

Graphical Abstract

BIR-Adapter: A parameter-efficient diffusion adapter for blind image restoration

Cem Eteke, Alexander Griessel, Wolfgang Kellerer, Eckehard Steinbach

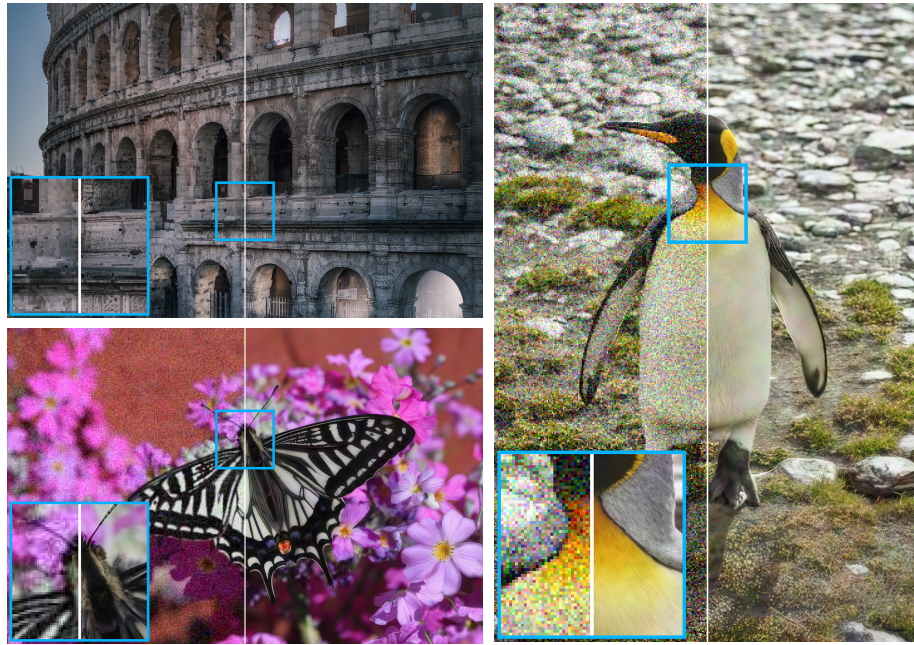


Figure 1: Example restoration results of BIR-Adapter on images degraded with $4\times$ downsampling. The lower-left image is further degraded with blur, white noise, and JPEG compression, while the right-hand image includes additional white noise.

Highlights

BIR-Adapter: A parameter-efficient diffusion adapter for blind image restoration

Cem Eteke, Alexander Griessel, Wolfgang Kellerer, Ekehard Steinbach

- Provides empirical evidence that diffusion models retain aligned representations under unknown image degradations.
- Leverages internal features to enable blind image restoration without auxiliary feature extractors.
- Introduces a parameter-efficient restoring attention mechanism that couples degraded and denoised latent features.
- Achieves strong performance with significantly fewer additional trained parameters.
- Preserves a plug-and-play adapter design that generalizes across diffusion-based restoration methods.

BIR-Adapter: A parameter-efficient diffusion adapter for blind image restoration

Cem Eteke^{a,c,*}, Alexander Griessel^b, Wolfgang Kellerer^b, Eckehard Steinbach^{a,c}

^a*Chair of Media Technology,*

^b*Chair of Communication Networks,*

^c*Munich Institute of Robotics and Machine Intelligence*

School of Computation, Information and Technology

Technical University of Munich, Munich, 80333, Germany

Abstract

We introduce the BIR-Adapter, a parameter-efficient diffusion adapter for blind image restoration. Diffusion-based restoration methods have demonstrated promising performance in addressing this fundamental problem in computer vision, typically relying on auxiliary feature extractors or extensive fine-tuning of pre-trained models. Motivated by the observation that large-scale pretrained diffusion models can retain informative representations under common image degradations, BIR-Adapter introduces a parameter-efficient, plug-and-play attention mechanism that substantially reduces the number of trained parameters. To further improve reliability, we propose a sampling guidance mechanism that mitigates hallucinations during the restoration process. Experiments on synthetic and real-world degradations demonstrate that BIR-Adapter achieves competitive, and in several settings superior, performance compared to state-of-the-art methods while requiring up to $36\times$ fewer trained parameters. Moreover, the adapter-based design enables seamless integration into existing models. We validate this generality by extending a super-resolution-only diffusion model to handle additional unknown degradations, highlighting the adaptability of our approach for broader image restoration tasks.

Keywords: Diffusion models, blind image restoration, parameter efficiency

*Corresponding author

Email address: cem.eteke@tum.de (Cem Eteke)

1. Introduction

Blind image restoration is the process of restoring the clean version of a degraded observation where the underlying degradation process is unknown. This problem is fundamentally challenging due to the wide range of degradations, including noise, blur, downsampling, compression artifacts, and complex cascades thereof. Supervised deep neural networks demonstrated promising performance by learning a direct mapping between degraded and clean images [1]. Furthermore, with the advent of generative methods such as Generative Adversarial Networks (GAN) [2] and Diffusion Models [3], the restoration performance has been further improved as generative modeling enables the synthesis of perceptually realistic details even when the observation is severely degraded [4].

The formulation of the denoising diffusion process in the image space [5] or latent space [6], in combination with the large-scale training of attention-based neural architectures [7], enables Diffusion Models to be utilized as versatile image priors for various computer vision tasks, including blind image restoration [8]. Extending these large-scale models for the image restoration task is achieved either by training task-specific architectures from scratch or by including auxiliary feature extractors that condition the pretrained models on degraded observations. While effective, the former sacrifices the strong image prior learned from large-scale data, and the latter incurs substantial training and memory costs due to the large number of additional trained parameters.

Large pretrained diffusion models belong to the broader class of generalizable large-scale models, *i.e.*, foundation models [9]. One prominent example from this class of models, CLIP [10], has been shown to exhibit a degree of robustness to distribution shifts that occur due to degradations. Cheng *et al.* demonstrated that CLIP features remain informative under out-of-distribution noise, enabling effective denoising by training only parameter-efficient decoders [11].

Inspired by this and to provide intuition for our approach, we examine how a pretrained latent diffusion model responds to degraded inputs. As shown

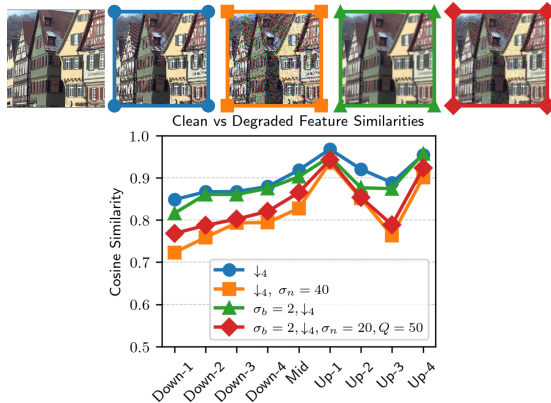


Figure 2: Cosine similarity between latent representations of clean and degraded images across different layers of a U-Net-based latent diffusion model. Similarities are measured at the outputs of the Downsampling (Down), Middle (Mid), and Upsampling (Up) blocks of the U-Net. The degradations are combinations of $4\times$ downsampling ($\downarrow_4$), additive white noise (σ_n), Gaussian blur (σ_b), and JPEG compression (Q).

in Fig. 2, we analyze the discrepancy in latent representations by measuring the cosine similarity between latent features extracted from a clean image and its degraded counterpart. The results hint that feature similarity is preserved to varying degrees across layers and degradation types. While this analysis does not constitute proof of robustness, it provides empirical evidence that pre-trained diffusion models retain informative representations under degradations. This observation suggests that auxiliary feature extractors, commonly used to guide diffusion-based restoration models, may be unnecessary, and that features extracted directly from the diffusion model itself can be leveraged for image restoration, enabling parameter-efficient conditioning of pretrained diffusion backbones.

Based on this insight, we introduce *BIR-Adapter*, a parameter-efficient, adapter attention mechanism for diffusion-based blind image restoration. Rather than training an auxiliary network to extract degradation-aware features, BIR-Adapter reuses features extracted by the frozen diffusion backbone itself and integrates them into lightweight self-referential restoring attention modules. To further

improve reliability during inference, BIR-Adapter additionally incorporates a guided sampling strategy that mitigates hallucinations, particularly in low-frequency regions. We evaluate BIR-Adapter on both synthetic and unknown real-world degradation benchmarks. Our results demonstrate that the proposed method achieves competitive and, in several cases, superior performance compared to state-of-the-art approaches, while requiring up to $36\times$ fewer additional trained parameters. Moreover, the adapter-based design enables seamless integration into existing diffusion models. As a proof of concept, we extend a super-resolution-only diffusion model to handle additional unknown degradations, highlighting the flexibility and generality of our approach.

2. Related work

2.1. Deep image restoration

Image restoration is a long-standing problem in low-level computer vision, with early methods relying on handcrafted priors tailored to specific degradations [12]. With the rise of deep learning, CNN-based models significantly improved restoration performance by learning direct mappings from degraded images to clean targets, addressing tasks such as denoising [13], super-resolution [14], and deblurring [15]. However, these methods typically assumed known or single degradations and showed limited generalization to unknown or compounded cases [16]. To mitigate this, later works introduced more realistic degradation modeling [17] and architectural enhancements, including multi-scale representations [18] and enlarged receptive fields [19]. Transformer-based models further improved performance by capturing long-range dependencies, as demonstrated by SwinIR [20], Restormer [21], and Uformer [22]. More recent efforts also explored frequency-domain modeling [23], subbands [24], and parameter-efficient attention mechanisms [25]. Despite these advances, most deep learning-based restoration methods remain regression-driven and struggle to recover perceptually realistic details under severe ambiguity, thereby limiting their ability to recover details in blind settings.

2.2. Generative image restoration

Restoration via generative modeling was initially explored using GAN-based methods such as BSRGAN [17], and Real-ESRGAN [26], which introduced realistic synthetic degradation pipelines to enable blind restoration under complex and mixed degradations. In contrast to regression-based approaches, generative models primarily emphasize perceptual quality rather than strict pixel-level fidelity. While diffusion models do not change this fundamental trade-off, they have been shown to outperform GANs in image synthesis [3]. Moreover, large-scale pretrained diffusion models have emerged as particularly strong priors over natural images, making them well-suited for generative image restoration [4].

ControlNet introduced a way to steer these powerful priors by fine-tuning a pretrained copy of the model that injects features into the original model [27]. This prior-based control is also applied to restoration. PASD introduced pixel-wise cross-attention and a ControlNet-based module for super-resolution [28], while StableSR introduced feature warping to improve image restoration [29]. Despite their effectiveness, such models still face challenges when dealing with unknown or compounded degradations. DiffBIR addressed this by incorporating pretrained restoration models, such as SwinIR, into a ControlNet-based pipeline to better generalize across degradation types [30]. SUPIR extended this two-stage approach to large-scale degradation estimation [31], and Kong *et al.* extended it with Transformers [32]. AdaptBIR divides the degradation estimation into various levels via quality estimation [33]. Furthermore, task-based restoration is utilized to better preserve the semantics by SeeSR [34], and EDTR [35]. Though very effective in restoring perceptually high-quality and realistic images, these approaches require a feature extractor, namely ControlNet.

2.3. Parameter-efficient diffusion models

To address the high complexity associated with large ControlNet-based models, Cheng *et al.* proposed lightweight architectures trained from scratch for super-resolution, introducing DiT-SR [36]. However, this eliminates the strong prior associated with large diffusion models. One efficient approach is low-rank

adaptation [37], which is not applicable to image restoration as it assumes the model is static, but for restoration, the models need to be extended to take degraded images into account. On the other hand, Adapter-based alternatives to complete fine-tuning emerged as a low-complexity way to leverage diffusion priors [38]. Liang et al. introduced the Diffusion Restoration Adapter, which combines adapters and lower rank training as a low-complexity approach to blind image restoration [39]. However, similar to ControlNet, adapter-based methods, although more parameter-efficient, still rely on external feature extractors. As discussed in the previous section, the diffusion prior itself removes the need for such auxiliary modules. Leveraging this insight, we introduce BIR-Adapter, illustrated in Figure 3.

3. Background

In this section, we provide the necessary background that underpins our work. We divided the background into three main parts, in which we discuss the latent diffusion models, sampling, and attention mechanisms.

3.1. Latent diffusion models

Diffusion models aim to revert a known T -step forward diffusion process defined by the decaying scheduling parameters β_t where $t \in \{0, \dots, T\}$ [5]. This process begins with a known vector $\mathbf{x}_0$ and gradually introduces noise up to timestep t . This defines a Markovian forward process q as

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \quad (1)$$

We can directly compute the noisy state $\mathbf{x}_t$ by the reparameterization trick where $\bar{\alpha}_t$ represents the cumulative scheduling parameter and formulated as $\bar{\alpha}_t = \prod_{i=0}^{t-1} \alpha_i$ where $\alpha_t = 1 - \beta_t$

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}). \quad (2)$$

A diffusion model aims to reverse this process by modeling the reverse process as a parametric function p_θ

$$p_{\theta}(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mu_{\theta}(\mathbf{x}_t, t), \Sigma_{\theta}(\mathbf{x}_t, t)), \quad (3)$$

where μ_{θ} and Σ_{θ} are controlled by a denoising network $\epsilon_{\theta}(\mathbf{x}_t, t)$. This model is trained to minimize

$$\mathbb{E}[(\boldsymbol{\epsilon} - \epsilon_{\theta}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, t))^2], \quad (4)$$

where the data is generated using the forward model in Eq. (2) reformulated as

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \quad (5)$$

Although diffusion models are shown to outperform GANs, it has been challenging to perform stable synthesis [3]. Rombach *et al.* resolved this issue by introducing the Latent Diffusion Models (LDM) [6]. Instead of supplying a noise-free image I_0 as $\mathbf{x}_0$, they perform the diffusion process on a latent space, dictated by an autoencoder with an encoder $\mathbf{x}_0 = \mathcal{E}(I_0)$ and a decoder $I_0 = \mathcal{D}(\mathbf{x}_0)$.

Finally, for controllable diffusion models, the denoising model is extended to include a control signal $\mathbf{y}_0$ as $\epsilon_{\theta}(\mathbf{x}_t, \mathbf{y}_0, t)$. This $\mathbf{y}_0$ could be used to train the LDM together with it or fed through a fine-tuned approach, such as ControlNet [27] or adapters [38]. The former requires large datasets to train billion-scale models from scratch, while the latter enables the utilization of these hyper-scale LDMs as a prior. ControlNet fine-tunes a pretrained copy of ϵ_{θ} to extract features from $\mathbf{y}_0$ and inject them into the original model. However, as discussed above, this approach also requires fine-tuning of a large set of parameters, thereby increasing the overall model’s training complexity. Adapter-based approaches extract features from $\mathbf{y}_0$ more efficiently and, instead of using residual connections, they utilize cross-attention at the layers of the original model. This approach reduces complexity; however, it still requires a large number of parameters due to the feature extraction. In Section 1, we empirically show that auxiliary feature extraction is inefficient for diffusion-based blind image restoration, as the diffusion model’s internal representations for clean and degraded inputs exhibit similarity to some degree.

3.2. Sampling

During inference, the goal is to go from noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ to a clean image $\mathbf{x}_0$. Song *et al.* introduced Denoising Diffusion Implicit Models, which utilize the estimated noise-free latents $\mathbf{x}_{t \rightarrow 0}$ to solve the inverse process

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_{t \rightarrow 0} + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_\theta(\mathbf{x}_t, \mathbf{y}_0, t). \quad (6)$$

The noise-free latents are estimated by simply inverting the forward model in Eq. (5) by replacing the initial noise with the model prediction

$$\mathbf{x}_{t \rightarrow 0} = \left(\frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, \mathbf{y}_0, t)}{\sqrt{\bar{\alpha}_t}} \right). \quad (7)$$

3.3. Attention mechanism

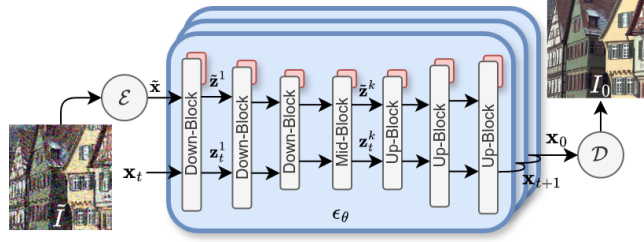
Attention, introduced in the Transformer architecture [7], allows models to dynamically weight input features based on their relevance to each other. This is achieved through learned similarity scores between query and key vectors. Let $\mathbf{z}_t^k$ be the output of the layer k of $\epsilon_\theta(\mathbf{x}_t, \mathbf{y}_0, t)$. Self-attention is computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (8)$$

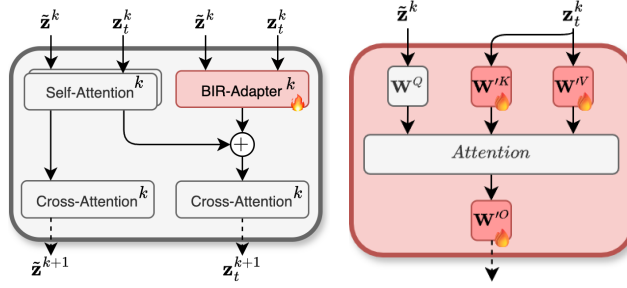
where $Q = \mathbf{z}_t^k \mathbf{W}^Q$, $K = \mathbf{z}_t^k \mathbf{W}^K$, and $V = \mathbf{z}_t^k \mathbf{W}^V$, and $\{\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V\}$ constitute the trainable attention weights. Key and value can be replaced with some other modality to form cross-attention. For example, it can be text embeddings of prompts to enable prompt-based image synthesis via $K = \mathbf{y}_0 \mathbf{W}^K$, and $V = \mathbf{y}_0 \mathbf{W}^V$. A typical LDM consists of self-attention and cross-attention layers attached to one another consecutively, as shown in Fig. 3. Both attention types finally transform the attention output via $\mathbf{W}^O$.

4. Methodology

In blind image restoration an unknown degradation function $\mathcal{F}$ and noise η were applied to an image such that $\tilde{I} = \mathcal{F}(I_0) + \eta$, and the goal is to restore I_0



(a) Latent diffusion model with BIR-Adapter



(b) Restoring attention

(c) BIR-Adapter

Figure 3: BIR-Adapter in a denoising diffusion model ϵ_θ . BIR-Adapter (c) introduces a self-referential restoring attention mechanism within an attention block (b) of the denoising model (a). The model incorporates intermediate features $\bar{\mathbf{z}}^k$ of the degraded latents $\tilde{\mathbf{x}}$, which are extracted by ϵ_θ itself. This design leverages the observation that diffusion models retain informative representations under degradation and enables a parameter-efficient adaptation without auxiliary feature extractors. For clarity, the U-Net-based LDM in (a) is simplified by visualizing fewer blocks and omitting residual connections, and the number of attention layers is reduced. In (b), self-attention is applied in a cascaded manner with parallel processing of degraded and diffused features.

from $\tilde{I}$. We introduce the BIR-Adapter, which comprises a self-referential restoring attention-based adapter for LDMs and a diffusion guidance mechanism. As motivated in Sec. 1 we utilize LDMs as natural image priors for parameter efficiency. Since we utilize LDMs, we treat this problem as restoration on the latent space, *i.e.* obtaining $\mathbf{x}_0$ from degraded latents $\tilde{\mathbf{x}} = \mathcal{E}(\tilde{I})$.

4.1. BIR-Adapter

Intuitively, attention determines how query features aggregate information from value features based on key-query relevance. In blind image restoration,

this corresponds to refining degraded features based on their relevance to the structures of the clean image. While this ideal formulation would require access to clean features as keys and values, such features are unavailable in the blind setting.

Parallel to the diffused latents $\mathbf{x}_t$, we pass the degraded latents $\tilde{\mathbf{x}}$ through ϵ_θ in the same batch and end up with, for each layer k , the features $\tilde{\mathbf{z}}^k$. As we discussed in Sec. 3.2, the reverse diffusion process solved by a denoising model ϵ_θ , denoises, *i.e.* cleans an image. Inspired by this, we discussed in Sec. 1 that one can form the hypothesis that when ϵ_θ is pretrained at scale, degraded features should exhibit similarity to clean ones. Fig. 2 shows, for an example image under various degradations, a proof of concept for this idea. Please note that we introduced Figure 2 as a visual motivation for the idea behind our approach and do not claim robustness of the diffusion models. There are factors such as the diffusion timestep and the text prompt that affect the latent representations. However, the degraded features should contain enough information about the original image, as there are similarities between the features $\mathbf{z}_t^k$ and $\tilde{\mathbf{z}}^k$. As such, we treat the currently diffused features $\mathbf{z}_t^k$ of a layer k at diffusion timestep t as the *clean* features that were missing in the formulation above. Using these definitions, we introduce the restoring attention mechanism as follows

$$Attention(Q, K, V)\mathbf{W}^O + Attention(\tilde{Q}, K', V')\mathbf{W}'^O, \quad (9)$$

where $\tilde{Q} = \tilde{\mathbf{z}}^k \mathbf{W}^Q$, $K' = \mathbf{z}_t^k \mathbf{W}'^K$, and $V' = \mathbf{z}_t^k \mathbf{W}'^V$. This formulation allows us to fine-tune the parameter set $\{\mathbf{W}'^K, \mathbf{W}'^V, \mathbf{W}'^O\}$ of each self-attention layer, initialized via their self-attention counterparts, resulting in parameter efficiency. Figure 3 visualizes this architecture inside a simplified LDM. We do not fine-tune the query weights since we want to preserve the features of the degraded image.

4.2. Guided sampling

As the resolution of an image increases and goes beyond the supported resolution of an LDM, a common practice is to tile the latent space with overlaps

and execute the diffusion process on the tiles. Finally, the tiles are merged with certain weighting functions [29]. In our approach, we observed that if a tile only consists of low-frequency information, BIR-Adapter leads to some severe hallucinations. A common practice to avoid this issue is to use an initial restoration step, either via an off-the-shelf method [30] or a custom-trained one [31], and combine them with frequency estimation [35]. We assume an initial restoration I_0^{init} exists.

We use this information to guide the diffusion-based sampling discussed in Sec. 3.2. However, we do not apply this guidance to high-frequency regions. To achieve this, we utilize the downsampled and normalized gradients $\mathcal{G}_\downarrow$ estimated via the Sobel filter [30]. We would like to note that the filter choice does not affect the results, as we are only searching for a general estimate. We define the downsampled and normalized gradient operator $\mathcal{G}_\downarrow(I)$ applied to an input image I as

$$\mathcal{G}_\downarrow(I) = \frac{\text{Sobel}(I) \downarrow s}{\|\text{Sobel}(I) \downarrow s\|}, \quad (10)$$

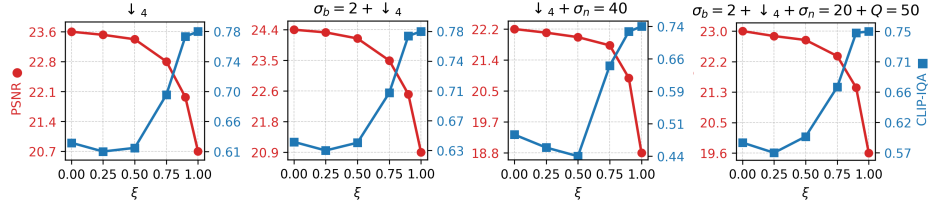
where $\downarrow s$ indicates downsampling by a factor s , and division with the norm $\|\cdot\|$ normalizes the result to the range $[0, 1]$. Here, the downsampling factor s is dictated by the encoder $\mathcal{E}$ of the latent diffusion model. Using this operator, we compute the corresponding guidance weights similar to DiffBIR [30] as

$$\mathcal{W} = 1 - \mathcal{G}_\downarrow(I_0^{init}). \quad (11)$$

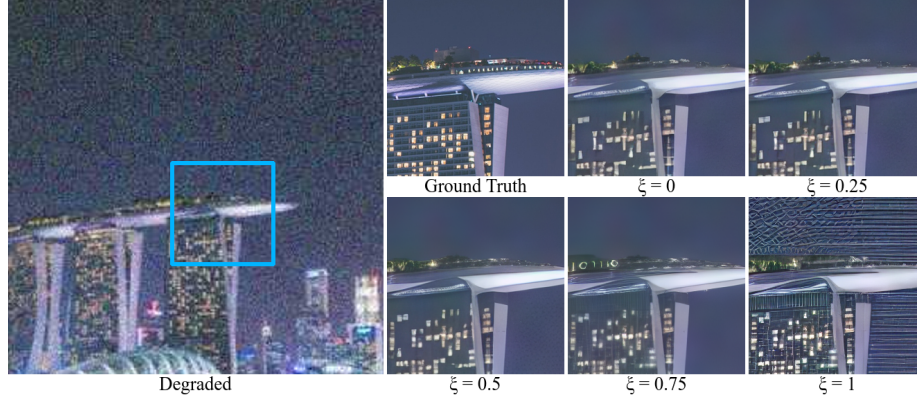
We then use the estimated noise-free latent $\mathbf{x}_{t \rightarrow 0}$ found in Eq. (6) and update it to include the clean low-frequency information that the diffusion process will follow, reducing the likelihood of hallucination

$$\mathbf{x}_{t \rightarrow 0} \leftarrow \begin{cases} \mathbf{x}_{t \rightarrow 0} + \mathcal{W} \odot (\mathcal{E}(I_0^{init}) - \mathbf{x}_{t \rightarrow 0}), & \text{if } \frac{t}{T} > \xi \\ \mathbf{x}_{t \rightarrow 0}, & \text{o/w.} \end{cases} \quad (12)$$

Figure 4 provides quantitative and visual analysis of the proposed guidance mechanism, with quantitative results computed over the evaluation dataset. We



(a) Quantitative effect of ξ in terms of PSNR and CLIP-IQA under various combinations of 4× downsampling ($\downarrow_4$), white noise (σ_n), Gaussian blur (σ_b), and JPEG (Q). Reported values are averaged over the DIV2K validation dataset.



(b) Visual effect of the guidance parameter ξ on the restoration of a single example.

Figure 4: Quantitative and visual analysis of the effect of ξ . An increase in CLIP-IQA indicates higher-quality images, while a sudden drop in PSNR suggests inconsistencies.

examine the effect of ξ on fidelity, measured by PSNR, and perceptual quality, measured by CLIP-IQA [40]. Stronger guidance (lower ξ) increases PSNR but reduces perceptual quality, while weaker guidance (higher ξ) improves CLIP-IQA at the cost of PSNR. Notably, disabling guidance ($\xi = 1$) results in a sharp drop in PSNR, indicating increased inconsistency. Overall, $0.75 \leq \xi \leq 0.9$ yields the best trade-off. We emphasize that, for generative restoration methods, a higher PSNR does not necessarily imply better perceptual restoration quality; similarly, a higher CLIP-IQA does not imply consistency. Therefore, we explicitly report the trade-off between PSNR and CLIP-IQA.

5. Experiments

We validated our approach through a series of experiments involving simulated and real-world degradations that were previously unknown. In this section, we provide the details of the implementation and evaluation setup.

5.1. Training

We utilized the Stable Diffusion 1.5 (SD-1.5) backbone. It has 860×10^6 frozen parameters, and BIR-Adapter introduces only an additional 37×10^6 extra trainable parameters. We initialized $\mathbf{W}'^K$ and $\mathbf{W}'^V$ from their counterparts and $\mathbf{W}'^O$ as zeros; thus, BIR-Adapter has no initial effect. We discussed how we utilized these weights in Sec. 4.1. Fig. 3 illustrates the architecture.

For training, we utilized a mixture of datasets, consisting of the training sets of DIV2K [41], DIV8K [42], Flickr2K [43], WED, OutdoorSceneTrain [44], and the first 5,000 images from the FFHQ dataset [45], creating a training set of size 25,000 images. During training, we selected a random square patch of size 512×512 . We degraded the patch using the stochastic degradation model utilized by Real-ESRGAN [26].

Once we have the clean and degraded image pair I_0 and $\tilde{I}$, we encode them to the latent space by the encoder of the latent diffusion model $\mathcal{E}$, resulting in $\mathbf{x}_0$ and $\tilde{\mathbf{x}}$ respectively. For each training step, we sampled an $\mathbf{x}_t$ by using the formulation in Eq. (2). We passed $\tilde{\mathbf{x}}$ and $\mathbf{x}_t$ together to the denoising model ϵ_θ and executed the BIR-Adapter-extended model as shown in Fig. 3. We finally used the loss formulation in Eq. (4) to train the parameters of the BIR-Adapter. Notice how $\tilde{\mathbf{x}}$ is always the same for each diffusion step, and it is not utilized in the loss.

We utilized the AdamW optimizer with learning rate 10^{-5} and weight decay of 10^{-2} [46]. We used $3 \times$ NVIDIA A40 GPUs with 48GB VRAM each. We set the batch size and gradient accumulation to 4. This resulted in an effective batch size of $3 \times 4 \times 4 = 48$. We trained the model for 150×10^3 steps, which took a week to complete.

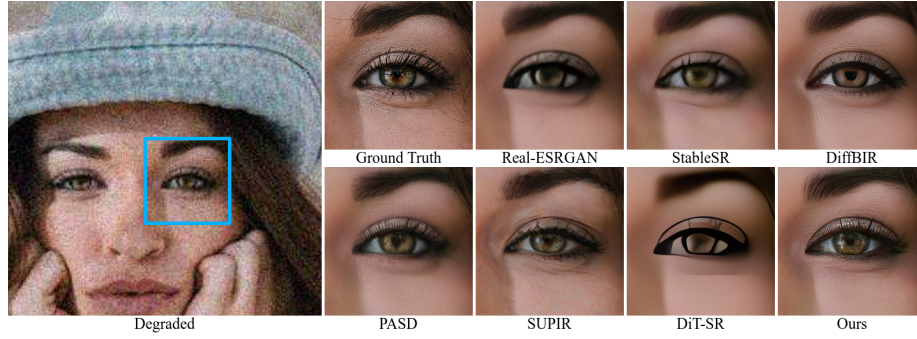
5.2. Evaluation

We utilized two main datasets to present both quantitative and qualitative results. First, we used the validation set of the DIV2K dataset. We measured the performance on various synthetic degradation settings: (i) downsampling ($\downarrow$), (ii) Gaussian blur and downsampling ($\sigma_b + \downarrow$), (iii) downsampling and white noise ($\downarrow + \sigma_n$), and (iv) Gaussian blur, downsampling, white noise, and JPEG compression ($\sigma_b + \downarrow + \sigma_n + Q$). These settings represent common and challenging mixed degradations in blind restoration. We further evaluated our method using RealSR [47]. While degradations such as motion blur or spatially varying effects are not explicitly modeled in the synthetic setting, the RealSR dataset includes unknown real-world degradations. Figure 5 displays examples from both datasets. We always included super-resolution as a part of the restoration of these two datasets. We tiled the latent space to a size of 64×64 with a stride of 32×32 to support high-resolution images. We combined these tiles after each diffusion step using Gaussian weights and then re-tiled them. This iterative approach avoids blocking [29].

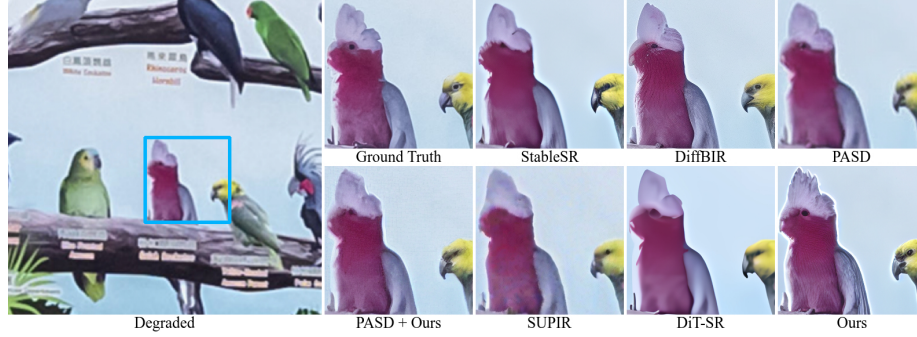
We ran the synthesis for 20 diffusion steps, which provides a good balance between restoration quality and computational cost and is commonly adopted in recent diffusion-based image restoration works. We used class-free guidance with a guidance weight of 9.0 [48]. We set the guidance parameter $\xi = 0.9$. We qualitatively and quantitatively discussed the choice in Sec. 4.2. We used Real-ESRGAN for the initial restoration I_0^{init} , which serves as a lightweight and widely used baseline capable of handling complex and mixed degradations, while contributing only marginally to the overall computational cost.

5.2.1. Baseline methods

We compared our method with non-generative SwinIR [20], GAN-based Real-ESRGAN [26], and diffusion prior-based methods that utilize ControlNet, such as StableSR [29] and PASD [28], which are specifically designed for super-resolution. The central focus of our method is to address unknown and complex degradations. Hence, we included DiffBIR [30] and SUPIR [31] in our com-



(a) DIV2K: Synthetic degradation $\sigma_b = 2 + \downarrow_4 + \sigma_n = 20 + Q = 50$.



(b) RealSR: unknown real degradation. Restoration of $\downarrow_4$ and unkown degradations.

Figure 5: Example degraded and restored images using the baselines and our method. We used synthetic degradation on the DIV2K dataset (5a) while RealSR contains $4\times$ downsampled images with further unknown degradations (5b).

Table 1: Comparison of the number of additional *trainable* parameters, number of training images, and restoration time. The reported parameter counts exclude the frozen backbone parameters, which are nevertheless involved during inference.

	StableSR	DiffBIR	PASD	SUPIR	DiT-SR	DR-Adapter	Ours
Trainable Parameters ($\times 10^6$)	490	363	609	1,332	61	80	37
Training Images ($\times 10^3$)	15	15,000	45	20,000	98	300	25
Restoration time (s)	3.03	3.86	3.60	7.50	4.42	-	7.42

parisons as universal restoration methods. Finally, we compared our approach with DiT-SR, which does not utilize a pretrained diffusion prior [36]. We set up the same synthetic degradation experimentation as SUPIR. We used the source codes provided by the authors to run the baseline models. We used static prompts. We compare the number of additional trained parameters and the number of images used for training in Tab. 1 and provide qualitative comparisons in Fig. 5.

5.2.2. Parameter efficiency

We introduce a parameter-efficient approach that leverages degraded features extracted by the diffusion model itself, rather than relying on auxiliary feature extractors such as ControlNet. Table 1 compares the number of additional trained parameters of the baseline methods, along with restoration times, measured for 512×512 inputs using 20 diffusion steps on an A40 GPU. Our approach requires significantly fewer parameters. We note that parameter efficiency does not directly translate to lower inference cost: one limitation of our approach is the increased runtime, as the degraded image is processed in parallel with the diffused sample. We discuss this in more detail below.

5.2.3. Adapter baseline

We also compare our approach with a recent adapter-based blind image restoration method, namely DR-Adapter, which we discussed in Sec. 2 [39]. However, the approach is closed-source. The authors reported performance on the DIV2K dataset under synthetic degradations, but the degradation parameters are unknown. However, they report real-life performance on the Re-

alPhoto60 dataset [31]. We compare the performances on the RealPhoto60 dataset in Tab. 4, including DR-Adapter performance as reported by the authors. We do not use guidance in this setting, as tiling is not necessary.

5.2.4. Metrics

We utilized a wide range of metrics to measure the performance. These metrics include pixel-level reference-based metrics such as PSNR and MS-SSIM, a perceptual reference-based metric LPIPS [49], and reference-free metrics CLIP-IQA [40], ManIQA [50], and MUSIQ [51] that are shown to correlate with human preferences. The reference-free metrics are shown to correlate more with human perception, as pixel-level reference-based metrics do not reflect the perceptual quality of an image. However, reference-free metrics do not necessarily imply consistency, as we discussed in the context of hallucination-quality trade-off in Sec. 4.2. We report qualitative results in Fig. 5 and additional ones in the appendix. We report performance of the methods in terms of the aforementioned metrics in Tab. 2, Tab. 3, and Tab. 4.

5.3. Plug & play

We selected PASD to evaluate the plug-and-play nature of the BIR Adapter. The reasoning is twofold: (i) PASD shares the same backbone, and (ii) it is super-resolution-only, and we want to show that the performance on unknown degradations can be improved. This experiment serves as proof of concept that our method could enable new applications. We disabled the guidance mechanism as PASD already serves as a guiding restoration. We report the performance of PASD extended with BIR-Adapter on RealSR in Tab. 5.

5.4. Ablation studies

In Sec. 4.2, we introduced a guidance mechanism to mitigate hallucinations and additionally report results without guidance ($\xi = 1$); its quantitative and qualitative effects are shown in Fig. 4b. We further evaluate two non-fine-tuned variants to assess our design choices at the attention and BIR-Adapter levels.

Experiment	Metric	Real-ESRGAN (ICCV '21)	SwinIR (ICCV '21)	StableSR (IJCV '24)	DiffBIR (CVPR '24)	PASD (ECCV '24)	SUPIR (CVPR '24)	DiT-SR (AAAI '25)	Ours	Ours w/o guidance
$\downarrow_4$	CLIP-IQA	0.566	0.583	0.608	0.788	0.534	0.532	0.552	0.772	0.779
	ManIQA	0.367	0.378	0.373	0.581	0.358	0.359	0.397	0.529	0.548
	MUSIQ	63.597	64.487	64.798	69.83	60.882	63.824	63.49	69.212	69.232
	LPIPS $\downarrow$	0.243	0.233	0.245	0.322	0.267	0.264	0.264	0.29	0.37
	PSNR	23.873	24.014	24.018	22.921	24.314	23.589	24.921	22.188	20.827
	MS-SSIM	0.894	0.899	0.894	0.855	0.881	0.863	0.91	0.839	0.783
$\sigma_b = 2$ $\downarrow_4$	CLIP-IQA	0.57	0.581	0.582	0.788	0.476	0.568	0.563	0.776	0.782
	ManIQA	0.39	0.39	0.362	0.579	0.326	0.38	0.403	0.529	0.554
	MUSIQ	64.811	65.269	63.169	70.296	56.439	66.602	63.302	69.617	69.51
	LPIPS $\downarrow$	0.231	0.215	0.252	0.303	0.281	0.247	0.267	0.276	0.363
	PSNR	25.149	25.665	25.213	23.667	25.811	24.438	26.18	22.778	21.072
	MS-SSIM	0.923	0.933	0.915	0.879	0.914	0.894	0.946	0.864	0.8
$\downarrow_4$ $\sigma_n = 40$	CLIP-IQA	0.466	0.568	0.451	0.655	0.585	0.532	0.51	0.728	0.74
	ManIQA	0.286	0.307	0.27	0.363	0.399	0.363	0.431	0.473	0.512
	MUSIQ	50.21	57.649	50.377	49.157	62.501	65.0	61.042	68.043	69.686
	LPIPS $\downarrow$	0.444	0.404	0.436	0.648	0.449	0.359	0.457	0.399	0.488
	PSNR	22.016	21.725	22.788	18.39	20.733	21.87	22.012	21.224	19.006
	MS-SSIM	0.778	0.804	0.794	0.699	0.721	0.765	0.798	0.757	0.635
$\sigma_b = 2$ $\downarrow_4$ $\sigma_n = 20$ $Q = 50$	CLIP-IQA	0.556	0.572	0.481	0.805	0.479	0.563	0.513	0.752	0.754
	ManIQA	0.369	0.371	0.298	0.592	0.328	0.374	0.453	0.5	0.524
	MUSIQ	60.471	62.186	54.195	70.11	56.65	65.921	62.542	69.003	69.755
	LPIPS $\downarrow$	0.348	0.328	0.352	0.419	0.391	0.35	0.462	0.374	0.469
	PSNR	23.267	23.928	24.152	22.272	23.073	22.242	22.467	21.774	19.832
	MS-SSIM	0.833	0.836	0.839	0.797	0.812	0.78	0.811	0.78	0.68

Table 2: Evaluation results on dataset DIV2K under various combinations of synthetic degradations of $4\times$ downsampling ($\downarrow_4$), white noise (σ_n), Gaussian blur (σ_b), and JPEG compression (Q). We highlight **best**, *second-best*, and **top-3** performances. Lower LPIPS values signify better results, denoted by $\downarrow$. We present the performance of our method both with ($\xi = 0.9$) and without guidance ($\xi = 1$).

Metric	Real-ESRGAN (ICCV '21)	SwinIR (ICCV '21)	StableSR (IJCV '24)	DiffBIR (CVPR '24)	PASD (ECCV '24)	SUPIR (CVPR '24)	DiT-SR (AAAI '25)	Ours w/o guidance
CLIP-IQA	0.49	0.469	0.505	<i>0.751</i>	0.38	0.517	0.529	0.767
ManIQA	0.367	0.343	0.351	<i>0.58</i>	0.295	0.386	0.508	0.609
MUSIQ	59.668	59.595	58.173	<i>67.188</i>	47.608	62.35	64.484	68.265
LPIPS ↓	0.273	0.251	<i>0.272</i>	0.414	0.286	0.343	0.326	0.469
PSNR	24.446	<i>25.02</i>	24.345	23.494	26.057	24.126	22.658	20.347
MS-SSIM	0.886	0.895	0.879	0.855	<i>0.89</i>	0.851	0.87	0.76

Table 3: Evaluation results on dataset RealSR, containing 4× downsampling and unknown degradations. We present the performance of our method both with ($\xi = 0.9$) and without guidance ($\xi = 1$)

Metric	SwinIR (ICCV '21)	StableSR (IJCV '24)	DiffBIR (CVPR '24)	PASD (ECCV '24)	SUPIR (CVPR '24)	DR-Adapter (Reported)	Ours w/o guidance
CLIP-IQA	0.646	0.523	0.817	0.637	0.798	0.706	<i>0.811</i>
ManIQA	0.551	0.287	0.646	0.402	<i>0.642</i>	0.529	0.618
MUSIQ	57.459	31.173	<i>71.946</i>	63.903	71.109	71.62	73.141

Table 4: Evaluation results on dataset RealPhoto60, containing unknown degradations. The performance of Diffusion Restoration Adapter (DR-Adapter) is directly reported as the code is not available [39]. We omit guidance as we do not supersample in this experiment. We highlight **best**, *second-best*, and *top-3* performances. Lower LPIPS values signify better results, denoted by ↓.

	CLIP-IQA	ManIQA	MUSIQ	LPIPS ↓	PSNR	MS-SSIM
PASD	0.38	0.295	47.608	0.286	26.057	0.89
PASD + Ours	0.612	0.376	56.463	0.371	25.159	0.87

Table 5: The performance of PASD and its BIR-Adapter extension on the RealSR dataset.

The first replaces self-attention with cross-attention, while the second retains the original attention structure but copies self-attention weights into the BIR-Adapter. Both variants perform poorly; thus, we provide results only in the appendix. These experiments confirm the necessity of both the BIR-Adapter formulation and fine-tuning of the key and value weights.

6. Results & discussion

6.1. Restoration quality

Tab. 2 presents the performance of the baselines and our approach under synthetic degradations. When reference-free metrics are considered, our approach is comparable to the state-of-the-art in all scenarios, both with and without guidance. Notably, in the challenging case of $\downarrow_4 + \sigma_n$, our method achieves state-of-the-art performance. In terms of reference-based metrics, our method remains competitive. This trend also holds for real-world degradations, as shown in Tab. 3 and Tab. 4.

A notable observation is the sudden drop in reference-based metrics when guidance is disabled due to hallucinations. Although the perceptual quality is higher, as indicated by reference-free metrics, inconsistencies remain. Figure 4b displays this for an example image. Conversely, better reference-based performance with worse reference-free performance would not imply better restoration, as confirmed by the overly smooth or artifact-prone results of SwinIR. Qualitative results in Figure 5 further illustrate how our method produces finer details. Furthermore, we show in Table 5 the performance of PASD with our approach on the RealSR dataset. The plug-and-play achieves improved image quality and produces sharper restorations, as illustrated by the example output in Figure 5. For more visual results, please refer to the appendix. Some baseline results differ

from those originally reported; we provide codebase references and evaluation details in the appendix. To compare with the closed-source DR-Adapter [39], we evaluate our model on a real-world dataset and report the results in Tab. 4.

6.2. *Parameter efficiency & inference time*

Importantly, as shown in Table 1, BIR-Adapter achieves competitive performance while requiring $10\times$ and $36\times$ fewer additional trained parameters than DiffBIR and SUPIR, respectively. While the diffusion backbone remains frozen during training, all backbone parameters are naturally involved during inference; thus, the reported reduction primarily reflects the training complexity rather than the total model size.

BIR-Adapter exhibits inference times comparable to SUPIR despite delivering stronger perceptual quality in most settings. A standard SD-1.5 ControlNet-based pipeline requires approximately 486 GMacs per image, compared to 481 GMacs for the BIR-Adapter, indicating that the runtime overhead stems primarily from the batched processing, rather than architectural complexity. This design enables the reuse of the pretrained diffusion prior without requiring the training of auxiliary feature extractors or full fine-tuning. Degraded features remain static, hence caching can eliminate redundant computation; preliminary experiments have reduced inference time to approximately two seconds under identical conditions.

7. Conclusion

We argued that the latent representations of a diffusion prior can be directly leveraged for restoration without large auxiliary feature extractors. To that extent, we introduced a few-parameter self-referential restoring attention to a frozen backbone, along with a diffusion guidance mechanism. Extensive experiments on synthetic and real-world benchmarks demonstrate that BIR-Adapter achieves competitive and, in several cases, superior restoration quality compared to state-of-the-art diffusion-based approaches, while reducing the number

of additional trainable parameters by up to $36\times$. In addition, the proposed guidance mechanism effectively mitigates hallucinations by explicitly controlling the trade-off between perceptual quality and fidelity. The adapter-based design further enables seamless plug-and-play integration into existing diffusion pipelines, as demonstrated by extending a super-resolution-only model to handle unknown degradations without the costly retraining of the backbone.

Despite these advantages, the proposed approach has several limitations. Although BIR-Adapter significantly reduces the number of additional trained parameters, it introduces overhead during inference because degraded latents are processed in parallel to enable self-referential restoring attention. Furthermore the guidance parameter is selected empirically and is an additional hyperparameter. Third, while the method generalizes well to unknown real-world degradations, it has not been explicitly evaluated on motion blur or low-light conditions, which limits the scope of the current experimental validation.

These limitations also suggest promising directions for future research. Since degraded features remain constant across diffusion steps, caching strategies would improve inference efficiency without architectural changes. Moreover, few-step diffusion via distillation would further reduce runtime. Automatically adapting the guidance strength based on image content is another important avenue to reduce manual hyperparameter tuning. Finally, extending BIR-Adapter to recent pretrained architectures, such as diffusion transformers, and evaluating it on a broader range of degradations would further strengthen its applicability to real-world blind image restoration tasks.

8. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to refine sentences and perform a grammar check. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Appendix A. More Visual Results

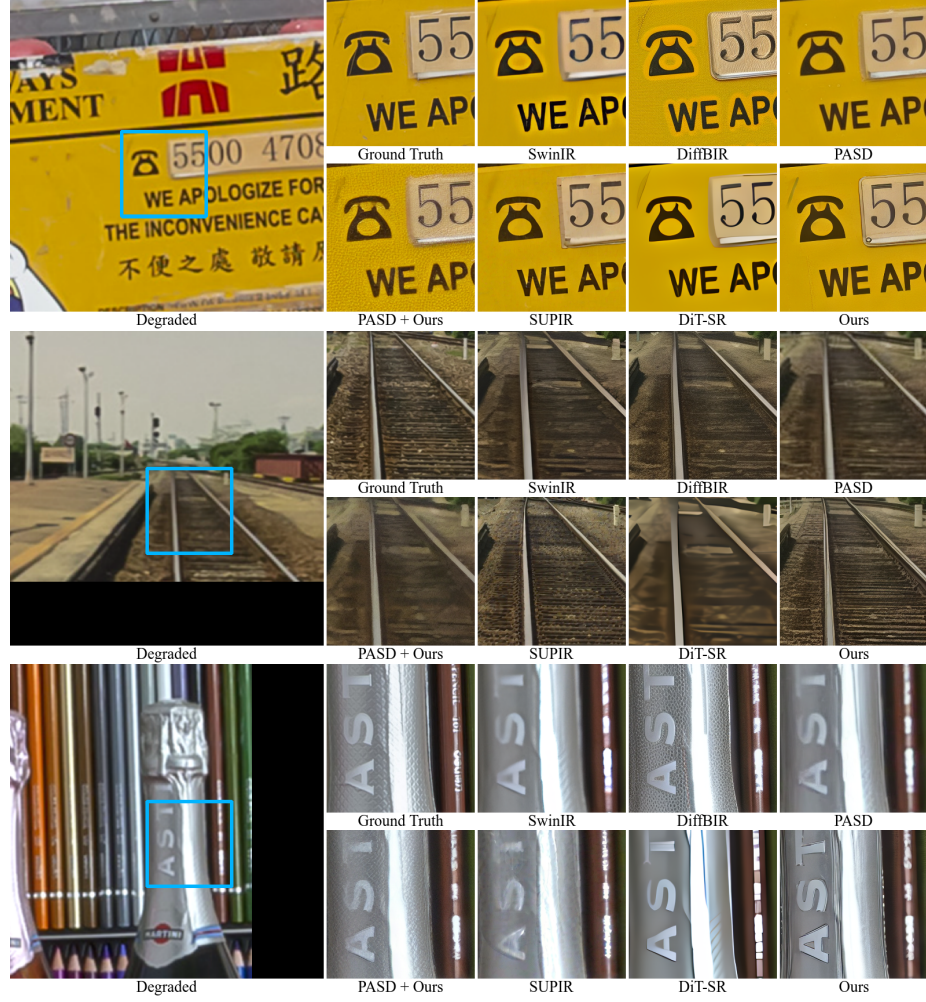


Figure A.6: Visual results from RealSR with real-world unknown degradations [47]. All images are restored with $4\times$ upsampling using baseline methods and ours. We also include results from the extended PASD.

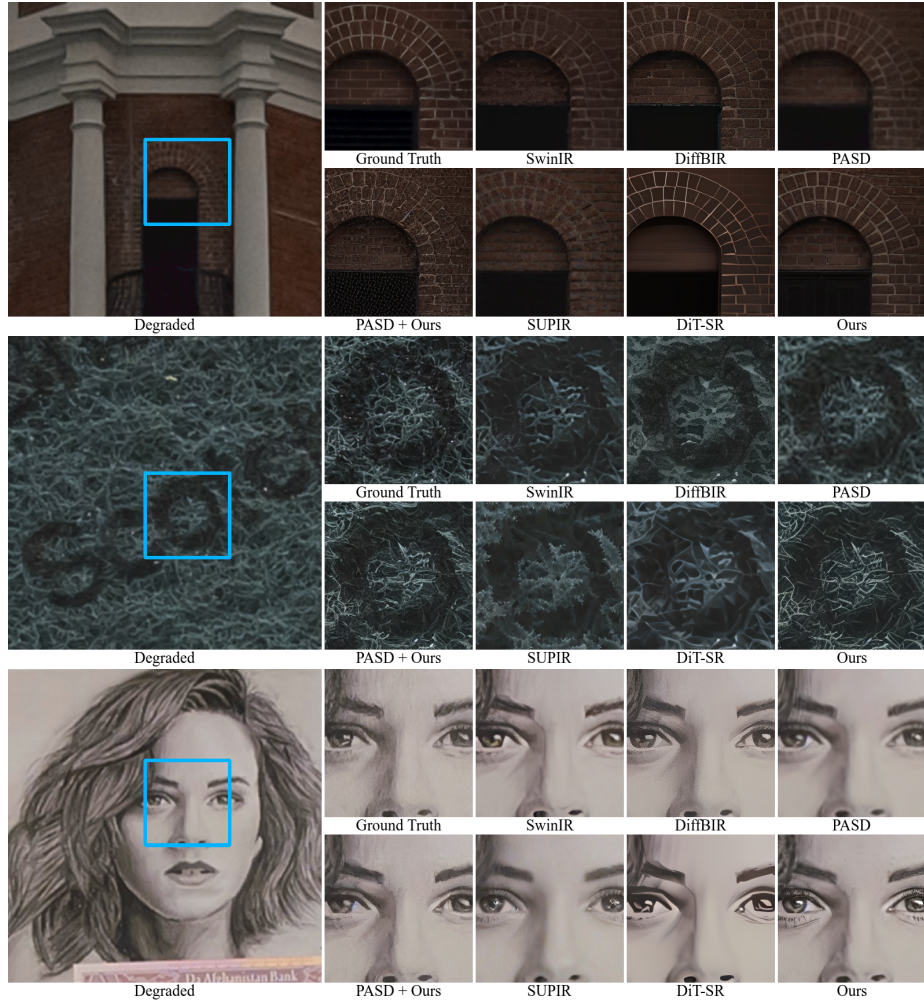


Figure A.7: Visual results from RealSR with real-world unknown degradations [47]. All images are restored with $4\times$ upsampling using baseline methods and ours. We also include results from the extended PASD.

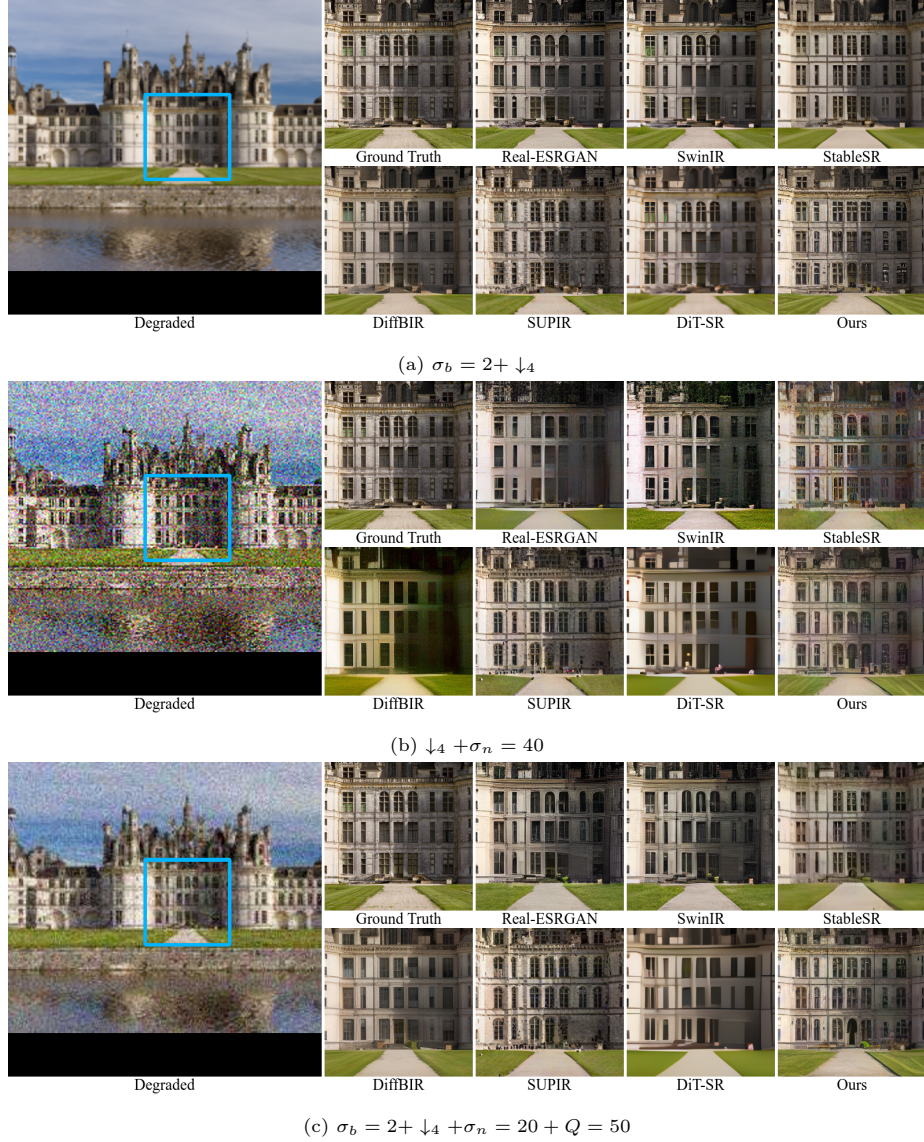
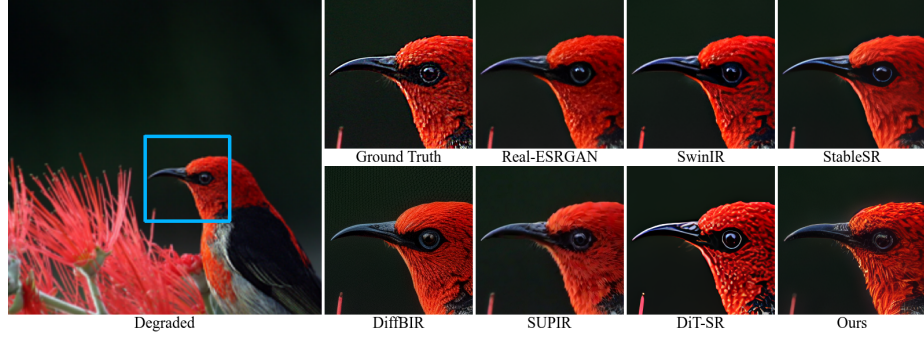
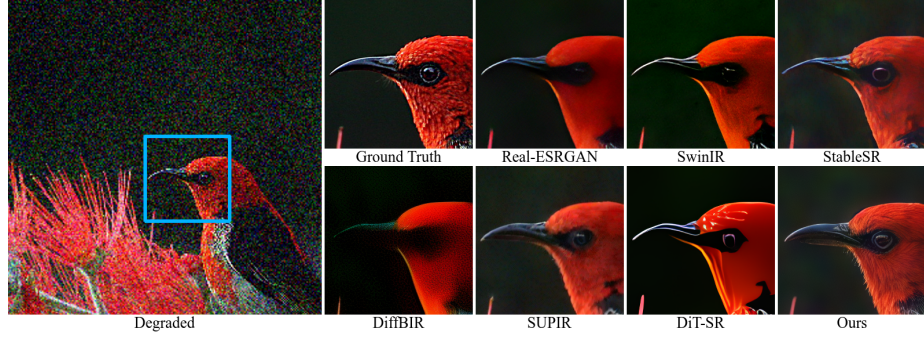


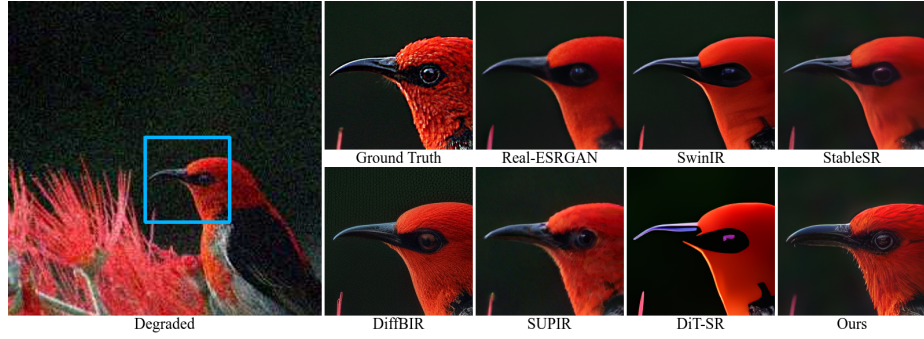
Figure A.8: Visual results of baseline methods and ours on a sample frame from the DIV2K validation set under synthetic degradations [41].



(a) $\sigma_b = 2 + \downarrow_4$



(b) $\downarrow_4 + \sigma_n = 40$



(c) $\sigma_b = 2 + \downarrow_4 + \sigma_n = 20 + Q = 50$

Figure A.9: Visual results of baseline methods and ours on a sample frame from the DIV2K validation set under synthetic degradations [41].

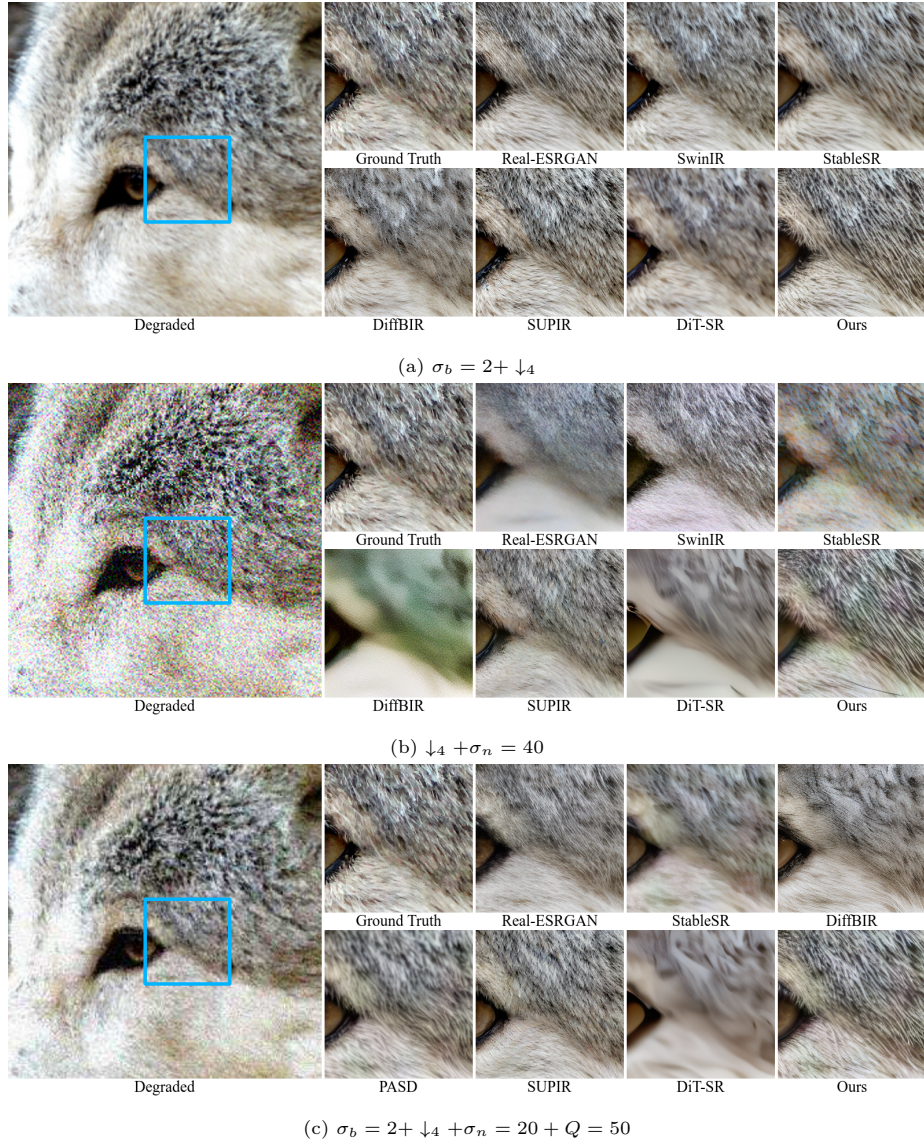


Figure A.10: Visual results of baseline methods and ours on a sample frame from the DIV2K validation set under synthetic degradations [41].

Appendix B. Training-free Variants and Pseudocodes

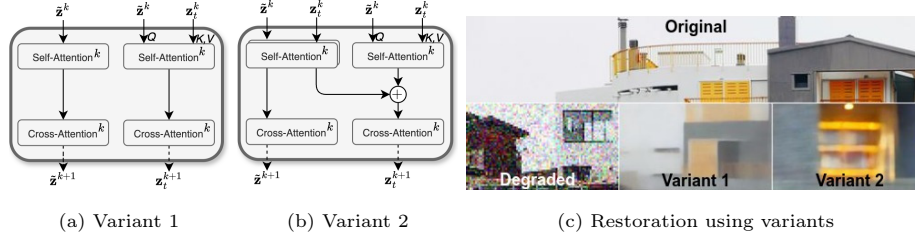


Figure B.11: Variant 1 utilizes the degraded features in the self-attention layer as query. Variant 2 does not fine-tune the weights of BIR-Adapter. Restoration results in c show how the variants fail.

Algorithm 1 Pseudocode of BIR-Adapter and its variants.

```

procedure  $\epsilon_\theta(\tilde{\mathbf{x}}, \mathbf{x}_t, y^+)$ 
 $\mathbf{X} \leftarrow \text{concat}[\tilde{\mathbf{x}}, \mathbf{x}_t]$ 
BIR-Adapter
for  $k \leftarrow 1$  to last layer  $K$  do
    Extract hidden state  $\mathbf{Z}^k$  of layer  $k$  from  $\mathbf{Z}^{k-1}$ , where  $\mathbf{Z}^0$  is  $\mathbf{X}$  // Through convolution, pooling, etc.
     $\tilde{\mathbf{z}}^k, \mathbf{z}_t^k \leftarrow \text{chunk}(\mathbf{Z}^k, 2)$ 
     $\tilde{Q} \leftarrow \tilde{\mathbf{z}}^k \mathbf{W}^Q, \quad K' \leftarrow \mathbf{z}_t^k \mathbf{W}'^K, \quad V' \leftarrow \mathbf{z}_t^k \mathbf{W}'^V, \quad Q \leftarrow \mathbf{z}^k \mathbf{W}^Q, \quad K \leftarrow \mathbf{z}^k \mathbf{W}^K, \quad V \leftarrow \mathbf{z}^k \mathbf{W}^V$ 
     $\mathbf{Z}^{k+1} \leftarrow \text{Attention}(Q, K, V) \mathbf{W}^O + \text{Attention}(\tilde{Q}, K', V') \mathbf{W}'^O$ 
    Execute cross-attention between  $\mathbf{Z}^{k+1}$  and  $y^+$ .
end for

Variant 1: Degraded features as the query at Self-Attention
for  $k \leftarrow 1$  to last layer  $K$  do
    Extract hidden state  $\mathbf{Z}^k$  of layer  $k$  from  $\mathbf{Z}^{k-1}$ , where  $\mathbf{Z}^0$  is  $\mathbf{X}$  // Through convolution, pooling, etc.
     $\tilde{Q} \leftarrow \tilde{\mathbf{z}}^k \mathbf{W}^Q, \quad Q \leftarrow \mathbf{z}^k \mathbf{W}^Q, \quad K \leftarrow \mathbf{z}^k \mathbf{W}^K, \quad V \leftarrow \mathbf{z}^k \mathbf{W}^V$ 
     $\mathbf{Z}^{k+1} \leftarrow \text{Attention}(\tilde{Q}, K, V) \mathbf{W}^O$ 
    Execute cross-attention between  $\mathbf{Z}^{k+1}$  and  $y^+$ .
end for

Variant 2: Non-trained BIR-Adapter
for  $k \leftarrow 1$  to last layer  $K$  do
    Extract hidden state  $\mathbf{Z}^k$  of layer  $k$  from  $\mathbf{Z}^{k-1}$ , where  $\mathbf{Z}^0$  is  $\mathbf{X}$  // Through convolution, pooling, etc.
     $\tilde{Q} \leftarrow \tilde{\mathbf{z}}^k \mathbf{W}^Q, \quad Q \leftarrow \mathbf{z}^k \mathbf{W}^Q, \quad K \leftarrow \mathbf{z}^k \mathbf{W}^K, \quad V \leftarrow \mathbf{z}^k \mathbf{W}^V$ 
     $\mathbf{Z}^{k+1} \leftarrow \text{Attention}(Q, K, V) \mathbf{W}^O + \text{Attention}(\tilde{Q}, K, V) \mathbf{W}^O$ 
    Execute cross-attention between  $\mathbf{Z}^{k+1}$  and  $y^+$ .
end for
return final prediction  $\mathbf{z}_t^K$ 
end procedure

```

Appendix C. Baselines

```
export DEFAULT_PROMPT = "Cinematic, High Contrast, highly detailed, taken using a Canon EOS R camera, hyper
    ↳ detailed photo - realistic maximum detail, 32k, Color Grading, ultra HD, extreme meticulous
    ↳ detailing, skin pore detailing, hyper sharpness, perfect without deformations."
export NEGATIVE_PROMPT = "painting, oil painting, illustration, drawing, art, sketch, oil painting, cartoon,
    ↳ CG Style, 3D render, unreal engine, blurring, dirty, messy, worst quality, low quality, frames,
    ↳ watermark, signature, jpeg artifacts, deformed, lowres, over-smooth"
```

```
git clone https://github.com/XPixelGroup/DiffBIR.git

python inference.py --task sr --version v2 --upscale 4 --sampler spaced --cfg_scale 4.0 --steps 20
--captioner none --input /input --output /output --precision fp16 --cleaner_tiled
--vae_encoder_tiled --vae_decoder_tiled --cldm_tiled --seed 1234
--pos_prompt DEFAULT_PROMPT --neg_prompt NEGATIVE_PROMPT
```

```
git clone https://github.com/yangxy/PASD.git

python test_pasd.py --pretrained_model_path /path/to/sd --pasd_model_path /pasd/model
--num_inference_steps 20 --upscale 4 --high_level_info none --process_size 512
--image_path /input --output_dir /output --added_prompt DEFAULT_PROMPT
--negative_prompt NEGATIVE_PROMPT --seed 1234
```

```
git clone https://github.com/Fanghua-Yu/SUPIR.git

python test.py --edm_steps 20 --upscale 4 --no_llava --s_stage2 0.93 --s_cfg 6.0 --spt_linear_CFG 3.0
--s_noise 1.02 --img_dir /input --save_dir /output --ae_dtype bf16 --diff_dtype bf16
--use_tile_vae --encoder_tile_size 1024 --decoder_tile_size 128 --a_prompt DEFAULT_PROMPT
--n_prompt NEGATIVE_PROMPT
```

```
git clone https://github.com/IceClear/StableSR.git

from predict import Predictor
predictor = Predictor()
for degraded_image in /path/to/degraded/images:
    out_file = predictor.predict(input_image=degraded_image, ddp_steps=20, fidelity_weight=0.5,
                                upscale=4, tile_overlap=32, colorfix_type='wavelet', seed=1234)
```

```
git clone https://github.com/kunncheng/DiT-SR.git

python inference.py --task realer --scale 4 --config_path configs/realer_DiT.yaml
--ckpt_path /path/to/DiT-SR --chop_size 512 --chop_stride 256 -i /input -o /output
```

Listing 1: Source codes and the commands used to create the results of the baselines.

References

- [1] L. Zhai, Y. Wang, S. Cui, Y. Zhou, A comprehensive review of deep learning-based real-world image restoration, *IEEE Access* 11 (2023) 21049–21067.
- [2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, *Communications of the ACM* 63 (11) (2020) 139–144.
- [3] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, *Advances in neural information processing systems* 34 (2021) 8780–8794.
- [4] X. Li, Y. Ren, X. Jin, C. Lan, X. Wang, W. Zeng, X. Wang, Z. Chen, Diffusion models for image restoration and enhancement: a comprehensive survey, *International Journal of Computer Vision* (2025) 1–31.
- [5] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, *Advances in neural information processing systems* 33 (2020) 6840–6851.
- [6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [8] C. He, Y. Shen, C. Fang, F. Xiao, L. Tang, Y. Zhang, W. Zuo, Z. Guo, X. Li, Diffusion models in low-level vision: A survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).
- [9] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, F. S. Khan, Foundation models defining a new era in vision: a

survey and outlook, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

- [10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.
- [11] J. Cheng, D. Liang, S. Tan, Transfer clip for generalizable image denoising, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25974–25984.
- [12] K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian, Image denoising by sparse 3-d transform-domain collaborative filtering, *IEEE Transactions on image processing* 16 (8) (2007) 2080–2095.
- [13] K. Zhang, W. Zuo, S. Gu, L. Zhang, Learning deep cnn denoiser prior for image restoration, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3929–3938.
- [14] C. Dong, C. C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, *IEEE transactions on pattern analysis and machine intelligence* 38 (2) (2015) 295–307.
- [15] S. Nah, T. Hyun Kim, K. Mu Lee, Deep multi-scale convolutional neural network for dynamic scene deblurring, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3883–3891.
- [16] J. Su, B. Xu, H. Yin, A survey of deep learning approaches to image restoration, *Neurocomputing* 487 (2022) 46–65.
- [17] K. Zhang, J. Liang, L. Van Gool, R. Timofte, Designing a practical degradation model for deep blind image super-resolution, in: *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4791–4800.

- [18] Y. Cui, W. Ren, S. Yang, X. Cao, A. Knoll, Irnext: Rethinking convolutional network design for image restoration (2023).
- [19] Y. Cui, W. Ren, A. Knoll, Omni-kernel network for image restoration, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 38, 2024, pp. 1426–1434.
- [20] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, Swinir: Image restoration using swin transformer, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1833–1844.
- [21] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5728–5739.
- [22] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, H. Li, Uformer: A general u-shaped transformer for image restoration, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 17683–17693.
- [23] L. Kong, J. Dong, J. Ge, M. Li, J. Pan, Efficient frequency domain-based transformers for high-quality image deblurring, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5886–5895.
- [24] Y. Cui, Y. Tao, Z. Bing, W. Ren, X. Gao, X. Cao, K. Huang, A. Knoll, Selective frequency network for image restoration, in: The eleventh international conference on learning representations, 2023.
- [25] Y. Cui, M. Liu, W. Ren, A. Knoll, Modumer: Modulating transformer for image restoration, IEEE Transactions on Neural Networks and Learning Systems (2025).
- [26] X. Wang, L. Xie, C. Dong, Y. Shan, Real-esrgan: Training real-world blind super-resolution with pure synthetic data, in: Proceedings of the

- IEEE/CVF international conference on computer vision, 2021, pp. 1905–1914.
- [27] L. Zhang, A. Rao, M. Agrawala, Adding conditional control to text-to-image diffusion models, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 3836–3847.
 - [28] T. Yang, R. Wu, P. Ren, X. Xie, L. Zhang, Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization, in: European Conference on Computer Vision, Springer, 2024, pp. 74–91.
 - [29] J. Wang, Z. Yue, S. Zhou, K. C. Chan, C. C. Loy, Exploiting diffusion prior for real-world image super-resolution, *International Journal of Computer Vision* 132 (12) (2024) 5929–5949.
 - [30] X. Lin, J. He, Z. Chen, Z. Lyu, B. Dai, F. Yu, Y. Qiao, W. Ouyang, C. Dong, Diffbir: Toward blind image restoration with generative diffusion prior, in: European Conference on Computer Vision, Springer, 2024, pp. 430–448.
 - [31] F. Yu, J. Gu, Z. Li, J. Hu, X. Kong, X. Wang, J. He, Y. Qiao, C. Dong, Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 25669–25680.
 - [32] D. Kong, F. Li, Z. Wang, J. Xu, R. Pei, W. Li, W. Ren, Dual prompting image restoration with diffusion transformers, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 12809–12819.
 - [33] Y. Liu, J. He, Y. Liu, X. Lin, F. Yu, J. Hu, Y. Qiao, C. Dong, Adaptbir: Adaptive blind image restoration with latent diffusion prior for higher fidelity, *Pattern Recognition* 155 (2024) 110659.
 - [34] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, L. Zhang, Seesr: Towards semantics-aware real-world image super-resolution, in: Proceedings of the

- IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 25456–25467.
- [35] J. Kim, J. Oh, K. M. Lee, Exploiting diffusion prior for task-driven image restoration, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 10151–10161.
 - [36] K. Cheng, L. Yu, Z. Tu, X. He, L. Chen, Y. Guo, M. Zhu, N. Wang, X. Gao, J. Hu, Effective diffusion transformer architecture for image super-resolution, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, 2025, pp. 2455–2463.
 - [37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models. arxiv 2021, arXiv preprint arXiv:2106.09685 10 (2021).
 - [38] Z. Ma, Y. Zhang, G. Jia, L. Zhao, Y. Ma, M. Ma, G. Liu, K. Zhang, N. Ding, J. Li, et al., Efficient diffusion models: A comprehensive survey from principles to practices, IEEE Transactions on Pattern Analysis and Machine Intelligence (2025).
 - [39] H. Liang, Z. Wang, W. Deng, Diffusion restoration adapter for real-world image restoration, arXiv preprint arXiv:2502.20679 (2025).
 - [40] J. Wang, K. C. Chan, C. C. Loy, Exploring clip for assessing the look and feel of images, in: AAAI, 2023.
 - [41] E. Agustsson, R. Timofte, Ntire 2017 challenge on single image super-resolution: Dataset and study, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 126–135.
 - [42] S. Gu, A. Lugmayr, M. Danelljan, M. Fritsche, J. Lamour, R. Timofte, Div8k: Diverse 8k resolution image dataset, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, 2019, pp. 3512–3516.

- [43] B. Lim, S. Son, H. Kim, S. Nah, K. M. Lee, Enhanced deep residual networks for single image super-resolution, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017.
- [44] X. Wang, K. Yu, C. Dong, C. C. Loy, Recovering realistic texture in image super-resolution by deep spatial feature transform, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 606–615.
- [45] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.
- [46] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).
- [47] J. Cai, H. Zeng, H. Yong, Z. Cao, L. Zhang, Toward real-world single image super-resolution: A new benchmark and a new model, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 3086–3095.
- [48] J. Ho, T. Salimans, Classifier-free diffusion guidance, arXiv preprint arXiv:2207.12598 (2022).
- [49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: CVPR, 2018.
- [50] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, Y. Yang, Maniqa: Multi-dimension attention network for no-reference image quality assessment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1191–1200.
- [51] J. Ke, Q. Wang, Y. Wang, P. Milanfar, F. Yang, Musiq: Multi-scale image quality transformer, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 5148–5157.