Restabilizing Diffusion Models with Predictive Noise Fusion Strategy for Image Super-Resolution

Luoqian Jiang[1*], Yong Guo[1*], Bingna Xu[1], Haolin Pan[1], Jiezhang Cao[2], Wenbo Li[3], Jian Chen[1†]
[1]South China University of Technology
[2]Harvard University
[3]The Chinese University of Hong Kong
seluoqianjiang@mail.scut.edu.cn, guoyongcs@gmail.com, sexbn@mail.scut.edu.cn, mr.haolinpan@qq.com, jcao16@bwh.harvard.edu, fenglinglwb@gmail.com, ellachen@scut.edu.cn

Abstract

Diffusion models are prominent in image generation for producing detailed and realistic images from Gaussian noises. However, they often encounter instability issues in image restoration tasks, e.g., super-resolution. Existing methods typically rely on multiple runs to find an initial noise that produces a reasonably restored image. Unfortunately, these methods are computationally expensive and time-consuming without guaranteeing stable and consistent performance. To address these challenges, we propose a novel Predictive Noise Fusion Strategy (PNFS) that predicts pixel-wise errors in the restored image and combines different noises to generate a more effective noise. Extensive experiments show that PNFS significantly improves the stability and performance of diffusion models in super-resolution, both quantitatively and qualitatively. Furthermore, PNFS can be flexibly integrated into various diffusion models to enhance their stability.

Code — https://github.com/Rosiekk/PNFS-main

Introduction

Diffusion models (DMs) have significantly advanced generative modeling in recent years (Song and Ermon 2019; Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Dhariwal and Nichol 2021; Nichol and Dhariwal 2021; Li et al. 2022). Unlike generative adversarial networks (GANs) (Goodfellow et al. 2020), they avoid mode collapse and training instability while offering more compact latent representations that mitigate blurry in variational autoencoders (VAEs) (Kingma and Welling 2013). Their robust generative power has been demonstrated in tasks like image super-resolution (SR) (Batzolis et al. 2021; Rombach et al. 2022; Saharia et al. 2022b; Sun et al. 2023), image inpainting (Chung, Sim, and Ye 2022; Esser et al. 2021; Jing et al. 2022), and image editing (Avrahami, Lischinski, and Fried 2022; Choi et al. 2021; Meng et al. 2021).

Diffusion models face unique challenges in SR task. SR is an ill-posed inverse problem (Wang, Chen, and Hoi 2020), where a single low-resolution (LR) image can correspond to multiple high-resolution (HR) counterparts, yet the content of each restored HR image must keep consistency with the LR image. Current DM-based SR approaches leverage LR images as conditions to restore HR images from Gaussian noise. This randomness enhances diversity but introduces uncertainty detrimental to SR. Most traditional methods tend to use DDPM (Ho, Jain, and Abbeel 2020) sampler. However, DDPM conducts multiple steps e.g. 1000, and becomes very expensive. To address this, DDIM (Song, Meng, and Ermon 2020) is proposed to accelerate by reducing steps from 1000 to 100 or 20. Nevertheless, DDIM requires a random initial noise which inevitably introduces randomness and instability. In practice, DDIM often produces unexpected artifacts. Thus, addressing the instability of DM-based SR methods becomes an important problem.

To investigate the performance of diffusion models under varied initial Gaussian noises, we compare SR results between SD-Upscaler, LDM, and our method on a single LR image from the Manga109 dataset. We conducted experiments with a set of 10 Gaussian noises. As shown in the Figure 1, SD-Upscaler and LDM exhibit significant performance fluctuations under different initial noises, with a PSNR difference of 4.42 dB between the best and worst results. The visual results corresponding to low PSNR values contain noticeable errors and artifacts. We obtained these low-quality results using a few random noises without deliberate selection, indicating that the instability in diffusion models is far from coincidental. Therefore, it is crucial to analyze the impact of Gaussian noise on diffusion models and to address the instability issues.

Research on addressing instability issues in diffusion models remains limited. Sun et al. (Sun et al. 2023) propose a non-uniform time-step learning strategy to retrain diffusion models, aiming to improve efficiency and stability. Ma et al. (Ma et al. 2023) stabilize the sampling process of pretrained diffusion models by solving diffusion partial differential equations using optimal boundary conditions. However, these methods have limitations. Retraining requires significant computational resources and may not be easily adaptable to different datasets. Solving diffusion partial differential equations is complex and does not guarantee the authenticity of HR images. These approaches primarily focus on tuning model structures or optimizing diffusion sampling strategies without adequately addressing the impact of

*Authors contributed equally.
†Corresponding author.

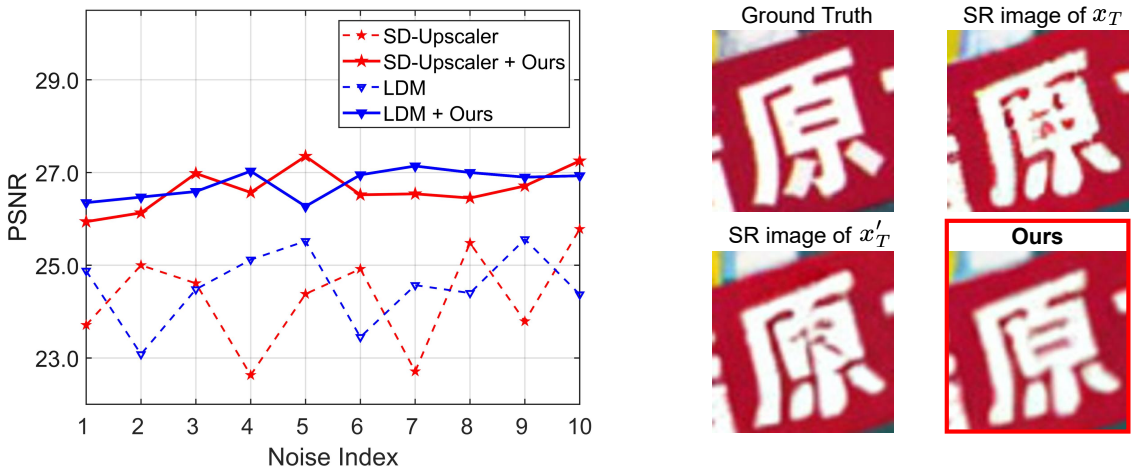

Figure 1: Comparisons of SR results between SD-Upscaler, LDM, and our PNFS with one same LR image. Left: PSNR results on different initial Gaussian noises. SD-Upscaler and LDM show significant fluctuations across different noises, while our PNFS significantly uplifts PSNR by 2.0-4.4 dB and ensures strong stability. Right: Visual comparisons between SD-Upscaler and PNFS based on two different initial Gaussian noises x_T and x'_T. PNFS produces realistic details without severe artifacts.

initial noise on model performance.

Unlike these methods, we analyze the relationship between initial noise and SR instability, leveraging noise diversity to stabilize DM-based SR methods. We propose a Predictive Noise Fusion Strategy (PNFS), comprising two modules, to predict and integrate optimal noise elements. Moreover, our PNFS can be easily integrated into various models to enhance restoration quality and stability.

Our key contributions are summarized below.

- We explore the instability of diffusion models in SR tasks and find that this instability is significantly related to the randomness of the initial Gaussian noise. Additionally, we observe that the noise elements affect the corresponding elements of the final restored images. By integrating the better-performing elements from these different noises, restoration performance can be improved.
- We propose the Predictive Noise Fusion Strategy (PNFS) to mitigate the instability of diffusion models by optimizing the initial noise. We highlight that, our PNFS can enhance DM-based SR methods in a plug-and-play manner.
- Experiments on multiple benchmark datasets demonstrate that, our PNFS yields significantly superior results both quantitatively and qualitatively, consistently producing high-quality images.

Related Work

Super-resolution (SR) aims to recover high-resolution (HR) images from low-resolution (LR) inputs. Diffusion models have recently emerged as a powerful generative paradigm, delivering impressive results in SR (Chung, Lee, and Ye 2022; Gao et al. 2023; Chen et al. 2024), image restoration (Guo et al. 2023; Wang et al. 2023c), image editing (Hertz et al. 2022; Brooks, Holynski, and Efros 2023; Kim, Kwon, and Ye 2022; Saharia et al. 2022a), and coloring (Yue et al. 2023; Lin et al. 2023; Croitoru et al. 2023). These methods fall into two categories. The first category involves retraining diffusion models conditioned on degraded images (Saharia et al. 2022b; Rombach et al. 2022; Niu et al. 2024; Bansal et al. 2024; Wang et al. 2023b). These methods diffuse a clear image into Gaussian noise and then recover it conditioned on the degraded input. For instance, StableSR (Wang et al. 2023a) uses a time-aware encoder and a feature wrapping module to maintain generated priors and adjust quality. LDM and SD-Upscaler (Rombach et al. 2022) operate in a lower-dimensional latent space, significantly reducing computational complexity while maintaining high-quality image synthesis, making it scalable for various generative tasks. The second category uses pre-trained diffusion models to restore degraded images (Wang et al. 2023a; Feng et al. 2023; Fei et al. 2023), avoiding the need for retraining and reducing computational overhead. Additionally, zero-shot methods (Chung et al. 2022; Rout et al. 2024; Fei et al. 2023; Cao et al. 2024) utilize pre-trained models as prior to restore LR images. In each iteration, they modify the intermediate outputs based on the degraded object or a trained classifier. However, modifications to intermediate outputs are uncontrollable and cause irreversible effects on subsequent sampling.

Stability of Diffusion Models. Current SR models encounter challenges in stability due to varying initial noises, affecting their real-world reliability. To address this, Sun et al. (Sun et al. 2023) introduce a non-uniform time step strategy for model retraining, while Ma et al. (Ma et al. 2023) optimize SR through differential equations. This characteristic is suppressed in certain scenarios of text-to-image tasks, which inspires enhancing stability in SR tasks. ControlNet (Zhang, Rao, and Agrawala 2023) and SpaText (Avrahami et al. 2023), developed for text-to-image tasks, offer advanced control and scene description capabilities. These innovations in controlling and stabilizing diffusion models could lead to more reliable SR performance.

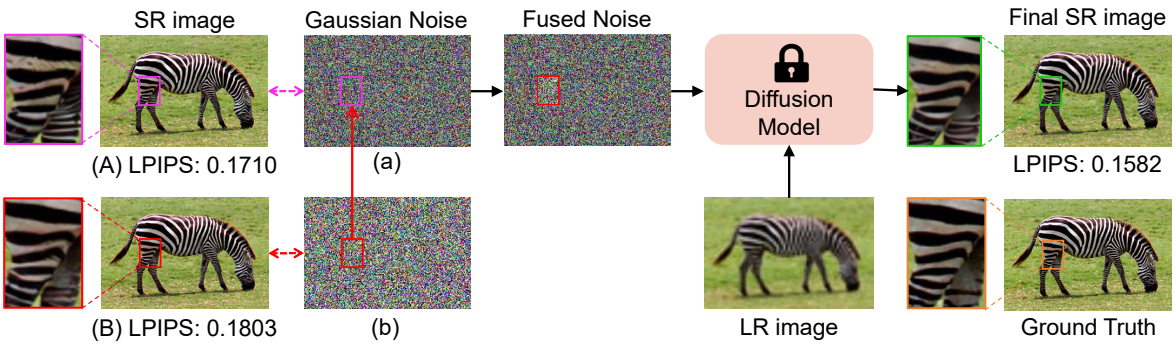

Figure 2: Super-resolution reconstruction results under random noise and fused noise. In the presence of random noise, certain regions suffer from poor reconstruction quality (e.g., the pink area in image A is worse than the red area in B). Fusing noise can enhance the visual quality of SR result.

Predictive Noise Fusion for Diffusion Models

In this paper, we observe that some regions in a restored image suffer from poor reconstruction quality. As shown in Figure 2, the pink area in image (A) is worse than the red area in (B). By replacing the noise in the pink area with the corresponding noise from the red area, the LPIPS value decreases by 0.012-0.022 compared to the original noises (a) and (b), effectively enhancing the visual quality. This finding inspired us to propose a novel Predictive Noise Fusion Strategy (PNFS), which optimizes the initial Gaussian noise for the pre-trained diffusion model to produce higher-quality restored images, significantly improving the stability of DM-based SR methods. The architecture of PNFS in Figure 3, includes the Discrepancy Prediction Module (DPM) and the Probabilistic Fusion Module (PFM). Specifically, DPM predicts pixel-level discrepancies between the restored and the ground truth images, while PFM uses a novel sampling strategy to reconstruct a new initial noise by selecting elements from various noises based on the predicted discrepancies.

Discrepancy Prediction Module

As illustrated in Figure 3, DPM receives the initial Gaussian noise and the LR image as inputs. It employs a lightweight Transformer to extract features and outputs a discrepancy matrix that reflects the quality of the noise. The prediction process is formulated as:

$$\widehat{D} = f(y, x_T), \tag{1}$$

where $f(\cdot)$ represents the prediction network. To train the network, we generate pseudo labels based on two key criteria: content fidelity and perceptual quality. First, we assess the content discrepancy of Y channel in YCbCr space between the restored HR image $\hat{x}_0$ and the ground truth HR image x_0. The process begins with generating an initial Gaussian noise $x_T \sim \mathcal{N}(0, I)$. For each run, the LR image and initial noise are fed into the pre-trained diffusion model through all T reverse steps to produce the restored HR image $\hat{x}_0$. The content discrepancy is then calculated as:

$$D_{content} = \text{AvgPool}(|x_0 - \hat{x}_0|), \tag{2}$$

where average pooling is applied to match the latent space dimensions. Next, we refer to the implementations of LPIPS (Zhang et al. 2018) and employ a feature extraction network across multiple scales. The features from each scale are resized and then summed to compute the perceptual difference between $\hat{x}0$ and x_0:

$$D_{percept} = \sum_s resize((\Phi_s(x_0) - \Phi_s(\hat{x}_0))^2), \tag{3}$$

where $\Phi_s(\cdot)$ denotes the feature extraction network at each scale s, and $resize(\cdot)$ aligns the feature dimensions before summing the differences. To construct the final pseudo label matrix D, we calculate a weighted combination of $D_{content}$ and $D_{percept}$, with weights α and β reflecting the importance of content fidelity and perceptual quality, respectively:

$$D = \alpha D_{content} + \beta D_{percept}, \tag{4}$$

where $\alpha + \beta = 1$. The training objective is formulated as:

$$\mathcal{L}_{DP} = \|D - \widehat{D}\|_2^2. \tag{5}$$

Upon completion of training, DPM can effectively predict the optimal noise elements for restoring each image pixel, providing essential prior information for the probabilistic fusion strategy described in the subsequent section.

Probabilistic Fusion Module

Given two different Gaussian noises x_T and x'_T as inputs, DPM produces discrepancy matrices $\widehat{D}$ and $\widehat{D}'$ using Eq. (1). A smaller discrepancy value indicates better noise quality, so the corresponding noise element should be more likely selected, resulting in a mask matrix:

$$M = \mathbb{1}\left(\widehat{D} \leq \widehat{D}'\right), \tag{6}$$

where the index of the smallest discrepancy value is retained. The deterministic selection of noise elements in M may lead to deviations in the fused noise from the standard Gaussian distribution. As shown in Figure 6, our experiments show that M induces shifts in the mean and variance, causing the fused noise x_M to stray from the expected Gaussian distribution $\mathcal{N}(0, I)$. In contrast, probabilistic sampling

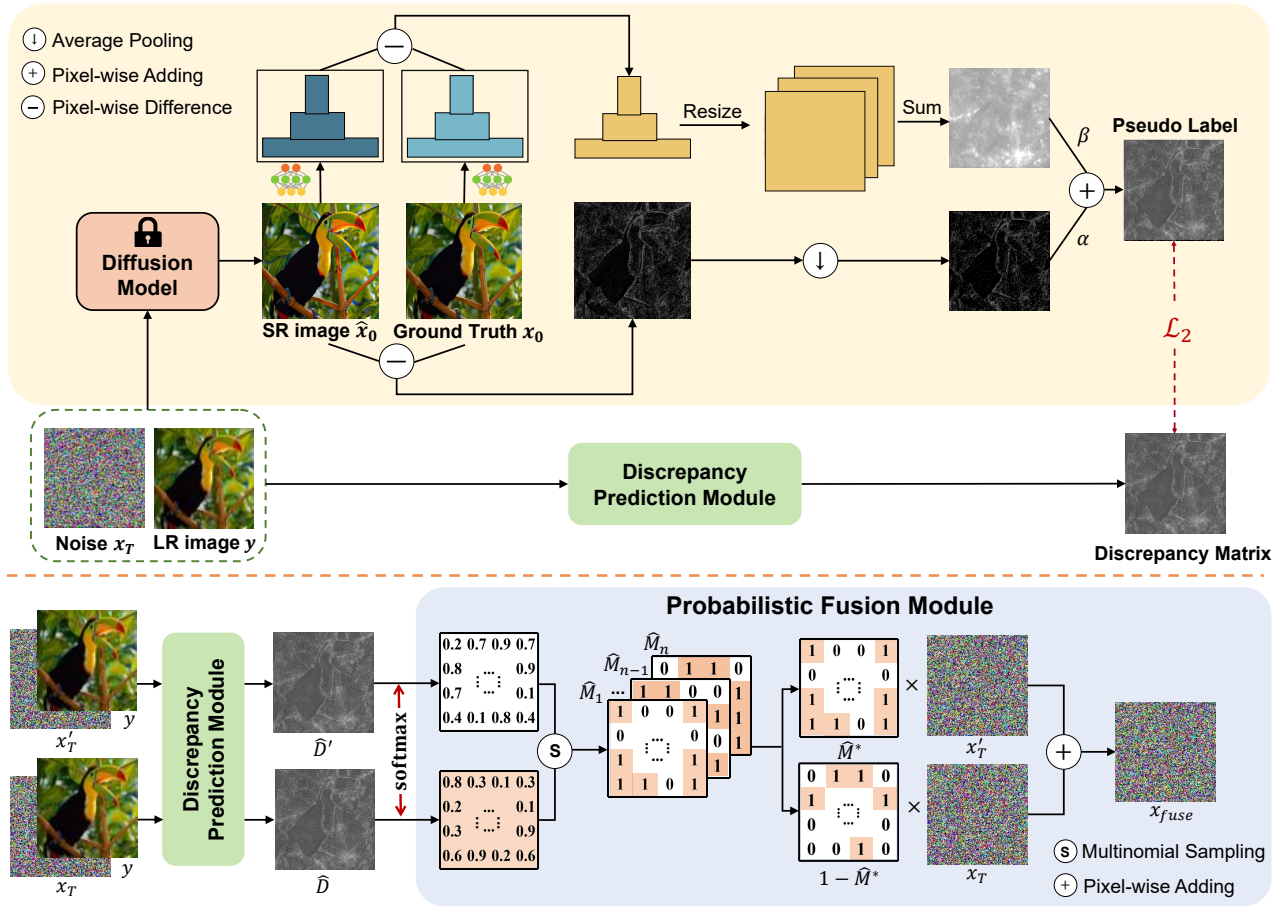

Figure 3: The training process of the Discrepancy Prediction Module (DPM) and the workflow of the Probabilistic Fusion Module (PFM). Top: Compute the pseudo labels based on fidelity and perceptual quality between the ground truth HR image and the restored image. DPM is trained to minimize the $\mathcal{L}_2$ loss between its predicted discrepancy matrix and the pseudo label matrix. Bottom: PFM workflow starts by inputting two initial noises into the DPM, producing corresponding discrepancy matrices. A probabilistic matrix is then computed for multinomial sampling, generating a series of fused mask matrices. Finally, the two initial noises are combined using the optimal fused mask $\widehat{M}^*$, resulting in a fused noise output.

preserves the adherence of the fused noise matrix x_{fuse} to the standard Gaussian distribution. According to the sampling criterion, each pixel of the whole Gaussian noise map is independently and identically distributed (i.i.d.). In other words, there is no correlation between any two noise pixels. Thus, the entire noise map that uses a binary mask to fuse pixels still follows the Gaussian distribution. We propose a probabilistic sampling strategy in Algorithm 1. Specifically, we obtain a probabilistic matrix using $\mathrm{softmax}(\cdot)$ function:

$$P = \mathrm{softmax}\left(\mathrm{concat}\left(\frac{1}{\widehat{D}}, \frac{1}{\widehat{D}'}\right)\right). \tag{7}$$

Based on the probability matrix P, we perform multinomial sampling to generate a fused mask matrix $\widehat{M}$, with each element indicating the index of the noise element with the highest probability at that position. This probabilistic matrix ensures a higher likelihood of selecting better noise elements (with lower discrepancies), leveraging the information in $\widehat{D}$ and $\widehat{D}'$. We repeat this sampling process n times to create a set of fused mask matrices $\{\widehat{M}_1, \widehat{M}_2, \ldots, \widehat{M}_n\}$. Then, we select five matrices that produce the fused noise with the smallest absolute mean value. This approach helps in reducing the potential fluctuations and conforming closely to the properties of a standard Gaussian distribution. These selected matrices are compared against the mask matrix M to calculate the abstract difference, ensuring that the noise remains well-behaved and aligns with the expectation of DPM. The final fused noise is computed as:

$$x_{fuse} = \widehat{M}^* \odot x_T + (1 - \widehat{M}^*) \odot x'_T. \tag{8}$$

Experiment

In this section, we outline the implementation details, followed by quantitative and qualitative comparisons with

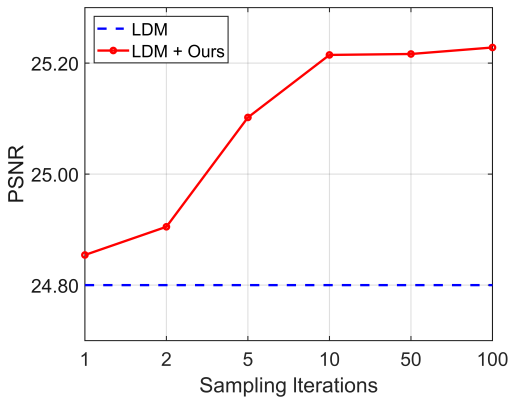

Figure 4: Ablation study on the number of sampling iterations to generate fused masks. Optimal performance is achieved with 10 iterations.

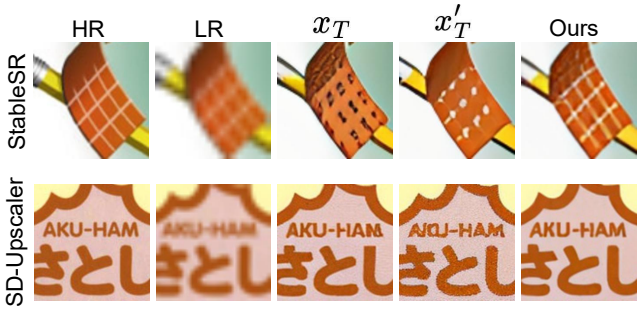

Figure 5: Visual comparisons for 4× image SR. PNFS achieve superior details and attractive visual quality. The SR results of x_T and x'_T often contain noticeable visual artifacts and incorrect details (zoom in for details).

state-of-the-art DM-based SR methods. Finally, we present ablation studies and discuss the results.

Implementation Details

All the experiments are conducted on a server with NVIDIA GeForce RTX 3090 24G GPU. DPM is trainable and trained for 160K iterations with a batch size of 100. We use the Adam optimizer with an initial learning rate of $6e^{-5}$, adjusted using a poly learning rate schedule with a default factor of 1.0. The training and validation sets consist of 900 and 100 images randomly selected from the DIV2K and Flickr2K (Agustsson and Timofte 2017) datasets. For each image, 100 Gaussian noises are sampled to generate 90,000 training samples. Random crops of size 512 × 512 are used for training, and evaluation is performed on full-size images.

During SR task inference, we freeze all parameters of the pre-trained diffusion model and the DPM, using DDIM (Song, Meng, and Ermon 2020) as the sampling strategy for all experiments. We evaluate different methods on well-known SR datasets, including Set5 (Bevilacqua et al. 2012), Set14 (Zeyde, Elad, and Protter 2010), BSD100 (Martin et al. 2001), Urban100 (Huang, Singh, and Ahuja 2015), Manga109 (Matsui et al. 2017) and RealSR (Cai et al. 2019). For the quantitative evaluation of SR task, we employ the widely used PSNR, SSIM (Wang et al. 2004), LPIPS (Zhang et al. 2018) and FID (Heusel et al. 2017) metrics to measure the fidelity and perceptual quality of the restored images. Additionally, we refer to local standard deviation (L-STD) metric (Sun et al. 2023) to evaluate the stability of diffusion models.

Algorithm 1: Scheme of the **Predictive Noise Fusion Strategy (PNFS)**.

Require: LR image y, random gaussian noise $x_T \sim \mathcal{N}(0, I)$ and $x'_T \sim \mathcal{N}(0, I)$, Discrepancy Prediction Module $f(\cdot)$, Empty set $\mathcal{S}$ and $\widetilde{\mathcal{S}}$.
Ensure: Fused noise x_{fuse}.

```
Get the discrepancy matrices with the Discrepancy Prediction Module as Eq. (1).
Compute the mask matrix M as Eq. (6).
Compute the probabilistic matrix P as Eq. (7).
Initialize S ← ∅, S̃ ← ∅.
for i = 1 to n do
    Conduct multinomial sampling based on P to generate a fused mask matrix M̂_i.
    Compute fused noise x_i = M̂_i ⊙ x_T + (1 − M̂_i) ⊙ x'_T.
    Let S ← S ∪ {(mean(x_i), M̂_i)}.
end for
Sort S by |mean(x_i)| to obtain the indices of the five smallest values.
Extract fused masks corresponding to the indices and put them in S̃.
Get optimal fused mask M̂* = arg min_{M̂_i ∈ S̃} |M̂_i − M|.
Obtain the final fused noise x_fuse = M̂* ⊙ x_T + (1 − M̂*) ⊙ x'_T.
```

Comparison with State-of-the-art Methods

We highlight that our proposed PNFS is generally applicable to existing DM-based SR methods. For comparison, we use advanced methods including LDM (Rombach et al. 2022), SD-Upscaler (Rombach et al. 2022) and StableSR (Wang et al. 2023a). The results are obtained using publicly released codes and models. Table 1 presents quantitative comparisons for 4× image SR. Our PNFS significantly surpasses the backbone methods on PSNR, SSIM, and L-STD across all datasets. For instance, PNFS improves PSNR by 1.37 dB on Set5 compared with SD-Upscaler, and by 0.52 dB on Manga109 compared with StableSR, validating its effectiveness across different backbone methods and datasets. Additionally, the low L-STD values in PNFS reflect the strong stability. For example, L-STD values for LDM vary from 0.031 to 0.306 on Set5 and Manga109, respectively. However, our PNFS consistently maintains L-STD values around 0.02, indicating greater consistency and stability under varying initial noises. As shown in Table 2, based on StableSR, our method obtains superior performance across multiple scaling factors (2×, 3×, and 4×) on RealSR. No-

Dataset	Method	LDM			SD-Upscaler			StableSR		
		PSNR↑	SSIM↑	L-STD↓	PSNR↑	SSIM↑	L-STD↓	PSNR↑	SSIM↑	L-STD↓
Set5	Baseline	26.67	0.777	0.031	25.05	0.723	0.208	24.28	0.712	0.208
	Ours	**27.06**	**0.791**	**0.024**	**26.42**	**0.741**	**0.028**	**24.81**	**0.727**	**0.048**
Set14	Baseline	24.80	0.664	0.038	24.60	0.628	0.235	24.22	0.666	0.232
	Ours	**25.20**	**0.681**	**0.028**	**25.87**	**0.688**	**0.026**	**24.54**	**0.674**	**0.033**
BSD100	Baseline	24.56	0.617	0.037	24.03	0.597	0.211	24.10	0.627	0.207
	Ours	**24.91**	**0.631**	**0.028**	**24.61**	**0.608**	**0.024**	**24.35**	**0.635**	**0.033**
Urban100	Baseline	22.76	0.672	0.230	23.41	0.679	0.229	22.74	0.689	0.226
	Ours	**23.24**	**0.693**	**0.033**	**23.60**	**0.686**	**0.024**	**22.93**	**0.692**	**0.031**
Manga109	Baseline	25.28	0.810	0.306	25.38	0.796	0.305	24.75	0.822	0.304
	Ours	**25.72**	**0.830**	**0.026**	**26.27**	**0.823**	**0.025**	**25.27**	**0.829**	**0.024**

Table 1: Quantitative comparison (PSNR/SSIM/L-STD) with state-of-the-art methods. Each method is sampled 20 times to report the average results based on the backbone methods. The bold black values indicate the best result. Our PNFS surpasses all the backbone methods on PSNR and SSIM metrics and achieves a significant improvement in the stability of all datasets.

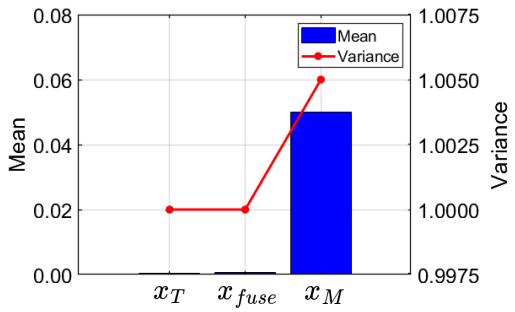

Figure 6: Statistical features of Gaussian noise. x_M denotes the fused noise with mask matrix M. x_M is deviated from the standard Gaussian distribution $\mathcal{N}(0, I)$.

Dataset	Method	PSNR↑	LPIPS↓	FID↓
RealSR 4×	BCs	22.13	0.299	–
	CCSR	25.86	0.294	126.1
	StableSR	24.69	0.309	127.2
	Ours	**26.39**	**0.292**	**120.1**
RealSR 3×	StableSR	27.16	0.325	108.4
	Ours	**27.37**	**0.322**	**104.7**
RealSR 2×	StableSR	27.45	0.288	87.7
	Ours	**27.69**	**0.284**	**83.0**

Table 2: Quantitative comparison on RealSR dataset.

tably, our method consistently improves the performance in terms of diverse metrics, surpassing BCs (Ma et al. 2023) and CCSR (Sun et al. 2023). Furthermore, PNFS introduces minimal computational overhead, with an inference time of just 0.03 seconds and a parameter size of only 13.69 MB, accounting for just 0.1% of the backbone method.

Qualitative Comparisons. Figure 5 and 7 presents the visual results of 4× image SR, demonstrating that our PNFS achieves superior detail and visual quality. For example, in the first row of Figure 5, our PNFS successfully recovers the grid lines, whereas StableSR produces completely incorrect shapes. Moreover, PNFS improves the structure preservation and clearness of the recovered images as shown in Figure 7, superior to the traditional SR method SwinIR (Liang et al. 2021). However, the other competing methods often contain noticeable visual artifacts and incorrect details.

Ablation Studies

Effect of the sampling iteration number n in the Probabilistic Fusion Module. Figure 4 illustrates how different sampling iteration numbers n affect the generation of fused mask matrices on Set14 using LDM as the backbone method. We observe that PNFS performance improves as the number of iterations increases, with stable and satisfactory results achieved after just a few iterations. When $n < 5$, random fluctuations may lead to deviations from the intended probability distribution. Beyond 10 iterations, benefits become marginal. $n = 10$ provides significant improvement.

Effect of the candidate noise number k. Table 3 shows the impact of different candidate noise numbers on Set5 using LDM. When $k > 2$, the probabilistic matrix is computed as $\widetilde{P} = \mathrm{softmax}(\mathrm{concat}(\frac{1}{\widehat{D}_1}, \frac{1}{\widehat{D}_2}, \ldots, \frac{1}{\widehat{D}_k}))$. The mask matrix predicted by DPM is calculated as $\widetilde{M} = \arg\min\{\widehat{D}_i\}_{i=1}^{k}$. A series of fused mask matrices is then generated via multinomial sampling as outlined in Algorithm 1. While increasing k can lead to further improvements, the gains are limited and the deviation of the fused noise from the standard Gaussian distribution becomes more significant (with a higher mean). The optimal candidate noise number for all experiments is $k = 2$.

Effect of conducting PNFS at different reverse steps t. Table 4 presents the performance of PNFS at various reverse steps on Set14 using LDM. When $t < T$, independent reverse steps for each initial noise increase time complexity without consistent performance gains. Randomness in the

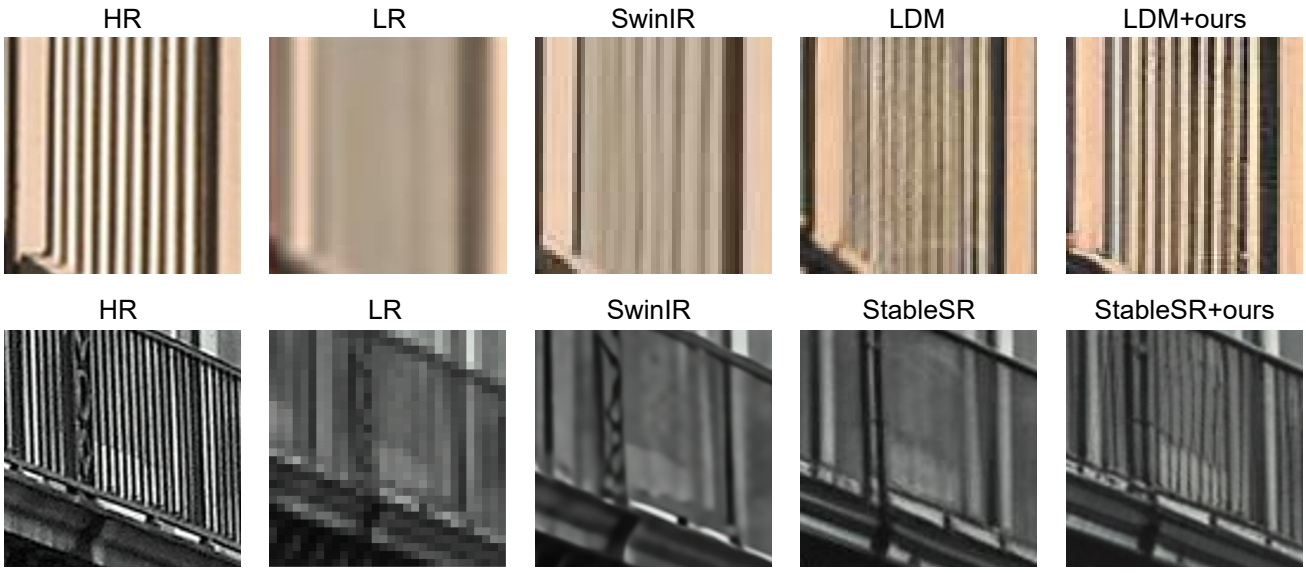

Figure 7: Visual comparisons for 4× image SR of different methods. Our PNFS is able to produce more realistic HR images compared with the backbone methods (zoom in for details).

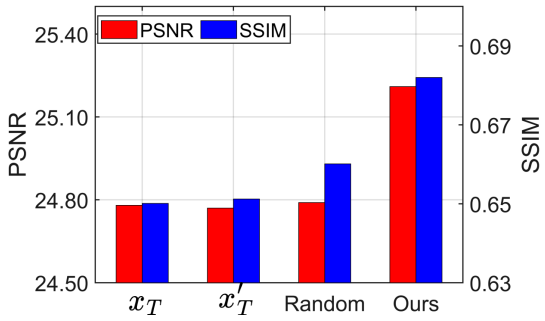

Figure 8: Comparison of different fused masks on Set14 based on LDM. Our fused noise achieves superior performance compared to the original and randomly fused noise.

reverse process may disrupt initial noise and LR image features, reducing the predictive accuracy of DPM. Thus, generating fused noise at the T-th step optimally balances performance and time complexity.

Effect of the fusion strategy. Figure 8 demonstrates the impact of PFM on Set14 using LDM. The random fused mask is generated from a uniform distribution. We observe that the performance of the reconstructed noise with the random fused mask is similar to the performance of the other random noises, which is significantly inferior to our method. This suggests that PFM effectively utilizes the insights of DPM to merge the better-performing noise elements, thereby enhancing SR performance.

Conclusion

In this paper, we examine the instability of restored images under different initial Gaussian noises, revealing that certain noise regions can be enhanced by merging with other sampled noises. Based on this observation, we propose the Predictive Noise Fusion Strategy (PNFS). PNFS first predicts pixel-wise errors in the restored image based on the current noise, then merges different noises into a superior one. Extensive experiments demonstrate that PNFS outperforms existing DM-based SR methods both quantitatively and qualitatively. Our strategy is broadly applicable to various diffusion models in a plug-and-play manner, significantly enhancing the consistency and stability of image restoration.

# Noise	mean(x_T)	PSNR↑	SSIM↑
1	0.0001	26.67	0.777
2	0.0001	27.06	0.791
5	0.01	27.26	0.810
10	0.05	27.69	0.821
50	0.08	27.60	0.805
100	0.12	27.65	0.809

Table 3: Ablation study on the number of candidate noises.

Step	PSNR	SSIM	Time(s)
None	26.67	0.777	4.70
T	27.06	0.791	4.73
T-5	27.14	0.810	5.19
T-10	27.55	0.818	5.55
T-30	28.01	0.827	6.46
T-50	27.90	0.812	7.35

Table 4: Ablation study on conducting PNFS at different reverse steps. T denotes the overall sampling steps.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62376099) and Natural Science Foundation of Guangdong Province, China (Grant No. 2024A1515010989).

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 126–135.

Avrahami, O.; Hayes, T.; Gafni, O.; Gupta, S.; Taigman, Y.; Parikh, D.; Lischinski, D.; Fried, O.; and Yin, X. 2023. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18370–18380.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18208–18218.

Bansal, A.; Borgnia, E.; Chu, H.-M.; Li, J.; Kazemi, H.; Huang, F.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2024. Cold diffusion: Inverting arbitrary image transforms without noise. *Advances in Neural Information Processing Systems*, 36.

Batzolis, G.; Stanczuk, J.; Schönlieb, C.-B.; and Etmann, C. 2021. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*.

Bevilacqua, M.; Roumy, A.; Guillemot, C.; and Alberi-Morel, M.-L. 2012. Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. In *BMVC*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. 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*, 3086–3095.

Cao, J.; Shi, Y.; Zhang, K.; Zhang, Y.; Timofte, R.; and Gool, L. V. 2024. Deep Equilibrium Diffusion Restoration with Parallel Sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chen, H.; Hao, J.; Zhao, K.; Yuan, K.; Sun, M.; Zhou, C.; and Hu, W. 2024. CasSR: Activating Image Power for Real-World Image Super-Resolution. *arXiv preprint arXiv:2403.11451*.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*.

Chung, H.; Kim, J.; Mccann, M. T.; Klasky, M. L.; and Ye, J. C. 2022. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*.

Chung, H.; Lee, E. S.; and Ye, J. C. 2022. MR image denoising and super-resolution using regularized reverse diffusion. *IEEE Transactions on Medical Imaging*, 42(4): 922–934.

Chung, H.; Sim, B.; and Ye, J. C. 2022. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12413–12422.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850–10869.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Esser, P.; Rombach, R.; Blattmann, A.; and Ommer, B. 2021. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in neural information processing systems*, 34: 3518–3532.

Fei, B.; Lyu, Z.; Pan, L.; Zhang, J.; Yang, W.; Luo, T.; Zhang, B.; and Dai, B. 2023. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9935–9946.

Feng, B. T.; Smith, J.; Rubinstein, M.; Chang, H.; Bouman, K. L.; and Freeman, W. T. 2023. Score-based diffusion models as principled priors for inverse imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10520–10531.

Gao, S.; Liu, X.; Zeng, B.; Xu, S.; Li, Y.; Luo, X.; Liu, J.; Zhen, X.; and Zhang, B. 2023. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10021–10030.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Guo, L.; Wang, C.; Yang, W.; Huang, S.; Wang, Y.; Pfister, H.; and Wen, B. 2023. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14049–14058.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control.(2022). *URL https://arxiv.org/abs/2208.01626*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, J.-B.; Singh, A.; and Ahuja, N. 2015. Single image super-resolution from transformed self-exemplars. *2015*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5197–5206.

Jing, B.; Corso, G.; Berlinghieri, R.; and Jaakkola, T. 2022. Subspace diffusion generative models. In *European Conference on Computer Vision*, 274–289. Springer.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2426–2435.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, X.; Thickstun, J.; Gulrajani, I.; Liang, P. S.; and Hashimoto, T. B. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35: 4328–4343.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Lin, J.; Xiao, P.; Wang, Y.; Zhang, R.; and Zeng, X. 2023. Diffcolor: Toward high fidelity text-guided image colorization with diffusion models. *arXiv preprint arXiv:2308.01655*.

Ma, Y.; Yang, H.; Yang, W.; Fu, J.; and Liu, J. 2023. Solving diffusion odes with optimal boundary conditions for better image super-resolution. *arXiv preprint arXiv:2305.15357*.

Martin, D. R.; Fowlkes, C. C.; Tal, D.; and Malik, J. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, 2: 416–423 vol.2.

Matsui, Y.; Ito, K.; Aramaki, Y.; Fujimoto, A.; Ogawa, T.; Yamasaki, T.; and Aizawa, K. 2017. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76: 21811–21838.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, 8162–8171. PMLR.

Niu, A.; Pham, T. X.; Zhang, K.; Sun, J.; Zhu, Y.; Yan, Q.; Kweon, I. S.; and Zhang, Y. 2024. ACDMSR: Accelerated conditional diffusion models for single image super-resolution. *IEEE Transactions on Broadcasting*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rout, L.; Raoof, N.; Daras, G.; Caramanis, C.; Dimakis, A.; and Shakkottai, S. 2024. Solving linear inverse problems provably via posterior sampling with latent diffusion models. *Advances in Neural Information Processing Systems*, 36.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022a. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022b. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4713–4726.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Sun, L.; Wu, R.; Zhang, Z.; Yong, H.; and Zhang, L. 2023. Improving the Stability of Diffusion Models for Content Consistent Super-Resolution. *arXiv preprint arXiv:2401.00877*.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2023a. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint arXiv:2305.07015*.

Wang, Y.; Yang, W.; Chen, X.; Wang, Y.; Guo, L.; Chau, L.-P.; Liu, Z.; Qiao, Y.; Kot, A. C.; and Wen, B. 2023b. SinSR: Diffusion-Based Image Super-Resolution in a Single Step. *arXiv preprint arXiv:2311.14760*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Chen, J.; and Hoi, S. C. 2020. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3365–3387.

Wang, Z.; Zhang, Z.; Zhang, X.; Zheng, H.; Zhou, M.; Zhang, Y.; and Wang, Y. 2023c. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1704–1713.

Yue, J.; Fang, L.; Xia, S.; Deng, Y.; and Ma, J. 2023. Dif-fusion: Towards high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing*.

Zeyde, R.; Elad, M.; and Protter, M. 2010. On Single Image Scale-Up Using Sparse-Representations. In *Curves and Surfaces*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.