DREAM: Diffusion Rectification and Estimation-Adaptive Models

Supplementary Material

In this supplementary material, we begin by describing more details of the evaluation metrics and experiment setup in Section 6. In following Section 7, we present more quantitative comparisons and visualization results on various baselines and datasets, which further demonstrates the effectiveness of our DREAM strategy. We conclude with a discussion of the ethical implications in Section 8.

6. Metrics and setups

We provide a more comprehensive explanation of the metrics and the experiment settings employed in the main text of the paper.

6.1. Metrics

In this section, we will detail the metrics applied to measure image distortion and perception quality. The distortion metrics encompass Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), as well as Consistency the the perception measurement include the Learned Perceptual Image Patch Similarity (LPIPS) and the Fréchet Inception Distance (FID).

Peak Signal-to-Noise Ratio (PSNR). PSNR is an indicator of image reconstruction quality. However, its value in decibels (dB) presents certain constraints when assessing super-resolution tasks [36]. Thus, it acts merely as a referential metric of image quality, comparing the maximum possible signal to the level of background noise. Generally, a higher PSNR suggests a lower degree of image distortion.

Structure Similarity Index Measure (SSIM). Building on the image distortion modeling framework [58], the SSIM applies the principles of structural similarity, mirroring the functionality of the human visual system. It is adept at detecting local structural alterations within an image. SSIM measures image attributes such as luminance, contrast, and structure by employing the mean for luminance assessment, variance for contrast evaluation, and covariance to gauge structural integrity.

Consistency. Consistency is measured by calculating the MSE ($\times 10^{-5}$) between the low-resolution inputs and their corresponding downsampled super-resolution outputs.

Learned Perceptual Image Patch Similarity (LPIPS). LPIPS evaluates the perceptual resemblance between generated images and their authentic counterparts by analyzing deep feature representations.

Fréchet Inception Distance score (FID). FID [18] assesses image quality by emulating human judgment of image resemblance. This is achieved by utilizing a pre-trained Inception-V3 network [51] to contrast the distribution patterns of the generated images against the distributions of the original, ground-truth images.

6.2. Setups

In this section, we will provide detailed descriptions of the configurations for various baseline models as well as the datasets utilized in our experiments.

SR3 model on face dataset. We train the SR3 [44] model on an upscaled $8\times$ FFHQ dataset for 1M iterations and evaluate on 100 images from the CelebA [23] validation dataset. During training, the LR images are consistently resized to 16×16 pixels, while the HR counterparts are scaled to 128×128 pixels. For the SR image generation, the LR images are first upscaled to 128×128 pixels using bicubic interpolation and serve as the conditioning input. In alignment with the DDPM [19], the Adam optimizer is utilized with a fixed learning rate of 1e-4 through the training phase. The training employs a batch size of 4, incorporates a dropout rate of 0.2, and utilizes a linear beta scheduler over 2000 steps with a starting value of 10^{-6} and a final value of 10^{-2}. A single 24GB NVIDIA RTX A5000 GPU is used under this situation.

IDM model on face dataset. Adhering to the offical implementation of the IDM [14], the model is trained on a $8\times$ FFHQ dataset for 1M iterations and evaluated on 100 images from the CelebA [23] validation dataset. Specifically, throughout training, LR images are consistently resized to 16×16 pixels, while their HR counterparts are scaled to 128×128 pixels. These LR images are then processed through a specialized LR conditioning network, which is stacked with a series of convolutional layers, bilinear downsampling filtering, and leaky ReLU activation to extract a hierarchy of multi-resolution features. These features are then employed as the conditioning input for the denoising network. The training employs the Adam optimizer with a constant learning rate of 10^{-4}, a batch size of 32, and a dropout rate of 0.2. We implement a linear beta scheduler that advances over 2000 steps, starting from 10^{-6} and escalating to 10^{-2}. This setup is supported by two 24GB NVIDIA RTX A5000.

SR3 model on general scene dataset. We train the SR3 [44] model on upscaled $4\times$ the training dataset comparising DIV2K [1] and Flicker2K [52] for 1M iterations. Consistent with the SRDiff [29], each image is cropped into patches of 160×160 as the HR ground truths. To produce the corresponding LR image patches of 40×40 pixels, the HR image patches are downscaled using a bicubic kernel. These LR image patches are then resized back to the HR dimensions using bicubic interpolation and are used as

Table 5. Comparison of training speed and memory usage. The values denote the ratio of DREAM/standard.

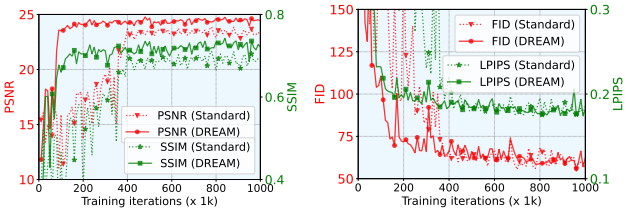
	Face		DIV2K	
	SR3	IDM	SR3	ResShift
Training time	1.38	1.21	1.24	1.08
Training memory	1.06	1.11	1.09	1.13

the conditioning input for the super-resolution process. For evaluation, the entire DIV2K validation set, consisting of 100 images, is utilized. The HR images are downsampled using a bicubic kernel to generate LR images, which are then cropped into 40×40 pixel patches with a 5-pixel overlap between adjacent patches. The SR3 model is applied to these LR patches to yield the SR predictions which are subsequently merged to form the final SR images. The model's training utilizes the Adam optimizer with a steady learning rate of 10^{-4}, a batch size of 32 patches, and a dropout rate of 0.2. A linear beta scheduler is applied over 1000 steps, initiating at 10^{-6} and culminating at 10^{-2}. This configuration is executed on two 24GB NVIDIA RTX A5000 GPUs.

ResShift on general scene datatset. Training the ResShift model [62]uses a $4\times$ dataset, combining the training sets from DIV2K [1] and Flickr2K [52] over 0.5M iterations. Similar as data process in the previous SR3 setting, each image is partitioned into patches of 256x256 pixels to serve as HR ground truths. The LR image patches, resized to 64x64 pixels, are derived by downscaling the HR patches with a bicubic kernel. The VQGAN encoder, pre-trained on the ImageNet dataset, processes these LR patches to distill salient features, furnishing the necessary conditioning input for the following latent denoiser network. For performance evaluation, we use the entire DIV2K validation set, which comprises 100 images. The HR images are downsampled to LR with a bicubic kernel, and then segmented into 64x64 pixel patches, maintaining an 8-pixel overlap between adjacent patches. The latent denoiser model is applied to the LR patches to generate the corresponding SR latent codes. These latent codes are subsequently processed by the VQGAN decoder to reconstruct the SR patches, thereby producing the final high-resolution super-resolution images. The training regimen employs the Adam optimizer with a consistent learning rate of 5×10^{-5} and a batch size of 32 patches. A linear beta scheduler is utilized over 50 steps, selected evenly from a linearly spaced 2000-steps schedule beginning at 10^{-6} and increasing to 10^{-2}. The training is conducted using two 24GB NVIDIA RTX A5000.

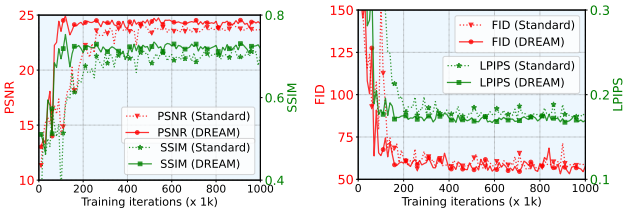
7. Additional experimental results

In this section, we begin by providing additional results on the acceleration of training and sampling across various baselines and datasets in Section 7.1. Lastly, in Section 7.2, we offer a more comprehensive visual comparison on the general scene dataset, using the SR3 [44] and ResShift [62] models as baselines.



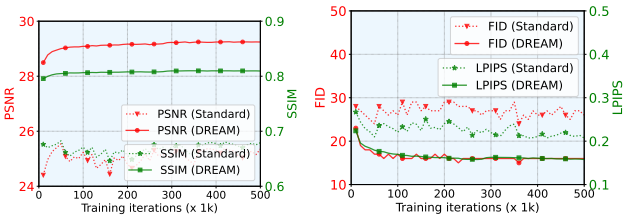
(a) Distortion (b) Perception

Figure 10. Evolution of distortion metrics (left) and perceptual metrics (right) using SR3 as a baseline on the face dataset.



(a) Distortion (b) Perception

Figure 11. Evolution of distortion metrics (left) and perceptual metrics (right) using IDM as a baseline on the face dataset.

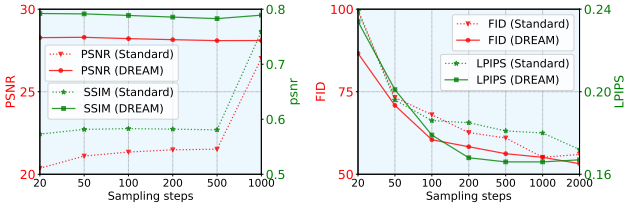


(a) Distortion (b) Perception

Figure 12. Evolution of distortion metrics (left) and perceptual metrics (right) using ResShift as a baseline on the DIV2K dataset.

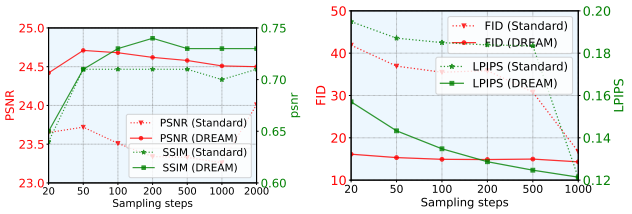
7.1. Training and sampling acceleration

Training efficiency. In Table 5, we detail the relative ratio of training speed and memory usage between our DREAM methodology and standard training approaches across a variety of baselines and datasets. Our DREAM method, which includes only a single additional forward computation, results in a marginal increase in training time (around $1.1 \sim 1.4\times$) and memory usage (approximately $1.05 \sim 1.15\times$). However, it offers a considerable advantage by significantly accelerating training convergence. We further illustrate the evolution of training through distortion metrics, namely PSNR and SSIM, as well as perception metrics such as LPIPS and FID. Utilizing SR3 and IDM as baselines for the face dataset, the improvements are evident in Figure 10 and Figure 11. The ResShift model, used as a baseline for the DIV2K dataset, demonstrates similar enhancements in Figure 12. Notably, DREAM not only facilitates quicker convergence but also outperforms the final outcomes of several baselines after they fully converge. For example, with the face dataset, the SR3 model using DREAM achieves a PSNR of 24.49 and an FID of 61.02 in just 490k iterations, whereas the standard diffusion baseline reaches a PSNR of



(a) Distortion (b) Perception

Figure 13. Comparison of distortion metrics (left) and perception metrics (right) with varying sampling steps, using IDM as a baseline on the CelebA-HQ dataset.



(a) Distortion (b) Perception

Figure 14. Comparison of distortion metrics (left) and perception metrics (right) with varying sampling steps, using SR3 as a baseline on the DIV2K dataset.

23.85 and an FID of 61.98 after 880k iterations. This underlines a substantial training speedup by roughly 2× with DREAM. Similarly, the IDM model with DREAM reaches a PSNR of 23.54 and an FID of 55.81 in only 330k iterations, compared to the baseline achieving a PSNR of 23.85 and an FID of 61.98 after 760k iterations, reinforcing the significant efficiency of DREAM.

Sampling acceleration. Furthermore, DREAM significantly enhances the efficiency of the sampling process, surpassing the performance of standard diffusion training with a reduced number of sampling steps. Figure 13 showcases the capabilities of DREAM using the IDM model on the CelebA-HQ dataset. It compares super-resolution images generated with different numbers of sampling steps, evaluating them against both distortion and perception metrics. While the conventional baseline necessitates up to 2000 sampling steps, DREAM attains superior distortion metrics (an SSIM of 0.73 compared to 0.71) and comparable perceptual quality (an LPIPS of 0.179 versus 0.172) with merely 100 steps, leading to an impressive 20× increase in sampling efficiency. In a similar vein, Figure 14a illustrates the impact of DREAM using the SR3 model on the DIV2K dataset. Here, the images produced with varying sampling steps are again evaluated using both sets of metrics. Standard baselines typically require 1000 sampling steps, but with DREAM, improved distortion metrics (an SSIM of 0.79 versus 0.76) and similar perceptual quality (an LPIPS of 0.127 versus 0.121) are achieved with just 100 steps, resulting in a substantial 10× sampling speedup.

(a) LR (b) Standard (c) DREAM (d) HR

Figure 15. Qualitative comparison for 8× SR using SR3 [44] on the CelebA-HQ dataset [23]. Results highlight DREAM's superior fidelity and enhanced identity preservation, leading to more realistic details, such as eye and teeth.

(a) LR (b) Standard (c) DREAM (d) HR

Figure 16. Qualitative comparison for 8× SR using IDM [14] on the CelebA-HQ dataset [23]. Results highlight DREAM's superior fidelity and enhanced identity preservation, leading to more realistic detail generation in features like nose, and wrinkles.

7.2. Visualization

Face dataset. In Figure 15 and Figure 16, we provide more representative examples from CelebA-HQ [23], employing SR3 and IDM as baselines, respectively.

General scene dataset. To further illustrate the effectiveness of our DREAM strategy, we present selected examples from the DIV2K [1]. These examples showcase complex image elements such as intricate textures, repeated symbols, and distinct objects. We conduct a comparative visualization of our DREAM strategy against standard training practices, employing the SR3 model as a baseline in Figure 17, Figure 18 and Figure 19. Similarly, we use the ResShift model as a baseline in Figure 20, Figure 21 and Figure 22.

All these comparisons unequivocally demonstrate the superior performance of our DREAM strategy.

8. Ethic impact

This research is applicable to the task of enhancing human facial resolution, a frequent requirement in mobile photography. It does not inherently contribute to negative social consequences. However, given personal security concerns, it is crucial to safeguard against its potential misconduction.

(a) Standard (b) DREAM

Figure 17. Qualitative comparison for $4\times$ SR on DIV2K [1] using SR3 [44] model as baseline. **Left Image:** standard training; **Right Image:** DREAM training. The model trained under DREAM framework exhibits enhanced fine-grained details and rendering more realistic results, as indicated by the magnified section of the synthesized SR images.

(a) Standard (b) DREAM

Figure 18. Qualitative comparison for $4\times$ SR on DIV2K [1] using SR3 [44] model as baseline. **Left Image:** standard training; **Right Image:** DREAM training. The model trained under DREAM framework exhibits enhanced fine-grained details and rendering more realistic results, as indicated by the magnified section of the synthesized SR images.

(a) Standard (b) DREAM

Figure 19. Qualitative comparison for $4\times$ SR on DIV2K [1] using SR3 [44] model as baseline. **Left Image:** standard training; **Right Image:** DREAM training. The model trained under DREAM framework exhibits enhanced fine-grained details and rendering more realistic results, as indicated by the magnified section of the synthesized SR images.

(a) Standard (b) DREAM

Figure 20. Qualitative comparison for 4× SR on DIV2K [1] using ResShift [62] model as baseline. **Left Image:** standard training; **Right Image:** DREAM training. The model trained under DREAM framework exhibits enhanced fine-grained details and rendering more realistic results, as indicated by the magnified section of the synthesized SR images.

(a) Standard (b) DREAM

Figure 21. Qualitative comparison for 4× SR on DIV2K [1] using ResShift [62] model as baseline. **Left Image:** standard training; **Right Image:** DREAM training. The model trained under DREAM framework exhibits enhanced fine-grained details and rendering more realistic results, as indicated by the magnified section of the synthesized SR images.

(a) Standard (b) DREAM

Figure 22. Qualitative comparison for 4× SR on DIV2K [1] using ResShift [62] model as baseline. **Left Image:** standard training; **Right Image:** DREAM training. The model trained under DREAM framework exhibits enhanced fine-grained details and rendering more realistic results, as indicated by the magnified section of the synthesized SR images.