alpha_i = \begin{cases} \alpha & \text{if } i \le n, \\ s\alpha & \text{if } i > n, \end{cases} \tag{17}$$

where n is the number of true experts and s is the degradation score of the input sample. It can be seen that the degradation score s (higher values indicate higher degradation) modulates the penalization weight for zero experts. Specifically, higher degradation (larger s) increases the penalty for zero expert activation, thereby incentivizing the model to prioritize real experts. Conversely, for mildly degraded inputs (lower s), the reduced penalty encourages zero expert usage, efficiently allocating computational resources without compromising restoration quality.

Figure 5: Visual comparisons of different Real-ISR methods. Please zoom in for a better view.

Experiments

Experiments Setup

Training Details. Following SeeSR (Wu et al. 2024b), we utilize LSDIR (Li et al. 2023) dataset and 10k face images from FFHQ (Karras, Laine, and Aila 2019) dataset for training. To generate low-resolution (LR) and high-resolution (HR) training pairs, we apply the degradation pipeline proposed in Real-ESRGAN. We employ Stable Diffusion 2.1 as the base model and fine-tune it to adapt Real-ISR tasks. The rank parameter for LoRA is set to 4 in both the VAE encoder and the variational score network. The Mixture-of-Ranks (MoR) module comprises 40 ranks, including 8 shared and 32 routed ranks, with a top-8 routing strategy implemented during training. The model undergoes 25,000 iterations with a batch size of 16 and a learning rate of $5e-5$.

Compared Methods. We compare MoR-DASR with state-of-the-art one-step Real-ISR methods, including SinSR (Wang et al. 2023a), AddSR (Xie et al. 2024), OSEDiff (Wu et al. 2024a), and S3Diff (Zhang et al. 2024). Additionally, to comprehensively evaluate the performance of MoR-DASR, we also compare it with multi-step Real-ISR methods (Lin et al. 2023; Wang et al. 2024b; Yue, Wang, and Loy 2024; Wu et al. 2024b) and GAN-based Real-ISR approaches (Zhang et al. 2021; Wang et al. 2021; Liang, Zeng, and Zhang 2022a; Chen et al. 2022). Experimental details are provided in the appendix.

Testing Details. We evaluate the performance of Real-ISR algorithms on two real datasets, RealSR (Cai et al. 2019) and DRealSR (Wu et al. 2024b), and one synthetic dataset, DIV2K-Val (Agustsson and Timofte 2017). We adopt non-reference metrics, including CLIP-IQA (Wang, Chan, and Loy 2023), MUSIQ (Ke et al. 2021), MANIQA (Yang et al. 2022), TOPIQ (Chen et al. 2024), and TRES (Golestaneh, Dadsetan, and Kitani 2022) to evaluate the performance of various Real-ISR methods. In Real-ISR tasks, these metrics are crucial as they better align with human visual perception (Wang et al. 2024b; Xie et al. 2024). Additionally, we also report traditional metrics such as PSNR, SSIM (Wang et al. 2004), and LPIPS (Zhang et al. 2018) for reference.

Comparison with State-of-the-Arts

Quantitative Comparison Tables 1 present a quantitative comparison of state-of-the-art methods on three benchmark datasets. Table 1 primarily evaluates MoR-DASR against existing one-step Real-ISR approaches. The following key observations can be drawn from the table: (1) MoR-DASR consistently ranks among the top-performing methods across all metrics and datasets. Notably, on the DrealSR dataset, it achieves the highest scores for nearly every metric, underscoring its superior performance. (2) While SinSR distills a multi-step Real-ISR model trained from scratch and delivers competitive results in PSNR and SSIM, its non-reference metrics are markedly inferior to other single-step methods. (3) S3diff attains the best LPIPS scores, particularly on the synthetic DIV2K dataset, but its performance on other metrics was suboptimal. This may be attributed to its substantial emphasis on minimizing LPIPS loss during training. (4) Compared to other single-step methods, our approach demonstrates significant improvements

Datasets	Methods	PSNR↑	SSIM↑	LPIPS ↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑	TRES↑
DIV2K-Val	AddSR	23.26	0.590	0.362	0.573	63.39	0.405	0.573	73.23
	SinSR	**24.41**	0.602	0.324	0.648	62.80	0.424	0.571	72.24
	OSEDiff	23.92	**0.614**	0.310	0.659	67.71	0.435	0.606	78.40
	S3Diff	23.53	0.593	**0.258**	**0.699**	67.92	0.452	0.633	80.72
	MoR-DASR	24.01	0.606	0.289	0.681	**68.09**	**0.475**	**0.663**	**84.14**
RealSR	AddSR	23.12	0.655	0.309	0.552	67.14	0.488	0.599	79.91
	SinSR	**26.16**	**0.739**	0.308	0.631	60.96	0.399	0.512	59.92
	OSEDiff	25.26	0.728	0.301	0.651	68.41	0.468	0.614	80.18
	S3Diff	25.18	0.727	**0.272**	0.672	67.82	0.459	0.616	78.82
	MoR-DASR	25.32	0.728	0.291	**0.691**	**69.78**	**0.512**	**0.662**	**84.97**
DRealSR	AddSR	26.71	0.738	0.321	0.593	62.13	0.458	0.569	71.41
	SinSR	28.32	0.747	0.372	0.642	55.36	0.388	0.512	59.92
	OSEDiff	28.29	**0.792**	**0.302**	0.673	64.47	0.469	0.616	76.76
	S3Diff	27.54	0.767	0.311	0.702	63.94	0.452	0.604	75.41
	MoR-DASR	**28.37**	0.776	0.307	**0.717**	**65.94**	**0.509**	**0.652**	**81.78**

Table 1: Quantitative comparison with state of the arts one-step Real-ISR methods. The best and second best results are highlighted in **bold** and underline.

Model	CLIPIQA ↑	MANIOA ↑	TRES ↑
LoRA	0.670	0.481	78.81
LoRA+MoE	0.689	0.484	79.36
MoR-v1	0.704	0.491	80.32
MoR-v2	0.699	0.479	79.41
MoR-full	**0.717**	**0.509**	**81.78**

Table 2: Ablation study of our proposed MoR architecture.

in non-reference metrics (e.g., MANIOA, TOPIQ). These non-reference metrics prioritize perceptual quality and are closely aligned with human perceptual evaluations, further validating MoR-DASR's ability to generate reconstructions that align with human visual preferences.

Qualitative Comparison Figure 5 present qualitative comparisons of reconstruction results. Our analysis demonstrates that existing methods frequently struggle to recover fine-grained details, whereas our method generates clear and accurate textures. While some baselines produce sharp details, these often deviate from ground-truth structures—for instance, synthesizing erroneous shoe logos. In contrast, our method not only reconstructs complex details but also avoids generating unnatural artifacts, achieving a balance between high-quality texture generation and visual fidelity. Extended visualizations are provided in the appendix.

Ablation Studies

In this section, we evaluate the contributions of different components in our method. Specifically, we focus on analyzing the impact of the MoR architecture and the degradation-aware load balancing loss on model performance. Additional ablation studies are provided in the appendix.

Table 2 presents an ablation study evaluating the efficacy of our MoR module. We compare five configurations on the real-world SR dataset. (1) Vanilla LoRA. (2) LoRA+MoE: Integration of LoRA with a standard MoE architecture. (3) MoR-v1: MoR without zero experts. (4) MoR-v2: MoR with zero experts but excluding degradation-aware load balancing loss. (5) MoR-full: our proposed MoR framework. According to the Table, we have key findings as follows. The MoE architecture (LoRA+MoE) improves baseline performance, validating the effectiveness of applying MoE in Real-ISR tasks. By designing each rank in LoRA as an independent expert (MoR-v1), the model is able to flexibly decompose and recombine knowledge, further enhancing performance. MoR-v2 introduces zero experts but still relies on a traditional load balancing loss to guide expert selection, resulting in a significant drop in performance. In contrast, the MoR-full model incorporates a degradation-aware load balancing loss while introducing zero experts, achieving SOTA results: $+7\%$ CLIPIQA, $+5.8\%$ MANIQA, and $+3.8\%$ TRES over baselines. It demonstrates that our method significantly enhances the quality of reconstructed images.

Conclusion

We present MoR-DASR, a novel MoE architecture for one-step Real-ISR tasks. MoR-DASR captures common knowledge through shared rank in a mixture-of-ranks architecture, decomposing and recombining features via routed ranks to enhance model performance without sacrificing computational efficiency. To guide the MoR module in selecting input-relevant experts, we design a degradation estimation module that leverages CLIP's cross-modal alignment capability, mapping sample-specific degradation scores to the router's gating mechanism. Additionally, we introduce a degradation-aware load-balancing loss combining zero experts, which dynamically adjusts the number of activated ranks (experts) to optimize reconstruction quality across inputs with diverse degradation levels. Extensive experiments demonstrate that MoR-DASR achieves state-of-the-art performance while retaining practical inference efficiency.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants U22A2096 and 62576261, in part by Scientific and Technological Innovation Teams in Shaanxi Province under grant 2025RS-CXTD-011, in part by the Shaanxi Province Core Technology Research and Development Project under grant 2024QY2-GJHX-11, in part by the Fundamental Research Funds for the Central Universities under GrantQTZX23042.

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 126–135.

Aljundi, R.; Chakravarty, P.; and Tuytelaars, T. 2017. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3366–3375.

Bao, Q.; Gang, B.; Yang, W.; Zhou, J.; and Liao, Q. 2022. Attention-driven graph neural network for deep face super-resolution. *IEEE Transactions on Image Processing*, 31: 6455–6470.

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. 3086–3095.

Chen, C.; Mo, J.; Hou, J.; Wu, H.; Liao, L.; Sun, W.; Yan, Q.; and Lin, W. 2024. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*.

Chen, C.; Shi, X.; Qin, Y.; Li, X.; Han, X.; Yang, T.; and Guo, S. 2022. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1329–1338.

Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; and Dong, C. 2023. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22367–22377.

Collobert, R.; Bengio, S.; and Bengio, Y. 2001. A parallel mixture of SVMs for very large scale problems. *Advances in Neural Information Processing Systems*, 14.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5): 2567–2581.

Ding, X.; Yu, L.; Li, X.; Tu, Z.; Chen, H.; Hu, J.; and Chen, Z. 2025. RaSS: Improving Denoising Diffusion Samplers with Reinforced Active Sampling Scheduler. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12923–12933.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, 184–199. Springer.

Dou, S.; Zhou, E.; Liu, Y.; Gao, S.; Zhao, J.; Shen, W.; Zhou, Y.; Xi, Z.; Wang, X.; Fan, X.; et al. 2023. Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. *arXiv preprint arXiv:2312.09979*, 4(7).

Fei, Z.; Fan, M.; Yu, C.; Li, D.; and Huang, J. 2024. Scaling diffusion transformers to 16 billion parameters. *arXiv preprint arXiv:2407.11633*.

Golestaneh, S. A.; Dadsetan, S.; and Kitani, K. M. 2022. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1220–1230.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

He, X.; Tang, H.; Tu, Z.; Zhang, J.; Cheng, K.; Chen, H.; Guo, Y.; Zhu, M.; Wang, N.; Gao, X.; et al. 2024. One step diffusion-based super-resolution with time-aware distillation. *arXiv preprint arXiv:2408.07476*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Jordan, M. I.; and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2): 181–214.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. 4401–4410.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. 5148–5157.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1646–1654.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Li, Y.; Zhang, K.; Liang, J.; Cao, J.; Liu, C.; Gong, R.; Zhang, Y.; Tang, H.; Liu, Y.; Demandolx, D.; et al. 2023. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1775–1787.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Liang, J.; Zeng, H.; and Zhang, L. 2022a. 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*, 5657–5666.

Liang, J.; Zeng, H.; and Zhang, L. 2022b. Efficient and degradation-adaptive network for real-world image super-resolution. In *European Conference on Computer Vision*, 574–591. Springer.

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Fei, B.; Dai, B.; Ouyang, W.; Qiao, Y.; and Dong, C. 2023. DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior. *arXiv preprint arXiv:2308.15070*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, Q.; Wu, X.; Zhao, X.; Zhu, Y.; Xu, D.; Tian, F.; and Zheng, Y. 2023. Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications. *CoRR*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tu, Z.; Du, K.; Chen, H.; Wang, H.; Li, W.; Hu, J.; and Wang, Y. 2024. IPT-V2: Efficient Image Processing Transformer using Hierarchical Attentions. *arXiv preprint arXiv:2404.00633*.

Tu, Z.; Hu, J.; Chen, H.; and Wang, Y. 2023. Toward accurate post-training quantization for image super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5856–5865.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring CLIP for Assessing the Look and Feel of Images.

Wang, J.; Fan, Q.; Zhang, Q.; Liu, H.; Yu, Y.; Chen, J.; and Ren, W. 2024a. Hero-SR: One-Step Diffusion for Super-Resolution with Human Perception Priors. *arXiv preprint arXiv:2412.07152*.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2024b. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 1–21.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1905–1914.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, 0–0.

Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.-P.; Liu, Z.; Qiao, Y.; Kot, A. C.; and Wen, B. 2023a. SinSR: Diffusion-Based Image Super-Resolution in a Single Step. *arXiv preprint arXiv:2311.14760*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. 13(4): 600–612.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023b. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36: 8406–8441.

Wu, R.; Sun, L.; Ma, Z.; and Zhang, L. 2024a. One-Step Effective Diffusion Network for Real-World Image Super-Resolution. *arXiv preprint arXiv:2406.08177*.

Wu, R.; Yang, T.; Sun, L.; Zhang, Z.; Li, S.; and Zhang, L. 2024b. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25456–25467.

Xie, R.; Tai, Y.; Zhang, K.; Zhang, Z.; Zhou, J.; and Yang, J. 2024. AddSR: Accelerating Diffusion-based Blind Super-Resolution with Adversarial Diffusion Distillation. *arXiv preprint arXiv:2404.01717*.

Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1191–1200.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2023. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*.

Yue, Z.; Wang, J.; and Loy, C. C. 2024. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36.

Zhang, A.; Yue, Z.; Pei, R.; Ren, W.; and Cao, X. 2024. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*.

Zhang, K.; Liang, J.; Van Gool, L.; and Timofte, R. 2021. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4791–4800.

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

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. 586–595.

Zhao, Z.; Zhou, Y.; Zhu, D.; Shen, T.; Wang, X.; Su, J.; Kuang, K.; Wei, Z.; Wu, F.; and Cheng, Y. 2025. Each Rank Could be an Expert: Single-Ranked Mixture of Experts LoRA for Multi-Task Learning. *arXiv preprint arXiv:2501.15103*.