Bi-Noising Diffusion: Towards Conditional Diffusion Models with Generative Restoration Priors

Kangfu Mei
Johns Hopkins University
kmei1@jhu.edu

Nithin Gopalakrishnan Nair
Johns Hopkins University
ngopala2@jhu.edu

Vishal M. Patel
Johns Hopkins University
vpatel36@jhu.edu

Abstract

Conditional diffusion probabilistic models can model the distribution of natural images and can generate diverse and realistic samples based on given conditions. However, oftentimes their results can be unrealistic with observable color shifts and textures. We believe that this issue results from the divergence between the probabilistic distribution learned by the model and the distribution of natural images. The delicate conditions gradually enlarge the divergence during each sampling timestep. To address this issue, we introduce a new method that brings the predicted samples to the training data manifold using a pretrained unconditional diffusion model. The unconditional model acts as a regularizer and reduces the divergence introduced by the conditional model at each sampling step. We perform comprehensive experiments to demonstrate the effectiveness of our approach on super-resolution, colorization, turbulence removal, and image-deraining tasks. The improvements obtained by our method suggest that the priors can be incorporated as a general plugin for improving conditional diffusion models. Our demo is http://bi-noising.demohub.cc.

1. Introduction

In recent years, conditional image generation has received significant attention in the computer vision community. Some applications that make use of conditional image generation include text-to-image generation (*e.g*. DALLE-2 [28]) and image restoration (*e.g*. SR3 [31]). The most challenging part of these restoration applications comes from the ill-posedness, *i.e*., the same degraded images may come from multiple different ground truth images. The ill-posedness affects the performance of traditional methods like sparse coding [20, 21] and makes it difficult for the learning-based algorithms to solve this problem. Although recent learning-based methods have made impressive progress [17], there remains a significant quality gap between the prediction and natural images.

Recent works that utilize pretrained generative networks have shown the superior visual quality of conditional generation compared to the aforementioned end-to-end learning methods. Generative models have shown impressive image generation results in terms of sample quality and diversity, indicating their capacity for encapsulating rich photo-realistic priors. Some representative methods include Generative Adversarial Networks (GANs) [7], Variational Autoencoders (VAEs) [14], and Autoregressive models [15]. Their generation process generally starts from the standard normal distribution from which diverse high-fidelity images sampled [12,13]. Recent work [29] has shown that the *continuity* in the normal distribution remains preserved in the sampled results. For example, the results produced from two different Gaussian noises with the same model will be close to each other if the two noises are close to each other in Euclidean space. The continuity allows one to perform conditional image generation in an inversion manner that inverts degraded images into standard noises. This inverted noise can then generate clear images by projecting the noise with generative models. Following the protocol, multiple GAN-based generative priors, including optimization-based [23] and learning-based [29] schemes have been proposed for various real-world tasks [37].

Denoising diffusion probabilistic models [9, 32] are the most recent deep generative models. They have shown comparable and even better performance at image synthesis than GANs with delicate guidance [6]. These models learn to sequentially denoise stochastic noise map starting from the normal distribution $\mathcal{N}(0, \mathbf{I})$ to clean images. However, the generation process is stochastic, and the continuity cannot be preserved from the initial sampled noise. For instance, two sampled noises from the same normal distribution with a small divergence may generate significantly different clear images. Such a noncontinuous generation process prevents the generative priors from being applied along with the denoising process like the inversion GANs [29]. Hence, despite their impressive synthesis capacity, diffusion model-based priors have not been explored before.

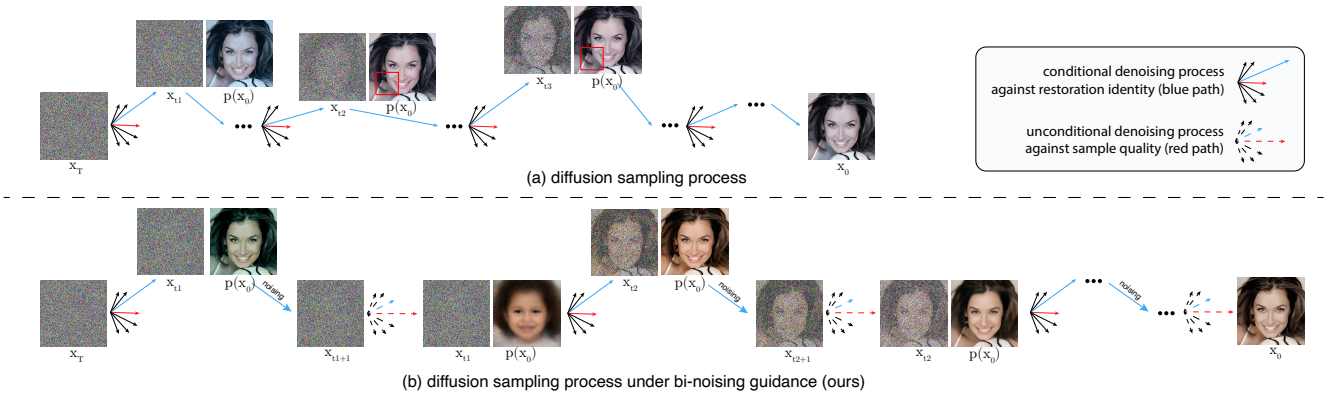

Figure 1. The graphical model showing the difference between the previous diffusion sampling process and ours with bi-noising guidance for colorization, where $\mathbf{x}_t$ is the noise of each diffusion process at timestep t, $p(\mathbf{x}_0)$ is the predicted noise-free start point of $\mathbf{x}_0$, and arrows indicate the denoising results of the diffusion models at each denoising process. Top figure shows how the conditional denoising process for colorization gradually accumulates the incorrect noise and results in artifacts. Instead, as shown on the bottom figure, the proposed additional noising and denoising steps diminish the incorrect noise and help in achieving better results.

In this work, we introduce a new method, named bi-noising diffusion, for utilizing rich priors encapsulated in the unconditional pretrained diffusion models. Inspired by implicit sampling that was first developed in the denoising diffusion implicit models [33] for acceleration, we show that the implicit sampling using an unconditional pretrained diffusion model has a capacity for correcting the divergence of distributions modeled by the conditional diffusion models. Specifically, we make a coarse implicit prediction at each intermediate diffusion time step by sampling from the conditional model. We then sample the prediction back to the intermediate step with the forward diffusion process. Finally, we make a refined prediction by utilizing an unconditional model. Fig. 1 visualizes the bi-noising procedure and the error by predicting the noiseless start-point image $p(\mathbf{x}_0)$ of the noise image $\mathbf{x}_t$. Using this two-step procedure, one can utilize the embedded rich priors learned by the unconditional model and produce better-quality images. This hypothesis is further validated through extensive experiments demonstrating that the introduced method performs favorably against state-of-the-art conditional diffusion models.

2. Related Work

Iterative methods. Finding the corresponding latent code [1, 2, 5, 8, 23] or sampled noise [4, 19, 22] of distorted images for restoration is one of the most straightforward ways of utilizing the generative priors. The intuition is that the pretrained generative models tend to produce natural results from their initial distribution. Thus the corresponding latent code or sampled noise can be projected to the restored images without additional optimization or learning. Menon et al. [23] proposed optimizing the latent code based on the difference between the generative results and the distorted images. Gu et al. [8] proposed to optimize multiple latent codes and compose them together for better visual quality. Similar iterative methods based on diffusion models have also been explored. Choi et al. [4] proposed to refine the sampled noise at each reverse diffusion step with the residual of distorted images. However, the applied stochastic iterative process tends to produce significantly different results though slightly changing its input. Therefore, these methods can only be applied to applications that do not require preserving the image identity.

Learning-based methods. Employing additional encoders [3, 29, 37, 42] to predict the latent code is another promising way that can bypass the stochastic optimization issues. However, such a method is incompatible with the diffusion models since it is impossible to encode the distribution of each reverse diffusion process for models that employ many sampling timesteps. Existing works learn to model conditional generative restoration [31, 38] instead. Richardson et al. [29] proposed to encode images with a ResNet backbone into an extended $\mathcal{W}+$ latent space, which defines upon features of each input layer of the generative networks. Wang et al. [37] proposed to encode images with a U-Net backbone and modulate the features of each generative layer of the generative networks. Saharia et al. [31] proposed to learn the noise distribution with the distorted image as the condition. Whang et al. [38] proposed to learn the generative process of residual given restored images. Compared with these GAN-based learning methods, a large number of sampling timesteps significantly increases the complexity of designing the corresponding encoders and thus makes the priors difficult to be learned.

Classifier guidance. Diffusion models have been using class information heavily to perform truncated or low-temperature sampling to increase the sample quality. The initial attempt [6, 26, 32, 35] is to incorporate a pre-trained

classifier by using its gradients to guide the diffusion sampling process. However, it complicates the diffusion model because additional training is required for the classifier on noisy data. Classifier-free guidance [10, 36] is another approach for addressing the complexity issue. It alleviates the complexity by combining the existing network with the classifier for guidance, *e.g.*, Ho et al. [10] use conditional diffusion network with an empty condition, and Wang et al. [36] use pretrained segmentation with a null label. Nevertheless, the classifier fails at natural images, and its gradient is meaningless for restoration. Its strength parameters also become less reasonable for the almost definite restoration sampling process. In this paper, we are interested in incorporating the sampling quality superiority of the empty condition and the sampling guidance ability of degraded images. We show that the empty condition can bring the incorrect noisy image back into the high-quality manifold. Compared with the classifier and classifier-free guidance, our binoising guided diffusion process keeps the same complexity but better fits the restoration task.

3. Proposed Method

In this section, we discuss the proposed mechanism to add the embedded priors to diffusion models. For consistency, we denote the intermediate output of the unconditional diffusion model as $\boldsymbol{\epsilon_\theta}(\cdot)$, parameterized by θ in the upcoming discussions following Denoising Diffusion Probabilistic Models [9] (DDPM). The additional, conditional diffusion model is denoted by $\boldsymbol{f_\phi}(\cdot)$, the condition (*i.e.* degraded images) and natural image pairs are denoted by $\{\mathbf{x}_0, \mathbf{y}_0\}$, where the conditional diffusion model $\boldsymbol{f_\phi}(\cdot)$ with parameters ϕ denoises noisy image $\mathbf{x}_t$ at timestep t with the concatenated condition $\mathbf{y}_0$.

3.1. Preliminaries

Diffusion probabilistic models belong to a new family of generative models [6,9,27,32,34] that can effectively model intractable distributions [32]. A diffusion process consists of two parts, *i.e.*, the forward process and the reverse diffusion process. In the forward diffusion process, a clean image is sampled from its data distribution and destroyed in T timesteps by repetitive noising using Gaussians of very small variances. Specifically, the forward process can be formulated as

$$\begin{aligned} q(\mathbf{y}_t|\mathbf{y}_{t-1}) &= \mathcal{N}\left(\mathbf{y}_t; \sqrt{\beta_t}\mathbf{y}_0, (1-\beta_t)\,\mathbf{I}\right) \\ &= \sqrt{\beta_t}\mathbf{y}_0 + \epsilon\sqrt{1-\beta_t}, \epsilon \sim \mathcal{N}(0, \mathbf{I}), \end{aligned} \tag{1}$$

or

$$\begin{aligned} q(\mathbf{y}_t|\mathbf{y}_0) &= \mathcal{N}\left(\mathbf{y}_t; \sqrt{\bar{\alpha}_t}\mathbf{y}_0, (1-\bar{\alpha}_t)\,\mathbf{I}\right) \\ &= \sqrt{\bar{\alpha}_t}\mathbf{y}_0 + \epsilon\sqrt{1-\bar{\alpha}_t}, \epsilon \sim \mathcal{N}(0, \mathbf{I}), \end{aligned} \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$ come from the variance schedule $\{\beta_1, \ldots, \beta_T\}$. The key idea here is that for large values of T, repetitive noising using Gaussians of small variances lead to a standard Gaussian, *i.e.*,

$$q(\mathbf{y}_T|\mathbf{y}_0) = \mathcal{N}\left(\mathbf{y}_T; 0, \mathbf{I}\right). \tag{3}$$

Now at each reverse timestep t, we attempt to reconstruct the noisy y_{t-1} from y_t using a distribution p modeled by a neural network with parameters θ. The parameters of the distribution $p_\theta(\cdot)$, found by optimizing variational lower bound of log-likelihood of $p_\theta(y_0)$, which is simplified by Ho et al. [9] by claiming that the major component in the objective comes from L_{t-1}, and the simplified loss is

$$L_{t-1} = E_{t\sim[1,T],\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\left\|\epsilon - \epsilon_\theta\left(\mathbf{y}_t, t\right)\right\|^2\right]. \tag{4}$$

Here network $\epsilon_\theta(\cdot)$ models the noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ at each timestep t with the denoised one $\mathbf{y}_t$, which can be seen as the process of learning the gradient of distributions with score matching according to Song et al. [34]. Therefore, we can learn the impressive perceptual synthesizing capacity with the simplified loss function between noises.

3.2. Learning to Refine Diffusion Process

In our experiments, we denote the recent diffusion models [31, 38] that learn the diffusion process with conditions as the way of Learning to Refine Diffusion Process (LRDP). LRDP models the conditional distribution of a clean image given a degraded image for restoration learning, and thus it requires separate training for different tasks or datasets. The objective for this learning process is formulated as

$$L_{\text{vlb}} := E_{t\sim[1,T],\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\left\|\epsilon - \epsilon_\theta\left(\mathbf{x}_0, \mathbf{y}_t, t\right)\right\|^2\right], \tag{5}$$

where $\mathbf{y}_t \sim \mathcal{N}(\mathbf{y}_t|\sqrt{\bar{\alpha}_t}\mathbf{y}_0, (1-\bar{\alpha}_t)\mathbf{I})$. The network architecture in LRDP is a slightly changed version from the original U-Net found in DDPM, and the additional input $\mathbf{x}_0$ and $\mathbf{y}_0$ are concatenated and passed to the input layer. Similarly, the reverse diffusion process of LRDP is slightly changed from the original one and formulated as

$$\mathbf{y}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_0, \mathbf{y}_t, t\right)\right) + \sqrt{1-\alpha_t}\mathbf{z}, \tag{6}$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ and $\alpha_t, \bar{\alpha}_t$ is the variant of the predefined variance schedule $\{\beta_1, \ldots, \beta_T\}$, that is $\alpha_t := 1-\beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. Since the diffusion process conditions on the specific type of degradation $p(\cdot)$ that produces degraded image $\mathbf{x}_0$ given clear image $\mathbf{y}_0$ as $p(\mathbf{x}_0|\mathbf{y}_0)$, LRDP needs re-training from scratch for different restoration tasks, which further heightens the training cost.

We experimentally find that such a protocol degrades the visual quality of the generation compared with the one without $\mathbf{x}_0$ conditioned. The most straightforward assumption towards the performance drop is that the assumed posterior $p(\mathbf{x}_0|\mathbf{y}_0)$ contrasts with the diffusion process $p_\theta(\mathbf{y}_0|\mathbf{x}_0) \propto p_\theta(\mathbf{y}_0)p_\theta(\mathbf{x}_0|\mathbf{y}_0)$ due to the ambiguous property of the degradation models. Thus, we claim that decomposing the diffusion generation process into different protocols should be a more promising way to handle restoration tasks.

3.3. Conditioning on Diffusion Process

The recent work [4, 19] falls into another category of utilizing the diffusion process, which uses a pretrained DDPM and changes its reverse diffusion process with distorted images by Conditioning on Diffusion Process (CDP). A similar way was previously explored in the other generative models, *e.g.*, mGANprior [8] and PULSE [23] invert a trained GAN by optimizing its latent code. However, CDP does not require optimization compared with the previously mentioned GAN-based methods. In contrast, it ensembles the conditions during sampling as

$$\hat{\mathbf{y}}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\mathbf{y}_t, t\right)\right) + \sqrt{1-\alpha_t}\mathbf{z} \tag{7}$$

$$\mathbf{y}_{t-1} = \hat{\mathbf{y}}_{t-1} + \boldsymbol{\sigma}(x_0, \hat{\mathbf{y}}_{t-1}), \tag{8}$$

where $\boldsymbol{\sigma}(\cdot)$ is a handcrafted transformation which aims at combining $\mathbf{x}_0$ with $\hat{\mathbf{y}}_{t-1}$ for accurate restoration. For example, Choi et al. [4] proposed to downsample $\mathbf{x}_0$ and $\hat{\mathbf{y}}_{t-1}$ and take their residual as the conditioning, while Lugmayr et al. [19] proposed to sum the visible region of $\mathbf{x}_0$ with the invisible region $\hat{\mathbf{y}}_{t-1}$ for the inpainting task.

Though CDP avoids the heavy training cost and is suitable for some conditional generation tasks like restoration with minimal modifications, its performance highly depends on the amount of degradation in the conditioned images. For example, when the conditioned images suffer from high amounts of distortion for face image restoration, CDP cannot preserve the face identity and tends to generate pseudo-sharp results with fake details. These fake details introduce further ill-posedness to the restored images and greatly limit the applications of such methods. Therefore, we propose refining the denoised results for correcting such artifacts at each step.

3.4. Implicit Error-feedback Diffusion Priors

Since the diffusion models follow a time-sequential process, the error in each step and the visual artifacts propagate and add up, hence severely degrading the quality of some CDP results. However, such issues are rarely observed in the unconditional diffusion models. We argue that the difference comes from conditioning breaking the inherent probabilistic distribution of noises at each sampling timestep, causing them to deviate from the manifold of natural images. Therefore, we propose to apply generative priors embedded in a pretrained unconditional model to regularize the noise predicted at each timestep from the conditional model. The trained diffusion model with conditioning denoted as $\boldsymbol{f}_{\boldsymbol{\phi}}(\cdot)$ takes as input the predicted image of the previous timestep and makes an implicit prediction $\tilde{\mathbf{y}}_0$ defined by

$$\tilde{\mathbf{y}}_0 = (\mathbf{y}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{f}_{\boldsymbol{\phi}}(\mathbf{x}_0, \mathbf{y}_t, t))/\sqrt{\bar{\alpha}_t}. \tag{9}$$

Here y_t denotes the prediction at the previous timestep. We then estimate the noisy version of the implicit prediction, which undergoes further regularization from an unconditional diffusion model. Please note that the unconditional diffusion model that fits the inherent probabilistic distribution. The diffusion process $\mathbf{y}_t \sim q(\mathbf{y}_t|\tilde{\mathbf{y}}_0)$ with $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\cdot)$ is formulated as

$$\begin{aligned} q(\mathbf{y}_t|\tilde{\mathbf{y}}_0) &:= \mathcal{N}(\mathbf{y}_t|\sqrt{\bar{\alpha}_t}\tilde{\mathbf{y}}_0, (1-\bar{\alpha}_t)\mathbf{I}) \\ &= \sqrt{\bar{\alpha}_t}\tilde{\mathbf{y}}_0 + \epsilon\sqrt{1-\bar{\alpha}_t}, \epsilon \sim \mathcal{N}(0, \mathbf{I}), \end{aligned} \tag{10}$$

and

$$\mathbf{y}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\left(\mathbf{y}_t, t\right)\right) + \sigma_t\mathbf{z}. \tag{11}$$

Following this procedure brings in an inherent regularization to the output of the conditional model during the reverse diffusion process. Note that Equation (11) takes the noised version $\mathbf{y}_t$ sampled from $\tilde{\mathbf{y}}_0$ as input. It is similar to the original reverse diffusion process, which takes the noised version of natural images as input.

In summary, we utilize two diffusion models for conditional image generation. The unconditional diffusion model regularizes the predicted outliers at each prediction timestep of the conditional diffusion model in an error-feedback way. Moreover, for the complex real-world application like draining where domain gaps may exist, we further discuss the details of applying our bi-noising diffusion with slight modifications to achieve better performance.

3.5. Complex Application: Deraining

Here we discuss one of the applications where we apply our introduced bi-noising diffusion for further clarification. The diffusion model is trained for the task of deraining using the rainy image as a condition for generating rain-free results. Motivated by the power of diffusion models to learn the distribution of clean natural images, we train a diffusion model to learn the distribution of rain and, at the same time, make the model aware of the distribution of real-world rain-free images. For this, we first train a diffusion model for image generation using the ImageNet dataset, and we then train another diffusion model by ensuring that the weights of the model trained for deraining are aligned to the weights learned for real-world rain-free images. Let W_r denote the weights of the first model and θ denote the weights of the diffusion decoder estimated after each iteration through backpropagation. Then the weights of the second model are updated after each iteration of training according to,

$$W' = \alpha.\theta + (1-\alpha).W_r, \tag{12}$$

where α denotes the rate of Exponential Moving Average (EMA) for updating the decoder weights. The encoder weights are updated as such.

One observation from our experiments on image deraining while training by direct conditioning like in SR3 [31] was that the restored images suffered from artifacts and color channel shift which can be seen in Fig. 4. On further investigation, we found that this is due to incorrect condi-

Figure 2. **Colorization visual result comparisons corresponding to the CelebAHQ dataset.**

tioning of input during the training process. Specifically, for the task of image restoration with source-target pairs denoted as $(\mathbf{x}_0, \mathbf{y}_0)$, existing methods optimize the weights of the network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\cdot)$ modelling the reverse process of diffusion, by minimizing the L_{simple} function defined in [9] as

$$L_{\text{simple}} := E_{t\sim[1,T],\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\epsilon - \epsilon_\theta\left(\mathbf{x}_0, \mathbf{y}_t, t\right)\|^2\right]. \quad (13)$$

The training objective L_{simple} holds the inherent assumption that during inference time $\mathbf{y}_t$, (i.e. the reconstructed image at time t) is close to the clean target. But for extreme cases where the intermediate diffusion outputs are not accurate during the initial steps of diffusion, the rain streaks continue to propagate through the diffusion process, as can be seen in Fig. 4. This is because, inherently, the diffusion model works by predicting the noise present in $\mathbf{y}_t$ than the amount of degradation in it. To account for this, we add a correction prior L_{corr} so that the network can give equally good output for high distortion levels. This term is defined by,

$$L_{\text{corr}} := \alpha_t \left\|\epsilon_\theta\left(\mathbf{x}_t, \mathbf{x}_0, t\right) - \epsilon_\theta\left(\mathbf{y}_t, \mathbf{x}_0, t\right)\right\|^2. \quad (14)$$

The final objective for training the network is,

$$L_{\text{final}} = L_{\text{simple}} + \lambda_{corr} L_{\text{corr}}. \quad (15)$$

The value of λ_{corr} is empirically set equal to 0.001 for all experiments.

4. Experiments

To demonstrate the restoration capacity of our method, we evaluate our method with several experimental settings following the most representative diffusion models, *i.e.*, ILVR [4] and SR3 [31] based on the Guided-diffusion architecture [26]. Following the common practice that pixel-wise metrics, *i.e.*, PSNR and SSIM cannot comprehensively denote the visual quality of restored results, we utilize FID and LPIPS as the additional metrics for evaluation. The tasks in which we evaluate our method on are

- Conditional image restoration which is trained on the FFHQ [30] dataset (70000 images) and evaluated on the CelebA-HQ [11, 18] dataset (first 3000 images) with a resolution of 256×256 pixels.
- Conditional image restoration which is $4\times$ face super-resolution trained on the FFHQ [30] dataset and evaluated on the CelebA-HQ [11, 18] dataset (first 3000 images) with a resolution of 256×256 pixels.
- Image turbulence removal follows the turbulence simulation settings [25] on the FFHQ dataset and conducts evaluation on the real long-range imaging images [24].
- Image deraining which is conducted on the Rain800 [41] dataset and Jorder 200L [39] dataset with their respective train sets. The diffusion models conduct in a resolution of 256×256 pixels.

Note that for the first three tasks, the diffusion models are trained on the FFHQ dataset for face generation. For the last task, the diffusion models are trained on the ImageNet dataset for natural image generation. The unconditional model utilized has never seen the validation dataset during its training process for all of these cases.

4.1. Colorization

Colorization aims at reconstructing grayscale images with colors that are fitted to natural statistics and image semantics. The grayscale image is obtained by averaging the values at red, green, and blue channels of the corresponding colour image. We empirically observed that conditional denoising diffusion models fail at colorization. Even though they can preserve the fine-grained details, unnatural colors always exist in their reconstructed results. In contrast, the method that adopts our proposed bi-noising diffusion is capable of correcting the reconstruction with more semantics and accurate color descriptions. The quantitative performance comparison is shown in Tab. 1, where our method achieves 7.906dB higher PSNR than the one without pretraining. The visual results in Fig. 2 further clarify the improvements that come from more globally consistent colors and tones of our results, even though the pretraining had never seen the ground truth before. In contrast, a similar method, i.e., ILVR cannot deal with the colorization task even though it also utilizes a pretrained unconditional model, which demonstrates the superiority of our proposed DDRP in such tasks. Therefore, we argue that utilizing the priors plays a crucial role in ensuring the color naturalism.

Method	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
ILVR Diffusion [4]	18.3936	0.5674	86.2642	0.5008
SR3 Diffusion [31]	*19.1647*	*0.8680*	*13.8126*	*0.2959*
Bi-Noising (Ours)	**27.0707**	**0.9531**	**12.6796**	**0.1417**

Table 1. **Colorization results corresponding to the CelebAHQ dataset.** The best and second-best performace is indicated with **blod** and *italic* respectively. We use ↑ and ↓ to suggest high/lower score should be achieved by better methods.

4.2. Face Super-resolution

Face super-resolution is the other representative task in image restoration, and it is widely evaluated in the other denoising diffusion-based restoration works. We follow the experimental settings of SR3 and ILVR, i.e., restore 256×256 face images from 64×64 face images downsampled by Bicubic interpolation. The implementation details of PULSE [23], ILVR [4], SR3 [31] are presented in the supplementary file. From Fig. 3, one can notice that our method achieves the best visual quality compared with the other methods. Compared with the state-of-the-art face super-resolution method based on GAN priors, our method better preserves the identity of the restored face images. As can be seen from Tab. 2, our method significantly outperforms the other methods in terms of the distortion measures, i.e., PSNR and SSIM with 4.8316 dB and 0.04 better than the second one. Though our results in the FID metric are not better than ILVR, FID doesn't denote the reconstruction accuracy that is curcial for super-resolution. Therefore, the above results demonstrate our performance superiority.

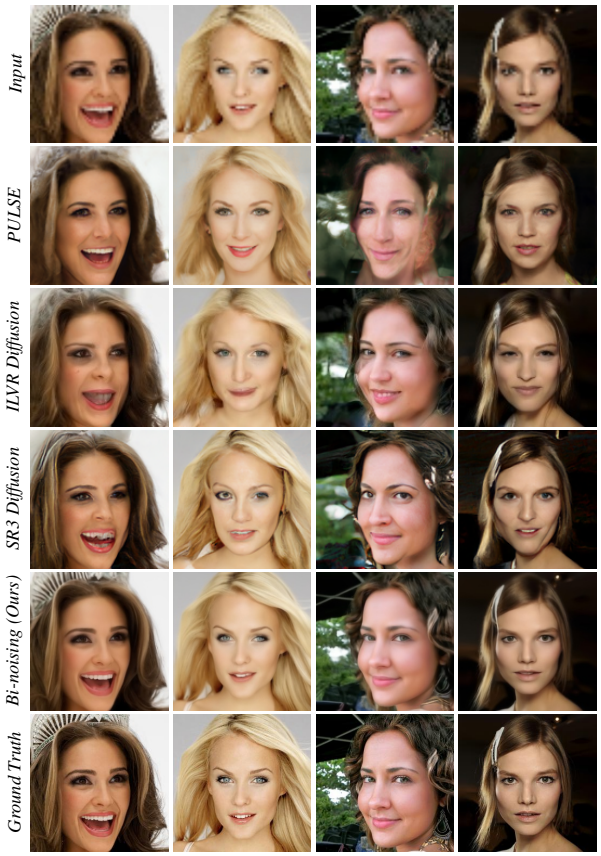

Figure 3. $4\times$ **super-resolution visual result comparisons corresponding to the CelebAHQ dataset.**

Method	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓
PULSE [23]	*23.5769*	*0.6794*	31.2309	0.3832
ILVR Diffusion [4]	22.5374	0.6150	**20.4621**	0.3393
SR3 Diffusion [31]	22.8290	0.6442	29.8932	*0.3350*
Bi-Noising (Ours)	**29.3996**	**0.8414**	*24.5632*	**0.1809**

Table 2. $4\times$ **super-resolution results corresponding to the CelebAHQ dataset.**

4.3. Image Deraining

We perform single image deraining on two popular draining datasets. Namely, the Jorder 200L dataset which contains large rain streaks, and the Rain800 dataset which contains realistic rain. Since no diffusion-based deraining method has been proposed in literature before, we perform comparisons after retraining the models proposed for super-resolution in the literature. Specifically, we perform comparisons with ILVR diffusion [4] and conditional diffusion models, and we include the improvements brought about by our modules. To evaluate the reconstruction quality, we use the PSNR and SSIM metrics. To assess the quality of

Figure 4. **Deraining visual result comparisons corresponding to the Rain 800 dataset.**

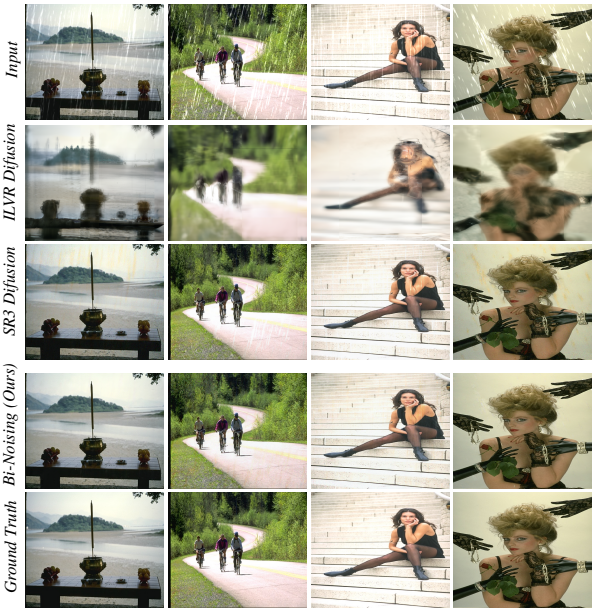

Figure 5. **Deraining visual result comparisons corresponding to the JORDER-200 dataset.**

images produced by various methods, we use LPIPS and NIQE as metrics. As we can see from Tab. 3, the proposed conditioning loss functions bring significant improvement for all metrics in the JORDER 200L dataset [39], obtaining about 2.45 dB PSNR over the exiting method as well as giving realistic natural images. The visual comparisons in Fig. 4 and Fig. 5 further demonstrate our method on the visual quality compared with the other methods.

Method	Jorder 200L dataset			
	PSNR ↑	SSIM ↑	LPIPS ↓	NIQE ↓
Rain Images (Input)	26.70	0.8439	0.2411	4.131
ILVR Diffusion [4]	21.22	0.4942	**0.0972**	6.467
SR3 Diffusion [31]	*31.45*	*0.9091*	*0.1779*	*3.588*
Bi-Noising (Ours)	**33.90**	**0.9555**	**0.0972**	**3.232**

Table 3. **Restoration results comparison on the Jorder 200L dataset with the other re-trained diffusion models.**

4.4. Turbulence Removal

We plug the proposed bi-noising approach into the recent diffusion restoration work [25] to demonstrate the applicability of our method on an extremely ill-posed atmospheric turbulence mitigation problem. Compared with the diffusion network with single noise conditioning, the results shown in Fig. 7 validate that our bi-noising method is able to remove the unnatural textures from the face images resulting from the incorrect denoising results.

4.5. Design Analysis

Nonparametric v.s. Parametric Priors. Inspired by Ho et al. [10], here we analysis the effect of alleviating complexity by parameterizing unconditional models into the conditional restoration model, denoted as *Nonparametric Prior* in Tab. 4. The compared methods use the same diffusion model but different guidance settings as the prior for fair comparisons. We showcase their performance and efficiency difference in the colorization task that was conducted using a single NVIDIA A6000 GPU. Specifically,

Settings	Methods				Ours		
	Ho et al. [9]	Dhariwal et al. [6]	Nichol et al. [26]	Ho et al. [10]	w/o parametric	w/o full guidance	Bi-Noising
classifier guidance [6]		✓					
CLIP guidance [26]			✓				
classifier-free guidance [10]				✓	✓		
alternative guidance						✓	
Bi-Noising					✓	✓	✓
PSNR ↑	19.16	20.10	23.14	25.91	26.46	*26.81*	**27.07**
Parameters (M) ↓	**93.6**	147.7	243.2	**93.6**	**93.6**	187.2	187.2
Running Time (s) ↓	**1.6**	4.9	3.4	3.1	3.1	*2.3*	3.1

Table 4. **Result comparisons between different prior parameterizations.**

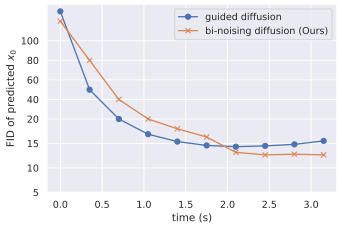

Figure 6. **FID v.s. Time.**

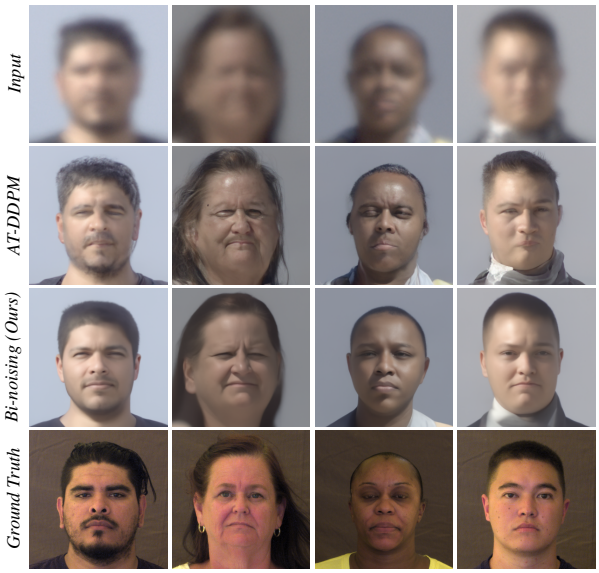

Figure 7. **Atmospheric turbulence mitigation results corresponding to the LRFID dataset [25].**

we use a single diffusion model that takes conditions for restoration, and it takes a null token ∅ for unconditional generation. From Tab. 4, we can conclude that the non-parametric prior, *i.e.*, w/o parametric, significantly reduces half of parameters for diffusion sampling, while the model suffers 0.26 dB performance drop compared with the parametric prior, *i.e.*, Bi-Noising, that is our final setting. The reason is that the null token increases the diffusion model training difficulty and thus the model fits worse than the unconditional model used in our final setting. Compared with other concurrent works that utilize classifier guidance [6] and clip guidance [26], our method outperforms them significantly with a slight increase in the number of parameters and running time. This clearly demonstrates the benefits of our parametric prior that can encapsulate the low-level information distribution for restoration.

Fig. 6 further demonstrates the efficiency of our method that achieves better FID score after 2.0s denoising process.

Priors Correlation. For the complex applications like deraining, we introduce additional correlation priors to further boost the final results. In Tab. 5, we present the ablation study of the introduced correlation priors to demonstrate its effectiveness. Since the rain streaks in the JORDER 200L dataset are relatively small, we choose the Rain800 dataset

Method	PSNR ↑	SSIM ↑	LPIPS ↓	NIQE ↓
Rain Images (Input)	26.70	0.8439	0.2411	4.131
SR3 Diffusion [31]	31.45	0.9091	0.1779	3.588
Ours	*33.23*	*0.9505*	*0.1043*	*3.285*
Ours + L_{corr}	**33.90**	**0.9555**	**0.0972**	**3.232**

Table 5. **Ablation study on the improvements brought by each introduced component.**

Figure 8. **Deraining visual result comparisons that demonstrate the improvement brought by our L_{corr} component.**

for a fair experiment. The ablation starts with the base model and then adds the two priors one by one to show the improvements. From the improved results due to L_{corr}, we can conclude that the introduced correlation priors allow our diffusion priors to better fit the probabilistic distribution of complex images, which ultimately benefits conditional generation with more realistic results. The comparisons presented in Fig. 8 also visually validate the conclusion.

5. Conclusion

We explored ways in which one can utilize denoising diffusion probability model priors for improving image enhancement and restoration tasks. The proposed way of integrating the stochastic priors into the deterministic conditioning denoising diffusion restoration model showed its superiority in colorization, face super-resolution, natural image super-resolution, and deraining tasks. Compared with similar denoising diffusion-based restoration methods, the restored results of our introduced method achieved better color consistency and contain more fine-grained details.

6. Acknowledgement

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The US. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, 2019. 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, 2020. 2

[3] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, 2021. 2

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *ICCV*, 2021. 2, 4, 5, 6, 7, 12

[5] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE TNNLS*, 2018. 2

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 1, 2, 3, 8

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 1

[8] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *CVPR*, 2020. 2, 4

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1, 3, 5, 8

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 7, 8

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 1

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 1

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[15] Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011. 1

[16] Chun Pong Lau, Hossein Souri, and Rama Chellappa. Atfacegan: Single face image restoration and recognition from atmospheric turbulence. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 32–39. IEEE, 2020. 12

[17] 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 *CVPR*, 2017. 1

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[19] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. *arXiv preprint arXiv:2201.09865*, 2022. 2, 4

[20] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *ICCV*, 2009. 1

[21] Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *IEEE TIP*, 2007. 1

[22] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 2

[23] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR*, 2020. 1, 2, 4, 6

[24] Kevin J Miller, Bradley Preece, Todd W Du Bosq, and Kevin R Leonard. A data-constrained algorithm for the emulation of long-range turbulence-degraded video. In *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXX*, volume 11001, pages 204–214. SPIE, 2019. 5, 12

[25] Nithin Gopalakrishnan Nair, Kangfu Mei, and Vishal M Patel. At-ddpm: Restoring faces degraded by atmospheric turbulence using denoising diffusion probabilistic models. In *WACV*, 2022. 5, 7, 8

[26] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2, 5, 8

[27] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 3

[28] 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. 1

[29] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021. 1, 2

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5

[31] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[32] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1, 2, 3

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[34] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 2019. 3

[35] 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. 2

[36] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 3

[37] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, 2021. 1, 2

[38] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. *arXiv preprint arXiv:2112.02475*, 2021. 2, 3

[39] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *CVPR*, 2017. 5, 7

[40] Rajeev Yasarla and Vishal M Patel. Learning to restore images degraded by atmospheric turbulence using uncertainty. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1694–1698. IEEE, 2021. 12

[41] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *IEEE TCSVT*, 2019. 5

[42] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016. 2

A. Demo

In order to provide a straightforward overview of our method, we have provided an online colorization demo and compared our bi-denoising process with the *naive-diffusion* by visualizing their intermediate predicted $\mathbf{x}_0$. The online demo is accessible at http://bi-noising.demohub.cc.

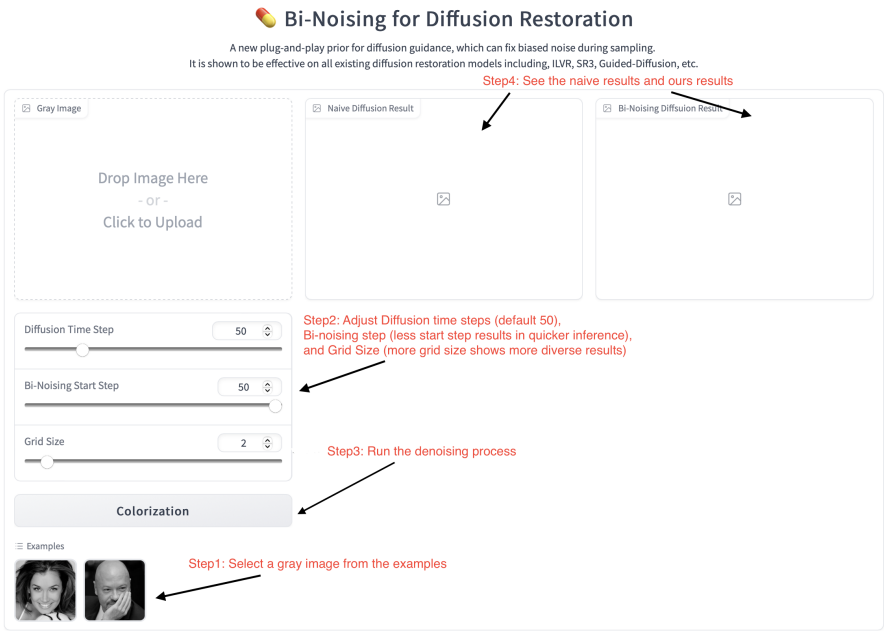

Figure 9. Screenshot of our online demo.

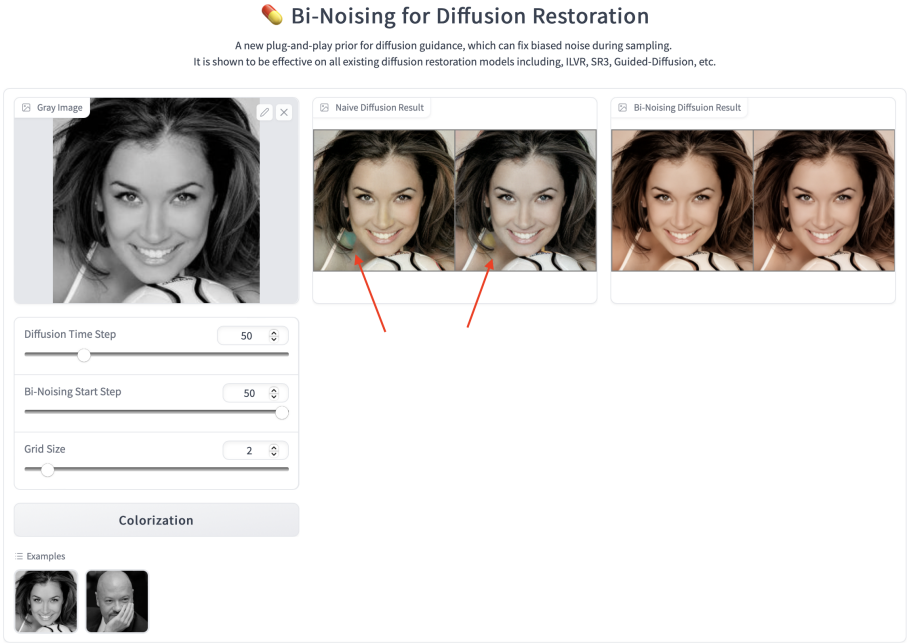

Figure 10. Screenshot of our online demo that shows artifacts in naive diffusion results, highlighted with red arrow.

B. Manifold Correction using Diffusion Priors

An alternate advantage of the proposed bi-noising diffusion is the effect of manifold correction. Consider any general restoration model$f(.))$ which maps from a domain $\mathcal{B}$ of the degraded image to a domain $\mathcal{A}$ of all natural images. The desired mapping function of this model for any restoration task is to learn the mapping

$$b \in \mathcal{B} \xRightarrow{h(.)} a \in \mathcal{A} \tag{16}$$

But in real scenarios, because the restoration task is ill-posed, rather than learning the natural image manifold, a deep network learns an inverse function that merely removes the degradation effect. Let this manifold be denoted by $\mathcal{C}$. The mapping function hence learned is

$$b \in \mathcal{B} \xRightarrow{f(.)} a \in \mathcal{C} \tag{17}$$

For example, for a restoration model trained for the task of face super-resolution, for an input (b), the network could create an output (c) that is the image of a disoriented face rather than an image in the manifold of faces. Theoretically, if we utilize any generative model for the restoration, the model should be able to achieve the ideal mapping. But often, the model learns the more complex problem of removing the degradation than learning to map to the domain of natural images. This is because it is difficult to reach the solution corresponding to the global optimum. In any generic restoration method, this deviation from the natural manifold can be corrected by adding a correction network that learns the mapping from domain $\mathcal{C}$ to $\mathcal{A}$. Unlike all other models, diffusion models contain a flexible model structure where intermediate latent variables can be accessed. This enables a manifold correction during inference time alone with explictly training a network to map from the generated manifold $\mathcal{C}$ to the natural manifold $\mathcal{A}$. Hence in our work, we exploit this property and perform the manifold correction to the domain of natural images through an additional step that utilizing an unconditional model. Consider a CDP trained for any restoration task denoted by $f_\phi(c_t, b, t)$. During inference, the restored sample c_T is generated through the cascade of steps

$$f_\phi(c_0, b, t) \rightarrow f_\phi(c_1, b, t), ... \rightarrow f_\phi(c_T, b, t) \tag{18}$$

or equivalently,

$$c_0 \xrightarrow{f_\phi(.)} c_1, ... \xrightarrow{f_\phi(.)} c_T \tag{19}$$

Here, $c_1, ..., c_T$ denotes the intermediate diffusion outputs of an image $c_T$in the manifold $\mathcal{C}$ that can be reconstructed from a degraded b. As mentioned before, the function f will not always map to the domain $\mathcal{A}$ of natural images. Hence we add an unconditional model $g_\theta(.)$ that does the task of aligning the manifold of the generated image to the manifold of natural images. The sequence of operations is as follows

$$f_\phi(c_0, b, t) \rightarrow g_\theta(c_1, t), ...f_\phi(a_{T-1}, b, t) \rightarrow g_\theta(c_T, t) \tag{20}$$

$$c_0 \xrightarrow{f_\phi(.)} c_1 \xrightarrow{g_\theta(.)} a_1 \xrightarrow{f_\phi(.)} c_2, ... \xrightarrow{f_\phi(.)} c_T \xrightarrow{g_\theta(.)} a_T \tag{21}$$

$a_1, ..., a_T$ denotes the intermediate diffusion outputs of an image a_T in manifold $\mathcal{A}$.

C. Turbulence Removal

Here we provide the quantitative evaluation of our method on the turbulence removal benchmark LRFID dataset. Compared with the other methods, ours not only has achieved better performance in the sample quality in terms of LPIPS but also better fidelity in terms of face recognition accuracy Top-1 and Top-3.

Dataset	LRFID dataset [24]		
Metric	LPIPS(↓)	Top-1(↑)	Top-3(↑)
degraded	0.6293	35.3	62.2
	CNN based models		
MPRNET [40]	0.5755	34.1	64.6
ATNet [40]	0.6128	36.5	64.6
	GAN based models		
ATFaceGAN [16]	0.6300	*47.5*	*65.8*
	Diffusion models		
ILVR Diffusion [4]	*0.5661*	31.7	59.7
Bi-Noising Diffusion (Ours)	**0.5500**	**48.7**	**73.1**

Table 6. Quantitative results on on real world turbulence degraded datasets: LRFID dataset [24]