Expert Weaving: Marrying Masked Auto-Regressive and Diffusion Models for Unified Image Restoration

Xin Lu, Jie Huang[‡], Jie Xiao, Dong Li, and Xueyang Fu[✉]

School of Information Science and Technology and MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China, Hefei, 230026, China.
Project page: `https://xin1u.github.io/MoSE_PAGE`
[‡] Project leader. [✉] Corresponding author.

Abstract. All-in-one image restoration aims to handle diverse degradations within a unified framework. While Mixture-of-Experts (MoE) architectures offer a scalable path, existing designs predominantly rely on vanilla end-to-end fidelity-oriented experts that lack task-adaptive generative priors, limiting their ability to recover realistic structures and textures in complex scenarios. In this paper, we first investigate complementary task-adaptive generative priors: Masked Auto-Regressive (MAR) experts are tailored for local degradations (e.g., shadow) by reconstructing semantic structures via clean local backgrounds; conversely, Diffusion experts excel at global degradations (e.g., noise) by recovering textural details through iterative denoising. Motivated by these insights, we propose Mixture-of-Synergy-Experts (MoSE), a novel framework that deeply integrates these complementary generative priors into the MoE architecture and synergizes MAR and Diffusion experts with vanilla backbones at both image and feature levels. MoSE is orchestrated by a DINOv3-guided semantic-aware router that adaptively regulates not only expert coordination but also generation control: mask sizes are adjusted to handle varying scales of local semantic corruption, while iteration steps are modulated according to degradation intensity. By synergistically "weaving" complementary experts, our approach achieves a controllable fidelity-perception trade-off. Comprehensive experiments demonstrate the effectiveness of MoSE, achieving competitive performance across various all-in-one image restoration benchmarks.

1 Introduction

Image restoration [5, 17, 74] aims to recover high-quality images from degraded observations caused by adverse conditions such as rain [12, 64], haze [19, 79],

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62422609, in part by the CAS Project for Young Scientists in Basic Research under Grant YSBR-122, and in part by the Fundamental Research Funds for the Central Universities under Grant WK2102026001.

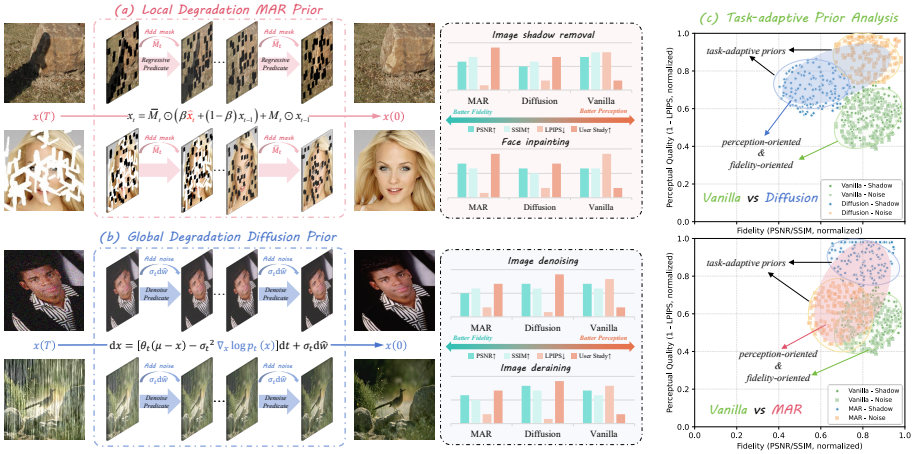


Fig. 1: *Motivation.* For each task in (a) and (b), the MAR, Diffusion, and vanilla models share an identical backbone and are trained individually, so that the performance gaps stem purely from the modeling paradigm rather than the architecture; all indicators are normalized. (a) Masked Auto-Regressive (MAR) [3] models effectively leverage local clean information to recover structurally coherent images from locally degraded inputs (*e.g.*, shadow removal and face inpainting). Quantitative results show that MAR achieves superior perceptual quality while maintaining fidelity comparable to vanilla models. (b) Diffusion models [54] excel in handling global degradations (*e.g.*, denoising and deraining) by iteratively reconstructing high-frequency details and textures. Experiments confirm that MAR and diffusion models exhibit complementary task-adaptive priors to different degradation types. This motivates our integration of both paradigms with vanilla restoration networks to achieve synergistic and high-quality results. (c) The data distribution sampled from the restoration results of the shadow and noise tasks further confirms the task-adaptive priors of MAR and Diffusion.

noise [24, 38], and shadow [46, 66]. It is widely applied in fields like autonomous driving, image editing, image synthesis, and scene understanding. Deep learning has significantly propelled the field of image restoration forward, achieving notable successes initially in task-specific image restoration [11, 62, 64, 65, 67]. Recently, all-in-one (AIO) image restoration [15, 21, 36, 43, 73], which accomplishes the restoration of various degradation types with a single model, provides a novel and practical solution for cross-task image restoration in real-world scenarios.

Vanilla AIO image restoration approaches typically build upon paradigms such as feature learning [6, 73], prompt learning [4, 43], or contrastive learning [20, 77]. They primarily strive to bridge the domain gap among various degradation types and have demonstrated remarkable performance in terms of reconstruction fidelity [4, 63]. In parallel, generative methods based on diffusion models [16, 35, 52, 68, 76] or autoregressive models [3, 22, 60, 81] attempt to learn more generalizable and cross-task restoration capabilities from the perspective of generative priors, thereby achieving superior perceptual quality [25, 26, 49].

Despite remarkable progress, existing methods have not fully explored the complementary task-adaptive priors of different generative models for different

degradation types. A key observation in this work is that different generative paradigms inherently provide task-adaptive restoration priors for different degradation types. Specifically, we find that Masked Auto-Regressive (MAR) [3, 38] modeling demonstrates unique advantages in handling **local degradations** such as shadow removal and inpainting—it effectively leverages local clean context to generate structurally coherent images by autoregressively predicting masked regions. In contrast, Diffusion [35, 49] modeling excels at addressing **global degradations** like denoising and deraining, where their multi-step iterative denoising process effectively restores high-frequency details and realistic textures, producing perceptually superior results [25, 63]. Fig. 1 illustrates the example of image space MAR and Diffusion models, these two paradigms inherently form complementary task-adaptive priors for handling different types of degradations.

In this paper, we leverage these complementary advantages from a Mixture-of-Experts (MoE) perspective and propose a novel AIO image restoration framework (see Fig. 2), termed Mixture-of-Synergy-Experts (MoSE). Unlike previous vanilla MoE methods [4, 73] that assign experts with specific parameter, receptive fields, or complexities, we deeply integrate **perception-oriented** MAR and Diffusion generative paradigms with a **fidelity-oriented** vanilla model to synergize their complementary capabilities for enhanced restoration. As shown in Fig. 2, we design a dual-level expert system operating at both image and feature levels to fully exploit expert complementarity: the image-level experts provide global anchors to guide the restoration, while the feature-level experts facilitate fine-grained feature synthesis and multi-scale integration. Specifically, the perception-oriented MAR and Diffusion experts are constructed with lightweight generative blocks via low-rank reparameterization to ensure efficient generation and the fidelity-oriented vanilla experts are empowered by spatial-frequency synergistic MoE blocks. The feature-level experts (MoSE blocks) are embedded within the decoder of the image-level vanilla expert. By employing a wavelet-guided frequency-division prior injection, they adaptively integrate low-frequency semantic priors and high-frequency textural details from image-level generative experts to enhance internal feature representations. By leveraging the robust semantic-aware representation space of DINOv3 to perceive semantic shifts across various degradations, our router effectively balances the optimization between semantic reconstruction for local corruptions and detail refinement for global degradations. This enables the precise control of expert coordination and generation: mask sizes are adaptively adjusted to address varying scales of local semantic loss, while iteration steps are modulated based on degradation intensity, ensuring a synergistic collaboration across the MoSE. During inference, we use a tunable parameter g to balance the weights of perception-oriented generative priors, achieving controllable fidelity-perception trade-off. Our main contributions are summarized as follows:

- We identify and analyze the complementary task-adaptive generative priors of MAR and Diffusion models for different degradation types.
- We propose MoSE, a unified mixture-of-experts framework that synergistically integrates perception-oriented generative experts with fidelity-oriented restoration networks to enhance restoration quality.

– We design a image-feature dual-level expert collaboration system with a DINOv3-guided semantic-aware router, enabling adaptive expert coordination and controllable fidelity–perception trade-offs.

2 Related Work

Vanilla Image Restoration. Advances in Vanilla Image Restoration have primarily focused on learning pixel-to-pixel mappings via neural networks [5, 17, 27–33, 55, 59, 69, 71, 74], with innovative architectures being increasingly applied to various degradation types, such as noise [24], rain [64], shadow [66], and haze [79]. Recently, All-in-One (AIO) Image Restoration [15, 23, 43, 73, 77, 82] has emerged as a new direction, which employs a single model to handle multiple degradation types, offering a practical and scalable solution for real-world cross-task restoration scenarios. Most existing AIO methods build upon paradigms such as feature learning [6, 73], prompt learning [4, 43], or contrastive learning [20, 77], aiming to bridge the domain gap among diverse degradations and achieving remarkable performance in terms of reconstruction fidelity. Among them, several leading approaches adopt the Mixture-of-Experts (MoE) architecture to accommodate multiple tasks. For instance, JarvisIR [23] designs specialized agents for different degradations and leverages vision-language models for collaborative restoration, while MoCE [73] introduces experts with varying receptive fields and complexities to adapt to different restoration needs. However, these methods remain limited to fidelity-oriented vanilla expert designs, lacking deliberately integrated task-adaptive priors tailored for cross-task restoration scenarios.

Generative Modeling for Image Restoration. Generative Modeling for Image Restoration presents a distinct paradigm from vanilla approaches, with diffusion models [48, 53, 54] and auto-regressive [13, 56, 72] models leading the trend. These methods learn more general cross-task restoration capabilities from an image generation perspective, achieving superior perceptual quality [25, 26, 35, 47, 49]. For instance, IR-SDE [35] proposes a mean-reversion diffusion process modeled via stochastic differential equations, excelling in handling global degradations like image deraining. RDDM [25] and ResShift [49] introduce residual prediction into the diffusion noise scheduling, enabling efficient image generation. On the other hand, [3] introduces an implicit masked auto-regressive generative model, demonstrating strong performance in local restoration tasks such as inpainting and outpainting. [38] further develops an image-level masked generation paradigm with a multi-step masking strategy for lightweight image restoration. We find that generative approaches exhibit complementary task-adaptive priors when dealing with global and local degradations, highlighting the potential of integrating such generative priors into vanilla restoration frameworks. Inspired by this complementarity, we design Mixture-of-Synergy-Experts (MoSE) that systematically incorporates MAR and diffusion experts, aiming to enhance the cross-task adaptation of all-in-one image restoration in a holistic manner.

3 Methodology

In this section, we detail the architecture of our Mixture-of-Synergy-Experts (MoSE) Framework . As illustrated in Fig. 2, MoSE is built upon a dual-level expert system that systematically integrates the complementary strengths of a fidelity-oriented vanilla model with perception-oriented generative priors at both the image and feature levels. The vanilla expert adopts a U-shaped encoder-decoder architecture, with 3×3 convolutions used in the input and output projection layers. The encoder is responsible for initially clarifying the degraded image into a latent feature representation L_i, and the decoder gradually restores the fidelity result. The generative expert comprises two complementary parts: a **Masked Auto-Regressive (MAR)** expert responsible for local structure reconstruction I_{mar} and a **Diffusion** expert responsible for generating high-frequency detail textures I_{diff}. To introduce generative priors into the decoding process of the latent representation L_i, we designed a novel MoSE Block, which also includes generative and vanilla parts, as shown in Fig. 2(a). The **spatial-frequency synergistic vanilla block** uses a frequency-domain attention module with variable complexity [73] and multi-kernel convolution with different receptive fields [9], while the **lightweight generative block** uses low-rank reparameterized convolution [80]. We also simplified MAR to mean generation from complementary masks [45] and diffusion to a two-step mean regression [49] to improve feature generation efficiency. As shown in Fig. 2(b), through a **Wavelet-guided frequency-division prior injection (WFPI)** mechanism, image-level generative priors are specifically decomposed into multi-scale low-frequency semantics and high-frequency textural details, which are then hierarchically injected into the decoder to enhance internal feature representations. As illustrated in Fig. 2(c), the outputs from the three image-level experts are effectively fused under the **DINOv3-guided Synergy Controllable Routing (SCR)**, which not only orchestrates expert coordination through adaptive weight assignment but also precisely regulates generation dynamics by adjusting mask sizes for varying scales of local semantic loss and modulating iteration steps based on degradation intensity.

3.1 Mixture-of-Synergy-Experts (MoSE)

Vanilla Expert Design As illustrated in Fig. 2, the vanilla expert serves as the backbone of our framework, primarily aimed at learning fidelity-oriented restoration capabilities. We specifically fortify the encoder's ability to extract clean semantic representations, while integrating perception-oriented generative priors during the decoding stage to further enhance restoration performance.

Image-level Vanilla Expert. Image-level vanilla expert utilizes a U-shaped architecture devoid of global skip connections to ensure deep feature refinement. The encoder comprises a four-stage stack of SfhFormer [17] blocks, which projects the degraded input X_i into a latent representation L_i. The decoder consists of three-level MoSE blocks that progressively reconstruct the high-fidelity output.

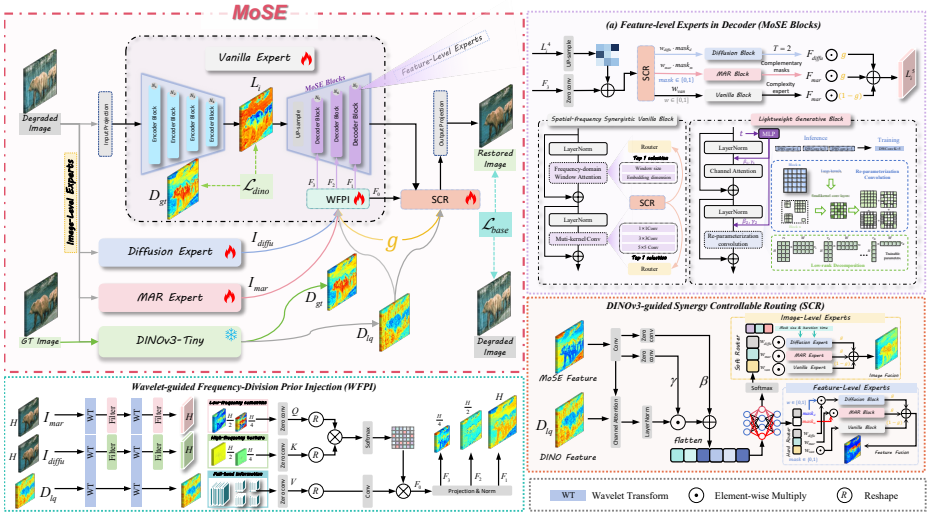


Fig. 2: Proposed Mixture-of-Synergy-Experts (MoSE) framework. We design a dual-level expert system under a mixture-of-experts architecture, achieving effective integration of perception-oriented generative and fidelity-oriented vanilla models through deep collaboration between image-level and feature-level experts. (a) Feature-level complementary experts are integrated in the decoder to achieve enhanced internal feature representations. (b) Complementary priors at the feature and image levels are fully integrated through wavelet-guided frequency-division injection. (c) DINOv3-guided semantic router facilitates synergistic collaboration between generative and vanilla experts by adaptively modulating expert weights and generation parameters.

Feature-level Vanilla Expert. We design the spatial-frequency synergistic vanilla block for Vanilla Expert. To enhance flexibility in handling diverse degradations, we incorporate window attention with adaptive complexities [73] and a multi-kernel convolution module within the Feed-Forward Network (FFN), effectively expanding the model's receptive field and structural adaptivity.

Semantic Guidance. To ensure the model learns robust, cross-task clean feature representations, we employ a pre-trained DINOv3 [51] model to guide the encoding process. The semantic alignment loss is formulated as:

$$\mathcal{L}_{\text{dino}} = \frac{1}{B}\sum_{i=1}^{B}\sqrt{\|\phi_{\text{DINOv3}}(Y_i) - L_i\|_2^2}, \tag{1}$$

where ϕ_{DINOv3} denotes the DINOv3 feature extractor, and here we use ConvNeXt-Tiny for efficiency. B is the batch size and training dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$.

Generative Experts Design To address diverse restoration challenges, we integrate two generative paradigms providing complementary task-adaptive priors. Specifically, both generative experts are built upon lightweight generative blocks

(Fig. 2a), which leverage low-rank reparameterized convolutions [80] to ensure structural flexibility and computational efficiency.

MAR Expert. The image-level MAR expert follows the MaskGIT [3] paradigm, iteratively recovering structural integrity through mask prediction. Given a degraded input X_i, the restoration process is formulated as:

$$Y_t = M_t \odot Y_0 + (1 - M_t) \odot f_\theta(Y_{t-1}), \quad t = 1, \ldots, T_m, \tag{2}$$

where M_t is the binary mask at step t. Crucially, both the mask size and the total iteration steps T_m are treated as adjustable parameters, precisely modulated by the SCR to adapt to the scale and intensity of local semantic loss. Following the training objective of generative masked modeling, the optimization is performed by sampling a time step $t \in [1, T_m]$ and minimizing the reconstruction error:

$$\mathcal{L}_{\text{mar}} = \mathbb{E}_{t,M_t}\Big[\, \|(1 - M_t) \odot f_\theta(Y_{t-1}) - (1 - M_t) \odot Y_i\|_2 \,\Big], \tag{3}$$

where Y_i is the ground truth image. This step-wise supervision ensures that the MAR expert learns to progressively recover clean semantic structures across varying masking ratios.

Diffusion Expert. The diffusion expert models the restoration as a Stochastic Differential Equation (SDE) [35]. It iteratively removes noise and residual components to recover detail-enhanced clean images. The forward degradation process is described by:

$$dx_t = f(x_t, t)dt + g(t)dw_t, \tag{4}$$

where x_t represents the degraded image at time t, $f(x_t, t)$ is the drift coefficient, $g(t)$ is the diffusion coefficient, and w_t is the Wiener process. The reverse process is learned by:

$$dx_t = [f(x_t, t) - g(t)^2 \nabla_{x_t} \log p_t(x_t)]dt + g(t)d\bar{w}_t, \tag{5}$$

where $\nabla_{x_t} \log p_t(x_t)$ is the score function estimated by the diffusion model, guiding the iterative denoising and detail restoration process. Similar to DDPM [16], the optimization objective is described by:

$$\mathcal{L}_{\text{diff}} = \sum_{t=1}^{T_d} \gamma_t \, \mathbb{E}\Big[\big\|\tilde{r}_\theta(X_t, X_0, t) - \epsilon_t\big\|\Big], \tag{6}$$

where $\gamma_1, \ldots, \gamma_{T_d}$ are shift weights and the iteration steps T_d are adaptively modulated by the SCR based on degradation intensity.

Feature-level Simplification. For computational efficiency, we simplify these generative priors for feature-level processing. The feature-level MAR prior P_{mar} is generated using dual complementary masks $M_1 \oplus M_2 = \mathbf{1}$:

$$P_{\text{mar}} = \frac{1}{2}[\mathcal{F}_\theta(M_1 \odot z) + \mathcal{F}_\theta(M_2 \odot z)], \tag{7}$$

where z represents the input feature map, $\mathcal{F}_\theta$ denotes the MAR transformation, and P_{mar} is the resulting generative prior.

The feature-level diffusion prior P_{diff} is approximated via a streamlined two-step regression following the ResShift [49] paradigm:

$$P_{\text{diff}} = \mathcal{G}_\phi(z, t=1) + \mathcal{G}_\phi(z + \mathcal{G}_\phi(z, t=1), t=2), \tag{8}$$

where $\mathcal{G}_\phi$ represents the diffusion transformation at reduced time step t, and P_{diff} is the approximated diffusion prior.

3.2 Wavelet-guided Frequency-division Prior Injection (WFPI)

To effectively integrate image-level insights into the feature-space decoding process, we propose the **WFPI** module (Fig. 2b). In this module, generative priors first pass through zero-initialized convolutional layers to ensure stable optimization. Subsequently, two cascaded wavelet transform layers are employed for lossless downsampling to obtain multi-scale representations. Specifically, we apply frequency-division filtering to decouple the complementary strengths of the two experts: the **MAR prior** is filtered to preserve **low-frequency semantic structures**, while the **diffusion prior** is filtered to retain **high-frequency textural details**. After performing cross-attention fusion with DINOv3 semantic features, the fused representations pass through projection and normalization layers to generate the final multi-scale priors. These priors are then hierarchically injected into the MoSE blocks of the decoder, achieving tight feature integration and synergistic complementary fusion.

3.3 Synergy Controllable Routing (SCR)

The collaboration between all experts is orchestrated by the **DINOv3-guided Synergy Controllable Routing (SCR)** module (Fig. 2c). Leveraging the robust representation space of DINOv3 to perceive semantic shifts across degradations, the SCR balances the optimization between semantic reconstruction and detail refinement. Notably, we generate $\{0, 1\}$ masks for feature-level expert routing weights, ensuring each feature is processed by only one feature-level expert (hard routing). For image-level experts, we employ flexible weighted fusion (soft routing) while simultaneously learning the optimal mask size and iteration steps for the current task.

Controllable Trade-off. To achieve a flexible fidelity-perception trade-off, we employ a dynamic integration strategy. During **training**, we adopt a stochastic selection approach with a probability $p = 0.2$ to choose between the generative-enhanced output and the vanilla reconstruction, ensuring the model excels in both modes, as illustrated in Algorithm 1. During **inference**, similar to Classifier-Free Guidance (CFG), we introduce a tunable parameter $g \in [0, 1]$ to perform linear interpolation:

$$\hat{Y}_{\text{final}} = (1 - g) \cdot \hat{Y}_{\text{vanilla}} + g \cdot \hat{Y}_{\text{gen}}, \tag{9}$$

Algorithm 1 Training Procedure of MoSE Framework

Input: Training dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$; Pre-trained ϕ_{DINOv3}
Parameters: θ_{vanilla} (Enc f_e, Dec f_d), θ_{mar}, θ_{diff}, θ_{SCR}, θ_{WFPI}; E_{wp}: Warmup epochs
Output: Trained unified restoration model $f(\cdot|\theta)$

Initialize all parameters θ; Set $T_m, T_d \leftarrow 20$
for $e = 1, 2, \cdots, E$ **do**
  Sample batch $\mathcal{B} = \{(X_i, Y_i)\}_{i=1}^{B} \sim \mathcal{D}$
  Phase 1: Experts Warmup (if $e \leq E_{wp}$)
    $L_i = f_e(X_i)$; $\hat{Y}_{\text{mar}} = f_{\text{mar}}(X_i; T_m)$; $\hat{Y}_{\text{diff}} = f_{\text{diff}}(X_i; T_d)$
    $\mathcal{L}_{wp} = \mathcal{L}_{\text{dino}}(L_i, \phi(Y_i)) + \mathcal{L}_{\text{mar}} + \mathcal{L}_{\text{diff}}$ ▷ Warm up individual experts
    Update $\{\theta_{\text{vanilla}}.\text{Enc}, \theta_{\text{mar}}, \theta_{\text{diff}}\}$ via $\nabla\mathcal{L}_{wp}$
  Phase 2: End-to-End Joint Training (if $e > E_{wp}$)
    /* Step A: Feature Extraction & Prior Injection */
    $L_i = f_e(X_i)$; $\hat{Y}_{\text{mar}}, \hat{Y}_{\text{diff}} = \text{GenerativeExperts}(X_i)$
    $L_i^{\text{enh}} = \text{WFPI}(L_i, \hat{Y}_{\text{mar}}, \hat{Y}_{\text{diff}})$
    /* Step B: Decoding & CFG Training Strategy */
    $\hat{Y}_{\text{vanilla}}^{\text{enh}} = f_d(L_i^{\text{enh}})$; $w = \text{SCR}(\phi(Y_i), L_i)$
    $\hat{Y}_{\text{gen}} = (w_m \hat{Y}_{\text{mar}} + w_d \hat{Y}_{\text{diff}})/(w_m + w_d)$ ▷ Synergistic Gen-output
    Sample $r \sim \text{Bernoulli}(p)$ ▷ Randomly select training path
    $\hat{Y}_{\text{train}} = \begin{cases} \hat{Y}_{\text{gen}}, & \text{if } r = 1 \\ \hat{Y}_{\text{vanilla}}^{\text{enh}}, & \text{if } r = 0 \end{cases}$
    /* Step C: Total Optimization */
    $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}}(\hat{Y}_{\text{train}}, Y_i) + \lambda_1 \mathcal{L}_{\text{dino}} + \lambda_2 \mathcal{L}_{\text{route}}$
    Update all parameters θ via $\nabla\mathcal{L}_{\text{total}}$
  if $e = E$ **then** ▷ Structural reparameterization for inference
    $f_{mar,diff}$.reparameterize()
return $f(\cdot|\theta)$

where $\hat{Y}_{\text{gen}}$ is the synergistic output of generative experts. This enables users to continuously navigate the trade-off between pixel-level fidelity ($g = 0$) and high-level perceptual quality ($g = 1$).

Routing Optimization. At the feature level, we apply hard routing via $\{0, 1\}$ masks to ensure computational efficiency. At the image level, we employ soft routing while simultaneously learning task-optimal mask sizes and iteration steps. To optimize the overall inference efficiency, the routing loss minimizes the total computational cost by aggregating the normalized complexities of the generative experts (iteration counts T_m, T_d) and the vanilla expert (parameter count $\|\theta_v\|_0$):

$$\mathcal{L}_{\text{route}} = \frac{1}{B} \sum_{i=1}^{B} \left(\hat{T}_m^{(i)} + \hat{T}_d^{(i)} + \|\hat{\theta_v^{(i)}}\|_0 \right)^2, \tag{10}$$

where $\hat{T}_m, \hat{T}_d$, and $\|\hat{\theta_v}\|_0$ represent the complexity metrics normalized to a unified scale. This constraint encourages the SCR module to adaptively allocate computational resources according to the degradation severity, selecting lighter

Table 1: *Comparison to state-of-the-art on three degradations.* PSNR (dB, ↑) , SSIM (↑) , LPIPS (↓) and our method achieves competitive performance.

Method	Venue	Params.	*Dehazing*		*Deraining*		*Denoising*						Average		
			SOTS		Rain100L		$BSD68_{\sigma=15}$		$BSD68_{\sigma=25}$		$BSD68_{\sigma=50}$				
MPRNet [75]	CVPR 21	36M	30.58	0.974	36.37	0.972	33.98	0.933	31.31	0.888	28.06	0.799	32.06	0.913	0.188
AirNet [20]	CVPR 22	9M	27.94	0.962	34.90	0.967	33.92	0.933	31.26	0.888	28.00	0.797	31.20	0.909	0.165
PromptIR [43]	NeurIPS 23	34M	30.37	0.970	37.15	0.972	33.93	0.931	31.37	0.887	28.11	0.801	32.19	0.912	0.194
Gridformer [8]	IJCV 24	33M	30.83	0.979	37.94	0.982	34.06	0.934	31.42	0.891	28.14	0.801	32.48	0.917	0.182
DA-CLIP [36]	ICLR 24	125M	29.46	0.963	36.28	0.968	30.02	0.821	24.86	0.585	22.29	0.476	28.58	0.763	0.159
UniProcessor [10]	ECCV 24	1002M	31.66	0.979	38.17	0.982	34.08	0.935	31.42	0.891	28.17	0.803	32.70	0.918	0.142
MoCE-IR [73]	CVPR 25	25M	31.34	0.979	38.57	0.984	34.11	0.932	31.45	0.888	28.18	0.800	32.73	0.917	0.184
JarvisIR [23]	CVPR 25	1300M	31.38	0.969	38.56	0.981	34.12	0.933	31.46	0.876	28.15	0.802	32.73	0.912	0.181
FoundIR [21]	ICCV 25	146M	30.89	0.961	38.33	0.967	33.24	0.909	30.32	0.768	27.67	0.765	32.09	0.874	0.176
MoSE (ours)	-	42M	31.65	0.980	38.58	0.984	34.14	0.934	31.47	0.894	28.11	0.803	32.79	0.919	0.132

Table 2: *Comparison to state-of-the-art on five degradations.* PSNR (dB, ↑) , SSIM (↑) , LPIPS (↓) and our method achieves competitive performance on average.

Method	Venue	Params.	*Dehazing*		*Deraining*		*Denoising*		*Deblurring*		*Low-Light*		Average		
			SOTS		Rain100L		$BSD68_{\sigma=25}$		GoPro		LoL				
AirNet [20]	CVPR 22	9M	21.04	0.884	32.98	0.951	30.91	0.882	24.35	0.781	18.18	0.735	25.49	0.847	0.241
DGUNet [40]	CVPR 22	17M	24.78	0.940	36.62	0.971	31.10	0.883	27.25	0.837	21.87	0.823	28.32	0.891	0.227
Restormer [74]	CVPR 22	26M	24.09	0.927	34.81	0.962	31.49	0.884	27.22	0.829	20.41	0.806	27.60	0.882	0.252
Transweather [58]	CVPR 22	38M	21.32	0.885	29.43	0.905	29.00	0.841	25.12	0.757	21.21	0.792	25.22	0.836	0.244
NAFNet [5]	ECCV 22	17M	25.23	0.939	35.56	0.967	31.02	0.883	26.53	0.808	20.49	0.809	27.77	0.881	0.215
IDR [77]	CVPR 23	15M	25.24	0.943	35.63	0.965	31.60	0.887	27.87	0.846	21.34	0.826	28.34	0.893	0.224
Gridformer [8]	IJCV 24	33M	26.79	0.951	36.61	0.971	31.45	0.885	29.22	0.884	22.59	0.831	29.33	0.904	0.206
MambaIR [14]	ECCV 24	27M	25.81	0.944	36.55	0.971	31.41	0.884	28.61	0.875	22.49	0.832	28.97	0.901	0.251
InstructIR-5D [7]	ECCV 24	17M	27.10	0.956	36.84	0.973	31.40	0.873	29.40	0.886	23.00	0.836	29.55	0.905	0.187
MoCE-IR [73]	CVPR 25	25M	30.48	0.974	38.04	0.982	31.34	0.887	30.05	0.899	23.00	0.852	30.58	0.919	0.217
JarvisIR [23]	CVPR 25	1300M	30.43	0.962	37.98	0.975	31.43	0.886	29.75	0.884	22.95	0.848	30.51	0.911	0.182
FoundIR [21]	ICCV 25	146M	30.32	0.954	37.97	0.970	31.42	0.887	29.93	0.896	22.98	0.850	30.52	0.911	0.193
MoSE (ours)	-	42M	30.61	0.978	38.42	0.989	31.65	0.889	30.24	0.899	23.14	0.860	30.81	0.923	0.154

expert configurations for simpler tasks while reserving high-intensity iterations and larger parameter sets for complex restoration challenges.

Total Objective. The framework is optimized via a multi-task loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{dino}} + \lambda_2 \mathcal{L}_{\text{route}}, \tag{11}$$

where $\mathcal{L}_{\text{recon}}$ denotes the L_1 reconstruction loss.

The optimization of the MoSE framework during training is shown in Algorithm 1. To ensure a stable convergence, we first warm up the generative experts and the vanilla encoder to establish reliable prior foundations, followed by an end-to-end joint optimization that orchestrates the synergistic collaboration across the entire framework

4 Experiments

Following the general settings of previous works [36,73], we conducted experiments under four All-in-One (AIO) image restoration settings: *(i) three degradations*

Fig. 3: Visual results for 10 types of degradation removal with other all-in-one methods. We successfully restored various degraded images and recovered more background details compared to other methods.

(**AIO-3**), *(ii) five degradations* (**AIO-5**), *(iii) ten degradations* (**AIO-10**) and *(iv) composited degradations.* In all AIO tasks, one model is trained to restore datasets containing different degradations.

Datasets and Metrics. We used BSD400 [2], WED [37], and BSD68 [39] with Gaussian noise at levels $\sigma \in [15, 25, 50]$ for denoising, Rain100L, Rain100H [70] and RainDrop [44] for rainy; SOTS [19] for hazy, GoPro [41] for blurry, LOL [62] for low-light, DIV2K [1], Flickr2K [57] and LIVE1 [50] for JPEG, RePaint [34] for Inpainting, SRD [46] for shadow, CDD11 [15] for the composite degradation. We use peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [61] for fidelity evaluation, and the learned perceptual image patch similarity (LPIPS) [78] for perceptual evaluation.

Implementation details. We implemented our method on the PyTorch platform using 4 NVIDIA RTX 3090 GPUs. The optimizer used is Adam [18] with an exponential decay rate of 0.9. The initial learning rate is set to 1e-4, and a cosine annealing strategy is employed for adjustment. For the architectural configuration, the vanilla expert employs a block distribution of $[1, 1, 1, 24]$ in the encoder and $[2, 4, 4]$ in the decoder, with an initial embedding dimension of $C = 32$. The generative experts (MAR and Diffusion) utilize a more lightweight configuration with block settings of $[1, 1, 1, 16]$ and $[1, 1, 1]$ for their respective components, starting with an initial embedding dimension of $C = 16$. During training, the input images are cropped into 128×128 patches with a total batch size of 16.

The loss balancing coefficients are empirically set as $\lambda_1 = 0.05$ and $\lambda_2 = 0.2$. More details are provided in the supplementary material.

Table 3: *Comparison to state-of-the-art on 10 degradations.* PSNR (dB, ↑) , SSIM (↑) , LPIPS (↓) and our method achieves competitive performance on average.

Method	Venue	Params.	Average (AIO-10)		
			PSNR (↑)	SSIM (↑)	LPIPS (↓)
AirNet [20]	CVPR 22	9M	25.62	0.844	0.182
Restormer [74]	CVPR 22	26M	26.44	0.850	0.157
NAFNet [5]	ECCV 22	17M	26.34	0.847	0.159
PromptIR [43]	NeurIPS 23	34M	27.14	0.859	0.148
DA-CLIP [36]	ICLR 24	125M	27.00	0.794	0.127
MoCE-IR [73]	CVPR 25	25M	27.29	0.806	0.162
JarvisIR [23]	CVPR 25	1300M	27.43	0.855	0.136
FoundIR [21]	ICCV 25	146M	27.36	0.843	0.144
MoSE (ours)	-	42M	28.46	0.861	0.126

Table 4: *Comparison to state-of-the-art on composited degradations.* PSNR (dB, ↑) , SSIM (↑) , LPIPS (↓) and our method achieves competitive results compared with even larger models on composited degradations.

Method	Venue	Params.	*CDD11-Single*								CDD11-Double										CDD11-Triple				Average	
			Low (L)		Haze(H)		Rain(R)		Snow(S)		L+H		L+R		L+S		H+R		H+S		L+H+R		L+H+S			
AirNet [20]	CVPR 22	9M	24.83	.778	24.21	.951	26.55	.891	26.79	.919	23.23	.779	22.82	.710	23.29	.723	22.21	.868	23.29	.901	21.80	.708	22.24	.725	23.75	.814	.524
PromptIR [43]	NeurIPS 23	34M	26.32	.805	26.10	.969	31.56	.946	31.53	.960	24.49	.789	25.05	.771	24.51	.761	24.54	.924	23.70	.925	23.74	.752	23.33	.747	25.90	.850	.471
WGWSNet [82]	CVPR 23	26M	24.39	.774	27.90	.982	33.15	.964	34.43	.973	24.27	.800	25.06	.772	24.60	.765	27.23	.955	27.65	.960	23.90	.772	23.97	.771	26.96	.863	.355
WeatherDiff [42]	TPAMI 23	83M	23.58	.763	21.99	.904	24.85	.885	24.80	.888	21.83	.756	22.69	.730	22.12	.707	21.25	.868	21.99	.868	21.23	.716	21.04	.698	22.49	.798	.424
OneRestore [15]	ECCV 24	6M	26.48	.826	32.52	.990	33.40	.964	34.31	.973	25.79	.822	25.58	.799	25.19	.789	29.99	.957	30.21	.964	24.78	.788	24.90	.791	28.47	.878	.253
MoCE-IR [73]	CVPR 25	25M	27.26	.824	32.66	.990	34.31	.970	35.91	.980	26.24	.817	26.25	.800	26.04	.793	29.93	.964	30.19	.970	25.41	.789	25.39	.790	29.05	.881	.215
JarvisIR [23]	CVPR 25	1300M	27.31	.828	32.68	.991	34.40	.972	35.92	.983	26.28	.825	26.29	.804	26.10	.796	29.84	.951	30.20	.966	25.32	.787	25.34	.791	29.06	.881	.203
FoundIR [21]	ICCV 25	146M	27.28	.812	32.43	.978	34.37	.966	35.85	.979	26.29	.824	26.17	.809	26.06	.791	29.95	.963	30.06	.951	25.44	.784	25.31	.783	29.02	.876	.311
MoSE (ours)	-	42M	27.43	.833	32.85	.992	34.47	.982	35.95	.984	26.47	.830	26.34	.816	26.17	.799	29.99	.966	30.34	.978	25.61	.791	25.74	.799	29.21	.888	.189

4.1 Comparison to State-of-the-Art Methods

AIO-3. As shown in Tab. 1, we achieve competitive performance on three degradations, performing favorably against the vanilla model MoCE-IR [73] in fidelity while approaching the generative model UniProcessor [10] in perceptual metrics with roughly 96% fewer parameters, validating our effectiveness.

AIO-5. As shown in Tab. 2, after expanding the degradations to 5, the performance gain of our method is further improved, achieving competitive results across the degradations. This verifies the robustness of MoSE for cross-task image restoration. It is worth noting that our method compares favorably with powerful VLM-based models [7, 23], suggesting that our architecture offers competitive robustness with far fewer parameters.

AIO-10. As shown in Tab. 3, by further increasing degradation types to 10, our method achieves a notable average gain of 1.03 dB, which demonstrates that MoSE learns more robust features when dealing with a variety of different degradations. The visual results in Fig. 3 validate this property, demonstrating

Table 5: Ablation study of different components in our proposed MoSE. The baseline is evaluated on the AIO-10 benchmark. Bold indicates the best performance.

Configuration	PSNR ↑	SSIM ↑	LPIPS ↓
MoSE (Ours)	**28.46**	**0.861**	**0.126**
w/o DINOv3	25.67	0.746	0.436
w/o $\mathcal{L}_{dino}$	26.85	0.797	0.341
w/o $\mathcal{L}_{route}$	26.79	0.797	0.334
w/o MAR	27.15	0.775	0.359
w/o Diffusion	26.58	0.713	0.346
w/o Experts Warmup	26.05	0.768	0.413
w/o Feature Expert	27.46	0.787	0.294
w/o Image Expert	27.65	0.808	0.264
w/o WFPI	26.85	0.769	0.284
w/o SCR	27.07	0.799	0.274

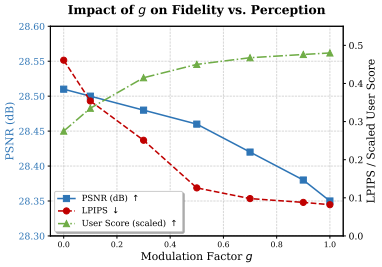
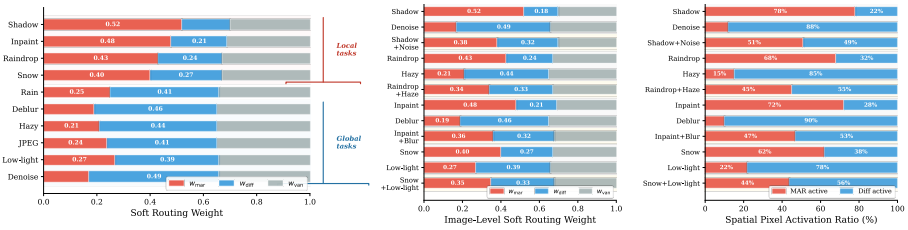


Fig. 4: Performance metrics under different g values. As g increases, the model prioritizes perceptual quality (LPIPS) over pixel-level fidelity (PSNR). We set $g = 0.5$ in the inference stage.



(a) Routing weights on single degradations. (b) Routing under mixed degradations.

Fig. 5: DINOv3-guided SCR routing behavior. (a) The router emergently assigns the MAR expert to local degradations and the Diffusion expert to global degradations across 10 single degradation types. (b) For mixed degradations, the router smoothly interpolates between the two experts without conflict.

that MoSE can fully utilize complementary robust priors to cope with both global and local degradations, significantly improving the restoration effect.

Composited Degradations. As shown in Tab. 4, MoSE achieves competitive performance on average across the 11 degradation settings when dealing with more realistic composite degradation scenarios, which further verifies the robustness.

4.2 Ablation Study

The ablations in this section were all conducted on **AIO-10**, with additional details available in **supplemental material**.

DINO Guidance. MoSE uses the semantic generalization of DINOv3 to guide the encoding of degraded images and complementary expert routing. Under the guidance of DINOv3's generalization prior, the model can more effectively remove common artifacts in the encoding stage. Considering the trade-off between effi-

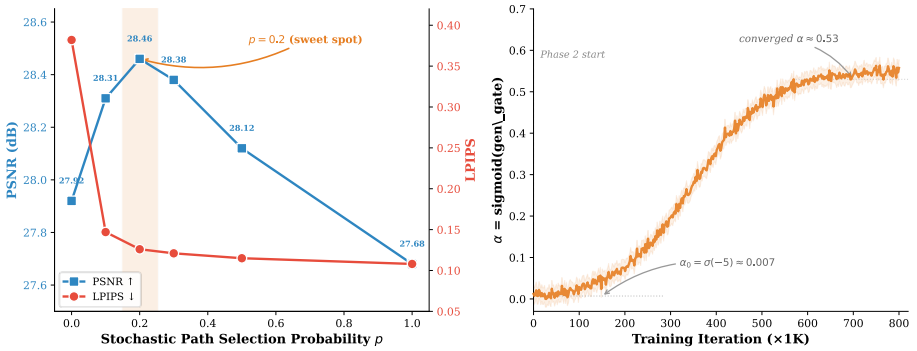


Fig. 6: Effect of the stochastic path selection probability p. $p = 0.2$ achieves the best fidelity-perception balance, and the generation gate α converges smoothly.

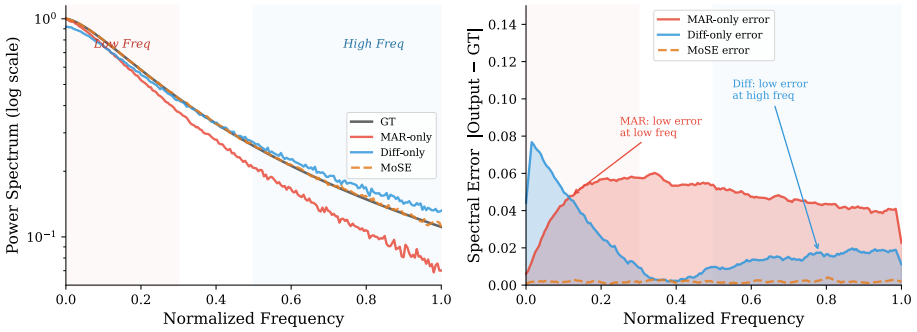


Fig. 7: Frequency-domain analysis of generative priors. MAR yields lower spectral error at low frequencies (structure), while Diffusion yields lower error at high frequencies (texture), supporting the WFPI frequency-division design.

ciency and effectiveness, MoSE adopts ConvNeXt-Tiny as the DINOv3 backbone. Leveraging DINOv3's robust semantic-aware representations, the model perceives distinct semantic shifts across various degradations. This enhanced discriminative capability allows the SCR module to achieve precise, task-adaptive control over the MoSE experts. As shown in Fig. 5(a), the learned SCR routing weights emergently converge to a local/global pattern through end-to-end training without explicit local/global labels, assigning higher weights to the MAR expert for local degradations and to the Diffusion expert for global degradations. Moreover, when multiple degradations co-occur in a single image, the router produces a smooth interpolation between the two experts rather than a hard switch (Fig. 5(b)): for a composite shadow+noise input, the MAR weight gradually decreases while the Diffusion weight increases without expert conflict, indicating that the DINOv3-guided routing generalizes gracefully to mixed degradations. Tab. 5 also shows the necessity of DINOv3 guidance in both the encoding and routing stages. Only when all stages are enabled can the best results be achieved by applying the model, this further confirms the effectiveness of our proposed method.

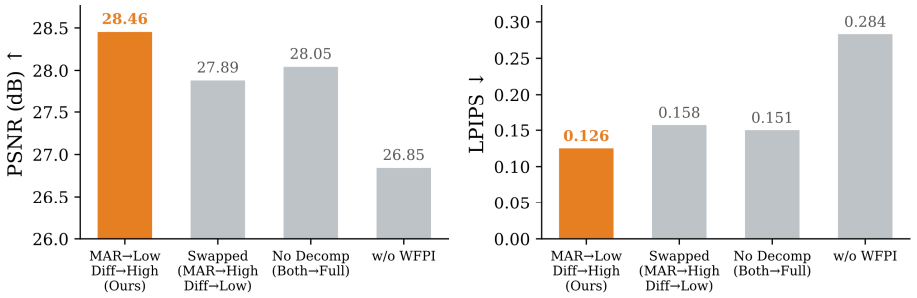


Fig. 8: Ablation of the WFPI frequency assignment on AIO-10. Our MAR→low, Diffusion→high design performs best.

Complementary Expert. MoSE introduces two generative models, MAR and diffusion, with complementary task-adaptive priors to enhance the vanilla model. Tab. 5 shows that the best results cannot be achieved without either of these generative models, verifying the necessity of complementary experts.

Fidelity-perception Trade-off. MoSE allows control over the contribution of generative priors during the inference phase by adjusting parameter g. Fig. 4 shows the impact of adjusting parameter g on fidelity and perceptual quality: increasing g increases the perceptual contribution of generative priors but sacrifices fidelity quality, which verifies the inference adjustability of our method. We further study the stochastic path selection probability p used during training (Algorithm 1). As shown in Fig. 6, $p = 0$ degenerates the model toward the vanilla path, while an overly large p degrades fidelity; $p = 0.2$ best balances fidelity and perceptual quality, and the generation gate converges smoothly during training.

WFPI Frequency Assignment. The WFPI module assigns the MAR prior to low-frequency (structural) features and the Diffusion prior to high-frequency (textural) features. Fig. 7 provides frequency-domain evidence: MAR outputs exhibit lower spectral error at low frequencies, while Diffusion outputs exhibit lower error at high frequencies, justifying our frequency-division design. As reported in Fig. 8, our MAR→low / Diffusion→high assignment outperforms the swapped assignment by +0.57 dB and the no-decomposition variant by +0.41 dB, confirming the effectiveness of the frequency-division injection.

5 Conclusion

All-in-one image restoration requires robust handling of diverse degradations within a single model. While Mixture-of-Experts architectures show promise, existing methods rely solely on vanilla experts, lacking specialized task-adaptive prior. We explore the complementary adaptation strengths of generative models: MAR experts handle local degradations better while diffusion models restore more global details. Our MoSE framework deeply integrates both generative experts via a dual-level architecture with DINOv3 guidance, enabling synergistic coordination and controllable fidelity-perception trade-offs. Extensive experiments validate its effectiveness.

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 (CVPR) Workshops. pp. 126–135 (2017)
2. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **33**(5), 898–916 (2010)
3. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11315–11325 (2022)
4. Chen, H., Li, W., Gu, J., Ren, J., Chen, S., Ye, T., Pei, R., Zhou, K., Song, F., Zhu, L.: Restoreagent: Autonomous image restoration agent via multimodal large language models. In: NeurIPS (2024)
5. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: Proceedings of the 17th European Conference on Computer Vision (ECCV). pp. 17–33. Springer (2022)
6. Chen, W.T., Huang, Z.K., Tsai, C.C., Yang, H.H., Ding, J.J., Kuo, S.Y.: Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17653–17662 (June 2022)
7. Conde, M.V., Geigle, G., Timofte, R.: Instructir: High-quality image restoration following human instructions. In: ECCV (2024)
8. dense transformer with grid structure for image restoration in adverse weather conditions, G.R.: Gridformer: Residual dense transformer with grid structure for image restoration in adverse weather conditions. International Journal of Computer Vision **132**(10), 4541–4563 (2024)
9. Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., Sun, J.: Repvgg: Making vgg-style convnets great again. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13733–13742 (2021)
10. Duan, H., Min, X., Wu, S., Shen, W., Zhai, G.: Uniprocessor: a text-induced unified low-level image processor. In: European Conference on Computer Vision. pp. 180–199. Springer (2024)
11. Fan, Z., Lu, X., Liu, Y., Huang, J., Li, D., Fu, X., Yin, B.: Bird-sr: Bidirectional reward-guided diffusion for real-world image super-resolution. arXiv preprint arXiv:2602.07069 (2026)
12. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3855–3863 (2017)
13. Gregor, K., Danihelka, I., Mnih, A., Blundell, C., Wierstra, D.: Deep autoregressive networks. In: International Conference on Machine Learning. pp. 1242–1250. PMLR (2014)
14. Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.T.: Mambair: A simple baseline for image restoration with state-space model. arXiv preprint arXiv:2402.15648 (2024)
15. Guo, Y., Gao, Y., Lu, Y., Liu, R.W., He, S.: Onerestore: A universal restoration framework for composite degradation. In: European Conference on Computer Vision (2024)

16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
17. Jiang, X., Zhang, X., Gao, N., Deng, Y.: When fast fourier transform meets transformer for image restoration. In: European Conference on Computer Vision. pp. 381–402. Springer (2024)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), https://api.semanticscholar.org/CorpusID:6628106
19. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1), 492–505 (2019)
20. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: CVPR (2022)
21. Li, H., Chen, X., Dong, J., Tang, J., Pan, J.: Foundir: Unleashing million-scale training data to advance foundation models for image restoration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12626–12636 (2025)
22. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems **37**, 56424–56445 (2024)
23. Lin, Y., Lin, Z., Chen, H., Pan, P., Li, C., Chen, S., Wen, K., Jin, Y., Li, W., Ding, X.: Jarvisir: Elevating autonomous driving perception with intelligent image restoration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22369–22380 (2025)
24. Liu, H., Fu, J., Xie, Q., Meng, D.: Rotation-equivariant self-supervised method in image denoising. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12720–12730 (2025)
25. Liu, J., Wang, Q., Fan, H., Wang, Y., Tang, Y., Qu, L.: Residual denoising diffusion models. In: CVPR (2024)
26. Liu, Z., Xu, Z., Ma, J., Li, W., Wang, R., Du, B., Chen, H.: Conditional visual autoregressive modeling for pathological image restoration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17828–17837 (2025)
27. Lu, X., Bao, Y., Yang, J., Hu, A., Xiao, J., Wang, K., Li, D., Xu, S., Liu, K., Fu, X., Zha, Z.J.: Evenformer: Dynamic even transformer for real-world image restoration. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops. pp. 1081–1091 (June 2025)
28. Lu, X., Fu, X., Xiao, J., Fan, Z., Zhu, Y., Zha, Z.J.: Elucidating and endowing the diffusion training paradigm for general image restoration. arXiv preprint arXiv:2506.21722 (2025)
29. Lu, X., Peng, Y., Ge, C., Sun, Z., Zhou, Z., Liao, Z., Li, Z., Li, D., Kang, Q., Fu, X., Zha, Z.J.: Boosting inverse tone mapping via diffusion regularization. In: 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 5654–5663 (2025). https://doi.org/10.1109/ICCVW69036.2025.00591
30. Lu, X., Sun, Z., Ge, C., Peng, Y., Zhou, Z., Li, Z., Liao, Z., Li, D., Kang, Q., Fu, X., Zha, Z.J.: Efficient high fps non-uniform motion deblurring via progressive learning. In: 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 5644–5653 (2025). https://doi.org/10.1109/ICCVW69036.2025.00590
31. Lu, X., Xiao, J., Zhu, Y., Fu, X.: Continuous adverse weather removal via degradation-aware distillation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 28113–28123 (June 2025)

32. Lu, X., Yang, J., Bao, Y., Fan, Z., Hu, A., Wang, K., Xiao, J., Wang, X., Liu, H., Fu, X., Zha, Z.J.: Advancing ambient lighting normalization via diffusion shadow generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops. pp. 1070–1080 (June 2025)
33. Lu, X., Zhu, Y., Wang, X., Li, D., Xiao, J., Zhang, Y., Fu, X., Zha, Z.J.: Hirformer: Dynamic high resolution transformer for large-scale image shadow removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 6513–6523 (June 2024)
34. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11461–11471 (2022)
35. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Image restoration with mean-reverting stochastic differential equations. In: ICML (2023)
36. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Controlling vision-language models for universal image restoration. In: ICLR (2024)
37. Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., Zhang, L.: Waterloo exploration database: new challenges for image quality assessment models. IEEE Transactions on Image Processing **26**(2), 1004–1016 (2016)
38. Ma, X., Wei, Z., Jin, Y., Ling, P., Liu, T., Wang, B., Dai, J., Chen, H.: Masked pre-training enables universal zero-shot denoiser. Proceedings of Advances in Neural Information Processing Systems (NeurIPS) **37**, 32570–32610 (2024)
39. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV (2001)
40. Mou, C., Wang, Q., Zhang, J.: Deep generalized unfolding networks for image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17399–17410 (2022)
41. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3883–3891 (2017)
42. Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. IEEE transactions on pattern analysis and machine intelligence **45**(8), 10346–10357 (2023)
43. Potlapalli, V., Zamir, S.W., Khan, S.H., Shahbaz Khan, F.: Promptir: Prompting for all-in-one image restoration. NeurIPS (2024)
44. Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J.: Attentive generative adversarial network for raindrop removal from a single image. In: CVPR (2018)
45. Qin, C.J., Wu, R.Q., Liu, Z., Lin, X., Guo, C.L., Park, H.H., Li, C.: Restore anything with masks: Leveraging mask image modeling for blind all-in-one image restoration (2024), https://arxiv.org/abs/2409.19403
46. Qu, L., Tian, J., He, S., Tang, Y., Lau, R.W.: Deshadownet: A multi-context embedding deep network for shadow removal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4067–4075 (2017)
47. Rajagopalan, S., Narayan, K., Patel, V.M.: Restorevar: Visual autoregressive generation for all-in-one image restoration (2025), https://arxiv.org/abs/2505.18047
48. Rissanen, S., Heinonen, M., Solin, A.: Generative modelling with inverse heat dissipation. In: Proceedings of International Conference on Learning Representations (ICLR) (2022)

49. Selikhanovych, D., Li, D., Leonov, A., Gushchin, N., Kushneriuk, S., Filippov, A., Burnaev, E., Koshelev, I., Korotin, A.: One-step residual shifting diffusion for image super-resolution via distillation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
50. Sheikh, H.: Live image quality assessment database release 2. http://live. ece. utexas. edu/research/quality (2005)
51. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
52. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=St1giarCHLP
53. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 12438–12448 (2020)
54. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (ICLR) (2021)
55. Tan, J., Yu, H., Huang, J., Yang, Z., Zhao, F.: Diffloss: unleashing diffusion model as constraint for training image restoration network. In: Proceedings of the Asian Conference on Computer Vision. pp. 1566–1584 (2024)
56. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
57. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: NTIRE 2017 challenge on single image super-resolution: methods and results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 114–125 (2017)
58. Valanarasu, J.M.J., Yasarla, R., Patel, V.M.: Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2353–2363 (2022)
59. Wang, S., Zhang, J., Huang, J., Zhao, F.: Image-free pre-training for low-level vision. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8825–8834 (2024)
60. Wang, S., Zheng, N., Huang, J., Zhao, F.: Navigating image restoration with var's distribution alignment prior. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7559–7569 (2025)
61. 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)
62. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. In: BMVC (2018)
63. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: Diffir: Efficient diffusion model for image restoration. ICCV (2023)
64. Xiao, J., Fu, X., Liu, A., Wu, F., Zha, Z.J.: Image de-raining transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(11), 12978–12995 (2023). https://doi.org/10.1109/TPAMI.2022.3183612

65. Xiao, J., Fu, X., Zhou, M., Liu, H., Zha, Z.: Random shuffle transformer for image restoration. In: International Conference on Machine Learning (2023), https://api.semanticscholar.org/CorpusID:260957206
66. Xiao, J., Fu, X., Zhu, Y., Li, D., Huang, J., Zhu, K., Zha, Z.J.: Homoformer: Homogenized transformer for image shadow removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25617–25626 (June 2024)
67. Xiao, J., Fu, X., Zhu, Y., Zha, Z.J.: Bayesian window transformer for image restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**(3), 2431–2444 (2026). https://doi.org/10.1109/TPAMI.2025.3626947
68. Xu, H., Huang, J., Huang, L., Li, D., Liu, Y., Zhao, F.: Freedna: Endowing domain adaptation of diffusion-based dense prediction with training-free domain noise alignment. arXiv preprint arXiv:2506.22509 (2025)
69. Xu, H., Huang, J., Yu, W., Tan, J., Zou, Z., Zhao, F.: Adaptive dropout: Unleashing dropout across layers for generalizable image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7513–7523 (2025)
70. Yang, W., Tan, R.T., Feng, J., Liu, J., Guo, Z., Yan, S.: Deep joint rain detection and removal from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
71. Yang, Z., Yu, H., Li, B., Zhang, J., Huang, J., Zhao, F.: Unleashing the potential of the semantic latent space in diffusion models for image dehazing. In: European Conference on Computer Vision. pp. 371–389. Springer (2024)
72. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Randomized autoregressive visual generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18431–18441 (2025)
73. Zamfir, E., Wu, Z., Mehta, N., Tan, Y., Paudel, D.P., Zhang, Y., Timofte, R.: Complexity experts are task-discriminative learners for any image restoration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12753–12763 (2025)
74. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022)
75. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: CVPR (2021)
76. Zhang, G., Tan, J., Huang, L., Yuan, Z., Yao, M., Huang, J., Zhao, F.: Infoscale: Unleashing training-free variable-scaled image generation via effective utilization of information. arXiv preprint arXiv:2509.01421 (2025)
77. Zhang, J., Huang, J., Yao, M., Yang, Z., Yu, H., Zhou, M., Zhao, F.: Ingredient-oriented multi-degradation learning for image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5825–5835 (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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018)
79. Zhang, X., Dong, H., Pan, J., Zhu, C., Tai, Y., Wang, C., Li, J., Huang, F., Wang, F.: Learning to restore hazy video: A new real-world dataset and a new method. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9235–9244 (2021). https://doi.org/10.1109/CVPR46437.2021.00912

80. Zhang, Y., Ding, X., Yue, X.: Scaling up your kernels: Large kernel design in convnets towards universal representations. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
81. Zheng, G., Li, Y., Pan, Y., Deng, J., Yao, T., Zhang, Y., Mei, T.: Hierarchical masked autoregressive models with low-resolution token pivots. ICML (2025)
82. Zhu, Y., Wang, T., Fu, X., Yang, X., Guo, X., Dai, J., Qiao, Y., hua Hu, X.: Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 21747–21758 (2023), https://api.semanticscholar.org/CorpusID:259144110