Unsupervised Diffusion-Based Degradation Modeling for Real-World Super-Resolution

Yuying Chen[1], Mingde Yao[2], Wenbo Li[3], Renjing Pei[3], Jinjing Zhao[4], Wenqi Ren[1]*

[1]Shenzhen Campus of Sun Yat-sen University
[2]The Chinese University of Hong Kong
[3]Huawei Noah's Ark Lab
[4]National Key Laboratory of Science and Technology on Information System Security, China
elvin.chenyuying@gmail.com, mingdeyao@foxmail.com, fenglinglwb@gmail.com, peirenjing@huawei.com, zhjj0420@126.com, renwq3@mail.sysu.edu.cn

Abstract

Single image super-solution (SR) aims to restore a high-resolution (HR) image from a degraded low-resolution (LR) image. However, existing SR models still face a significant domain gap between synthetic and real-world datasets due to the mismatched degradation distributions, hindering SR models from achieving optimal results. In this paper, we propose an unsupervised diffusion-based degradation modeling framework (UDDM) to effectively capture real-world degradation distributions. Specifically, given unpaired LR and HR images, a diffusion-based degradation module (DDM) first models the degradation distribution by diffusing real-world LR images to *downsampled* LR images, which does not require HR images. It then applies reverse diffusion to generate real-world LR images from *extremely downsampled* HR images. This approach allows DDM to model and generate real-world degradation distributions without requiring paired data, by using *extreme downsampling* to link unpaired LR and HR images. Additionally, we introduce a physics-based dynamic degradation module (P-DDM) that adaptively models content-aware degradation, ensuring both content and structural accuracy. Finally, the LR images generated by DDM and P-DDM are adaptively weighted to produce the final LR images, which are paired with the given HR images for training the SR network. Extensive experiments across multiple real-world datasets demonstrate that our framework achieves state-of-the-art performance in both qualitative and quantitative comparison.

Introduction

Single image super-resolution (SISR) aims to reconstruct high-resolution (HR) images from low-resolution (LR) ones, which is a fundamental problem in low-level vision. Previous SR methods (Dong et al. 2014, 2016; Ledig et al. 2017; Zhang et al. 2018b; Chen et al. 2021; Liang et al. 2021) leverage deep learning to learn mappings from LR to HR images, achieving good results on known LR degradations (e.g., Bicubic, Gaussian). However, these methods may not perform well on real-world images due to the complex and unknown characteristics of real-world degradation (Liu et al. 2022; Chen et al. 2022), which differ significantly from the synthetic conditions used in training.

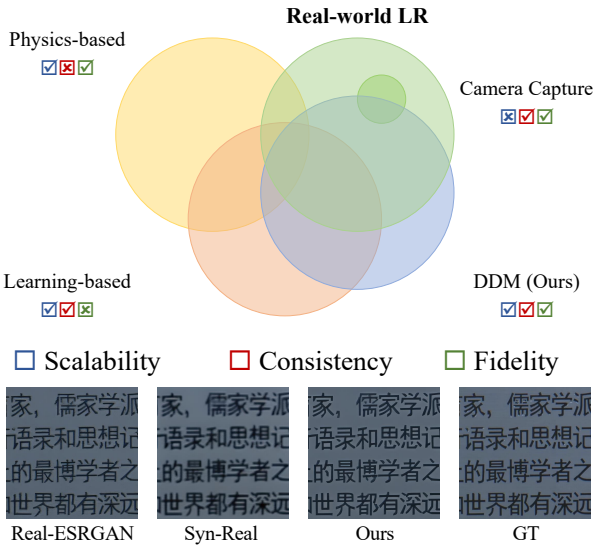


Figure 1: Degradation distribution of different generation methods and their SR results. Our method matches the real-world distribution and achieves the best SR result.

Existing methods for real-world SR can be broadly divided into four categories. The most straightforward way is capturing paired real-world datasets to train SR models (Cai et al. 2019; Wei et al. 2020). However, it is costly (Liu et al. 2022) and faces challenges such as spatial misalignment and color shift (Chen et al. 2022). Another way is to use unsupervised learning to model the distribution from LR to HR images (Yuan et al. 2018; Maeda 2020), which might mismatch the HR distribution due to unstable adversarial learning (Liu et al. 2022). Recently, generating LR data from existing HR images to create training pairs has shown promising results (Liu et al. 2023). Yet, these methods often rely on handcrafted degradation models (Zhang et al. 2021; Wang et al. 2021) or the unstable generative capabilities of GANs (Bulat, Yang, and Tzimiropoulos 2018; Wei et al. 2021), which fail to accurately model LR degradation distributions (Zhang et al. 2023), thereby limiting the performance of SR model

*Corresponding author.

in real-world circumstance.

Recently, diffusion model (Ho, Jain, and Abbeel 2020) has demonstrated remarkable capabilities in various domains. It excels at capturing complex and high-dimensional data distributions by iteratively refining noisy inputs into well-structured outputs (Wang, Yu, and Zhang 2022; Song et al. 2020; Rombach et al. 2022). This iterative denoising process allows us to model intricate real-world LR distributions with high accuracy. However, diffusion models typically require paired datasets (Yang et al. 2023a; Lin et al. 2024; Wang et al. 2023) to perform the HR-to-LR diffusion process and generate realistic real-world LR distributions, which poses a challenge due to the lack of LR-HR pairs.

In this paper, we propose a diffusion-based degradation module (DDM) to effectively capture real-world degradation distributions in an unsupervised manner. DDM is based on two main ideas. First, we leverage the diffusion model's powerful ability to accurately model real-world LR distributions. Second, rather than learning the mapping from HR to LR images (Bulat, Yang, and Tzimiropoulos 2018; Wei et al. 2021), we use *extreme downsampling* to obtain ultra-low-resolution (ULR) images from real-world LR images. By learning the diffusion process from extremely downsampled LR images to real-world LR images, we can model degradation distributions more effectively using pure real-world LR images, with scalability, consistency, and fidelity to real-world LR, as shown in Figure 1.

Specifically, as shown in Figure 2, given unpaired HR images and real-world LR images, our goal is to learn the distribution of real-world LR degradation and generate paired LR-HR data that matches this distribution. During training, DDM learns the real-world degradation by diffusing real-world LR images to extremely downsampled LR images. In the inference phase, DDM performs the reverse diffusion process to generate LR images from extremely downsampled HR images, ensuring that the generated LR images follow the real-world LR distribution. DDM leverages extreme downsampling as a bridge to connect unpaired LR and HR images, allowing the diffusion model to produce degradation distributions that closely align with real-world conditions.

Additionally, we propose a physics-based dynamic degradation module (P-DDM), which adaptively generates LR images based on the content of the HR image. P-DDM consists of a physics-based degradation module and a parameter generation module, where the former generates a wide range of highly complex degradation, and the latter produces content-aware parameters for the physics-based degradation module. Finally, the LR images obtained from DDM and P-DDM are fused together to form the final LR images following the real-world degradation distribution, paired with HR images to train SR models. Extensive experiments demonstrate that our method achieves state-of-the-art performance across multiple real-world datasets in both qualitative comparison and quantitative metrics.

Related Work

Single Image Super-resolution. Recent advancements in single image super-resolution (SISR) have primarily focused on deep learning techniques (Dong et al. 2014; Shi et al. 2016; Ledig et al. 2017; Goodfellow et al. 2020; Yao et al. 2024; Pan et al. 2022). SRCNN (Dong et al. 2014) makes significant strides by directly learning the mapping from LR to HR images. ESPCN (Shi et al. 2016) introduces sub-pixel convolution layers to enhance real-time SR performance, refines residual blocks, and removes batch normalization to boost image quality. To achieve photo-realistic results with detailed textures, SRGAN (Ledig et al. 2017) introduces the generative adversarial network (GAN) (Goodfellow et al. 2020) into the SR framework, and employs it as loss supervision to push the SR solutions closer to the natural manifold. While these methods excel with synthetic degradations, such as bicubic downsampling, they often struggle with complex and varied distortions in real-world scenarios.

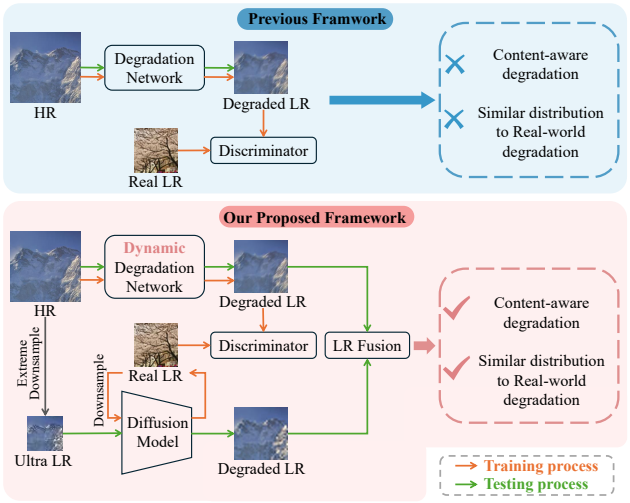


Figure 2: Illustration on motivation. Existing learning-based methods primarily rely on adversarial loss to adapt from the realistic domain. Our approach exploits diffusion priors to strengthen the details of synthetic LR.

Real-world Super-resolution. To alleviate the domain shift issues, GAN-based methods (Bulat, Yang, and Tzimiropoulos 2018; Yuan et al. 2018; Maeda 2020) tend to learn the real-world degradation model in the training stage implicitly. Besides, other methods (Zhang et al. 2021; Wang et al. 2021) try to simulate the real-world degradation distribution by enumerating the degradation operations. However, due to the complex and unknown processing of real-world images, it is hard to mimic all types of degradation. Instead, some methods (Gu et al. 2019; Hussein, Tirer, and Giryes 2020; Huang et al. 2020; Luo et al. 2022; Xu, Yao, and Xiong 2023) try to estimate the image-specific degradation model during the test time, which helps to reconstruct more plausible HR images. For instance, optimization-based methods (Gu et al. 2019; Hussein, Tirer, and Giryes 2020; Huang et al. 2020) estimate the blur kernel and SR image together in an iterative manner. However, these methods cannot generate satisfactory results when the test images contain different types of degradation (Liu et al. 2022). Thus, these methods still suffer from the domain shift on test images with unknown degradation.

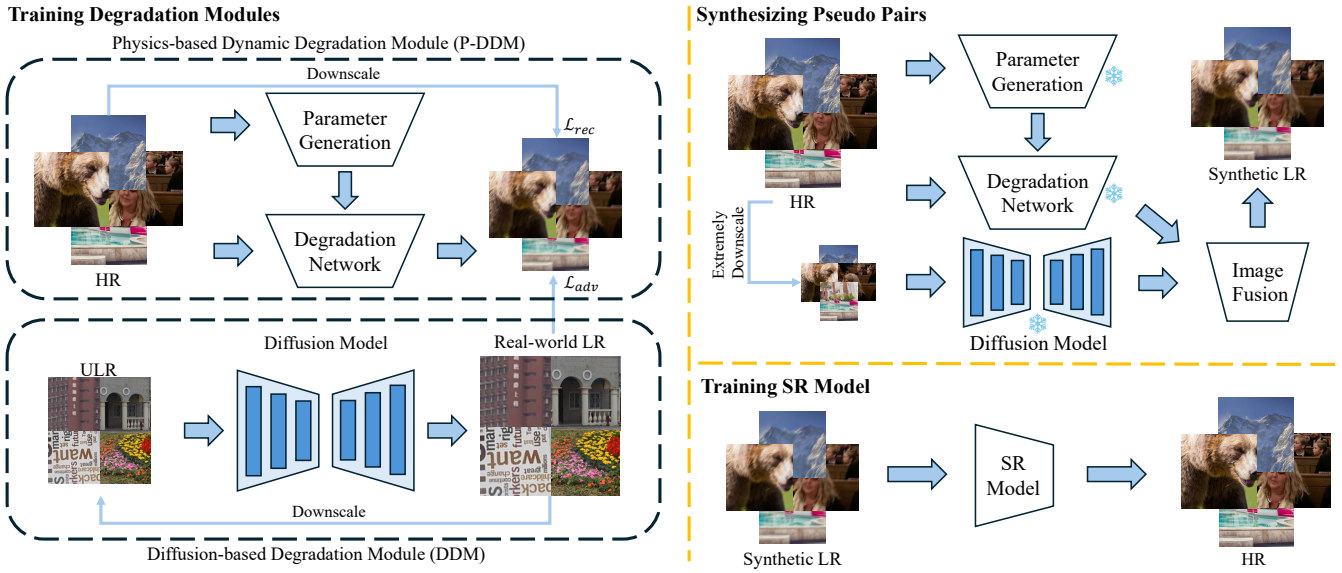


Figure 3: An overall illustration of the proposed degradation modeling framework. In the first stage, we train both the diffusion-based degradation module (DDM) and the physics-based dynamic degradation module (P-DDM) with unpaired HR and LR datasets. In the second stage, we apply the pre-trained models to synthesize well-degraded LR images. In the last stage, our synthetic HR-LR pairs can be utilized to train arbitrary SR models in a supervised manner.

Diffusion-based Image Super-resolution. Diffusion models (Dhariwal and Nichol 2021) are increasingly valued in SR for their powerful distribution modeling capabilities. One type is to use diffusion models to generate HR images directly from LR ones. For example, pre-trained priors are used to produce HR images (Kawar et al. 2022; Wang, Yu, and Zhang 2022; Fei et al. 2023; Yue and Loy 2022), leveraging the rich texture information in these models. However, these techniques often struggle with non-blind degradation or specific scenarios, such as facial images (Yue and Loy 2022). Recent advancements have seen the application of large-scale text-to-image diffusion models (Rombach et al. 2022; Podell et al. 2023; Ramesh et al. 2022; Saharia et al. 2022), trained on extensive high-definition image datasets. These models enhance the ability to handle diverse content (Wang et al. 2023; Lin et al. 2024; Yang et al. 2023b). However, despite their advantages, these methods can be affected by inaccurate priors and may introduce artifacts into the generated HR images.

Another type (Fei et al. 2023) concurrently estimates the degradation model to address real-world degradation. While effective for linear degradation, this strategy faces challenges when dealing with more complex, non-linear real-world degradations.

Method

Overview

Our goal is to reduce the degradation gap between training and testing datasets so that SR models trained on synthetic datasets can be applied to real-world LR images. As shown in Figure 3, our unsupervised diffusion-based degradation modeling (UDDM) framework has three stages: training degradation modules, synthesizing pseudo pairs, and training the SR model.

Training degradation modules is the core part of our method. It consists of two modules, i.e., a diffusion-based degradation module (DDM) and a physics-based dynamic degradation module (P-DDM). DDM learns the real-world degradation distribution, enabling the diffusion model to generate LR-HR image pairs that match the distribution. We achieve this by using the *extreme downsampling* operation to learn the diffusion process from an LR image to its *extreme downsampling* version (see next section for details).

For P-DDM, we employ a physics-based degradation model that uses CNN layers to simulate various degradation types. This enables the network to generate different degradations following the physics rule, e.g., downsampling, blur, color fading and noise. Additionally, we introduce a dynamic generator network to predict the parameters of the degradation model, allowing it to adaptively learn degradation based on the content of the image.

Once the degradation modules (DDM and P-DDM) are optimized, we create pseudo image pairs using both degradation modules. Specifically, we start by extremely downsampling the HR image to serve as input for DDM, while the HR image is directly fed into P-DDM. The LR outputs from DDM and P-DDM are then adaptively fused to form LR-HR image pairs as

$$\boldsymbol{x} = (1 - \boldsymbol{w})\, \mathcal{DDM}(I_{hr}) + \boldsymbol{w}\, \mathcal{PDDM}(I_{hr}), \quad (1)$$

where $\boldsymbol{x}$ represents the final synthesized LR image, I_{hr} is the HR image, and $\boldsymbol{w}$ denotes the weights of blending. Such a design ensures that the LR-HR pairs accurately reflect real-world degradation distributions, allowing the SR model trained on these pairs to perform effectively on real-world

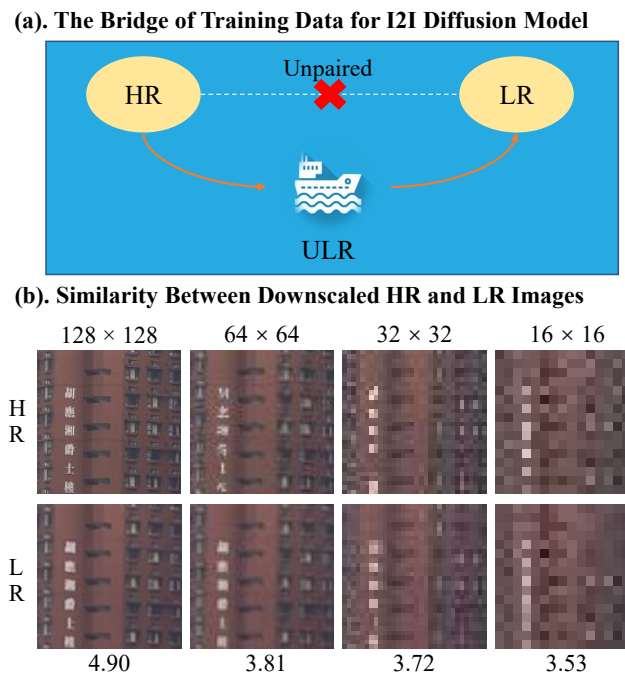


Figure 4: A. Ultra-LR (ULR) image performs a bridge between unpaired LR and HR images. B. MAE between HR images and their LR pairs decreases as the downscale level increases, thereby eliminating the domain gap.

LR images.

Diffusion-based Degradation

The Diffusion model has great potential to accurately model real-world degradation distributions. However, limited by the absence of paired data, it is difficult to directly learn the diffusing process from real-world LR and HR image pairs, which hinders the training of the diffusion model.

To overcome this challenge, we leverage the *extreme downsampling* operation to learn the real-world degradation distribution. Specifically, given unpaired LR image $I_{lr}^R \in \mathcal{R}^{H\times W\times 3}$ and HR image $I_{hr} \in \mathcal{R}^{rH\times rW\times 3}$, where H, W are the height and width of the LR image, and $r > 1$ is a scale factor. We first downsample the I_{lr}^R into its ultra LR version $I_{lr}^R \downarrow_s \in \mathcal{R}^{H/s\times W/s\times 3}$, where $\downarrow_s$ represents the bicubic downsampling with the scale factor $s > 1$.

In this way, we can learn the real-world degradation distribution by applying a forward diffusion process to the ultra LR image $I_{lr}^R \downarrow_s$ as

$$I_{lr}^R = \mathcal{DDM}(I_{lr}^R \downarrow_s). \tag{2}$$

This involves progressively adding Gaussian noise over several steps, creating a sequence of increasingly noisy images x_t. The forward process can be described by the equation $x_{t+1} = x_t + \sqrt{\beta_t} \cdot \epsilon$, where ϵ is Gaussian noise and β_t controls the noise variance at step t.

The training objective is to minimize the difference between the generated clean image and the original image, achieved by using the variational lower bound (VLB) of the data likelihood as the loss function

$$\mathcal{L}_{\text{d}} = \mathcal{E}_{x_0,\epsilon,t}\left[\|\epsilon - \hat{\epsilon}_\theta(x_t, t)\|^2\right]. \tag{3}$$

Here, $\hat{\epsilon}_\theta(x_t, t)$ is the model's prediction of the noise added at time t, and the squared L2 norm $\|\cdot\|^2$ measures the discrepancy between the true and predicted noise. By optimizing this loss function, the diffusion model learns to accurately capture and generate real-world degradation patterns, effectively bridging the gap between synthetic and real-world image distributions.

To utilize the DDM for generating low-resolution (LR) images I_{lr}^D paired with high-resolution (HR) images, we perform the reverse diffusion process. This process iteratively denoises the extremely downsampled HR image $I_{hr} \downarrow_{rs} \in \mathcal{R}^{H/s\times W/s\times 3}$. Specifically, the reverse diffusion can be represented as

$$I_{lr}^D = \mathcal{DDM}_{\text{reverse}}(I_{hr} \downarrow_{rs}). \tag{4}$$

Such a process enables us to generate paired LR-HR data (I_{lr}^D, I_{hr}) that aligns to real-world degradation distributions. These pairs can then be used to train the SR model, providing better performance on real-world images.

Analysis of *extreme downsampling*. As shown in Figure 4 extreme downsampling offers two main benefits. First, it enables the model to learn real-world degradation distributions with only real-world LR images. By significantly reducing the resolution of these images, the model captures the degradation characteristics, which allows for the generation of accurate LR-HR pairs in the following stage. Second, extreme downsampling bridges the gap between synthetic and real-world data. Figure 4 (b) shows that increasing the downsampling rate improves the similarity between synthetic and real-world images. This means that a model trained on extremely downsampled real-world LR images can effectively handle extremely downsampled HR images, producing LR-HR pairs that closely match real-world degradation patterns.

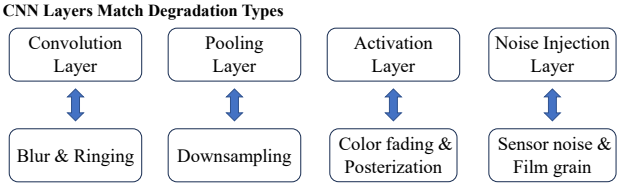


Figure 5: Observation of the relation between CNN layers and degradation types.

Physics-based Dynamic Degradation

Despite the effectiveness of DDM, diffusion models often produce unstable textures and excessive sharpening. To address these, we incorporate a physics-based dynamic degradation module (P-DDM) to ensure the texture quality of generated LR images. P-DDM comprises two main components: physics-based degradation generation and parameter generation. The former generates realistic content based on physical operations, while the latter generates parameters for the degradation model based on the input image.

Model (Generator)	RealSR(Cai et al. 2019)				DRealSR(Wei et al. 2020)			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
SwinIR (Real-ESRGAN)	24.395	0.776	0.3037	119.43	26.944	0.8308	0.3219	139.18
SwinIR (Syn-Real)	25.589	0.7687	0.3835	163.13	28.301	0.8309	0.3801	154.59
SwinIR (Ours)	**26.732**	**0.7913**	**0.2652**	**105.92**	**29.247**	**0.8386**	**0.2709**	**118.09**
Real-ESRGAN (Real-ESRGAN)	25.600	0.7587	0.2749	138.94	28.549	**0.8043**	0.2820	**146.94**
Real-ESRGAN (Syn-Real)	24.341	0.7370	0.3021	159.44	27.483	0.7899	0.3306	171.89
Real-ESRGAN (Ours)	**26.651**	**0.7769**	**0.2061**	**102.43**	**29.176**	0.8032	**0.2645**	150.44
StableSR (Real-ESRGAN)	24.629	0.7035	0.3014	133.92	27.846	0.7412	0.3337	152.62
StableSR (Syn-Real)	25.679	0.7302	0.3680	165.62	28.621	0.7952	0.3892	183.45
StableSR (Ours)	**26.820**	**0.7768**	**0.2514**	**128.11**	**29.678**	**0.8267**	**0.2567**	**140.550**

Table 1: Quantitative comparisons of the SR performance of representative models (trained with distinct data generation methods) on RealSR and DRealSR datasets. The best results are highlighted in **bold**.

Prior work (Luo, Wu, and Guo 2023) has shown that neural network operations can simulate general degradation types. As illustrated in Figure 5, convolution layers can mimic blurring and ringing effects, pooling layers for downsampling, activation layers reproduce color fading and posterization, and additional operations introduce noise. Hence, a lightweight convolutional network is sufficient for modeling the major degradation types in super-resolution tasks.

To more accurately model complex degradations, such as JPEG compression that affects different regions of an image, we introduce a dynamic network that generates degradation parameters based on the input image's content. This network, designed as a simple convolutional architecture, processes an HR image to produce a set of parameters for the physics-based degradation model. It maintains texture fidelity and improves the quality of the generated LR images, ensuring that the degradation process aligns more closely with real-world conditions.

The loss function of P-DDM is as

$$\mathcal{L}_{\mathcal{PDDM}} = \mathcal{L}_{rec} + \alpha \mathcal{L}_{adv}, \tag{5}$$

where α is a balance factor, $\mathcal{L}_{rec} = \| I_{lr}^{P} - I_{hr} \downarrow_r \|_2$ ensures the structure of output, and $\mathcal{L}_{adv}$ is the adversarial loss (Ledig et al. 2017) to learn the real-world distribution of LR image I_{lr}^{R}.

Experiments

Implementation Details

We adopt StableSR[1] to implement DDM, with minor modifications on loading the training dataset. To generate degraded images, we finetune StableSR (Wang et al. 2023) for 223 epochs with a batch size of 192, and the prompt is fixed as null. The training process is conducted on images cropped to 128×128 resolution. We adopt DDPM sampling (Ho, Jain, and Abbeel 2020) with 200 timesteps for inference.

After synthesizing fused degraded LR images, we still adopt StableSR (Wang et al. 2023) as our SR model. To super-resolve images, we finetune StableSR for 119 epochs with a batch size of 24, and the prompt is fixed as null. The training process is conducted on images cropped to 512×512 resolution. We still adopt DDPM sampling (Ho, Jain, and Abbeel 2020) with 200 timesteps for inference.

When cropping input images to our desired resolution, we crop the images in left to right and top to bottom direction without overlap, except for the last patch that is smaller than our desired resolution. To handle output images with arbitrary sizes, we adopt the aggregation sampling strategy proposed in StableSR (Wang et al. 2023) for images beyond 512×512. More details can be found in the appendix.

[1]https://github.com/IceClear/StableSR

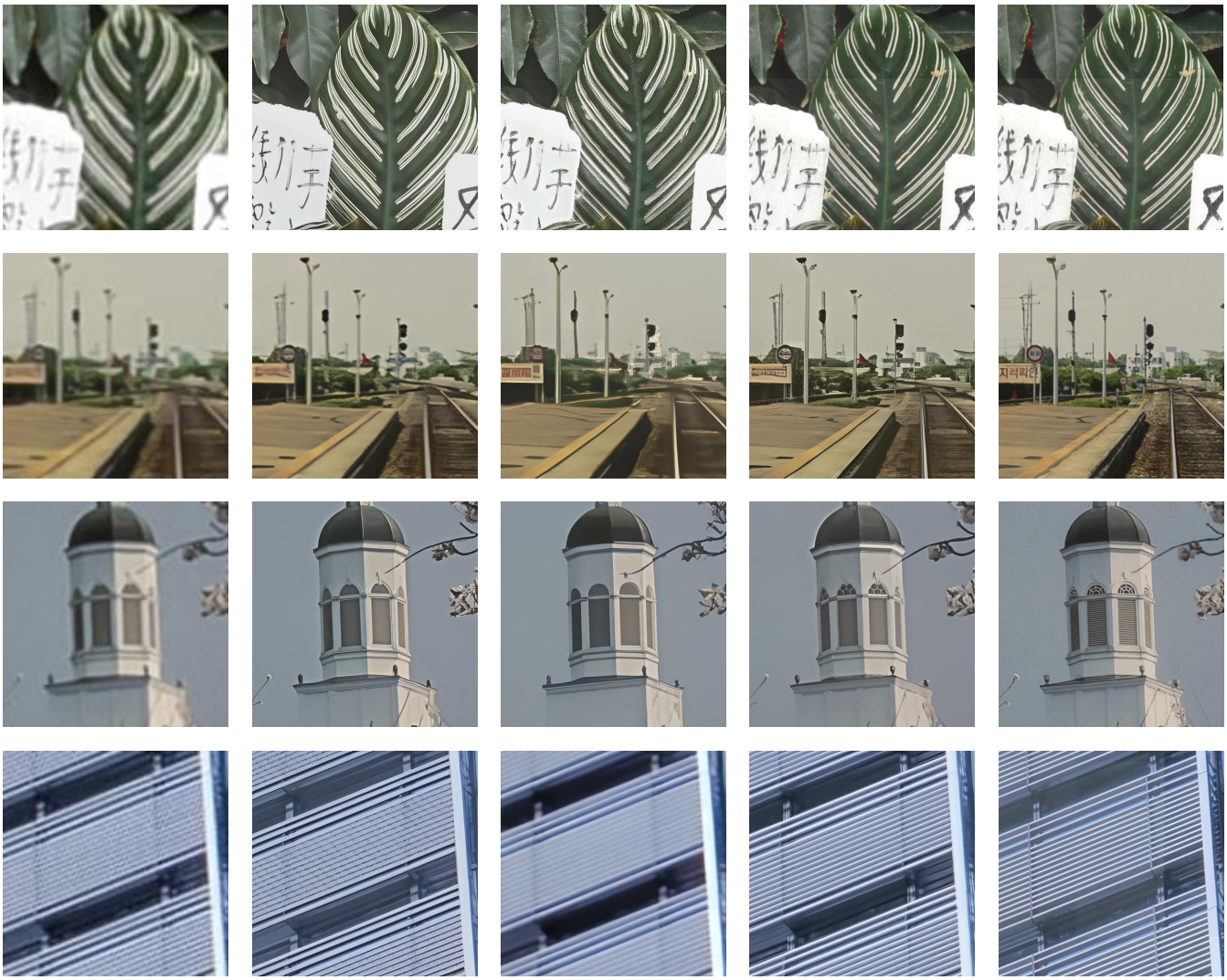
Experimental Settings

Training Datasets. We adopt the proposed dual branch degradation modeling framework to synthesize HR-LR pairs. The HR dataset consists of DIV2K (Agustsson and Timofte 2017), Flickr2K (Timofte et al. 2017) and OutdoorSceneTraining (Wang et al. 2018) datasets. The LR dataset is made up of the training sets from RealSR (Cai et al. 2019) and DRealSR (Wei et al. 2020).

Testing Datasets. We evaluate our approach on the testing sets of RealSR (Cai et al. 2019) and DRealSR (Wei et al. 2020). The resolution of LR and HR is 128×128 and 512×512, respectively. Note that for StableSR, the inputs are first upsampled to the same size as the outputs before inference.

Compared Methods. To verify the effectiveness of our framework, we apply our synthetic data pairs to train several state-of-the- art SR models in a supervised manner. We choose SwinIR (Liang et al. 2021), Real-ESRGAN (Wang et al. 2021), and StableSR (Wang et al. 2023) to represent transformers, GANs, and diffusion models respectively. We also train these SR models on data pairs generated by other degradation methods, i.e. Real-ESRGAN (Wang et al. 2021) and Syn-Real (Yang et al. 2023c), to compare the effects of distinct training data generators given their SR performances. For all methods mentioned above, we adopt the official code and models for training and testing.

Evaluation Metrics. Since benchmarks RealSR and DRealSR have paired HR-LR data, we employ various reference-based metrics including PSNR (Huynh-Thu and



(a) Zoomed LR (b) Real-ESRGAN (c) Syn-Real (d) Ours (e) HR

Figure 6: Qualitative comparisons of SR results trained on different synthetic data. Super-resolved images ($128 \rightarrow 512$) are collected from RealSR (Cai et al. 2019) and DRealSR (Wei et al. 2020) datasets. (**Zoom in for details**)

Ghanbari 2008), SSIM (Wang et al. 2004)[2], LPIPS[3] (Zhang et al. 2018a), FID (Heusel et al. 2017) to evaluate both the consistency and perceptual quality of generated images.

Comparison with Existing Methods

Quantitative Comparisons. We first show the quantitative comparison of two real-world benchmarks. As shown in Table 1, our approach achieves the best scores in most cases. It outperforms other synthesizing methods in both accurate (i.e. PSNR, SSIM) and perceptual metrics (i.e. LPIPS, FID). Although Real-ESRGAN achieves higher SSIM and FID scores on its own model, the differences between its scores and the scores of our approach are tiny and trivial.

[2]PSNR and SSIM scores are both evaluated on the luminance channel in YCbCr color space

[3]We use LPIPS-ALEX by default.

Moreover, existing methods fail to restore faithful textures and generate blurry results, as shown in Fig. 6. In contrast, our approach is capable of generating sharp images with realistic details. For instance, both b and c methods fail to restore the details on the windows in the third picture of Fig. 6, only our methods produce small circles similar to the windows' structure of the HR image.

Qualitative Comparisons. To demonstrate the effectiveness of our method, we present visual results on real-world images from both real-world benchmarks (Cai et al. 2019; Wei et al. 2020) in Fig. 6. It is observed that our approach outperforms previous methods in both artifact removal and detail generation. Specifically, our approach is able to generate faithful details, as shown in the first row of Fig. 6, while other methods either show blurry results (Syn-Real) or unnatural details (Real-ESRGAN). Moreover, as shown in the fourth row of Fig. 6, our approach generates sharp edges

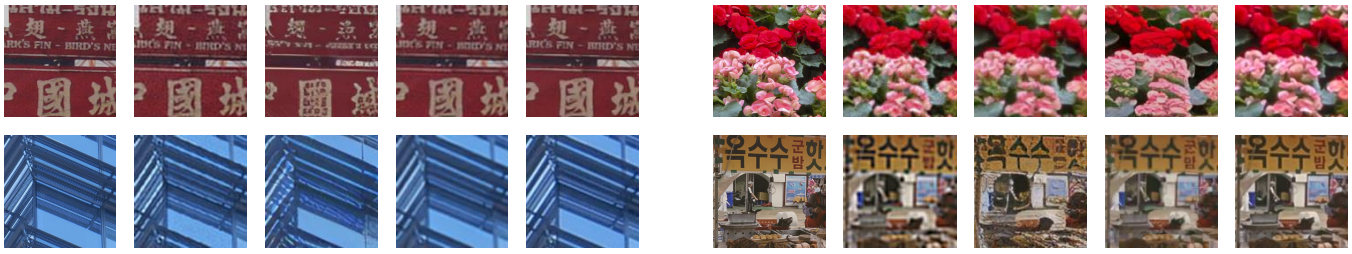


Figure 7: Qualitative comparisons of different data generation methods. Degraded images ($512 \rightarrow 128$) are from RealSR (Cai et al. 2019) and DRealSR (Wei et al. 2020) datasets. (**Zoom in for details**)

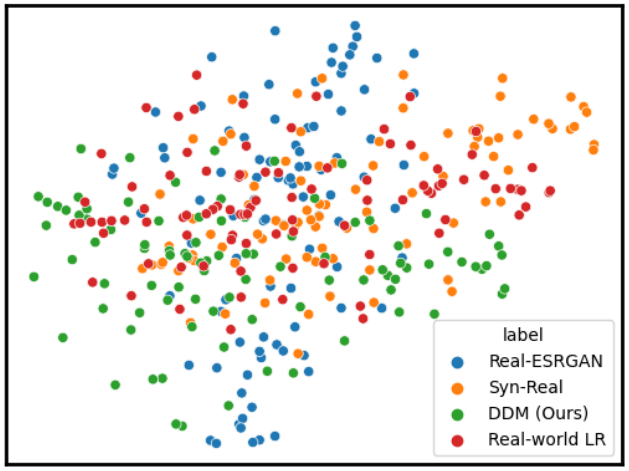


Figure 8: Distribution visualization of LR images.

without obvious degradations, whereas other state-of-the-art methods generate blurry results.

Visualization of Synthetic LR

According to Fig. 7, our synthetic LR images have the closest visual effect to the real LR images. LR images generated by Real-ESRGAN (Wang et al. 2021) often contain more blur and noise than real-world ones, while Syn-Real (Yang et al. 2023c) significantly destroy the content (e.g. letters, shapes, and colors) of LR images.

We also apply a t-distributed stochastic neighbor embedding (t-SNE) clustering algorithm to visualize the similarity between the distribution of synthetic LR images and real-world LR images. LR images are first reduced to 2 dimensions via the principal component analysis (PCA) algorithm. Then they are clustered by t-SNE and plotted in a scatter plot with distinct colors. According to Fig. 8, our synthetic LR images have a closer distribution to real-world LR images compared to Real-ESRGAN (Wang et al. 2021) and Syn-Real (Yang et al. 2023c).

Ablation Study

We use a fusion coefficient $w \in [0, 1]$ to control the weight of each branch in the final synthetic LR images. Specifically, a small w tends to utilize more diffusion prior while a larger w enhances the fidelity of generated images. As shown in Table 2, compared with DDM ($w = 0$), larger values of w achieve higher PSNR and SSIM on all benchmarks, indicating better fidelity. By contrast, DDM achieves better perceptual quality via higher CLIP-IQA and MUSIQ scores. We further find that a proper w can improve both fidelity and perceptual quality. UDDM ($w = 0.5$) shows comparable PSNR and SSIM with P-DDM ($w = 1$) but achieves better perceptual metric scores in CLIP-IQA scores and MUSIQ. Hence, we set the fusion coefficient w to 0.5 by default for trading between quality and fidelity.

Datasets	Metrics	DDM ($w = 0.0$)	**UDDM** ($w = 0.5$)	P-DDM ($w = 1.0$)
RealSR	PSNR ↑	26.10	**26.63**	26.44
	SSIM ↑	0.7626	**0.7821**	0.7716
	LPIPS ↓	0.3190	**0.2439**	0.2503
	CLIP-IQA ↑	0.6127	**0.6234**	0.5847
	MUSIQ ↑	65.81	**65.88**	64.05
DRealSR	PSNR ↑	27.43	**28.03**	27.97
	SSIM ↑	0.7341	**0.7936**	0.7540
	LPIPS ↓	0.3595	**0.2654**	0.3080
	CLIP-IQA ↑	0.6340	**0.6357**	0.5893
	MUSIQ ↑	**58.98**	58.51	56.77

Table 2: Ablation studies of the controllable branch fusion coefficient w on RealSR (Wang et al. 2021) and DRealSR (Wei et al. 2020) datasets.

Conclusion

Witnessed the rapid development of real-world image super-resolution, we attempt to develop a reliable and flexible approach to reduce the discrepancy between training and testing data. This paper proposes a novel framework to exploit dynamic networks, adversarial learning, and diffusion priors to produce realistic LR images for training SR models. We devote our efforts to tackling well-known problems such as the lack of paired training data for supervised learning and the distribution shift between training and testing data. We believe that our exploration would lay a good foundation in this field, and our proposed DDM and P-DDM could provide precious insights for future works. The most explicit limitation of the proposed framework is the high cost of time and computational resources due to the property of the diffusion model, hence hindering its application. We hope to enhance its efficiency in future work by changing the sampling strategy, refining the model structure, and distillation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.62322216), and the Shenzhen Science and Technology Program (Grant No. RCYX20221008092849068, JSGG20220831093004008).

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.

Bulat, A.; Yang, J.; and Tzimiropoulos, G. 2018. To Learn Image Super-Resolution, Use a GAN to Learn How to do Image Degradation First. In *European Conference on Computer Vision*, 185–200.

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.

Chen, H.; He, X.; Qing, L.; Wu, Y.; Ren, C.; Sheriff, R. E.; and Zhu, C. 2022. Real-world single image super-resolution: A brief review. *Information Fusion*, 79: 124–145.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-trained Image Processing Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12299–12310.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, 184–199.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2016. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295–307.

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. *arXiv preprint arXiv:2304.01247*.

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: 139–144.

Gu, J.; Lu, H.; Zuo, W.; and Dong, C. 2019. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1604–1613.

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, Y.; Li, S.; Wang, L.; Tan, T.; et al. 2020. Unfolding the alternating optimization for blind super resolution. In *Advances in Neural Information Processing Systems*, volume 33, 5632–5643.

Hussein, S. A.; Tirer, T.; and Giryes, R. 2020. Correction filter for single image super-resolution: Robustifying off-the-shelf deep super-resolvers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1428–1437.

Huynh-Thu, Q.; and Ghanbari, M. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 44(13): 800–801.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

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, X.; He, J.; Chen, Z.; Lyu, Z.; Dai, B.; Yu, F.; Ouyang, W.; Qiao, Y.; and Dong, C. 2024. DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior. arXiv:2308.15070.

Liu, A.; Liu, Y.; Gu, J.; Qiao, Y.; and Dong, C. 2022. Blind image Super-Resolution: A Survey and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Y.; Zhao, H.; Gu, J.; Qiao, Y.; and Dong, C. 2023. Evaluating the generalization ability of super-resolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Luo, F.; Wu, X.; and Guo, Y. 2023. AND: Adversarial Neural Degradation for Learning Blind Image Super-Resolution. In *Advances in Neural Information Processing Systems*, volume 36, 21255–21267. Curran Associates, Inc.

Luo, Z.; Huang, H.; Yu, L.; Li, Y.; Fan, H.; and Liu, S. 2022. Deep constrained least squares for blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17642–17652.

Maeda, S. 2020. Unpaired Image Super-Resolution using Pseudo-Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 291–300.

Pan, Z.; Li, B.; He, D.; Yao, M.; Wu, W.; Lin, T.; Li, X.; and Ding, E. 2022. Towards bidirectional arbitrary image rescaling: Joint optimization and cycle idempotence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17389–17398.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

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.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. 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*, 1874–1883.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.-H.; and Zhang, L. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 114–125.

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2023. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *arXiv preprint arXiv:2305.07015*.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1905–1914.

Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 606–615.

Wang, Y.; Yu, J.; and Zhang, J. 2022. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model. *arXiv preprint arXiv:2212.00490*.

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.

Wei, P.; Xie, Z.; Lu, H.; Zhan, Z.; Ye, Q.; Zuo, W.; and Lin, L. 2020. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 101–117. Springer.

Wei, Y.; Gu, S.; Li, Y.; Timofte, R.; Jin, L.; and Song, H. 2021. Unsupervised Real-World Image Super Resolution via Domain-Distance Aware Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13385–13394.

Xu, R.; Yao, M.; and Xiong, Z. 2023. Zero-Shot Dual-Lens Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9130–9139.

Yang, P.; Zhou, S.; Tao, Q.; and Loy, C. C. 2023a. PGDiff: Guiding Diffusion Models for Versatile Face Restoration via Partial Guidance. *arXiv preprint arXiv:2309.10810*.

Yang, T.; Ren, P.; Xie, X.; and Zhang, L. 2023b. Pixel-Aware Stable Diffusion for Realistic Image Super-resolution and Personalized Stylization. *arXiv preprint arXiv:2308.14469*.

Yang, T.; Ren, P.; Zhang, L.; et al. 2023c. Synthesizing realistic image restoration training pairs: A diffusion approach. *arXiv preprint arXiv:2303.06994*.

Yao, M.; Xu, R.; Guan, Y.; Huang, J.; and Xiong, Z. 2024. Neural degradation representation learning for all-in-one image restoration. *IEEE Transactions on Image Processing*.

Yuan, Y.; Liu, S.; Zhang, J.; Zhang, Y.; Dong, C.; and Lin, L. 2018. Unsupervised Image Super-Resolution Using Cycle-in-Cycle Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 701–710.

Yue, Z.; and Loy, C. C. 2022. DifFace: Blind Face Restoration with Diffused Error Contraction. *arXiv preprint arXiv:2212.06512*.

Zhang, K.; Liang, J.; Van Gool, L.; and Timofte, R. 2021. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4791–4800.

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

Zhang, W.; Li, X.; Chen, X.; Zhang, X.; Qiao, Y.; Wu, X.-M.; and Dong, C. 2023. SEAL: A Framework for Systematic Evaluation of Real-World Super-Resolution. In *The Twelfth International Conference on Learning Representations*.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018b. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision*, 286–301.