DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior

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

Abstract

We present DiffBIR, which leverages pretrained text-to-image diffusion models for blind image restoration problem. Our framework adopts a two-stage pipeline. In the first stage, we pretrain a restoration module across diversified degradations to improve generalization capability in real-world scenarios. The second stage leverages the generative ability of latent diffusion models, to achieve realistic image restoration. Specifically, we introduce an injective modulation sub-network – LAControlNet for finetuning, while the pre-trained Stable Diffusion is to maintain its generative ability. Finally, we introduce a controllable module that allows users to balance quality and fidelity by introducing the latent image guidance in the denoising process during inference. Extensive experiments have demonstrated its superiority over state-of-the-art approaches for both blind image super-resolution and blind face restoration tasks on 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.*, Gaussian noise and Bicubic downsampling). They have successfully promoted a vast number of excellent restoration algorithms [14; 65; 36; 7; 58; 63; 8], but are born to have limited generalization ability. To deal with real-world degraded images, 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).

The research of BIR is still in its primary stage, thus requiring more explanations of its current state. According to the problem settings, existing BIR methods can be roughly grouped into three research topics, namely blind image super-resolution (BSR), zero-shot image restoration (ZIR) and blind face restoration (BFR). They all have achieved remarkable progress, but also have apparent limitations. BSR is initially proposed to solve real-world super-resolution problems, where the low-resolution image contains unknown degradations. According to the recent BSR survey [37], the most popular solutions may be BSRGAN [64] and Real-ESRGAN [55]. 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 [31; 17; 56; 41; 49] is incorporated for learning the reconstruction process in an end-to-end manner. They have indeed removed most degradations on general images, but cannot generate

[*]Equal contribution
[†]Corresponding author

Preprint. Under review.

(a) Visual comparison of blind image super-resolution (BSR) methods on real-world low-quality images.

(b) Visual comparison of blind face restoration (BFR) methods on real-world low-quality face images.

Figure 1: Comparisons of DiffBIR and state-of-the-art BSR/BFR methods on real-world images. Compared to BSR methods, DiffBIR is more effective to 1) generate natural textures; 2) reconstruct semantic regions; 3) not erase small details; 4) overcome severe cases. Compared to BFR methods, DiffBIR can 1) handle occlusion cases; 2) obtain satisfactory restoration beyond facial areas (*e.g.*, headwear, earrings). (**Zoom in for best view**)

realistic details. Furthermore, their degradation settings are limited to $\times 4/\times 8$ super-resolution, which is not complete for the BIR problem. The second group ZIR is a newly emerged direction. Representative works are DDRM [26], DDNM [57], and GDP [16]. They incorporate the powerful diffusion model as the additional prior, thus having greater generative ability than GAN-base methods. With a proper degradation assumption, they can achieve impressive zero-shot restoration on classic IR tasks. However, the problem setting of 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. The third group is BFR, which focuses on human face restoration. State-of-the-art methods can refer to CodeFormer [68] and VQFR [18]. They have a similar solution pipeline as BSR methods, but are different in the degradation model and generation network. Due to a smaller image space, these methods can utilize VQGAN and Transformer to achieve surprisingly good results on real-world face images. Nevertheless, BFR is only a sub-domain of BIR. It usually

assumes a fixed input size and restricted image space, thus cannot be applied to general images. According to the above analysis, we can see that existing BIR methods cannot achieve (1) realistic image reconstruction on (2) general images with (3) general degradations, simultaneously. Therefore, we desire a new BIR method to overcome these limitations.

In this work, we propose DiffBIR to integrate the advantages of previous works into a unified framework. Specifically, DiffBIR (1) adopts an expanded degradation model that can generalize to real-world degradations, (2) utilizes the well-trained Stable Diffusion as the prior to improve generative ability, (3) introduces a two-stage solution pipeline to ensure both realness and fidelity. We also make dedicated designs to realize these strategies. First, to increase generalization ability, we combine the diverse degradation types in BSR and the wide degradation ranges in BFR to formulate a more practical degradation model. This helps DiffBIR to handle diverse and extreme degradation cases. Second, to leverage Stable Diffusion, we introduce an injective modulation sub-network – LAControlNet that can be optimized for our specific task. Similar to ZIR, the pre-trained Stable Diffusion is fixed during finetuning to maintain its generative ability. Third, to realize faithful and realistic image reconstruction, we first apply a Restoration Module (*i.e.*, SwinIR) to reduce most degradations, and then finetune the Generation Module (*i.e.*, LAControlNet) to generate new textures. Without this pipeline, the model may either produce over-smoothed results (remove Generation Module) or generate wrong details (remove Restoration Module). In addition, to meet users' diverse requirements, we further propose a controllable module that could achieve continuous transition effects between restoration result in stage one and generation result in stage two. This is achieved by introducing the latent image guidance during the denoising process without re-training. The gradient scale that applies to the latent image distance can be tuned to trade off realness and fidelity.

Equipped with the above components, the proposed DiffBIR demonstrates excellent performance in both BSR and BFR tasks on synthetic and real-world datasets. It is worth noting that DiffBIR achieves a great performance leap in general image restoration, outperforming existing BSR and BFR methods (*e.g.*, BSRGAN [64], Real-ESRGAN [55], CodeFormer [68], et.al). We can observe the differences of these methods in some aspects. For complex textures, BSR methods tend to generate unrealistic details, while DiffBIR can produce visually pleasant results, see Figure 1(first row). For semantic regions, BSR methods tend to achieve over-smoothed effects, while DiffBIR can reconstruct semantic details, see Figure 1(second row). For tiny stripes, BSR methods tend to erase those details, while DiffBIR can still enhance their structures, see Figure 1(third row). Moreover, DiffBIR is able to deal with extreme degradations and regenerate realistic and vivid semantic content, see Figure 1(the last row). All these show that DiffBIR has successfully broken the bottlenecks of existing BSR methods. For blind face restoration, DiffBIR shows superiority in dealing with some hard cases, such as maintaining good fidelity on facial area occluded by other objects (see first row in 1 (b)), achieving successful restoration beyond facial areas (see first row in 1 (b)). In conclusion, our DiffBIR could obtain competitive performance for both BSR and BFR tasks in a unified framework for the first time. Extensive and intensive experiments have demonstrated the superiority of our proposed DiffBIR over the existing state-of-the-art BSR and BFR methods.

2 Related Work

Blind Image Super-Resolution. Latest advances [37] on BSR have explored more complex degradation models to approximate real-world degradations. In particular, BSRGAN [64] aims to synthesize more practical degradations based on a random shuffling strategy, and Real-ESRGAN [55] exploits "high-order" degradation modeling. They both utilize GANs [17; 41; 49; 31; 56] to learn the image reconstruction process under complex degradations. SwinIR-GAN [36] uses the new prevailing backbone Swin Transformer [38] to achieve better image restoration performance. FeMaSR [6] formulates SR as a feature-matching problem based on pre-trained VQ-GAN [15]. Although BSR methods can be useful to remove degradations in the real world, they are not good at generating realistic details. In addition, they typically assume the low-quality image input is downsampled by some certain scales (e.g. $\times 4/\times 8$), which is limited for BIR problem.

Zero-shot Image Restoration. ZIR aims to achieve image restoration by leveraging a pre-trained prior network in an unsupervised manner. Earlier works [2; 10; 40; 44] mainly concentrate on searching a latent code within a pre-trained GAN's latent space. Recent advancements in this field embrace the utilization of Denoising Diffusion Probabilistic Models [21; 51; 52; 46; 45; 48]. DDRM [26] 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 [12], GDP [16] 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 real world.

Blind Face Restoration. As a specific sub-domain of general images, the face image typically carries more structural and semantic information. Early attempts utilize geometric priors (e.g. facial parsing maps [5], facial landmarks[9; 27], and facial component heatmaps [62]) or reference priors[34; 33; 32; 13] as auxiliary information to guide the face restoration process. With the rapid development of generative networks, many BFR approaches incorporate powerful generative-prior to reconstruct faces in great realness. Representative GAN-prior-based methods [54; 61; 19; 4] have demonstrated their capability in achieving both high-quality and high-fidelity face reconstruction. State-of-the-art works [68; 18; 59] introduce the HQ codebook to generate surprisingly realistic face details by exploiting Vector-Quantized (VQ) dictionary learning [53; 15].

3 Methodology

In this work, we aim to exploit a powerful generative prior – Stable Diffusion to solve blind restoration problems for both general and face images. Our proposed framework adopts a two-stage pipeline that is effective, robust, and flexible. First, we employ a Restoration Module to remove corruptions, such as noises or distortion artifacts, using regression loss. As the lost local textures and coarse/fine details are still absent, we then leverage Stable Diffusion to remedy the information loss. The overall framework is illustrated in Figure 2. Specifically, we first pretrain a SwinIR [36] on large-scale dataset to achieve the preliminary degradation removal across diversified degradations (Section 3.1). Then, the generative prior is leveraged for producing realistic restoration results (Section 3.2). In addition, a controllable module based on latent image guidance is introduced for trade-off between *realness* and *fidelity* (Section 3.3).

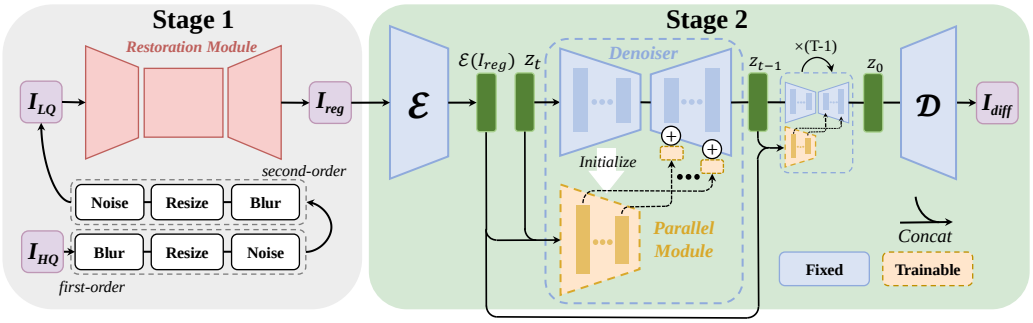


Figure 2: The two-stage pipeline of DiffBIR: 1) pretrain a Restoration Module (RM) for degradation removal to obtain I_{reg}; 2) leverage fixed Stable Diffusion through our proposed LAControNet for realistic image reconstruction and obtain I_{diff}. RM is trained across diversified degradations in a self-supervised manner, and is fixed during stage-two. LAControlNet contains a parallel module that is partially initialized with the denoiser's checkpoint and has several fusion layers. It uses VAE's encoder to project the I_{reg} to the latent space, and performs concatenation with the randomly sampled noisy z_t as the conditioning mechanism.

3.1 Pretraining for Degradation Removal

Degradation Model. BIR aims to restore clean images from low-quality (LQ) ones with unknown and complex degradations. Typically, blur, noise, compression artifacts, and low-resolution are often involved. In order to better cover the degradation space of the LQ images, we employ a comprehensive degradation model that considers *diversified degradation* and *high-order degradation*. Among all degradations, **blur**, **resize**, and **noise** are the three key factors in real-world scenarios [64]. Our *diversified degradation* involves **blur**: isotropic Gaussian and anisotropic Gaussian kernels; **resize**: area resize, bilinear interpolation and bicubic resize; **noise**: additive Gaussian noise, Poisson noise, and JPEG compression noise. Regarding *high-order degradation*, we follow [55] to use the second-order degradation, which repeats the classical degradation model: **blur**-**resize**-**noise** process

twice. Note that our degradation model is designed for image restoration, thus all the degraded images will be resized back to their original size.

Restoration Module. To build a robust generative image restoration pipeline, we adopt a conservative yet feasible solution by first removing most of the degradations (especially the noise and compression artifacts) in the LQ images, and then use the subsequent generative module to reproduce the lost information. This design will promote the latent diffusion model to focus more on textures/details generation without the distraction of noise corruption, and achieve more realistic/sharp results without wrong details (see Section 4.3). We modify SwinIR [36] as our restoration module. Specifically, we utilize the pixel unshuffle [50] operation to downsample the original low-quality input I_{LQ} with a scale factor of 8. Then, a 3×3 convolutional layer is adopted for shallow feature extraction. All the subsequent transformer operations are performed in low resolution space, which is similar to latent diffusion model. The deep feature extraction adopts several Residual Swin Transformer Blocks (RSTB), and each RSTB has several Swin Transformer Layers (STL). The shallow and deep features will be added for maintaining both low-frequency and high-frequency information. For upsampling the deep features back to the original image space, we perform nearest interpolation for three times, and each interpolation is followed by one convolutional layer as well as one Leaky ReLU activation layer. We optimize the parameters of the restoration module by minimizing the $\mathcal{L}_2$ pixel loss. The formulation is as follows:

$$I_{reg} = \texttt{SwinIR}(I_{LQ}), \;\; \mathcal{L}_{reg} = ||I_{reg} - I_{HQ}||_2^2, \tag{1}$$

where I_{HQ} and I_{LQ} denote the high-quality image and the low-quality counterpart, respectively. I_{reg} is obtained by regression learning and will be used for the finetuning on latent diffusion model.

3.2 Leverage Generative Prior for Image Reconstruction

Preliminary: Stable Diffusion. In this paper, we implement our method based on the large-scale text-to-image latent diffusion model – Stable Diffusion. Diffusion models learn to generate data samples through a denoising sequence that estimate the score of the data distribution. In order to achieve better efficiency and stabilized training, Stable Diffusion pretrains an autoencoder [29] that converts an image x into a latent z with encoder $\mathcal{E}$ and reconstructs it with decoder $\mathcal{D}$. This latent representation is learned by using hybrid objectives of VAE [30], Patch-GAN [23], and LPIPS [67]. The 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 $z = \mathcal{E}(x)$ for producing 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 latent diffusion model is defined as follows:

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

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.

LAControlNet. Although stage-one could remove most degradations, the obtained I_{reg} is often over-smoothed and still far from the distribution of high-quality natural images. We then leverage the pre-trained Stable Diffusion for image reconstruction with our obtained I_{reg}-I_{HQ} pairs. First, we utilize the encoder of Stable Diffusion's pretrained VAE to map I_{reg} into the latent space, and obtain the condition latent $\mathcal{E}(I_{reg})$. The UNet [47] denoiser performs latent diffusion, which contains an encoder, a middle block, and a decoder. In particular, the decoder receives the features from encoder and fuses them in different scales. Here we create a parallel module (denoted as orange in Figure 2) that contains the same encoder and the middle block as in the UNet denoiser. Then, we concatenate the condition latent $\mathcal{E}(I_{reg})$ with the randomly sampled noisy z_t as the input for the parallel module. Since this concatenation operation will increase the channel number of the first convolutional layer in the parallel module, we initialize the newly added parameters to zero, where all other weights are initialized from the pre-trained UNet denoiser checkpoints. The outputs of the parallel module are added to the original UNet decoder. Moreover, one 1×1 convolutional layer is applied before the addition operation for each scale. During finetuning, the parallel module and these

1×1 convolutional layers are optimized simultaneously, where the prompt condition is set to empty. We aim to minimize the following latent diffusion objective:

$$\mathcal{L}_{Diff} = \mathbb{E}_{z_t,c,t,\epsilon,\mathcal{E}(I_{reg})}[||\epsilon - \epsilon_\theta(z_t, c, t, \mathcal{E}(I_{reg}))||_2^2]. \tag{4}$$

The obtained result in this stage is denoted as I_{diff}. To summarize, only the skip-connected features in the UNet denoiser are tuned for our specific task. This strategy alleviates overfitting in small training dataset, and could inherit the high-quality generation from Stable Diffusion. More importantly, our conditioning mechanism is more straightforward and effective for image reconstruction task compared to ControlNet [66], which utilizes an additional condition network trained from scratch for encoding the condition information. In our LAControlNet, the well-trained VAE's encoder is able to project the condition images into the same representation space as the latent variables. This strategy significantly alleviates the burden on the alignment between the internal knowledge in latent diffusion model and the external condition information. In practice, directly utilizing ControlNet for image reconstruction leads to severe color shifts as shown in the ablation study (see Section 4.3).

3.3 Latent Image Guidance for Fidelity-Realness Trade-off

Although the above two-stage approach could already achieve good restoration results, a trade-off between *realness* and *fidelity* is still needed due to various users' preferences. Thus, we propose a controllable module that could guide the denoising process towards the obtained I_{reg} in stage-one, thus obtaining an adjustment between realistic and smooth results. Classifier guidance is proposed by Dhariwal and Nichol [12] which utilizes a classifier trained on noisy images to guide generation towards target class. While in most cases, the pre-trained models that serve as guidance are usually trained on clean images. To handle this situation, the work in [1; 16] turn to guide the intermediate variable $\tilde{x}_0$ to control the generation process of diffusion models. Specifically, in the sampling process, they estimate a clean image x_0 from the noisy image x_t by estimating the noise in x_t. In this work, the diffusion and denoising processes are based on the latent space. Thus, we aim to obtain a clean latent z_0 by the following equation:

$$\tilde{z}_0 = \frac{z_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, c, t, \mathcal{E}(I_{reg}))}{\sqrt{\bar{\alpha}_t}}. \tag{5}$$

Then, a latent-based loss $\mathcal{D}_{latent}$ is defined as the $\mathcal{L}_2$ distance between the latent image guidance $\mathcal{E}(I_{reg})$ and the estimated clean latent $\tilde{z}_0$:

$$\mathcal{D}_{latent}(x, I_{reg}) = \mathcal{L}(\tilde{z}_0, \mathcal{E}(I_{reg})) = \sum_j \frac{1}{C_j H_j W_j} ||\tilde{z}_0 - \mathcal{E}(I_{reg}))||_2^2. \tag{6}$$

The above guidance could iteratively force spatial alignment and color consistency between latent features, and guide the generated latent to preserve the content of the reference latent. Therefore, one can control how much information (such as structure, layout and color) is maintained from the reference image I_{reg}, thus achieving a transition from generated output to more smooth result. The whole algorithm of our latent image guidance is illustrated in Algorithm 1.

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

Input: Guidance image I_{reg}, 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**
  $\tilde{z}_0 \leftarrow \frac{z_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, c, t, \mathcal{E}(I_{reg}))}{\sqrt{\bar{\alpha}_t}}$
  $\mathcal{L} = \mathcal{L}(\tilde{z}_0, \mathcal{E}(I_{reg}))$
  Sample z_{t-1} by $\mathcal{N}\left(\mu_\theta(z_t) - s\nabla_{\tilde{z}_0}\mathcal{L}, \sigma_t^2\right)$
end for
return $\mathcal{D}(z_0)$

4 Experiments

4.1 Datasets, Implementation, Metrics

Datasets. We train DiffBIR on the ImageNet [11] dataset at 512×512 resolution for BIR. As for BFR, we use FFHQ [25] dataset and resize it to 512×512. To synthesize the LQ images, we utilize the proposed degradation pipeline to process the HQ images during training (please see Appendix A for details). For BSR, we utilize RealSRSet [3] dataset for comparison in a real-world setting. For a more thorough comparison in real-world scenarios, we collect 47 images from the Internet, denoted as Real47. It contains general images of diverse scenes, such as natural outdoor landscapes, old photos, architecture, humans from portraits to dense people crowds, plants, and animals, etc. For BFR task, we evaluate our method on a synthetic dataset CelebA-Test [39] and three real-world datasets: LFW-Test [54], CelebChild-Test [54], and WIDER-Test [68]. In particular, CelebA-Test contains 3,000 images selected from the CelebA-HQ dataset, where LQ images are synthesized under the same degradation range as our training settings.

Implementation. The restoration module adopts 8 residual Swin Transformer blocks (RSTB), and each RSTB contains 6 Swin Transformer Layers (STL). The head number is set to 6 and the window size is set to 8. We train the restoration module with a batch size of 96 for 150k iterations. We utilize Stable Diffusion 2.1-base [3] as the generative prior, and finetune the diffusion model for 25k iterations with a batch size of 192. We use Adam [28] optimizer and set the learning rate to 10^{-4}. The training process is conducted on 512×512 resolution with 8 NVIDIA A100 GPUs. For inference, we adopt spaced DDPM sampling [43] with 50 timesteps. Our DiffBIR is able to handle images with arbitrary sizes larger than 512×512. For images with sides < 512, we first upsample them with the short side enlarged to 512, and then resize them back.

Metrics. Regarding the evaluation with ground truth, we adopt the traditional metrics: PSNR, SSIM, and LPIPS [67]. To better evaluate the *realness* for BIR task, we also include several no-reference image quality assessment (IQA) metrics: MANIQA[4] [60] and NIQE. For BFR, we evaluate the identity preservation - IDS [68], and employ the widely used perceptual metric FID [20]. We also deploy a user study for a more thorough comparison.

4.2 Comparisons with State-of-the-Art Methods

For BSR, we compare our DiffBIR with state-of-the-art BSR methods: Real-ESRGAN+ [55], BSRGAN [64], SwinIR-GAN [36], and FeMaSR [6]. The recent state-of-the-art ZIR methods (DDNM [57] and GDP [16]) are also included [5]. Regarding BFR task, we compare with the most recent state-of-the-art methods: DMDNet [35], GFP-GAN [54], GPEN [61], GCFSR [19], VQFR [18], CodeFormer [68], RestoreFormer [59].

BSR on real-world dataset. We provide the quantitative comparison on real-world datasets in Table 1. It is observed that our DiffBIR obtains the best scores in MANIQA on both the widely used RealSRSet [24] and our collected Real47. While BSRGAN and Real-ESRGAN+ could achieve top-3 results in MANIQA on both two datasets. The visual comparison results are presented in Figure 3. It can be seen that DiffBIR is able to restore text information more naturally, while other methods tend to distort the characters or produce blurry output. On the other hand, our DiffBIR could also generate realistic texture details for natural images, where other methods produce over-smooth results. More visualization results can be found in Figure 11 and Figure 12.

Table 1: Comparison with state-of-the-art BSR and ZIR methods on real-world datasets with a $4\times$ upsampling scale. **Red** and blue indicate the best and second best performance. The top 3 results are marked as gray.

Dataset	Metric	DDNM [57]	GDP [16]	Real-ESRGAN+[55]	BSRGAN [64]	SwinIR-GAN [36]	FeMaSR [6]	**DiffBIR(Ours)**
RealSRSet	MANIQA↑	0.4535	0.4581	0.5376	0.5640	0.5295	0.5247	0.5906
	NIQE↓	6.8415	5.0626	5.7401	5.6074	5.6093	5.2353	6.0738
Real47	MANIQA↑	0.4813	0.5237	0.5900	0.5889	0.5721	0.5718	0.6293
	NIQE↓	6.4768	3.9866	3.9103	4.0338	3.9910	4.1731	3.9240

[3] https://github.com/Stability-AI/stablediffusion

[4] MANIQA (https://github.com/IIGROUP/MANIQA) won first place in the NTIRE2022 Perceptual Image Quality Assessment Challenge Track 2 No-Reference competition.

[5] DDNM and GDP are selected because they provide an approach to restore images with arbitrary sizes.

Figure 3: Visual comparison on real-world datasets with upsampling scale factor of 4.(**Zoom in for best view**)

To further compare DiffBIR with other state-of-the-art methods, we conduct a user study on our collected Real47 dataset. This user study compares DiffBIR, SwinIR-GAN, BSRGAN, and RealESRGAN+. For each image, users are asked to rank the results of the four methods and assign 1-4 points to different methods in an ascending order. To be more exact, better result obtains higher score. 31 users are recruited to conduct this user study under detailed instruction. The distribution of scores obtained by each method is shown in Figure 4. It can be observed that DiffBIR achieves the highest median score, and its upper quartile exceeds 3. This indicates that users tend to rank DiffBIR's results in the first place. The user study results again demonstrate that DiffBIR's visual results are superior to other methods, which aligns with its highest score on MANIQA.

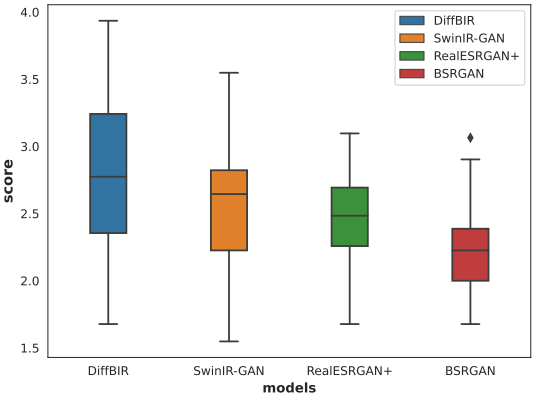


Figure 4: The distribution of scores obtained by SwinIR-GAN, Real-ESRGAN+, BSRGAN, and our DiffBIR in user study.

BFR on both synthetic and real-world datasets. We show the quantitative comparison on both synthetic and real-world datasets in Table 2. For the synthetic dataset CelebA-Test [39], our DiffBIR achieves the highest FID score. Meanwhile, it is also the top-3 methods regarding PSNR and IDS. This reveals that the proposed DiffBIR can successfully produce results with both high *realness* and high *fidelity*. For real-world datasets, DiffBIR obtains the best results on both LFW-Test [22] (mild degradation) and WIDER-Test [68] (heavy degradation) datasets, and comparable results with state-of-the-art methods on CelebChild-Test. Figure 5 depicts a visual comparison of various methods on synthetic dataset. The first example demonstrates that only DiffBIR succeeds in restoring extremely degraded cases while other methods fail. It can be seen from the second example that only DiffBIR

can successfully recover the occluded left eye. Figure 6 presents a visual comparison on real-world dataset. It can be observed from the first example that DiffBIR is able to accurately restore the hair, while other methods mistake the hair for a part of the facial area. The second example suggests that our DiffBIR is the only method that can generate realistic details on non-face area (*i.e.,* the decoration in the forehead). More visualization results can be found in Figure 9 and Figure 10.

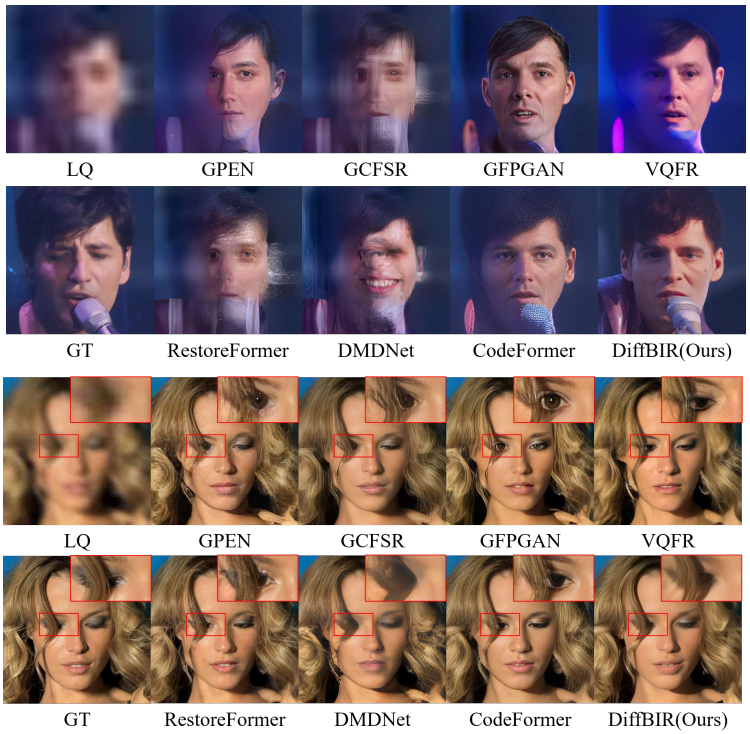


Figure 5: Qualitative comparison of different BFR methods on synthetic datasets.(**Zoom in for best view**)

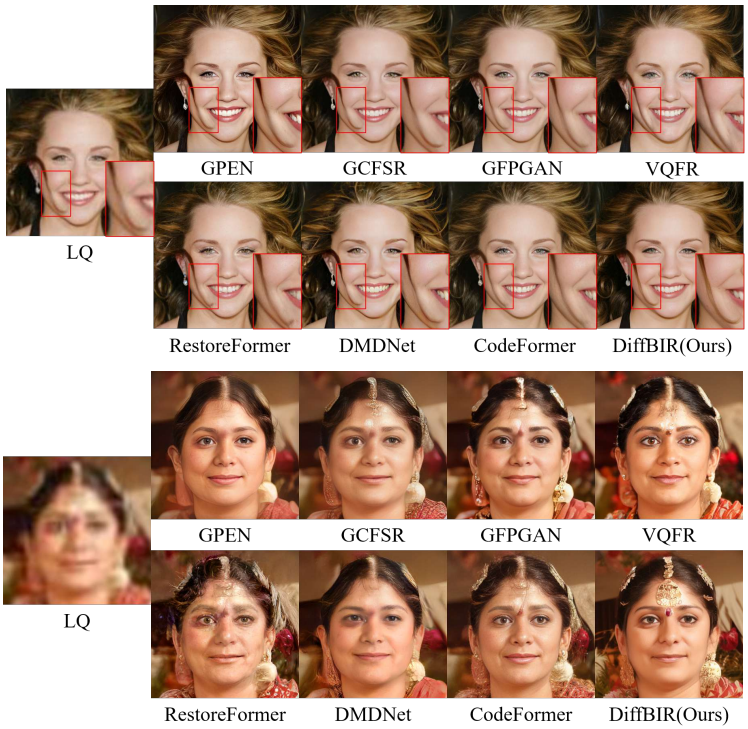


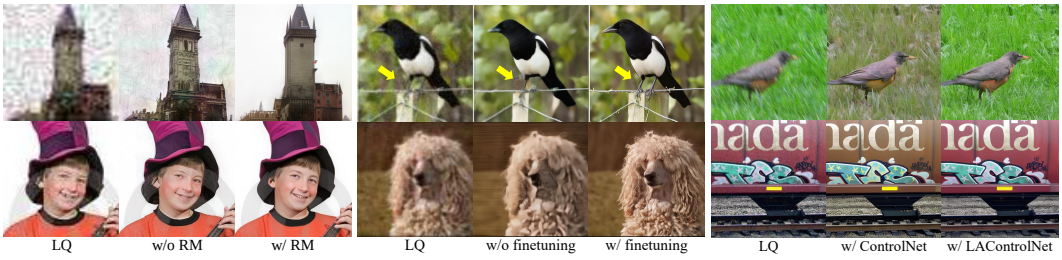
Figure 6: Qualitative comparison of different BFR methods on real-world datasets.(**Zoom in for best view**)

Table 2: Comparison with state-of-the-art methods for BFR on both synthetic and real-world face datasets. **Red** and blue indicate the best and second best performance. The top 3 results are marked as gray.

Dataset	Synthetic CelebA-Test					Wild LFW-Test	Wild WIDER-Test	Wild CelebChild-Test
Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	IDS↑	FID↓	FID↓	FID↓
GPEN [61]	21.3995	0.5742	0.4687	23.92	0.48	51.97	46.35	76.58
GCFSR [19]	21.8791	0.6072	0.4577	35.49	0.44	52.20	40.86	76.29
GFPGAN [54]	21.6953	0.6060	0.4304	21.69	0.49	52.11	41.70	80.69
VQFR [18]	21.3014	0.6132	0.4116	20.30	0.48	49.88	37.87	74.76
RestoreFormer [59]	21.0025	0.5283	0.4789	43.77	0.56	48.43	49.79	70.54
DMDNet [35]	21.6617	0.6000	0.4828	64.79	0.67	43.36	40.51	79.38
CodeFormer [68]	22.1519	0.5948	0.4055	22.19	0.47	52.37	38.78	79.54
DiffBIR(Ours)	21.7509	0.5971	0.4573	20.02	0.51	39.58	32.35	75.94

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. The removal of restoration module leads to a noticeable performance drop in FID/MANIQA across all real-world datasets (see Table 3). The visual comparison is presented in Figure 7(a). As seen from the first example, the one-stage model (w/o RM) regards the degradations as semantic information by mistake. This demonstrates that the restoration module contributes to preserving fidelity. The second example clearly illustrates that solely finetuning the Stable Diffusion without applying the RM cannot fully remove the real-world noise/artifacts. This indicates that the RM is indispensable in degradation removal.



(a) (b) (c)

Figure 7: Visual comparison of ablation studies. (a) DiffBIR w/o restoration module performs poorly in both fidelity maintaining (first row) and degradation removal (second row); (b) w/o finetuning Stable Diffusion, directly applying the image guidance technique [57; 16] is not able to produce realistic results. (c) ControlNet [66] has a color shift problem which can be addressed by our LAControlNet.(**Zoom in for best view**)

Table 3: The effectiveness of restoration module. The best results are denoted as Red.

Dataset	Face LFW-Test	Face WIDER-Test	Face CelebChild-Test	General RealSRSet	General Real47
Method	FID↓	FID↓	FID↓	MANIQA↑	MANIQA↑
DiffBIR(w/o RM)	40.78	33.22	75.98	0.582	0.624
DiffBIR(w/ RM)	39.58	32.35	75.94	0.591	0.629

The Necessity of Finetuning Stable Diffusion. Next, we illustrate the necessity of finetuning the latent diffusion model. Zero-shot IR methods [57; 16] provide an effective approach that guides the reverse diffusion process using the degraded image in the image space. Following their methodology, we employ the smoothed result I_{reg} to guide the original Stable Diffusion without finetuning. However, as depicted in Figure 7(b), this guidance strategy tends to generate unrealistic content (*i.e.,* a bird with one leg missing). This demonstrates that the widely used guidance in image space may not effectively generalize to the latent space, thus finetuning Stable Diffusion becomes indispensable for this image reconstruction task.

The Effectiveness of LAControlNet. Then we aim to emphasize the effectiveness of our proposed LAControlNet that encodes I_{reg} to the latent space. Here we compare with ControlNet [66], which adopts an additional condition network trained from scratch for conditioning the input information. As shown in Figure 7(c), ControlNet tends to output results with color shifts, as there is no explicit

regularization on color consistency during training. One might use non-uniform sampling to increase the probability of optimization in the early sampling stage and achieves better color controlling [42]. Nevertheless, our method is much more straightforward and fully exploits the latent diffusion prior.

The Flexibility of Controllable Module. Considering that generative restoration models may produce unexpected details, here we provide a controllable module for users to explore according to their personal preferences. The visualization result is shown in Figure 8. Our experiments suggest that a larger gradient scale s tends to produce a high-fidelity smooth result which is close to I_{reg}. As seen from the first row, DiffBIR's output I_{diff} has some blue artifacts in the dog's eyes, thus we set s to 200 and higher as well for obtaining a better result. Moreover, the background is also changing (tends to be more blurry) as the gradient scale s grows.

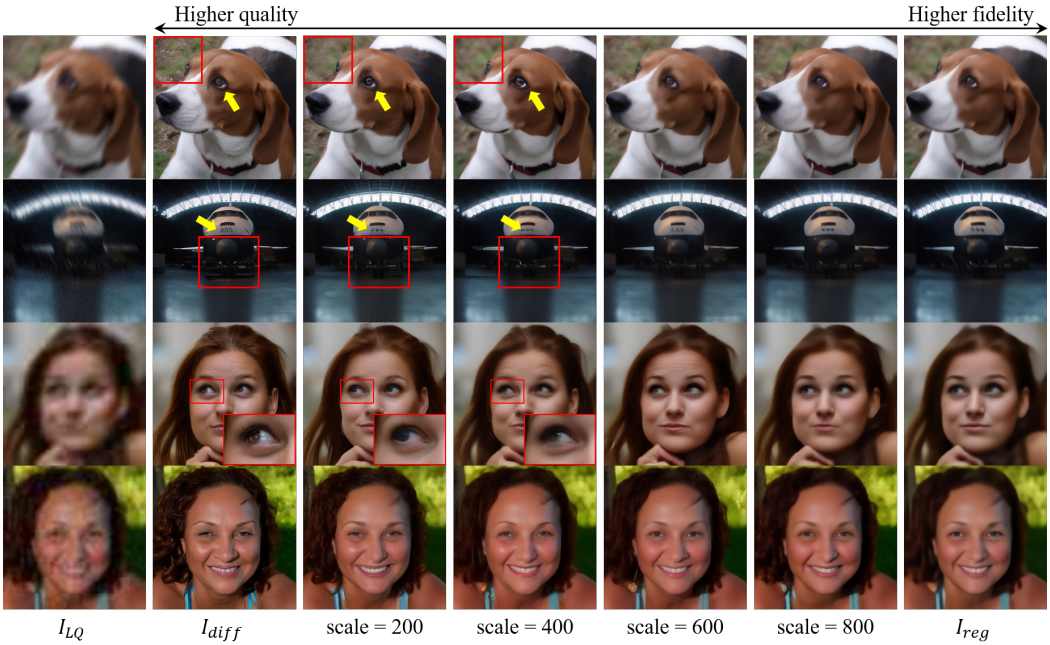


Figure 8: Our latent image guidance is able to achieve a trade-off between quality and fidelity. The gradient scale can be tuned to obtain transition effects between sharp I_{diff} and smooth I_{reg}.(**Zoom in for best view**)

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. It consists of two stages: the restoration and generation stage, which ensures both fidelity and realness. Extensive experiments have validated the superiority of DiffBIR over existing state-of-the-art methods for both BSR and BFR tasks. Although our proposed DiffBIR has shown promising results, the potential of text-driven image restoration is not explored. Further exploitation in Stable Diffusion for image restoration task is encouraged. On the other hand, our DiffBIR method requires 50 sampling steps to restore a low-quality image, resulting in much higher computational resource consumption and more inference time compared to other image restoration methods.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

[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] 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.

[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.

[5] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. Progressive semantic-aware style transformation for blind face restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11896–11905, 2021.

[6] 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.

[7] 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.

[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.

[9] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. *Cornell University - arXiv*, 2017.

[10] 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.

[11] 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.

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

[13] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. *Computer Vision and Pattern Recognition*, 2019.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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, October 2007.

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[24] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 466–467, 2020.

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

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

[27] Deokyun Kim, Minseon Kim, Gihyun Kwon, and Dae-Shik Kim. Progressive face super-resolution via attention to facial landmark. *arXiv preprint arXiv:1908.08239*, 2019.

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

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

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

[31] 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.

[32] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. *Springer International Publishing eBooks*, 2020.

[33] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2706–2715, 2020.

[34] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. *Cornell University - arXiv*, 2018.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[40] 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.

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

[42] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

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

[44] 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.

[45] 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.

[46] 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.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[48] 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.

[49] 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.

[50] 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.

[51] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[52] 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.

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

[54] 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.

[55] 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.

[56] 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.

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

[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.

[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.

[60] 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.

[61] 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.

[62] Xin Yu, Basura Fernando, Bernard Ghanem, Fatih Porikli, and Richard Hartley. Face super-resolution guided by facial component heatmaps. In *Proceedings of the European conference on computer vision (ECCV)*, pages 217–233, 2018.

[63] 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.

[64] 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.

[65] 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.

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

[67] 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.

[68] 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.

A Degradation Details

Degradation settings used for training our DiffBIR are introduced in this section. Following [55], we employ the second-order degradation process to enhance the robustness of the restoration module in real-world scenarios. Specifically, a degradation model in a certain stage consists of three operations: **blur**, **resize**, and **noise**. **Blur.** We utilize isotropic Gaussian blur or anisotropic Gaussian blur with equal probabilities. The size of the blur kernel follows a uniform distribution ranging from 7 to 21, and the blur sigma is uniformly sampled between 0.2 and 3 for the first degradation process and between 0.2 and 1.5 for the second degradation process. **Resize.** We consider multiple resize algorithms, including area resize, bilinear interpolation and bicubic resize. The scaling factor for resize follows a uniform distribution ranging from 0.15 to 1.5 for the first degradation process and from 0.3 to 1.2 for the second degradation process. **Noise.** We incorporate Gaussian noise, Poisson noise, and JPEG compression noise. The scale of Gaussian noise is uniformly sampled between 1 and 30 in the first degradation process and between 1 and 25 in the second degradation process. The scale of Poisson noise is randomly sampled from 0.05 to 3 and 0.05 to 2.5 for the first and second degradation processes, respectively. The quality of JPEG compression follows a uniform distribution ranging from 30 to 95.

Moreover, we combine the degradation settings adopted in blind face restoration. Specifically, we consider a large dowsampling range $[1, 12]$, and a large blur kernel range whose sigma is within $[0.1, 12]$. In this way, the generation module is trained to remedy the information loss within a wide range.

B More Qualitative Comparisons For BFR

Figure 9: More qualitative comparisons for BFR on synthetic dataset.(**Zoom in for best view**)

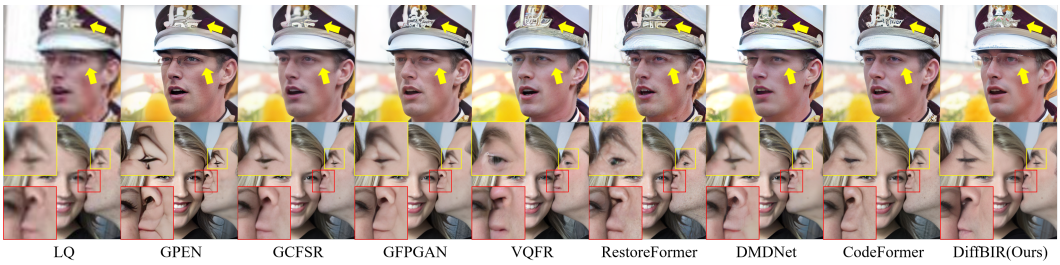
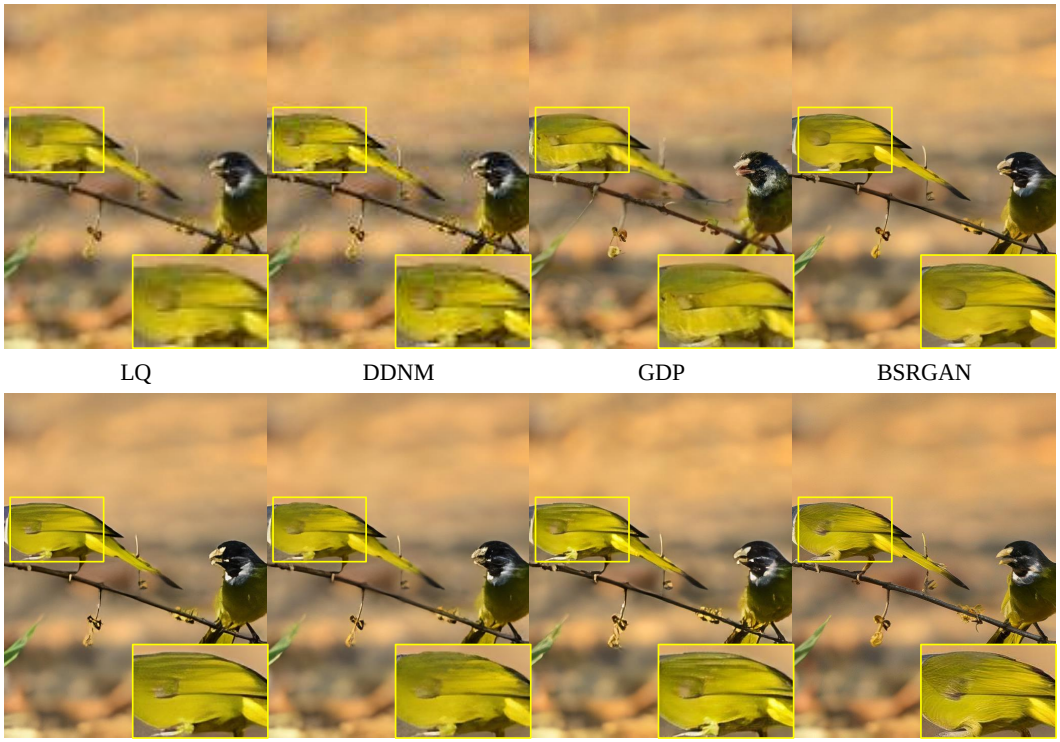
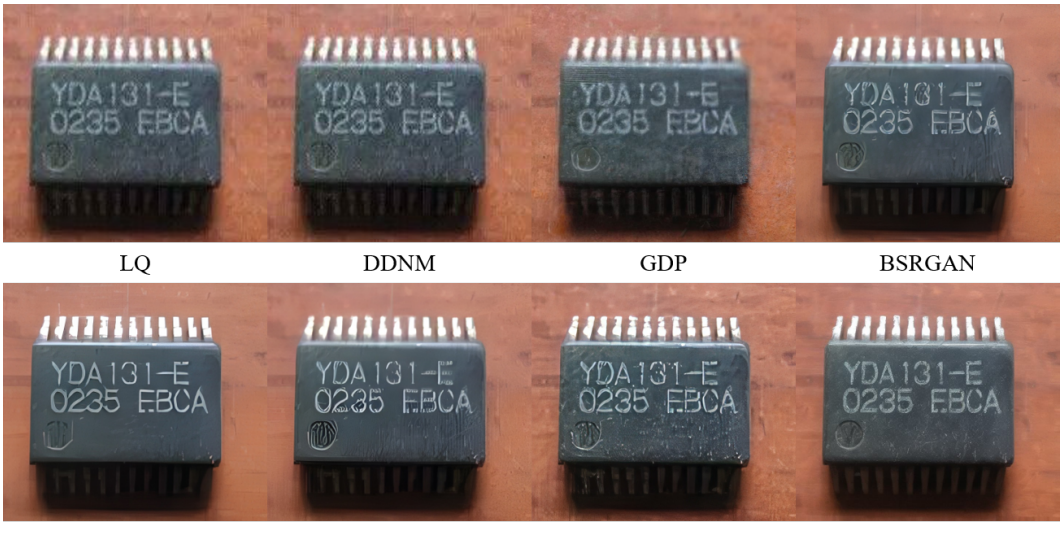


Figure 10: More qualitative comparisons for BFR on real-world dataset.(**Zoom in for best view**)

C More Qualitative Comparisons For BSR



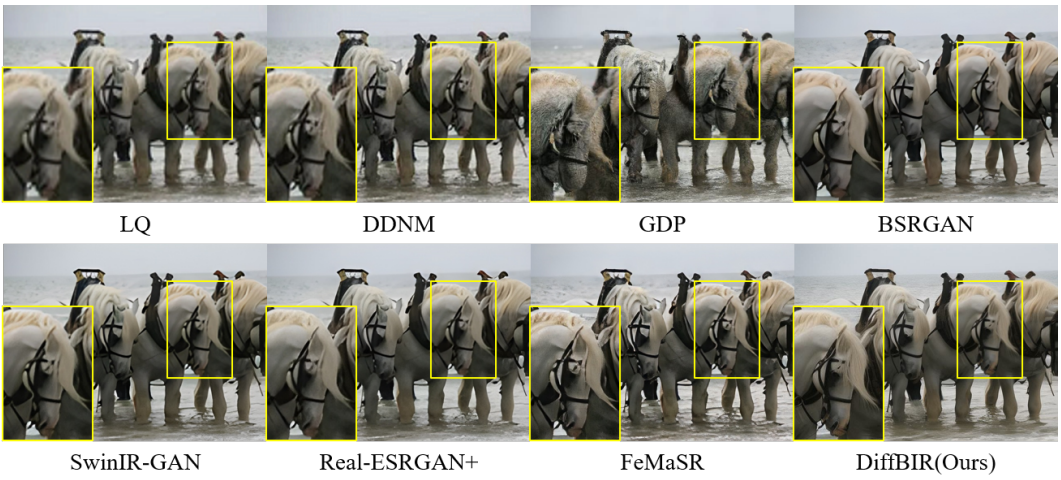
(a)



(b)

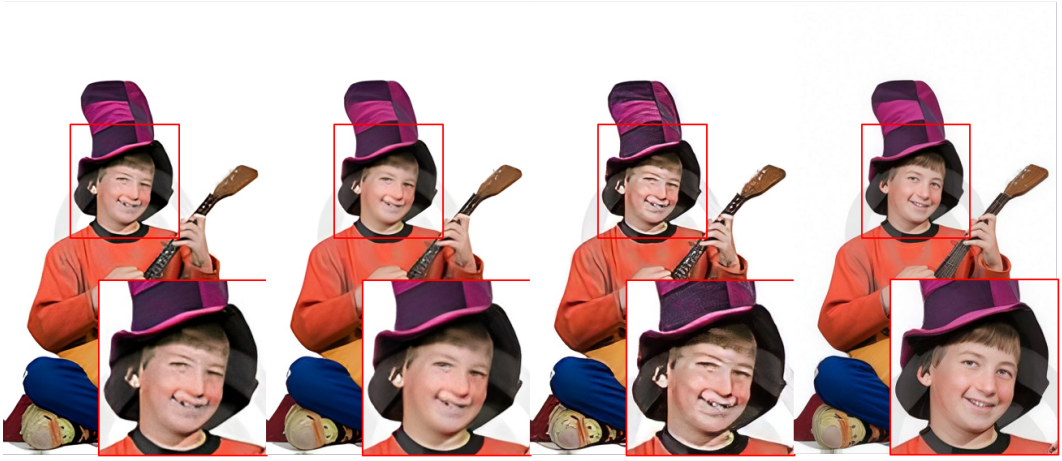
Figure 11: More qualitative comparisons on Real47 dataset.(**Zoom in for best view**)

(c)



(d)

Figure 11: More qualitative comparisons on Real47 dataset.(**Zoom in for best view**)



(e)

(f)

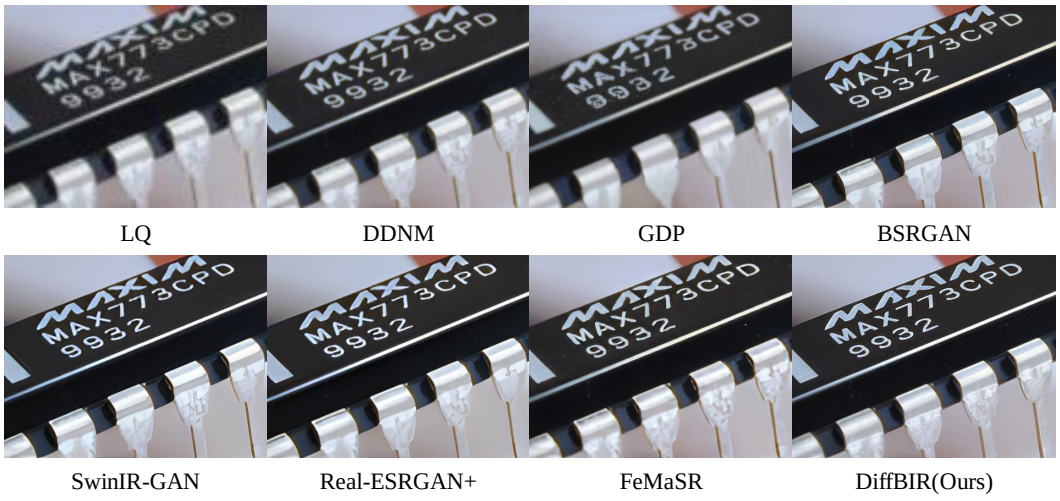
Figure 11: More qualitative comparisons on Real47 dataset.(**Zoom in for best view**)

(a)

(b)

Figure 12: More qualitative comparisons on RealSRSet [24] dataset.(**Zoom in for best view**)

(c)



(d)

Figure 12: More qualitative comparisons on RealSRSet [24] dataset.(**Zoom in for best view**)