DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior

Xinqi Lin[1,*] Jingwen He[2,3,*] Ziyan Chen[1,2] Zhaoyang Lyu[2] Bo Dai[2] Fanghua Yu[1]
Wanli Ouyang[2,3] Yu Qiao[2] Chao Dong[1,2,†]
[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences,
[2]Shanghai AI Laboratory
[3]The Chinese University of Hong Kong

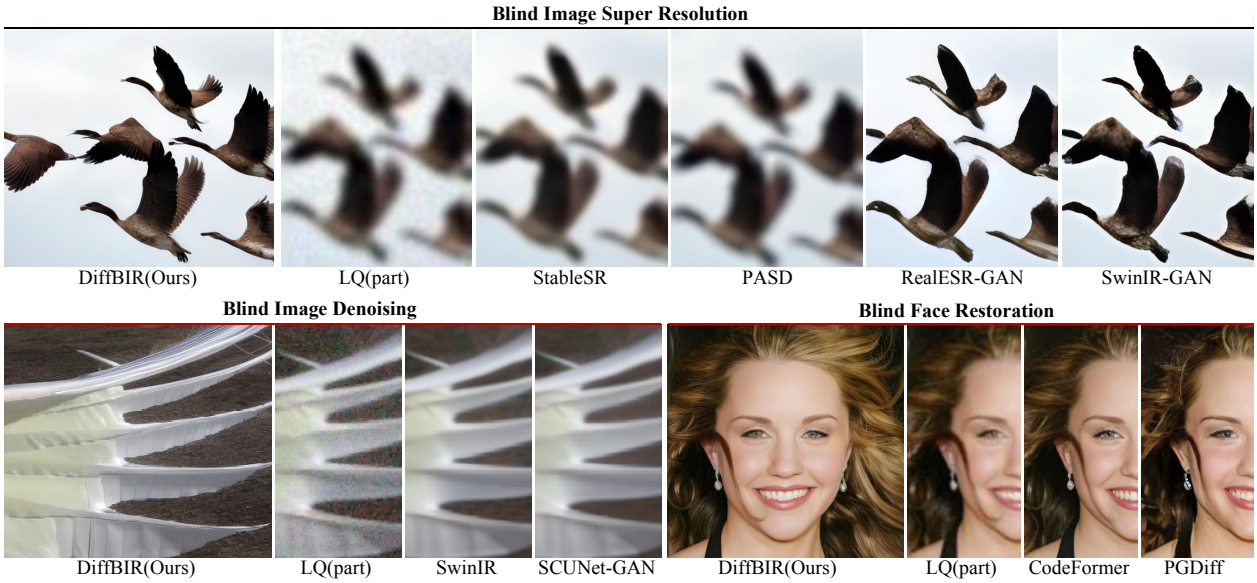

Figure 1. Comparisons of state-of-the-art methods and our DiffBIR for blind image super-resolution (BSR), blind image denoising (BID), and blind face restoration (BFR). (**Zoom in for best view**)

Abstract

We present DiffBIR, a general restoration pipeline that could handle different blind image restoration tasks in a unified framework. DiffBIR decouples blind image restoration problem into two stages: 1) degradation removal: removing image-independent content; 2) information regeneration: generating the lost image content. Each stage is developed independently but they work seamlessly in a cascaded manner. In the first stage, we use restoration modules to remove degradations and obtain high-fidelity restored results. For the second stage, we propose IRControlNet that leverages the generative ability of latent diffusion models to generate realistic details. Specifically, IRControlNet is trained based on specially produced condition images without distracting noisy content for stable generation performance. Moreover, we design a region-adaptive restoration guidance that can modify the denoising process during inference without model re-training, allowing users to balance realness and fidelity through a tunable guidance scale. Extensive experiments have demonstrated DiffBIR's superiority over state-of-the-art approaches for blind image super-resolution, blind face restoration and blind image denoising tasks on both synthetic and real-world datasets. The code is available at https://github.com/XPixelGroup/DiffBIR.

1. Introduction

Image restoration aims at reconstructing a high-quality image from its low-quality observation. Typical image restoration problems, such as image denoising, deblurring and super-resolution, are usually defined under a constrained setting, where the degradation process is simple and known (*e.g.*, bicubic downsampling). They have successfully promoted a vast number of excellent restoration algorithms [6, 8, 12, 29, 58, 69, 71], but are born to have limited generalization ability. To deal with real-world degraded images,

*Equal contribution
†Corresponding author

blind image restoration (BIR) comes into view and becomes a promising direction. The ultimate goal of BIR is to realize realistic image reconstruction on general images with general degradations. BIR does not only extend the boundary of classic image restoration tasks, but also has a wide practical application field (*e.g.*, old photo/film restoration).

Typical BIR problems are blind image super-resolution (BSR), blind image denoising (BID), blind face restoration (BFR), etc. BSR is initially proposed to solve real-world super-resolution problems, where the low-resolution image contains unknown degradations. The most popular solutions may be BSRGAN [73] and Real-ESRGAN [56]. They formulate BSR as a supervised large-scale degradation overfitting problem. To simulate real-world degradations, a degradation shuffle strategy and high-order degradation modeling are proposed separately. Then the adversarial loss [15, 27, 37, 45, 54] and reconstruction loss are incorporated to learn the reconstruction process in an end-to-end manner. They have demonstrated their great robustness in degradation removal for real-world super-resolution, but usually fail in generating realistic details due to the limited generative ability. BID aims to achieve blind denoising [17, 74] for real-world noisy photographs, which usually contain various noises (*e.g.*, dark current noise, short noise, and thermal noise) due to the processing in real camera system. SCUNet [74] is the state-of-the-art method, which designs a practical noise degradation model to synthesize the noisy images, and adopts L1 loss as well as optional adversarial loss for training a deep denoiser model. Its solution is similar as BSR methods and thus has the same weakness. BFR only focuses on blind restoration for face images. Due to a smaller image space, BFR methods (*e.g.*, CodeFormer [77], GFPGAN [55]) could incorporate powerful generative facial priors (*e.g.*, VQGAN [13], StyleGAN [22]) to generate faithful and high-quality facial details. They have achieved remarkable success in both academia and industry in recent years. Nevertheless, BFR assumes a fixed input size and restricted face image space, and thus cannot be applied to general images.

Recently, denoising diffusion probabilistic models (DDPMs [20]) have shown outstanding performance in image generation. DDRM [23], DDNM [57], and GDP [14] incorporate the powerful diffusion model as the additional prior, thus having greater generative ability than GAN-based methods. With a proper degradation assumption, they can achieve impressive zero-shot restoration on classic IR tasks. However, the problem setting of zero-shot image restoration (ZIR) is not in accordance with BIR. Their methods can only deal with clearly defined degradations (linear or non-linear), but cannot generalize well to unknown degradations. In other words, they can achieve realistic reconstruction on general images, but not on general degradations.

In this work, we aim to solve different BIR tasks in a unified framework. According to the review and analyses on recent progress in BIR tasks, we decouple the BIR problem into two stages: 1) *degradation removal*: removing image-independent content; 2) *information regeneration*: generating the lost image content. Considering that each BIR task corresponds to different degradation process and image dataset, we utilize different restoration modules to achieve degradation removal for each BIR task respectively. For the second stage, we utilize one generation module that leverages pre-trained text-to-image latent diffusion models [42] for generating faithful and visual-pleasing image content. By treating stage II as a conditional image generation problem, we have made some important observations that indicate bad conditions, the original LQ images with distracting noises/artifacts, will disturb the generation process, causing unpleasant artifacts. Thus, we additionally train a MSE-based restoration module using simple degradation model with wide degradation ranges to produce reliable and diversified conditions. Furthermore, we propose IRControlNet to control the generative diffusion prior based on our produced conditions. Specifically, we use the pre-trained VAE encoder for condition encoding and follows ControlNet [75] to adopt an auxiliary and copied encoder for efficient add-on controlling. Our trained generation module remains effective and stable when combined with different restoration modules for different BIR tasks. Moreover, a training-free controllable module is provided to trade-off between *fidelity* and *quality*. Specifically, we introduce a training-free region-adaptive restoration guidance, which minimizes our designed region-adaptive MSE loss between the generated result and the high-fidelity guidance image at each sampling step through gradient-descent algorithm. During guidance, the detected low-frequency regions are influenced more by the high-fidelity guidance image, while the high-frequency regions maintain more generative ability. Besides, a guidance scale can be tuned to achieve a smooth transition between two effects regarding *fidelity* and *quality*.

To sum up, the main contributions of this work are:

- DiffBIR decouples BIR problem into two stages: restoration module for degradation removal, and generation module for lost information regeneration. With the two-stage design, DiffBIR is able to achieve the state-of-the-art performance for BSR, BFR, and BID tasks in a unified framework for the first time.
- We propose IRControlNet that leverages text-to-image diffusion prior for realistic image reconstruction. Comprehensive exploration on main components for generation module has been conducted, and IRControlNet proves to be a solid backbone for generation module in BIR tasks.
- We introduce a training-free controllable module – region-adaptive restoration guidance that performs in sampling process, for achieving flexible trade-off between *quality* and *fidelity* for various user preferences.

2. Related Work

2.1. Blind Image Restoration

Blind Image Super-Resolution. Latest advances [31] on BSR have explored more complex degradation models to approximate real-world degradations. In particular, BSRGAN [73] aims to synthesize more practical degradations based on a random shuffling strategy, and RealESRGAN [56] exploits "high-order" degradation modeling. SwinIR-GAN [29] uses the prevailing backbone Swin Transformer [32] to achieve better image restoration performance. FeMaSR [5] formulates SR as a feature-matching problem based on pre-trained VQ-GAN [13]. Recently, the powerful Stable Diffusion has been leveraged for image restoration tasks. StableSR [52] designs a time-aware encoder to control the Stable Diffusion. PASD [66] has proposed a PACA module, which could effectively inject the pixel-level condition information into diffusion prior and achieve higher fidelity. Although they have achieved great performance in real-world super-resolution, these methods require re-training for handling other image restoration tasks.

Blind Face Restoration. As a specific sub-domain of general images, the face image typically carries more structural information. Recent BFR approaches mainly incorporate powerful generative priors to reconstruct faces with great realness. Representative GAN-prior-based methods [4, 18, 55, 65] have demonstrated their capability in achieving both high-quality and high-fidelity face reconstruction. State-of-the-art works [16, 59, 77] introduce the HQ codebook to generate surprisingly realistic face details by exploiting Vector-Quantized (VQ) dictionary learning [13, 50]. Latest advances [61, 63, 67] leverage the powerful generative capability of diffusion prior and achieve high-quality and robust face restoration. However, all these methods can only achieve good performance for face images, while the general images are beyond their scopes.

Blind Image Denoising. Blind image denoising (BID) aims to handle unknown noise levels/types. Several attempts have been made to solve the BID problem. Among them, DnCNN [71], as an end-to-end deep CNN, is proposed to handle Gaussian denoising with multiple noise levels. GCBD [7] leverages generative adversarial networks (GAN) for noise modeling. CBDNet [17] uses a more realistic noise model to synthesize low-quality data and incorporates real-world noisy-clean image pairs. VDNet [68] proposes to implement noise estimation and denoising simultaneously based on the variational denoising network. Although the above methods have shown great ability in removing unknown noises, they usually produce smooth results.

2.2. Zero-shot Image Restoration.

ZIR aims to achieve image restoration by leveraging a pre-trained prior network in an unsupervised manner. Earlier works [2, 9, 36, 39] mainly concentrate on searching a latent code within a pre-trained GAN's latent space. Recent advancements in this field embrace the utilization of DDPMs [20, 40, 42, 43, 48, 49]. DDRM [23] introduces an SVD-based approach to handle linear image restoration tasks efficiently. Meanwhile, DDNM [57] analyzes the range-null space decomposition of a vector theoretically and then designs a sampling schedule based on the null space. Inspired by classifier guidance [11], GDP [14] introduces a more convenient and effective guidance approach, in which the degradation model can be estimated during inference. Although these works contribute to the advancement of zero-shot image restoration techniques, ZIR methods still cannot achieve satisfactory restoration results in low-quality images from the real world.

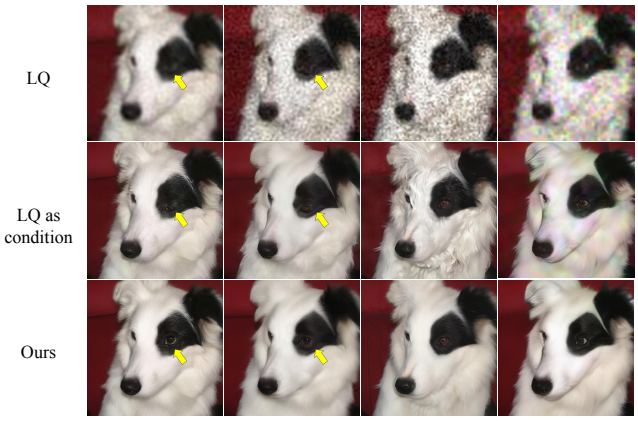

Figure 2. The effects of condition information on generated results. The 2nd row shows that directly using LQ images as conditions causes unpleasant artifacts induced by different degradations (Gaussian, speckle, Poisson, and JPEG compression noises). While our DiffBIR's two-stage pipeline is more stable (see 3rd-row).

3. Method

3.1. Motivation and Framework

In this work, we aim to exploit a powerful generative prior to solve BIR problem. Generative diffusion prior has demonstrated its effectiveness in conditional image generation [75] through enabling condition inputs, such as edge and segmentation maps. This provides a potential solution for BIR problem, that is to regard it as conditional image generation and directly utilize the LQ images as condition inputs. However, low-quality image domain is vast and complex, thus the corresponding condition information is extremely diversified. More importantly, as the degradation and content information of LQ images are entangled, directly treating them as control signals will cause instability and induce artifacts.

As presented in Fig. 2, the LQ images are degraded with different types of noises based on the same HQ image. We train a generation module with synthesized LQ images as

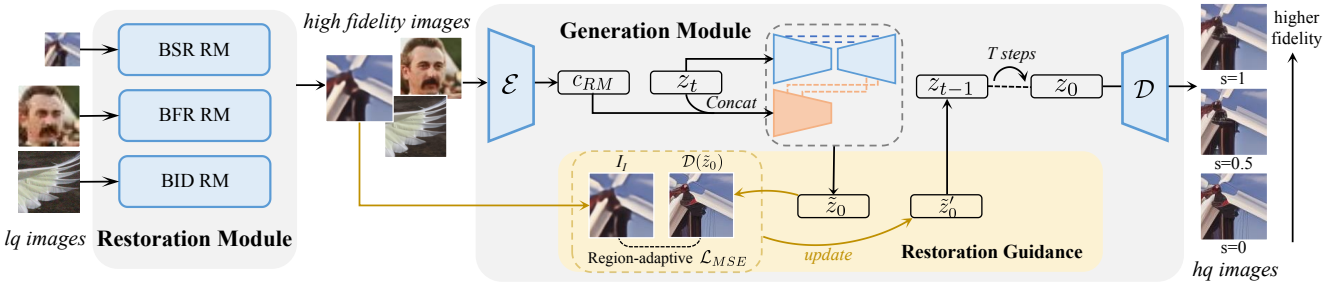

Figure 3. The two-stage pipeline of DiffBIR. 1) Restoration Module (RM) for degradation removal; 2) Generation Module (GM) for realistic image reconstruction with optional region-adaptive restoration guidance for a trade-off between *quality* and *fidelity*.

the conditions, and obtain the corresponding results. It is observed that the degradation indeed has an effect on the produced results: different unpleasant artifacts are generated due to the degradation difference. Since the training is not explicitly guided to distinguish the content information from the degraded image, the generation process is disturbed by the unreliable condition information.

According to our observations and analyses, we adopt a general two-stage pipeline for BIR tasks, which contains restoration modules for removing image-independent degradation, and one generation module that only focuses on image content regeneration. These two stages are decoupled and optimized independently. In this way, we can use any off-the-shelf/self-trained restoration module to address the challenging degradation removal for different BIR tasks. More importantly, the generation is only conditioned on the image content of LQ input, thus it will not be disturbed by degradation. This two-stage pipeline provides a flexible, stable, and unified solution to BIR problem. Besides, a training-free controllable module is introduced to achieve fidelity-quality trade-off by region-adaptive restoration guidance in the sampling process. The whole pipeline is illustrated in Fig. 3.

3.2. Restoration Module

In the first stage, we aim to remove distracting degradations of low-quality images without generating any new content for different BIR tasks. Note that each BIR task has its own characteristics in terms of degradation process and image dataset. For instance, BID methods should especially consider processed camera sensor noises, while BFR methods only focus on restoring low-quality face images. Therefore, we use separate restoration modules instead of a general one for different BIR tasks to maintain their expertise. In this work, we directly adopt the off-the-shelf BIR models trained with MSE loss as the restoration modules.

As mentioned in Section 3.1, training a stable generation module requires reliable conditions. To this end, we additionally train a restoration module (RM) to produce appropriate condition images for training generation module. Specifically, this RM is trained with classic degradation model and MSE loss:

$$I_{RM} = \mathrm{RM}(I_{lq}), \quad \mathcal{L}_{RM} = ||I_{RM} - I_{hq}||_2^2, \tag{1}$$

where I_{hq}, I_{lq}, and I_{RM} denote the high-quality image, the synthesized low-quality counterpart, and the restored image, respectively. Note that the degradation range is set to large since we desire to generate sufficiently diversified condition images. This will improve the overall generative capacity of the generation module (see Section 4.3). Please refer to Appendix for implementation details. This naively trained RM performs as a condition preprocessing for the generation module, and it will be discarded during inference as it cannot handle complex degradations in real-world scenarios.

3.3. Generation Module

Preliminary: Stable Diffusion. We implement our method based on the large-scale text-to-image latent diffusion model, Stable Diffusion. To achieve better efficiency and stabilized training, Stable Diffusion pretrains an autoencoder [26] that converts an image x into a latent z with encoder $\mathcal{E}$ and reconstructs it with decoder $\mathcal{D}$. Both diffusion and denoising processes are performed in the latent space. In diffusion process, Gaussian noise with variance $\beta_t \in (0, 1)$ at time t is added to the encoded latent $z = \mathcal{E}(x)$ to produce the noisy latent:

$$z_t = \sqrt{\bar{\alpha}_t} z + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. When t is large enough, the latent z_t is nearly a standard Gaussian distribution. A network ϵ_θ is learned by predicting the noise ϵ conditioned on c (*i.e.*, text prompts) at a randomly picked time-step t. The optimization of the latent diffusion model is defined as follows:

$$\mathcal{L}_{ldm} = \mathbb{E}_{z,c,t,\epsilon}[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} z + \sqrt{1 - \bar{\alpha}_t}\epsilon, c, t)||_2^2], \tag{3}$$

where x, c are sampled from the dataset and $z = \mathcal{E}(x)$, t is uniformly sampled and ϵ is sampled from the standard Gaussian distribution.

IRControlNet. Given the reliable condition image I_{RM}, we then leverage the pre-trained Stable Diffusion for our generation module (GM). To conclude, it mainly involves

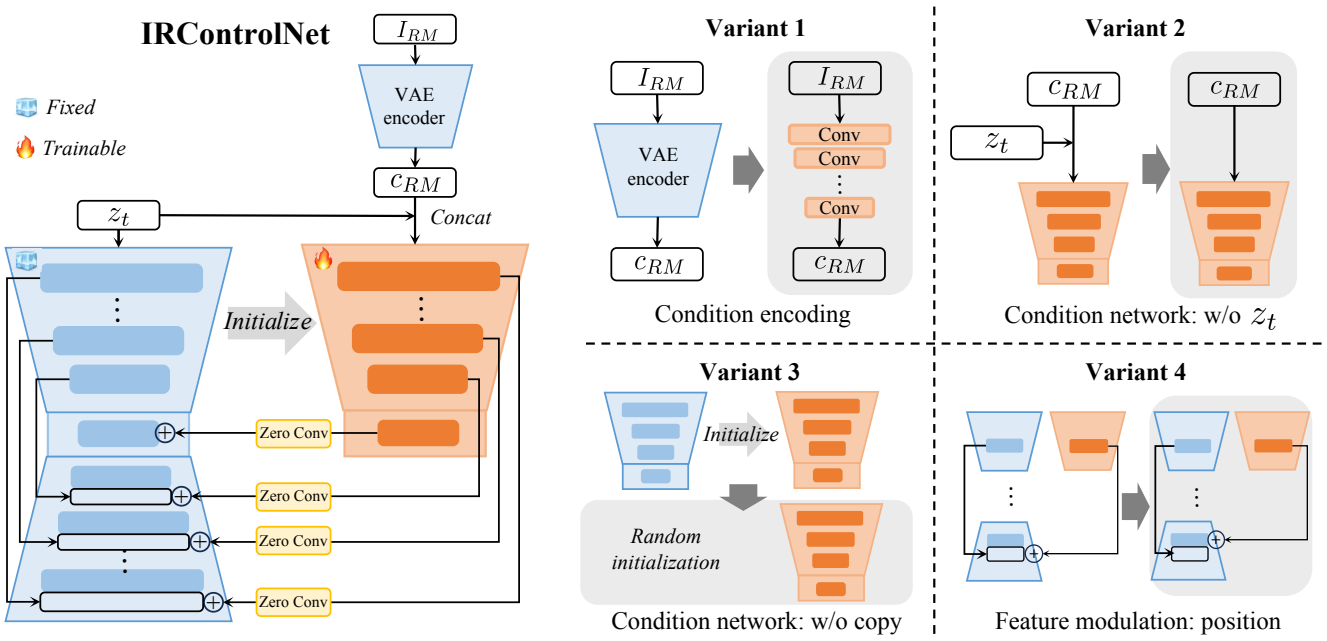

Figure 4. Architectures of our IRControlNet and four model variants.

three aspects: 1) condition encoding; 2) condition network; 3) feature modulation. Our IRControlNet has explored in-depth and provided effective modules for addressing each of them. The architecture is illustrated in Fig. 4.

1) condition encoding. In IRControlNet, we utilize the pretrained and fixed VAE encoder $\mathcal{E}$ to encode the condition image I_{RM} into the latent space for condition encoding: $c_{RM} = \mathcal{E}(I_{RM})$, where c_{RM} is obtained condition latent. Since the VAE is trained on large-scale datasets, the obtained c_{RM} is capable of preserving sufficient image information.

2) condition network. As for condition network, we follow ControlNet [75] and make a trainable copy of the pre-trained UNet encoder and middle block (denoted as $\mathbf{F}_{cond}$), which receives condition information and then outputs control signals. This copy strategy provides a good weight initialization for condition network. Then, we use the concatenation of the condition c_{RM} and the noisy latent z_t at time t as input of $\mathbf{F}_{cond}$, which is denoted as $z'_t = cat(z_t, c_{RM})$. As the concatenation operator $cat(\cdot)$ will increase the channel number, we introduce a few parameters to the first layer of $\mathbf{F}_{cond}$ and initialize them to zero. This zero initialization functions similarly to zero convolution in ControlNet, which is to avoid random noise as gradients in the early stage of training.

3) feature modulation. The previous condition network outputs multi-scale features, which will be used to modulate the intermediate features of the frozen UNet denoiser. Following ControlNet, we only modulate the middle block features and the skipped features through addition operation. Besides, zero convolutions are employed to connect the condition network with the fixed UNet denoiser for improving stability of model training.

During training, only the parameters of condition network and feature modulation will be updated. Specifically, we aim to minimize the following latent diffusion objective:

$$\mathcal{L}_{GM} = \mathbb{E}_{z_t,c,t,\epsilon,c_{RM}}[||\epsilon - \epsilon_\theta(z_t, c, t, c_{RM})||_2^2], \quad (4)$$

where the obtained result in this stage is denoted as I_{GM}.

Discussion. In this part, we aim to validate IRControlNet to be a solid backbone as a generation module in BIR tasks. Specifically, we construct four model variants (see Fig. 4) to obtain a comprehensive empirical analysis of the crucial components in IRControlNet.

Variant 1. Regarding condition encoding, we replace IRControlNet's condition encoder $\mathcal{E}$ by a tiny trained-from-scratch network, consisting of several stacked convolution layers and one zero convolution at the end. The encoded condition is added to the output features from the first layer of condition network. This model variant is identical to ControlNet.

Variant 2. Regarding condition network, we remove noisy z_t and only use the condition latent c_{RM} as the condition network input.

Variant 3. Regarding condition network, we do not copy the original weights from UNet denoiser but train the condition network from random initialization.

Variant 4. Regarding feature modulation, we control the middle block features and decoder features instead of skipped ones.

The comparison of Variant 1 (or ControlNet) and IRControlNet is in Fig.10 (right) and Table 7. We observe that Variant 1 cannot maintain the original color of input LQ

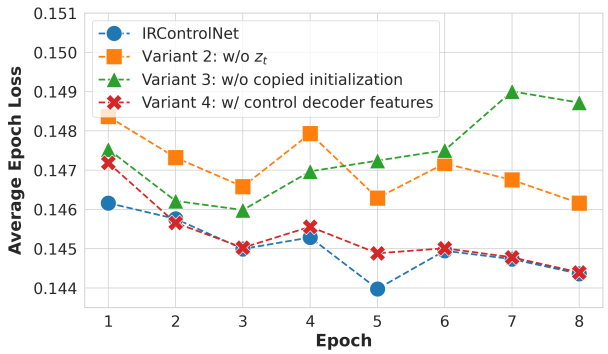

Figure 5. The training loss curves of IRControlNet and Variant 2,3,4 on ImageNet1k dataset under the same training setting.

Variants	PSNR↑	SSIM↑	MANIQA↑
IRControlNet	22.9865	0.5200	0.2689
2) w/o z_t	23.1461	0.5398	0.2611
3) w/o copied initialization	22.8818	0.5192	0.2384
4) w/ control decoder features	22.9721	0.5203	0.2686

Table 1. Quantitative comparisons of IRControlNet, Variant 2, 3 and 4 on ImageNet1k-Val with Real-ESRGAN[56] degradation.

images, and the quantitative results are significantly worse than IRControlNet in PSNR (3dB↓ on average). This observation reveals that condition encoding plays a vital role in controlling latent diffusion prior for IR tasks.

The explanation might be that the image generation process is performed in the latent space, thus the condition should be projected to the same space. IRControlNet identifies it and cleverly uses the pretrained VAE encoder $\mathcal{E}$ for effective encoding and has achieved prominent improvement over ControlNet.

Next, we compare our IRControlNet with Variant 2,3,4 in both training and testing aspects. In Fig. 5, we observe that IRControlNet achieves the fastest model convergence among all model variants, showing its superiority of architecture design. For Variant 2 (w/o z_t), its training losses are consistently higher than those of IRControlNet in all epochs. This indicates that z_t could facilitate convergence, as it makes the condition network aware of randomness at each timestep, thus improving the accuracy of model predictions. From the quantitative comparison in Table 1, Variant 2 achieves the best performance in metrics that measure fidelity, but its IQA score is worse than IRControlNet. From qualitative results in Appendix, we find that Variant 2 usually produces smooth results without sufficient texture details. To conclude, z_t in condition network can boost convergence and helps generate high-quality results, so it is important in generation module and should be incorporated. As shown in Fig. 5, Variant 3 struggles in training loss convergence. Besides, it achieves the worst performance in all metrics. Therefore, a good weight initialization for condition network is crucial in the generation module. As for Variant 4, it achieves comparable convergence speed and quantitative results to IRControlNet, thus applying control to skipped features or decoder features has similar effects. However, the channel numbers of decoder features are about twice the ones of corresponding skipped features, which will introduce more parameters and computation for feature modulation. Therefore, IRControlNet's feature modulation on skipped features is fairly enough.

In conclusion, IRControlNet proves to be a solid backbone for generative module in BIR tasks, as its main components are crucial for either model convergence or performance. We have compared more model variants in Appendix and our conclusion still stands.

3.4. Restoration Guidance

Here we design a controllable module to achieve trade-off between *quality* and *fidelity*. Note that users usually expect more generated details in high-frequency regions (*e.g.*, textures, edges) but less generated content in flat regions (*e.g.*, sky, wall). To this end, we present a region-adaptive restoration guidance, which guides the denoising process towards the restored result in stage I under a tunable guidance scale controlled by users. This restoration guidance is training-free and applied for every sampling step. The whole pipeline is in Fig. 6.

At time t, the UNet denoiser first predicts the noise ϵ_t of the noisy latent z_t. Then the predicted noise ϵ_t is removed from z_t to get the clean latent $\tilde{z}_0$:

$$\epsilon_t = \epsilon_\theta(z_t, c, t, c_{RM}), \tilde{z}_0 = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t}{\sqrt{\bar{\alpha}_t}}. \quad (5)$$

In this stage, we aim to guide $\mathcal{D}(\tilde{z}_0)$ towards the high-fidelity condition I_{RM}. Thus, we propose a region-adaptive MSE loss function that applies between them in pixel space and update the clean latent $\tilde{z}_0$ with gradient descent algorithm. First, we compute the gradient magnitude by applying sobel operators. As pixels with strong gradient signals are very rare in an image, we then divide I_{RM} into multiple non-overlapping patches and calculate patch-level gradient magnitude $\mathcal{G}(I_{RM})$ for better estimating the gradient density. More details can be found in Appendix. Finally, we calculate a weight map by $\mathcal{W} = 1 - \mathcal{G}(I_{RM})$. And the MSE loss is adjusted by the weight map $\mathcal{W}$, and is defined as follows:

$$\mathcal{L}(\tilde{z}_0) = \frac{1}{HWC}||\mathcal{W} \odot (\mathcal{D}(\tilde{z}_0) - I_{RM})||_2^2, \quad (6)$$

where H, W, C denotes the spatial size of I_{RM}. In this way, regions with weak gradients are assigned with larger weights, and vice versa. This indicates that low-frequency regions induce higher loss, thus they are influenced more by the high-fidelity condition I_{RM}. On the contrary, high-frequency regions are less affected and could maintain more generated content during sampling process. This analysis corresponds

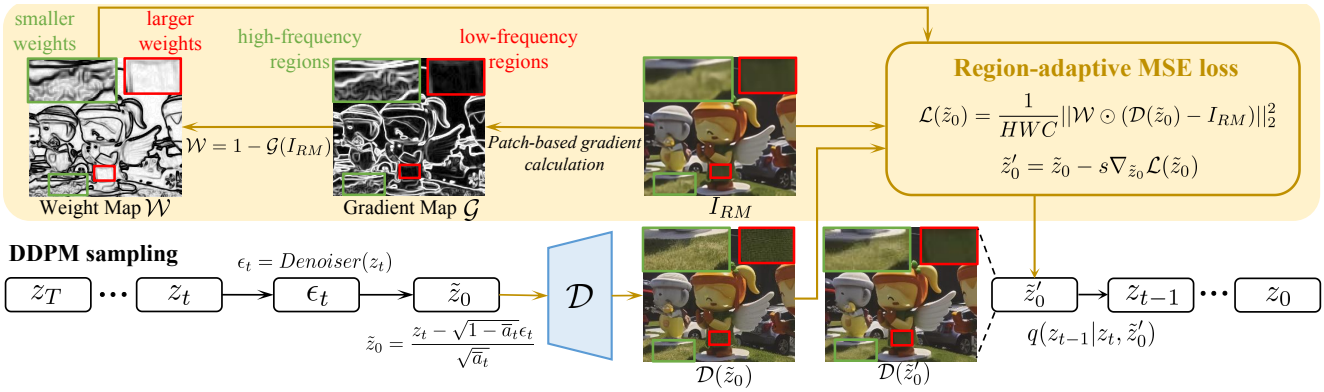

Figure 6. Region-adaptive restoration guidance. Given the high-fidelity guidance image I_{RM}, it aims to minimize the region-adaptive MSE loss between clean latent $\tilde{z}_0$ and I_{RM} at each sampling step through gradient-descent algorithm.

well to the illustration in Fig. 6, in which the noisy content in flat regions is largely eliminated while the generated textures for glass are maintained well after restoration guidance.

The gradient descent algorithm is applied for optimizing the region-adaptive MSE loss at each sampling step t by the following equation:

$$\tilde{z}_0' = \tilde{z}_0 - s\nabla_{\tilde{z}_0}\mathcal{L}(\tilde{z}_0), \tag{7}$$

where s denotes the guidance scale, which can be used to control how much information is maintained from the guidance image I_{RM}. For instance, larger guidance scale pushes $\mathcal{D}(\tilde{z}_0)$ closer to I_{RM}, indicating a higher fidelity. The whole algorithm of our restoration guidance is presented in Appendix.

4. Experiments

4.1. Datasets, Implementation, Metrics

Datasets. We train DiffBIR on our filtered laion2b-en [46] dataset that contains around 15M high-quality images. All images are randomly cropped to 512×512 during training. We evaluate our method for 1) BSR task on three synthetic datasets: DIV2K-Val [1], DRealSR [62], RealSR [3], and two real-world datasets: RealSRSet [73] and our collected real47, 2) BFR task on the real-world datasets LFW-Test [55] and WIDER-Test [77], and 3) BID task on a mixed real-world dataset, which contains images from real3 [74], real9 [74], and RNI15 [72].

Implementation. We train the restoration module for 150k iterations (batch size=96). Then, we adopt Stable Diffusion 2.1-base[1] as the generative prior, and finetune the proposed IRControlNet for 80k iterations (batch size=256). Adam [25] is used as the optimizer. The learning rate is set to 10^{-4} for the first 30k iterations and then decreased to 10^{-5} for the following 50k iterations. The training process is conducted on 512×512 resolution with 8 NVIDIA A100 GPUs.

During inference, we replace our trained restoration module with off-the-shelf task-specific restoration models: BSRNet [73][2] for BSR, SwinIR [29][3] used in DifFace [67] for BFR, and SCUNet-PSNR [74][4] for BID. While the trained IRControlNet remains unchanged for all tasks. The positive prompt is set to empty and we use texts like *"low quality"*, *"blurry"* as our negative prompt. We set the restoration guidance scale to 0, 0.5, and 1 for comparisons on synthetic datasets. As for real-world scenarios, the restoration guidance scale is set to 0 for higher quality. To accelerate the sampling process, we adopt a spaced DDPM sampling schedule [38] which requires 50 sampling steps. For images larger than 512, we directly feed them into DiffBIR. For images with sides < 512, we first upsample them with the short side enlarged to 512, and then resize them back after restoration.

Metrics. For synthesized data, we adopt the traditional metrics: PSNR, SSIM, and LPIPS [76]. To better evaluate the *quality*, we include several no-reference image quality assessment (IQA) metrics: MANIQA[64], MUSIQ [24] and CLIP-IQA [51]. For BFR, we employ the widely used perceptual metric FID [19].

4.2. Comparisons with State-of-the-Art Methods

DiffBIR is compared with state-of-the-art 1) BSR methods: FeMaSR [5], DASR [30], Real-ESRGAN+ [56], BSRGAN [73], SwinIR-GAN [29], StableSR [52] and PASD [66], 2) BFR methods: CodeFormer [77], DifFace [67], DMDNet [28], DR2 [61], GCFSR [18], GFP-GAN [55], GPEN [65], RestoreFormer++ [60], VQFR [16] and PGDiff [63], 3) BID methods: CBDNet [17], DeamNet [41], Restormer [69], SwinIR [29] and SCUNet-GAN [74].

BSR on synthetic datasets. Table 2 presents quantitative comparisons on DIV2K-Val [1] dataset. The LQ images are synthesized using the degradation model adopted in Real-ESRGAN [56]. It is observed that our DiffBIR

[1] Stable Diffusion v2.1: https://github.com/Stability-AI/stablediffusion

[2] https://github.com/cszn/BSRGAN

[3] https://github.com/zsyOAOA/DifFace

[4] https://github.com/cszn/SCUNet

Metrics	FeMaSR [5]	DASR [30]	Real-ESRGAN+ [56]	BSRGAN [73]	SwinIR-GAN [29]	StableSR [52]	PASD [66]	DiffBIR (s=0)	DiffBIR (s=0.5)	DiffBIR (s=1)
PSNR↑	20.1303	21.2141	21.0348	21.4531	20.7488	21.2392	20.7838	20.5824	21.5808	21.9154
SSIM↑	0.4451	0.4773	0.4899	0.4814	0.4844	0.4790	0.4727	0.4277	0.4794	0.4986
LPIPS↓	0.3971	0.4479	0.3921	0.4095	0.3907	0.3993	0.4353	0.3939	0.3935	0.4263
MUSIQ↑	62.7855	58.1591	64.6389	62.9271	65.4945	57.8069	63.8094	73.1019	68.6657	61.1476
MANIQA↑	0.1443	0.1531	0.2238	0.1833	0.2061	0.1648	0.2354	0.3836	0.3146	0.2466
CLIP-IQA↑	0.5674	0.5571	0.5905	0.5195	0.5779	0.5541	0.6125	0.7656	0.7158	0.6347

Table 2. Quantitative comparisons on synthetic dataset (DIV2K-Val) for BSR task. **Red** and blue indicate the best and second best. The top 3 results are marked as gray.

Datasets	Metrics	FeMaSR [5]	DASR [30]	Real-ESRGAN+ [56]	BSRGAN [73]	SwinIR-GAN [29]	StableSR [52]	PASD [66]	DiffBIR (s=0)
RealSRSet [73]	MUSIQ↑	64.6735	59.2695	63.2675	67.6705	64.2512	64.8372	67.4052	69.4208
	MANIQA↑	0.2142	0.1595	0.1963	0.2240	0.2054	0.2083	0.2370	0.3211
	CLIP-IQA↑	0.6879	0.5236	0.5772	0.6456	0.6008	0.6418	0.6761	0.7637
real47	MUSIQ↑	68.9384	62.2026	68.1098	69.4741	68.8467	68.3422	70.9712	73.1397
	MANIQA↑	0.2347	0.1454	0.2055	0.2063	0.2217	0.2264	0.2607	0.3682
	CLIP-IQA↑	0.6911	0.5445	0.6382	0.6111	0.6246	0.6574	0.6913	0.7781

Table 3. Quantitative comparisons on real-world datasets for BSR task. **Red** and blue indicate the best and second best performance. The top 3 results are marked as gray.

Datasets	Metrics	CodeFormer[77]	DifFace[67]	DMDNet[28]	DR2 [61]	GCFSR[18]	GFP-GAN[55]	GPEN[65]	RestoreFormer++[60]	VQFR[16]	PGDiff[63]	DiffBIR (s=0)
LFW-Test [21]	MUSIQ↑	75.4830	70.4957	73.4027	67.5357	71.3789	76.3779	76.6210	72.2492	74.3847	72.2175	76.4206
	MANIQA↑	0.3188	0.2692	0.2973	0.2830	0.2790	0.3688	0.3616	0.3179	0.3280	0.2927	0.4499
	CLIP-IQA↑	0.6890	0.5945	0.6467	0.5728	0.6143	0.7196	0.7181	0.7025	0.7099	0.6133	0.7948
	FID (ref. FFHQ)↓	52.8765	44.9201	43.5403	45.9420	52.6972	47.4717	51.9862	50.7309	50.1300	41.5814	40.9065
Wider-Test [77]	MUSIQ↑	73.4081	65.2397	69.4709	67.3163	69.9634	74.8308	75.6160	71.5155	71.4163	66.0014	75.3213
	MANIQA↑	0.2971	0.2403	0.2613	0.2795	0.2803	0.3508	0.3472	0.2905	0.3060	0.2406	0.4443
	CLIP-IQA↑	0.6984	0.5639	0.6335	0.5821	0.6266	0.7147	0.7039	0.7171	0.7069	0.5685	0.8085
	FID (ref. FFHQ)↓	39.2517	37.8440	38.9580	40.1202	41.1986	41.3247	46.4419	45.4686	38.1675	40.2700	35.8094

Table 4. Quantitative comparisons for BFR on real-world datasets. **Red** and blue indicate the best and second best performance. The top 3 results are marked as gray.

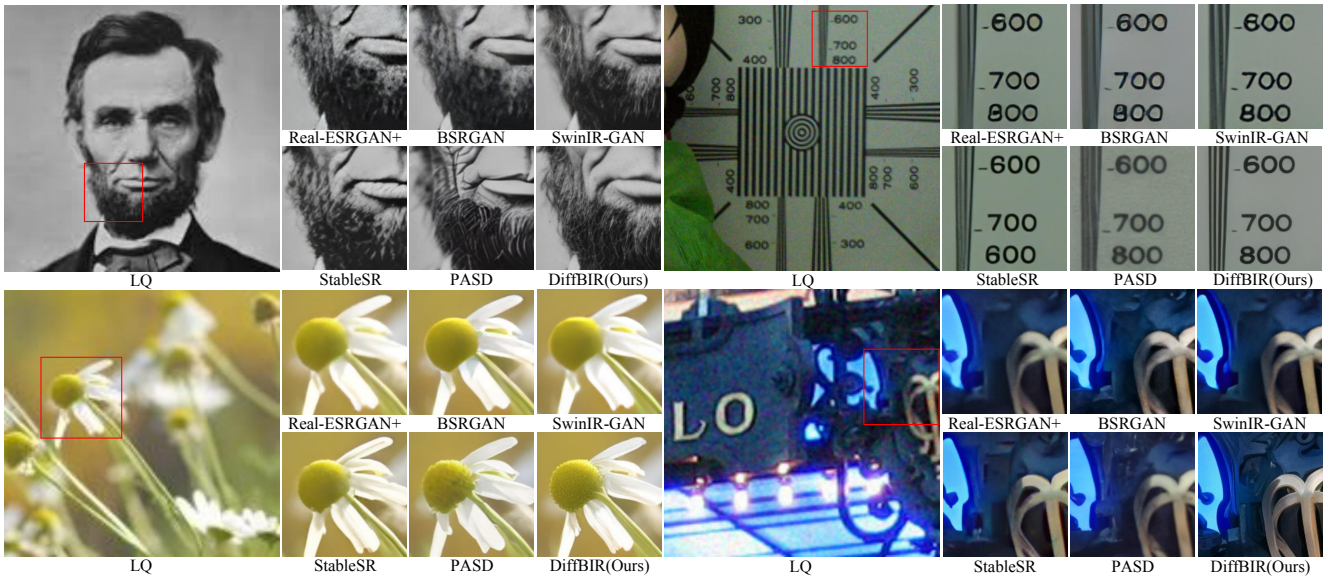

Figure 7. Visual comparison of BSR methods on real-world datasets.

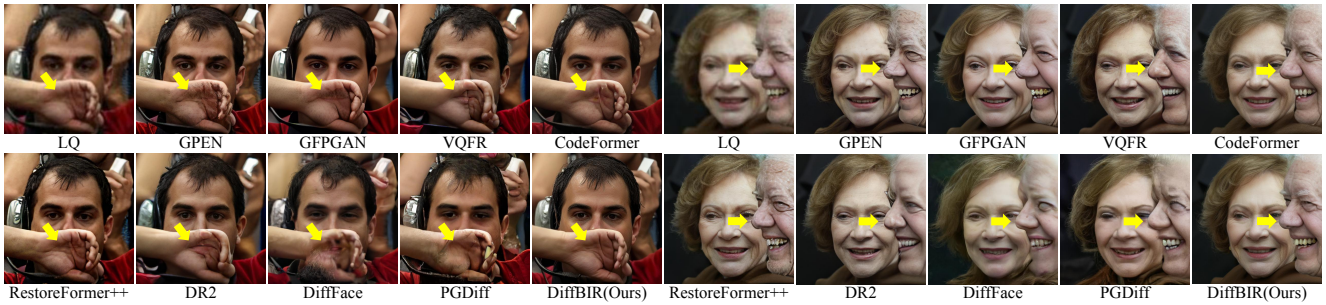

Figure 8. Visual comparison of BFR methods on real-world datasets.

($s = 0$) significantly outperforms all the baseline methods in terms of IQA metrics: MUSIQ, MANIQA, and CLIP-IQA. Moreover, our DiffBIR is able to obtain the best PSNR and SSIM when the restoration guidance scale is set to 1, where the IQA metrics (MANIQA, CLIP-IQA) still rank top-3. Users are recommended to control the restoration guidance scale to achieve a better balance between *quality* and *fidelity* (*e.g.*, setting $s = 0.5$). (Quantitative comparisons on DRealSR/RealSR and visual comparisons on DIV2K-Val are presented in Appendix.)

BSR on real-world datasets. We also provide the quantitative comparison on real-world datasets in Table 3. It is observed that our DiffBIR ($s = 0$) obtains the best scores across all metrics on both the widely used RealSRSet [73] and our collected Real47. This demonstrates DiffBIR's superiority in handling challenging real-world scenarios compared to the baseline methods. As for visual comparison shown in Fig. 7, DiffBIR is capable of producing sharper results than GAN-based methods, whose outputs tend to be over-smoothed. In contrast to diffusion-based methods, DiffBIR's restoration results are more realistic, such as the restored whiskers and lips, the pistil of flowers, texts, etc. More visual results can be found in Appendix.

BFR on real-world datasets. We show the quantitative comparison on real-world datasets in Table 4. DiffBIR has achieved the highest FID score on both the LFW and Wider datasets, demonstrating its ability to generate more realistic faces. Regarding IQA metrics, DiffBIR also obtains the highest scores in CLIP-IQA and MANIQA, while the MUSIQ scores are close to the highest ones. Although the IRControlNet is not finetuned on face dataset (*e.g.*, FFHQ), it outperforms all other baseline methods, indicating the excellent generalization ability of our proposed restoration pipeline. The visual comparisons are shown in Fig. 8. From the first example, it can be seen that only DiffBIR could restore the hand correctly, while other methods are influenced by facial priors thus distorting the hand area. In the second example, only DiffBIR successfully restores the side face, while other methods fail in restoring areas such as teeth, nose, and chin. Both two cases have demonstrated the superiority of using generative priors for general images rather than just face images.

BID on real-world datasets. The quantitative comparisons are shown in Table 5. We can see that DiffBIR significantly outperforms the baseline methods across all metrics. This remarkable difference can be attributed to DiffBIR's introduction of powerful generative diffusion prior, which allows for effective high-quality image restoration. Fig. 9 illustrates visual comparisons between DiffBIR and baseline methods. It is observed that only DiffBIR can remove noise as well as generate realistic textures. Although SwinIR and SCUNet-GAN could successfully remove the noises, they produce smoothed results without vivid texture details.

Methods	MUSIQ↑	MANIQA↑	CLIP-IQA↑
CBDNet [17]	48.1149	0.1103	0.4709
DeamNet [41]	45.9942	0.0949	0.4391
Restormer [69]	47.4605	0.0927	0.3857
SwinIR [29]	55.0493	0.1595	0.4130
SCUNet-GAN [74]	58.2170	0.1822	0.5045
DiffBIR (s=0)	69.7278	0.3404	0.7420

Table 5. Quantitative comparisons on real-world datasets for BID task. **Red** and blue indicate the best and second best. The top 3 results are marked as gray.

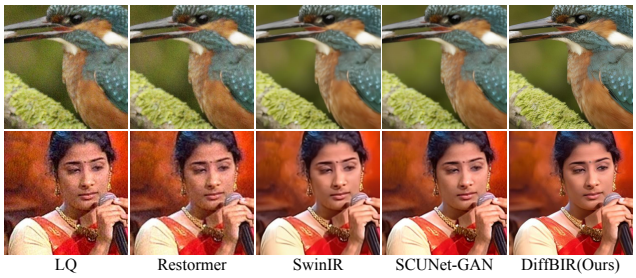

Figure 9. Visual comparisons for BID on real-world datasets.

Figure 10. Visual comparison of ablation studies. (Left) DiffBIR w/o RM regards degradations as image content and performs poorly in fidelity maintaining; (Right) ControlNet[75] has a color shift problem which can be addressed by our IRControlNet.

4.3. Ablation Studies

The Importance of Restoration Module. In this part, we investigate the significance of our proposed two-stage pipeline. Here, we remove the Restoration Module (RM) and directly finetune the diffusion model with synthesized training pairs. From Table 6, the removal of the restoration module leads to a noticeable performance drop in all IQA and reference-based metrics on real-world and synthetic datasets. The visual comparison is presented in Fig.10(left). From the first example, the one-stage model (w/o RM) causes severe distortion in facial generation. While the two-stage model could generate correct facial content. The second example shows that the one-stage model interprets the degradation as semantic information and produces a colorful background and unusual eye shapes. In contrast, the two-stage model produces more realistic results, demonstrating its superiority.

The Effectiveness of IRControlNet. We compare our proposed IRControlNet with ControlNet [75] for BSR. As shown in Fig. 10(right), ControlNet tends to output results with color shifts, as there is no explicit regularization of color consistency during training. The quantitative results pre-

Datasets	Metrics	w/o RM	w/ RM
RealSRSet [73]	MANIQA↑	0.2386	0.2477
	MUSIQ↑	62.5683	64.7319
	CLIP-IQA↑	0.6818	0.7075
ImageNet-Val-1k [10]	PSNR↑	22.8481	23.0078
	SSIM↑	0.5039	0.5198
	LPIPS↓	0.4076	0.4026

Table 6. Ablation study on RM. The best results are denoted as Red.

	Set14 [70]	BSD100 [34]	manga109 [35]	ImageNet-Val-1k [10]
w/ ControlNet	20.9435	22.4923	20.2692	22.2874
w/ IRControlNet	23.5193	23.8778	23.2439	24.2534

Table 7. Comparison of ControlNet and ours in PSNR. Red denotes the best results.

Degradation	MANIQA↑	MUSIQ↑	CLIP-IQA↑
RealESRGAN [56]	0.2351	64.1718	0.6936
Ours	0.2504	64.7319	0.7075

Table 8. Ablation study on degradation model evaluated on RealSRSet [73]. Red denotes the best results.

sented in Table 7 also show that our IRControlNet achieves higher PSNR scores than ControlNet.

The Effectiveness of Wide Degradation Range In this work, we employ a classic degradation model with a wide degradation range to obtain conditions for training generation module, aiming to improve the overall generative capability. One commonly used degradation model for BSR is proposed by RealESRGAN [56]. It adopts a very complex degradation process but uses much smaller degradation ranges. Here we compare these two degradation models and present the quantitative comparison in Table 8. It is observed that using our degradation model leads to better utilization of generative capabilities, thus enhancing the quality of the restored results.

5. Conclusion and Limitations

We propose a unified framework for blind image restoration, named DiffBIR, which could achieve realistic restoration results by leveraging the prior knowledge of pre-trained Stable Diffusion. Extensive experiments have validated the superiority of DiffBIR over existing state-of-the-art methods for BSR, BFR, and BID tasks. Although our proposed DiffBIR has shown promising results, it requires 50 sampling steps to restore one low-quality image, which is computationally expensive. Besides, our two-stage restoration pipeline might be feasible for other BIR tasks, so more exploration can be conducted.

References

[1] Eirikur Agustsson and Radu 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*, pages 126–135, 2017. 7, 15, 16

[2] Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International Conference on Machine Learning*, pages 537–546. PMLR, 2017. 3

[3] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2019. 7, 15

[4] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14245–14254, 2021. 3

[5] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1329–1338, 2022. 3, 7, 8, 15

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 1

[7] Jingwen Chen, Jiawei Chen, Hongyang Chao, and Ming Yang. Image blind denoising with generative adversarial network based noise modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3155–3164, 2018. 3

[8] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023. 1

[9] Giannis Daras, Joseph Dean, Ajil Jalal, and Alexandros G Dimakis. Intermediate layer optimization for inverse problems using deep generative models. *arXiv preprint arXiv:2102.07364*, 2021. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 10

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 184–199. Springer, 2014. 1

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 3

[14] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative

diffusion prior for unified image restoration and enhancement. *arXiv preprint arXiv:2304.01247*, 2023. 2, 3

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[16] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 126–143. Springer, 2022. 3, 7, 8

[17] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1712–1722, 2019. 2, 3, 7, 9

[18] Jingwen He, Wu Shi, Kai Chen, Lean Fu, and Chao Dong. Gcfsr: a generative and controllable face super resolution method without facial and gan priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1889–1898, 2022. 3, 7, 8

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[21] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 8

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[23] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022. 2, 3

[24] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 7

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[27] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 2

[28] Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning dual memory dictionaries for blind face restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 7, 8

[29] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1, 3, 7, 8, 9, 14, 15

[30] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *European Conference on Computer Vision*, pages 574–591. Springer, 2022. 7, 8, 15

[31] Anran Liu, Yihao Liu, Jinjin Gu, Yu Qiao, and Chao Dong. Blind image super-resolution: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[33] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 15

[34] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, pages 416–423. IEEE, 2001. 10

[35] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76:21811–21838, 2017. 10

[36] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. 3

[37] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 2

[38] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 7

[39] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7474–7489, 2021. 3

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[41] Chao Ren, Xiaohai He, Chuncheng Wang, and Zhibo Zhao. Adaptive consistency prior based deep network for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8596–8606, 2021. 7, 9

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[44] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 15

[45] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8207–8216, 2020. 2

[46] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 7

[47] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 14

[48] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[49] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[50] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[51] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 7

[52] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*, 2023. 3, 7, 8, 15

[53] Xintao Wang, Ke Yu, Chao Dong, and Chen Change 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*, pages 606–615, 2018. 13

[54] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 2

[55] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021. 2, 3, 7, 8

[56] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 2, 3, 6, 7, 8, 10, 13, 15

[57] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022. 2, 3

[58] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 1

[59] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022. 3

[60] Zhouxia Wang, Jiawei Zhang, Tianshui Chen, Wenping Wang, and Ping Luo. Restoreformer++: Towards real-world blind face restoration from undegraded key-value pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 7, 8

[61] Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1704–1713, 2023. 3, 7, 8

[62] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 101–117. Springer, 2020. 7, 15

[63] Peiqing Yang, Shangchen Zhou, Qingyi Tao, and Chen Change Loy. Pgdiff: Guiding diffusion models for versatile face restoration via partial guidance. *arXiv preprint arXiv:2309.10810*, 2023. 3, 7, 8

[64] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image

quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 7

[65] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021. 3, 7, 8

[66] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. *arXiv preprint arXiv:2308.14469*, 2023. 3, 7, 8, 15

[67] Zongsheng Yue and Chen Change Loy. Diface: Blind face restoration with diffused error contraction. *arXiv preprint arXiv:2212.06512*, 2022. 3, 7, 8, 14

[68] Zongsheng Yue, Hongwei Yong, Qian Zhao, Deyu Meng, and Lei Zhang. Variational denoising network: Toward blind noise modeling and removal. *Advances in neural information processing systems*, 32, 2019. 3

[69] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 1, 7, 9

[70] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 10

[71] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 1, 3

[72] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 7

[73] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 2, 3, 7, 8, 9, 10, 15

[74] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Deng-Ping Fan, Radu Timofte, and Luc Van Gool. Practical blind image denoising via swin-conv-unet and data synthesis. *Machine Intelligence Research*, pages 1–14, 2023. 2, 7, 9

[75] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 3, 5, 9

[76] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[77] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35:30599–30611, 2022. 2, 3, 7, 8

6. Appendix

6.1. Comparison with More Variants

More Variants for IRControlNet. For more comprehensive analysis, we construct another two variants. The architecture is illustrated in Fig. 11.

Variant 5. Regarding feature modulation, we simultaneously control the middle block features, decoder features and skipped features. We use concat features for simplified denotation.

Variant 6. Regarding feature modulation, we use SFT layer[53] to modulate the intermediate features. Specifically as follows:

$$SFT(\mathbf{F}|\boldsymbol{\gamma},\boldsymbol{\beta}) = \mathbf{F} \odot (1+\boldsymbol{\gamma}) + \boldsymbol{\beta} \tag{8}$$

where $\mathbf{F}$ denotes feature maps, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ denotes the element-wise scale and shift transformation. Both $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are produced by zero-conv, thus they are initialized to zero at the beginning of training.

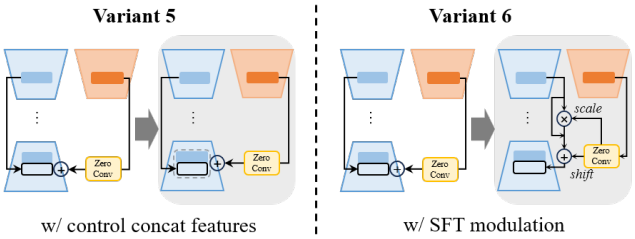

Figure 11. Architectures of our IRControlNet and two model variants.

Table 9 presents the quantitative results. We can observe that both Variant 5 and 6 achieve better performance in terms of PSNR and SSIM while their MANIQA scores are worse than IRControlNet. Variant 5 applies more control on the pretrained model, which enhances the fidelity but damages generation quality. As for Variant 6, it utilizes SFT layer to modulate the skipped features. As SFT layer brings more precise control, which also improves the fidelity. In conclusion, both Variant 5,6 trade the quality for fidelity. IRControlNet achieves such a trade-off through restoration guidance and utilizes the add-on control to preserve most of the generation capability.

Variants	PSNR↑	SSIM↑	MANIQA↑
IRControlNet	22.9865	0.5200	0.2689
Variant 5: w/ control concat features	23.0449	0.5261	0.2567
Variant 6: w/ SFT modulation	22.9974	0.5292	0.2622

Table 9. Quantitative comparisons of IRControlNet, Variant 5 and 6 on ImageNet1k-Val with Real-ESRGAN[56] degradation.

Qualitative Comparisons for Variant 2. We present the visual comparisons for Variant 2 in Fig. 12. It can be observed that IRControlNet can generate more vivid textures while Variant 2 tends to produce over-smoothed results.

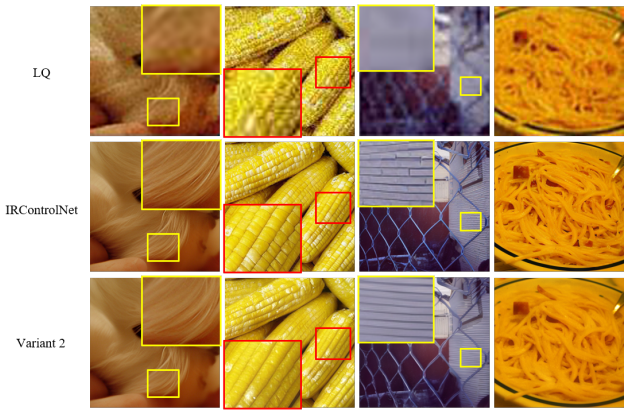

Figure 12. Visual comparisons of Variant 2 and IRControlNet.

6.2. More Details of Training RM

During the training of generation module, we follow [67] and modify a widely-used IR backbone, SwinIR [29], as our restoration module. Specifically, we utilize the pixel unshuffle [47] operation to downsample the original low-quality input I_{lq} with a scale factor of 8. For upsampling the deep features back to the original image space, we perform the nearest interpolation three times, and each interpolation is followed by one convolutional layer as well as one Leaky ReLU activation layer. This modified SwinIR will be trained on synthetic LQ-HQ image pairs. Here we adopt a classic first-order degradation model to synthesize the LQ images.

$$I_{lq} = \{[(I_{hq} \otimes k_\sigma)_{\downarrow_r} + n_\delta]_{\texttt{JPEG}_q}\}_{\uparrow_r}, \tag{9}$$

where the HQ image I_{hq} is first convolved with a Gaussian kernel k_σ, followed by a downsampling of scale r. After that, additive Gaussian noise n_δ is added to the images, and then JPEG compression with quality factor q is applied. Finally, the LQ image is resized back to the original size. Note that the downsampling and blurring contribute most to the information loss, thus we expand the degradation ranges of these two operations. Specifically, we randomly sample σ, r, δ and q from {0.1:12}, {1:12}, {0:15}, {30:100}, respectively.

6.3. More Details about Restoration Guidance

We provide a detailed explanation for our proposed restoration guidance in this section. Restoration guidance aims to achieve a trade-off between *quality* and *fidelity* through guiding the denoising process towards the high-fidelity I_{RM} obtained in the first stage. At time t, the UNet denoiser first predicts the noise ϵ_t of the noisy latent z_t. Then the predicted noise ϵ_t is removed from z_t to obatin the clean latent $\tilde{z}_0$ through the following equations:

$$\epsilon_t = \epsilon_\theta(z_t, c, t, c_{RM}), \tag{10}$$

$$\tilde{z}_0 = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t}{\sqrt{\bar{\alpha}_t}}. \tag{11}$$

This indicates that we could modify the clean latent $\tilde{z}_0$ in each time step, and then sample z_{t-1} according to the predefined distribution $q(z_{t-1}|z_t, \tilde{z}_0)$. In this way, we are able to achieve preferred restoration results without additional training. To modify $\tilde{z}_0$, we define a region-adaptive MSE loss in image space:

$$\mathcal{L}(\tilde{z}_0) = \frac{1}{HWC}||\mathcal{W} \odot (\mathcal{D}(\tilde{z}_0) - I_{RM})||_2^2, \tag{12}$$

$$\mathcal{W} = 1 - \mathcal{G}(I_{RM}), \tag{13}$$

where H, W, C denotes the spatial size of I_{RM}, and $\mathcal{W}$ is a weight map. $\mathcal{G}(I_{RM})$ is the normalized gradient magnitude of I_{RM}, which represents the gradient intensity of each pixel in I_{RM}. To obtain $\mathcal{G}(I_{RM})$, we first calculate the gradient magnitude for each pixel in I_{RM}:

$$M(I_{RM}) = \sqrt{G_x(I_{RM})^2 + G_y(I_{RM})^2} \tag{14}$$

where G_x and G_y denotes the sobel operator in x and y axis, respectively. As pixels with strong gradient signals are very rare in an image, we then use patch-level gradient signals for better estimate the gradient intensity. We divide I_{RM} into multiple equal-sized non-overlapping patches as follows:

$$\{I_{RM}^{(1)}, I_{RM}^{(2)}, ..., I_{RM}^{(k)}, ...\} \tag{15}$$
$$\forall i, j, I_{RM}^{(i)} \cap I_{RM}^{(j)} = \emptyset, \bigcup_i I_{RM}^{(i)} = I_{RM}$$

For patch $I_{RM}^{(k)}$, we calculate the sum of the gradient magnitudes of all pixels, and use the tanh function to map them into the range of $[0, 1)$:

$$S(I_{RM}^{(k)}) = \tanh\left(\sum_{i,j} M_{i,j}(I_{RM})\right), (i,j) \in I_{RM}^{(k)} \tag{16}$$

where (i, j) denotes a pixel in patch $I_{RM}^{(k)}$. As $S(I_{RM}^{(k)})$ is closer to 1, the corresponding gradient signal is stronger, and vice versa. The final gradient magnitude can be formulated as below:

$$\mathcal{G}_{i,j}(I_{RM}) = \sum_k \mathbb{I}\left[(i,j) \in I_{RM}^{(k)}\right] S(I_{RM}^{(k)}), \tag{17}$$

where $\mathbb{I}\left[(i,j) \in P^{(k)}\right]$ is an indicator function, denoting whether the pixel (i, j) is located in the patch $I_{RM}^{(k)}$. The whole algorithm is illustrated in Algorithm 1.

Datasets	Metrics	FeMaSR [5]	DASR [30]	Real-ESRGAN+ [56]	BSRGAN [73]	SwinIR-GAN [29]	StableSR [52]	PASD [66]	DiffBIR (s=0)	DiffBIR (s=0.5)	DiffBIR (s=1)
DRealSR [62]	PSNR↑	23.1977	26.3844	24.6878	25.6903	25.3898	23.8669	24.8735	24.2037	24.9891	25.6238
	SSIM↑	0.6239	0.7271	0.6705	0.6765	0.6962	0.6400	0.6529	0.5874	0.6246	0.6544
	LPIPS↓	0.2190	0.1793	0.2290	0.2308	0.2057	0.2355	0.2016	0.2448	0.2328	0.2350
	MUSIQ↑	68.7458	66.0651	67.4608	68.9388	68.1393	69.2621	70.7670	72.3514	71.5339	69.8821
	MANIQA↑	0.3073	0.2048	0.2315	0.2309	0.2375	0.2565	0.2889	0.3915	0.3847	0.3530
	CLIP-IQA↑	0.6327	0.5086	0.5022	0.5328	0.5244	0.5988	0.6151	0.6878	0.6761	0.6440
RealSR [3]	PSNR↑	23.1627	25.5503	24.2400	24.9717	24.6244	23.5627	24.5385	23.5237	24.2216	24.7531
	SSIM↑	0.6534	0.7183	0.6793	0.6839	0.7051	0.6549	0.6694	0.5989	0.6346	0.6615
	LPIPS↓	0.2520	0.2397	0.2556	0.2545	0.2340	0.2429	0.2317	0.2646	0.2544	0.2565
	MUSIQ↑	66.1208	59.5565	66.7333	68.0673	67.0964	68.4594	70.0043	72.3909	71.3969	69.5167
	MANIQA↑	0.2652	0.1713	0.2243	0.2329	0.2281	0.2407	0.2746	0.3820	0.3792	0.3504
	CLIP-IQA↑	0.5925	0.4300	0.4787	0.5233	0.4920	0.5852	0.5822	0.6868	0.6817	0.6478

Table 10. Quantitative comparisons on synthetic datasets (DRealSR [62] and RealSR [3]) for BSR task. **Red** and blue indicate the best and second best performance. The top 3 results are marked as gray.

Metrics	Real-ESRGAN+ [56]	BSRGAN [73]	SwinIR-GAN [29]	FeMaSR [5]	DASR [30]	StableSR [52]	PASD [66]	DiffBIR
Inference Time (ms)	46.19	46.42	126.44	89.01	12.69	19278.46	16951.08	10906.51
Model Size (M)	16.69	16.69	11.71	34.05	8.06	1409.11	1675.76	1716.7

Table 11. Quantitative comparisons of inference efficiency and model complexity.

Algorithm 1 Restoration guidance, given a diffusion model ϵ_θ, and the VAE's encoder $\mathcal{E}$ and decoder $\mathcal{D}$

Input: Guidance image I_{RM}, text description c (set to empty), diffusion steps T, gradient scale s
Output: Output image $\mathcal{D}(z_0)$
Sample z_T from $\mathcal{N}(0, \mathbf{I})$
for t from T to 1 **do**
$$\tilde{z}_0 \leftarrow \frac{z_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, c, t, \mathcal{E}(I_{RM}))}{\sqrt{\bar{\alpha}_t}}$$
$$\mathcal{W} = 1 - \mathcal{G}(I_{RM})$$
$$\mathcal{L}(\tilde{z}_0) = \frac{1}{HWC}\,||\mathcal{W} \odot (\mathcal{D}(\tilde{z}_0) - I_{RM})||_2^2$$
Sample z_{t-1} from $q(z_{t-1}|z_t, \tilde{z}_0 - s\nabla_{\tilde{z}_0}\mathcal{L}(\tilde{z}_0))$
end for
return $\mathcal{D}(z_0)$

6.4. More Quantitative and Qualitative Comparisons for BSR on Synthetic Datasets

The quantitative results on DRealSR [62] and RealSR [3] are presented in Table 10. The comparisons on these two datasets lead to similar observations. When the guidance scale s is set to 0, DiffBIR significantly outperforms baseline methods in terms of all IQA metrics. When the guidance scale s is set to 1, DiffBIR still surpasses the baseline methods in MANIQA and CLIP-IQA. As for evaluation in PSNR, DiffBIR performs better than diffusion-based methods and shows comparable performance to GAN-based methods, indicating that DiffBIR can achieve a good balance between *quality* and *fidelity*. Visual comparisons on DIV2K-Val[1] are presented in Figure 13. We can observe that only DiffBIR is able to produce restored results with correct semantic information. For example, it correctly recovers details such as the eyes behind the helmet, the lines of fireworks, and the wings of the penguin. GAN-based methods shows a lack of generation capability, thus producing over-smoothed results. In comparison, diffusion-based baseline methods are usually affected by the severe degradation and fail to generate correct semantics.

6.5. Quantitative Comparisons for Efficiency

We present a quantitative comparison regarding inference speed and model complexity for both diffusion-based and GAN-based methods in Table 11. This comparison is performed on a super-resolution task with an input size of 128×128 and a scale factor of 4. We conduct multiple inferences and calculate the average inference time. It can be observed that DiffBIR is the most efficient among DM-based baselines. It's about 1.8x faster than StableSR and about 1.6x faster than PASD. Although GAN-based methods are more efficient, they perform significantly worse than DM-based methods. The development of diffusion models is extremely fast. There're works [33, 44] that can already achieve satisfactory generation performance with only 1$\sim$4 steps, thus the time-consuming problem can be solved in the future.

6.6. More Real-world Visual Comparisons

We provide more visual comparisons for BSR, BID, BFR tasks in Figure 14, Figure 15 and Figure 16, respectively.

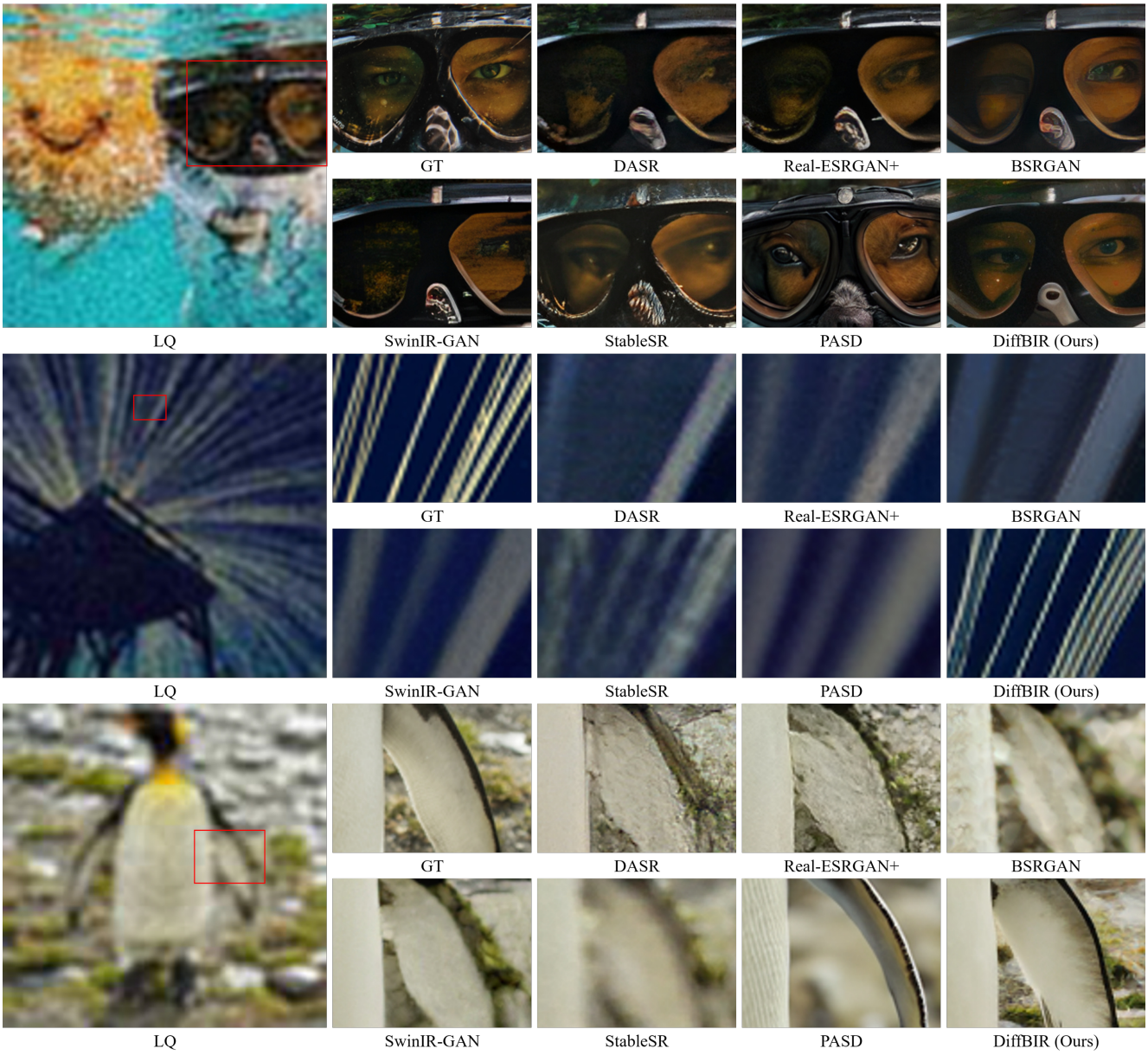

Figure 13. Visual comparisons of BSR methods on synthetic dataset (DIV2K-Val [1]).

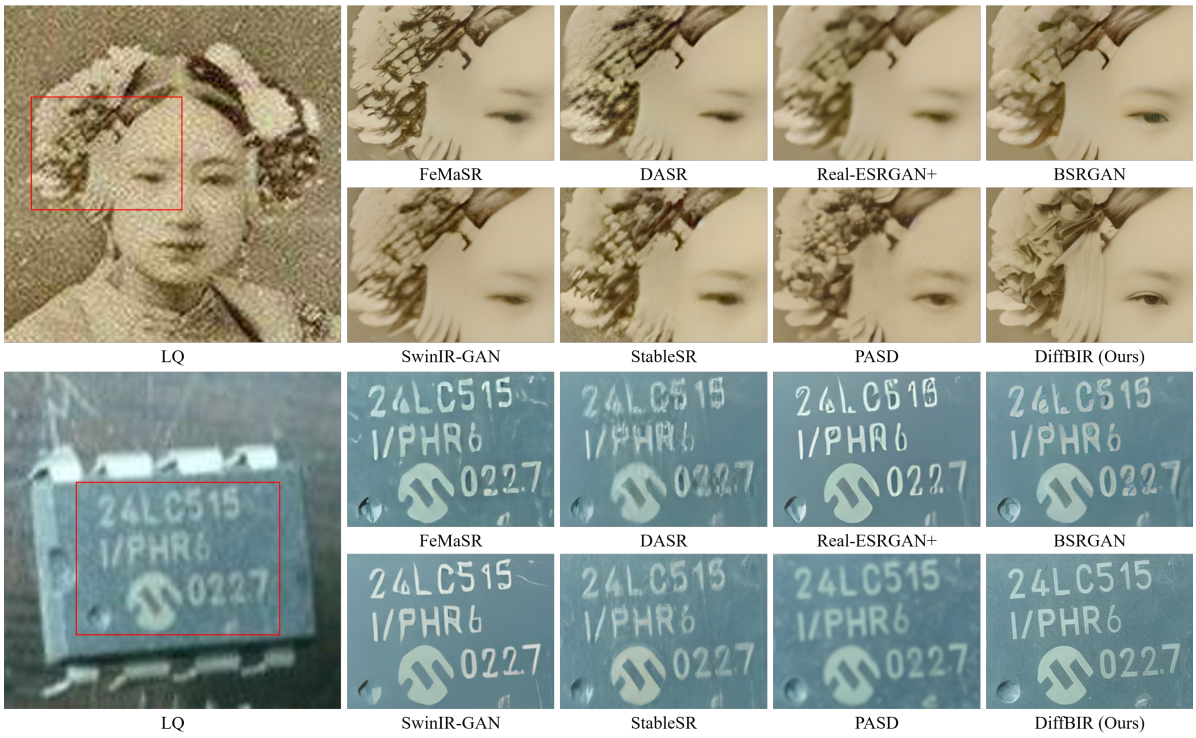

Figure 14. More visual comparisons for BSR on real-world datasets.

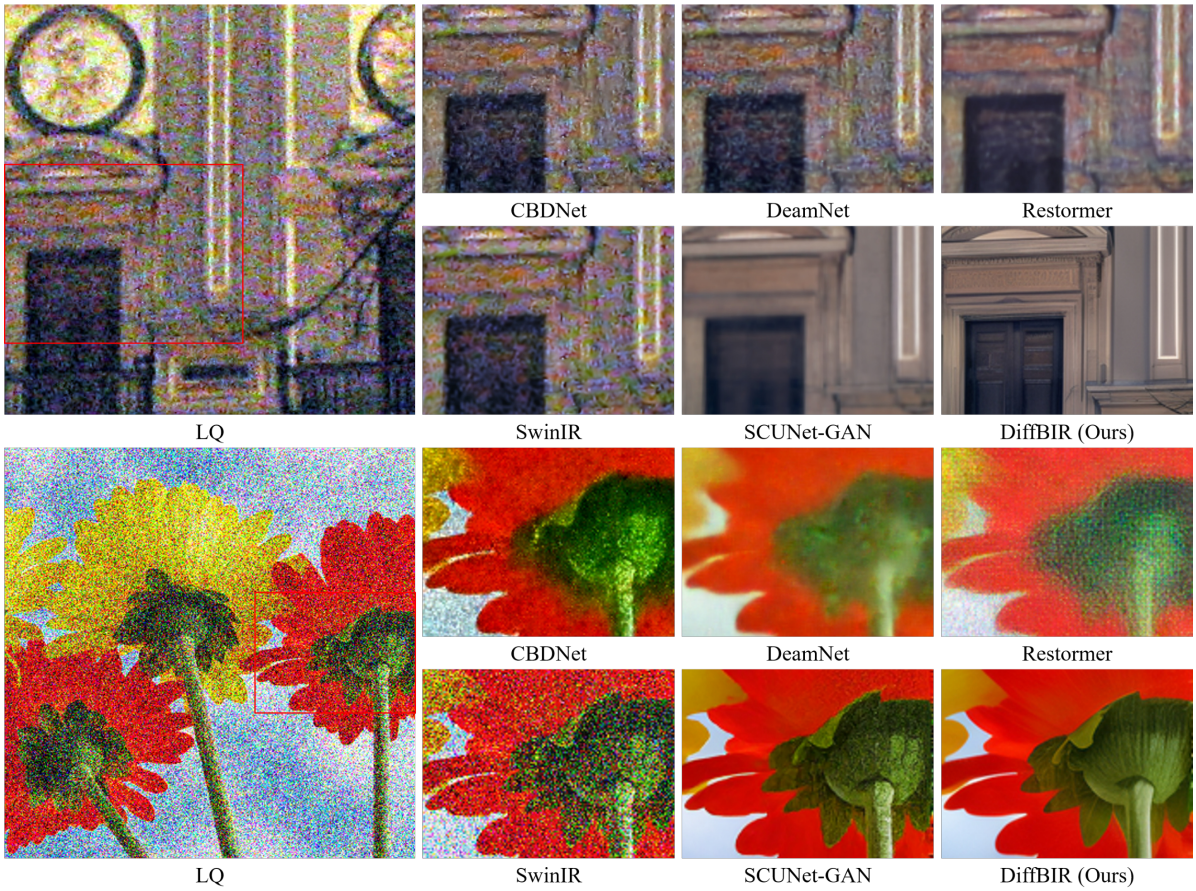

Figure 15. More visual comparisons for BID on real-world datasets.

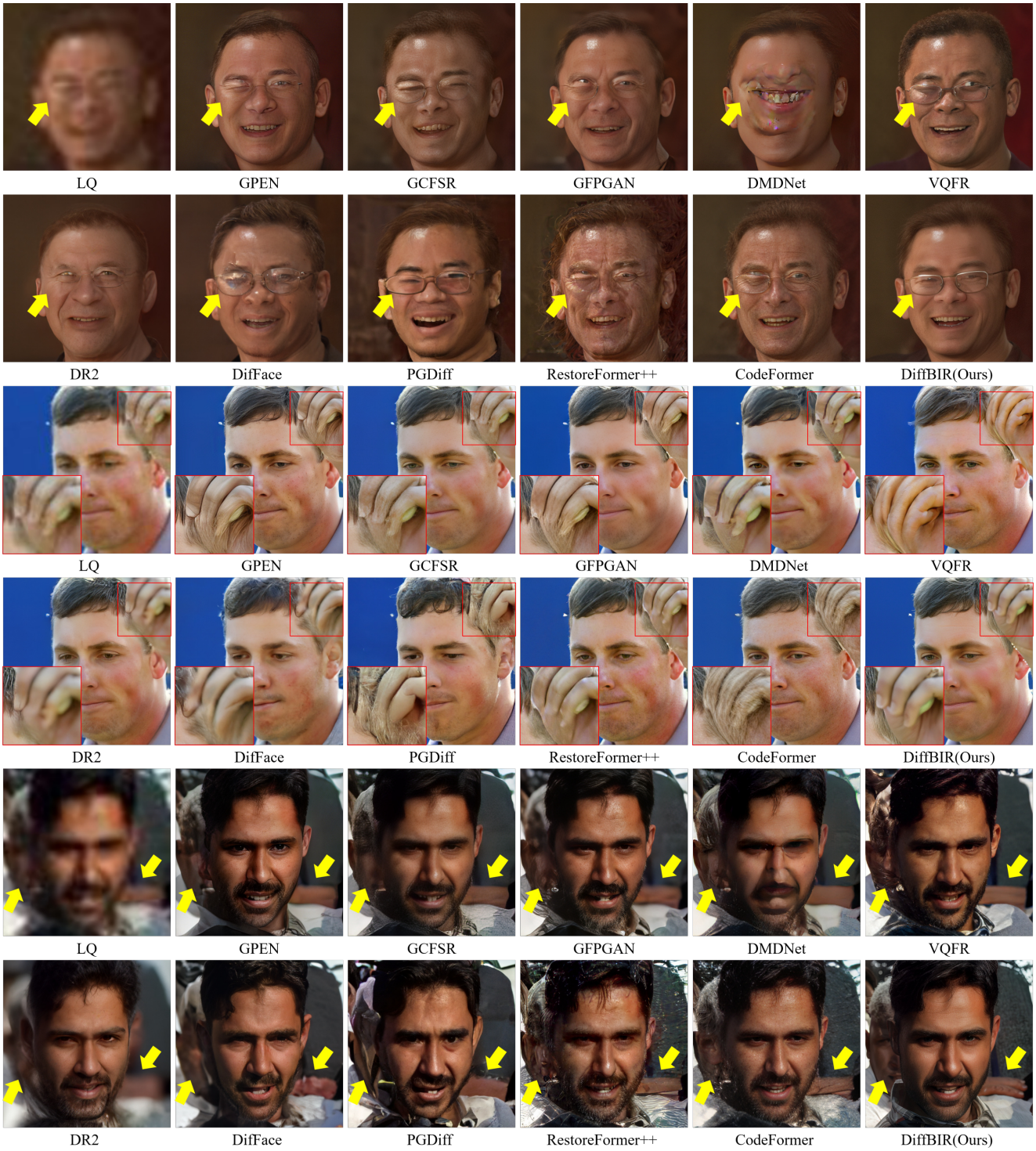

Figure 16. More visual comparisons for BFR on real-world datasets.