Diffusion Model for Image Restoration

Chih-Ying Liu

Abstract—We experiment three different diffusion based methods on image restoration tasks, including a supervised method SRLDM [1], an unsupervised method DDRM [2], and a combination of two DDRM-SRLDM. We find that DDRM outperforms other methods at most of the tasks. SRLDM is applicable to image restoration tasks other than super resolution, and is much faster than DDRM. In addition, we discuss different possibilities of combining the DDRM pipeline with latent diffusion models.

Index Terms—diffusion model

1 Introduction

IMAGE restoration tasks, such as super resolution, deblurring, denosing, inpainting, colorization, are instances of inverse problems, where we want to recover original images given degraded measurements. These problems are inherently ill-posed because multiple output images may be consistent with the input image. With the rapid development of deep learning techniques, various deep learning based super-resolution models have been proposed. Previous work explored techniques ranging from simple regression-based methods with feedforward convolutional network [3] to advanced generative techniques including generative adversarial nets (GAN) [4], variational autoencoders (VAEs) [5] and normalizing flows (NFs) [6]. However, existing techniques often suffer from various limitations: regression-based method is computationally costly for high-resolution image generation, GANs suffer instability training and mode collapse, and NFs and VAEs yield suboptimal generation quality. Diffusion model, which iteratively refine noisy input image into a high-quality image, has been shown to have the capability to generate high quality samples [7] and achieve impressive results in various image synthesis tasks [1], [8], [9].

One of the limitations of diffusion models is that it is computationally demanding, since it requires repeat computation in the high-dimensional pixel space. Evaluating a trained model requires a large number of refinement steps. Latent Diffusion Models [1] reduces the computational complexity by using an autoencoder to project images into lower-dimensional latent space, and perform diffusion processes there. The model can take inputs of different modalities and be supervised fine-tuned for each image restoration task.

On the other hand, unsupervised approaches based on prior diffusion models, which are applicable to any inverse problems is another interesting topic. DDRM [2] falls in this category. The authors derived a variational inference objective that can be solved by an unconditional denoising diffusion generative model. By using this objective, the pipeline can be applied to any degradation models without retraining.

The goal of this project is to explore several diffusion model based methods on image restoration tasks, including super resolution, denoising, and deblurring. We experiment with a supervised method: SRLDM (a LDM supervised trained for super resolution) and an unsupervised method: DDRM. We also merge the two methods by using a supervised fine-tuned SRLDM as a denoising diffusion generative model in the DDRM pipeline.

2 Related Work

2.1 Diffusion Model

Diffusion model [10] is a probabilistic modeling that is able to convert a simple known distribution into a target distribution and is widely used in image generation. The algorithm assumes that there is a Markov chain that iteratively adds some noise to its input in the forward diffusion process, and the diffusion model is trained to reverse this chain.

2.2 Supervised Diffusion Model

Using pairs of original and degraded images, diffusion models can be supervised trained end-to-end for a specific image restoration task. For example, SR3 [8] applies the diffusion framework to image super-resolution tasks. It adopts UNet to model a stochastic iterative denoising process and iteratively refine input images by predicting the addition noise at each step. Since a default inference process requires thousands of refinement steps, SR3 introduces a noise schedule, allowing a trade-off between image quality and efficiency.

To further speed up training and testing, Latent diffusion Model (LDM) [1] applies diffusion to the latent space of a powerful pre-trained autoencoder. It also designs a general-purpose conditioning mechanism based on cross-attention, which enables multiple-modal training. It shows success in multiple image synthesis tasks, including inpainting, class-conditional image synthesis, unconditional image generation, text-to-image synthesis, and super resolution.

2.3 Unsupervised Diffusion Model

Real world applications often require flexibility to deal with multiple degradation models. Unsupervised approaches, which are applicable to all image restoration tasks, are desirable in this case. In the unsupervised setting, we have a set of clean images during training, but the degradation

models are only known during inference. This setup is inherently general to all linear inverse problems.

Common unsupervised approaches learn a prior-related model based on the clean images and a likelihood term from the degradation model. They combine the two terms to form the posterior of the samples. SNIPS [11] and RED [12] use denoisers as a part of their iterative optimization method. DGP [13] captures more high-level semantics by exploiting the latent space of a Generative Adversarial Network (GAN) trained on large-scale natural images, resulting in more faithful image restoration.

The aforementioned unsupervised approaches use iterative optimization algorithms to solve the posterior, which is time-consuming and requires hyper-parameter tuning. DDRM [2] overcomes this drawback by introducing a variational inference objective for learning the posterior distribution and shows that the objective can be solved by using a general denoising diffusion generative model. DDRM produces high-quality and diverse samples while being more efficient than other iterative unsupervised methods.

3 METHOD

To understand how different supervised and unsupervised settings and diffusion mechanism differs in performance and speed, we experiment with a supervised method: SRLDM (a LDM supervised fine-tuned on super resolution tasks) [1], an unsupervised method: DDRM [2], and a combined of two methods: DDRM-SRLDM.

3.1 SRLDM

Diffusion Models

Diffusion Models are probabilistic models that learn the data distribution by gradually denoise the noisy inputs. The learnt joint distribution $p_\theta(x_{0:T})$ is called a reverse process, where x_0 is the final clean image prediction and $x_{1:T}$ are intermediate results. The reverse process is defined by a fixed posterior $q(x_{1:T}|x_0)$, called forward process or diffusion process, that gradually adds a Gaussian with predefined variance $\beta_1, \ldots, \beta_T$ to the inputs. Most image synthesis diffusion models predict the noise ϵ_t given a noisy input x_t, and recover the denoise output x_{t-1} by pre-defined noise scheduler and simple computation. The corresponding objective can be expressed as

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\big[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\big],$$

where ϵ_t is the learned model and t is uniformly sampled from $\{1, ..., T\}$.

Latent Diffusion Models

Diffusion process in high-dimensional pixel space is computationally heavy. Latent Diffusion Models (LDM) [1] leverages a pre-trained autoencoder consisting of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ to compress images into a lower-dimensional latent space. This space focuses more on semantic meaningful data and the diffusion process in this space is more computationally efficient. The objective becomes

$$L_{LDM} = \mathbb{E}_{\mathcal{E},\epsilon\sim\mathcal{N}(0,1),t}\big[\|\epsilon - \epsilon_\theta(z_t, t)\|^2\big],$$

Where the latent vector z_t is obtained by $z_t = \mathcal{E}(x_t)$ and the diffusion model ϵ_t is a time conditional UNet.

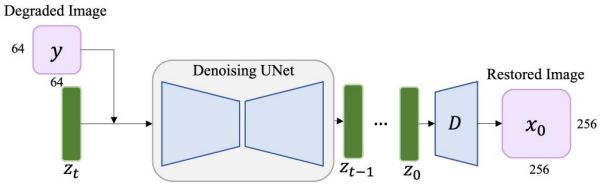

Fig. 1. SRLDM architecture. This is a LDM fine-tuned for 64 $\rightarrow$ 256 super resolution task. The denoising diffusion model takes the concatenation of the down-scaled image y and the latent vector z_t as input to predict the denoised latent vector z_{t-1}. The diffusion process is repeated until we get the final clean latent vector z_0, then the restored image is obtained from $x_0 = \mathcal{D}(z_0)$

Supervised Training

Diffusion models can solve image restoration tasks by modeling conditional distributions $p_\theta(z|y)$, where y is an input modality such as text, semantic maps, and degraded images. LDM [1] augments underlying UNet backbone with a cross-attention mechanism, which is effective for different modalities inputs. An attention-based domain specific encoder projects various input modalities y into an intermediate representation space. The resulting modality vector $\tau_\theta(y)$ is concatenated with the latent vector z_t and togetherly fed into the denoising diffusion model ϵ_θ.

For image-to-image translation tasks, such as super resolution, the conditioning method is more straightforward. The low-resolution image is directly concatenated with the latent vector z_t and togetherly fed into the denoising model. Figure 1 illustrates the framework for LDM fine-tuned for super resolution tasks. We use the weights of a fine-tuned SRLDM (an LDM model fine-tuned on a 4x super resolution task) from the LDM official repository. For other image restoration tasks, such as denoising and deblurring, we down-scale the degraded image four times and follow the same pipeline.

3.2 DDRM

Inverse Linear Problem

DDRM considers image restoration tasks as linear inverse problems:

$$y = Hx + z,$$

where we aim to recover the original image x from measurements y, where H is a known linear degradation model and z is some noise.

Variational Objective

DDRM learns a conditional reverse process $p_\theta(x_{0:T}|y)$, where x_0 is the final clean image prediction, $x_{1:T}$ are intermediate results, and y is the degraded image. The underlying reverse and forward Markov chain are similar with the original diffusion models, with an extra conditioning on the degraded image y.

Diffusion Process for Image Restoration

To solve the variational objective efficiently, DDRM considers the singular decomposition (SVD) of H, and performs a

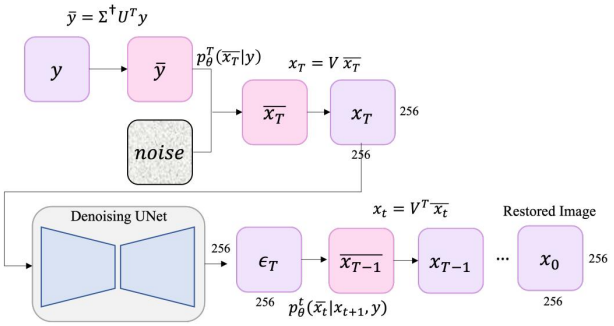

Fig. 2. DDRM architecture. The degraded image y is projected into the spectral space by $\bar{y} = \Sigma^\dagger U^T y$. Then, the initial iterative variable $\bar{x_T}$ is obtained form the variational distribution $p_\theta^T(\bar{x_T}|y)$. Project $\bar{x_T}$ back to pixel space and feed x_T into a pre-trained denosier to get the predicted noise ϵ_t. Finally, We get the refined spectral variable $\bar{x_{T-1}}$ by the variational distribution $p_\theta^t(\bar{x_t}|x_{t+1}, y)$, project it to pixel space, and repeat the process.

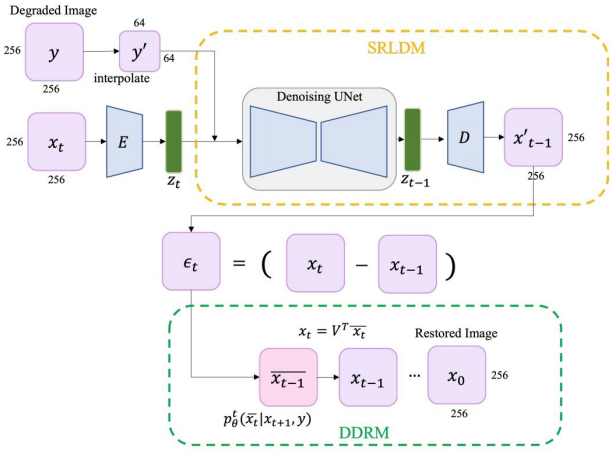

Fig. 3. DDRM-SRLDM architecture. The degraded image y is down-scaled to resoluton 64x64, and the iterative variable x_t is projected to latent space by the encoder $\mathcal{E}$. The SRLDM takes the concatenation of the down-scaled degraded image y' and latent vector z_t as input and predict the refined latent vector z_{t-1}. The error prediction ϵ_t is obtained by $x_t - \mathcal{D}(z_{t-1})$. Finally, using ϵ_t, we follow the DDRM pipeline to get the refined variable x_{t-1}. The variable intialization of DDRM is ignored in this figure.

diffusion process in its spectral space. Given a degradation matrix,

$$H = U\Sigma V^T$$

We project the iterative variable x_t and measurement y into its spectral space by

$$\bar{x_t} = V^T x_t,$$

and

$$\bar{y} = \Sigma^\dagger U^T y$$

We have $\bar{x_t} = \bar{y}$, and can derive the variational distribution as follows:

$$p_\theta^{(T)}(\bar{x_T}^{(i)}|y) = \begin{cases} \mathcal{N}(\bar{y}^{(i)}, \sigma_T^2 - \frac{\sigma_y^2}{s_i^2}) & s_i > 0 \\ \mathcal{N}(0, \sigma_T^2) & s_i = 0 \end{cases}$$

$$p_\theta^{(t)}(\bar{x_t}^{(i)}|x_{t+1}, y)$$
$$= \begin{cases} \mathcal{N}(\bar{x_{\theta,t}}^{(i)} + \sqrt{1-\eta^2}\sigma_t \frac{\bar{x_{t+1}}^{(i)} - \bar{x_{\theta,t}}^{(i)}}{\sigma_{t+1}}, \eta^2\sigma_t^2) & s_i = 0 \\ \mathcal{N}(\bar{x_{\theta,t}}^{(i)} + \sqrt{1-\eta^2}\sigma_t \frac{\bar{y}^{(i)} - \bar{x_{\theta,t}}^{(i)}}{\sigma_y/s_i}, \eta^2\sigma_t^2) & \sigma_t < \frac{\sigma_y}{s_i} \\ \mathcal{N}((1-\eta_b)\bar{x_{\theta,t}}^{(i)} + \eta_b\bar{y}^{(i)}, \sigma_t^2 - \frac{\sigma_y^2}{s_i^2}\eta_b^2) & \sigma_t \leq \frac{\sigma_y}{s_i} \end{cases},$$

where $\bar{x_{\theta,t}}^{(i)}$ is an uncoditional denoising diffusion model.

DDRM is unsupervised in the sense that it can use the same denoising diffusion model for different image restoration tasks. It only need to know the degradation model H during inference. Figure 2 illustrates DDRM pipeline. We uses the H functions implementation from the official DDRM repository and pre-trained denosing weight from OpenAI.

3.3 DDRM-SRLDM

Since the SRLDM model has knowledge about image distribution and our preliminary experiments show that it works for simple denoising tasks, we combine the two methods by using the SRLDM model as a diffusion generative model in the DDRM pipeline.

The challenge of this method is how to get ϵ_t, the noise prediction in pixel space. We found that decoding noise using decoder $\mathcal{D}$ and using this term in the DDRM pipeline results in complete noise generation. A possible reason is that the decoder is trained for decoding natural images, and thus it can't recognize noise. To overcome this issue, we first convert the predicted noise in latent space into the refined latent variable z_{t-1} by a noise scheduler, obtain x'_{t-1} by decoding z_{t-1} into pixel space, then compute ϵ_t by taking the difference between x_t and x'_{t-1}. Finally, we use the variational distribution defined by the DDRM pipeline to get the refined iterative variable x_{t-1} and repeat the process. It is noteworthy this method requires encoding and decoding at every refinement step and computing variational distribution in the pixel space, which loses the efficiency of performing diffusion in the latent space. In addition, following the design of SRLDM, we downscale degraded images to resolution 64x64 and encode the iterative variable x_t into latent vector z_t by encoder $\mathcal{E}$.

4 Comparison of Methods

The three methods have two different components to compare. First, SRLDM performs diffusion in latent space, while DDRM and DDRM-SRLDM performs diffusion in pixel space. Therefore, SRLDM should be more computationally efficient. Secondly, SRLDM is supervised fine-tuned for super resolution tasks, while DDRM and DDRM-SRLDM are unsupervised, and thus can be applied to any inverse linear problems. This makes DDRM more flexible to applications. In this project, we also explore how SRLDM performs on image restoration tasks other than super resolution.

5 Results

5.1 Experiment Details

Data

We evaluated the three methods on 2000 randomly sampled images from ImageNet-1K validation dataset. We crop images at their longer sides and resize into resolution 256x256.

Task

We experiment with four image restoration tasks: super resolution ($64 \rightarrow 256$), deblurring ($\sigma = 1$), denosing ($\sigma = 0.1$), and noisy super resolution ($64 \rightarrow 256$, $\sigma = 0.05$).

Experiment Setting

We use diffusion models from LDM repository for SRLDM, which is trained on pair low-resolution and high-resolution images from ImageNet-1K dataset. For the DDRM pipeline, we use fine-tuned weights from OpenAI guided diffusion repository for diffusion denoising model, which is trained on images with resolution 256x256 from ImageNet-1k. If not specially mentioned, number of inference steps is set to 100 for each experiment. All experiments are conducted on an A4000 GPU.

5.2 Quantitative Results

We report the average peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) to measure faithfulness to the original image, and Fréchet inception distance (FID) to measure the resulting image quality. We also measure the inference time, and report throughput (sample/ second). In addition, we define the baseline as bicubic upscaling interpolation for super-resolution tasks, and the degraded image itself for other restoration tasks, and report quantitative scores of the baseline.

Table 1, 2, 3, 4 presents the quantitative results. DDRM outperforms SRLDM and DDRM-SRLDM at super resolution, denosing, and noisy super resolution tasks. It is especially good at denoising, due to the fact that this task is exactly what the denoising model in the DDRM pipeline does.

For deblurring tasks, SRLDM surprisingly outperforms DDRM , while DDRM produces ripple artifacts as we will show in Section 5.3. Though SRLDM doesn't achieve the best score, it produces reasonable outputs on all tasks. It is noteworthy that SRLDM is about 16x faster than DDRM, because it performs diffusion processes at lower-dimensional space and doesn't require variational distribution computation. Also, DDRM has access to the exact degradation model while SRLDM does not.

Lastly, DDRM-SRLDM has the worst performance on all tasks. Further analysis of this method will be presented in Section 6.

5.3 Qualitative Results

Figure 4 shows the generation results as well as original and degraded images of different tasks and methods. Both SRLDM and DDRM produce high-quality results over all tasks. Following the quantitative result on deblurring, SRLDM successfully deblur images, while DDRM produces small ripple artifacts. On the other hand, DDRM-SRLDM generates obvious artifacts and color distortion.

TABLE 1
Results of 4x super resolution ($64 \rightarrow 256$).

Method	PSNR ↑	SSIM ↑	FID ↓	Throughput ↑
Baseline	25.8	0.745	66.3	-
SRLDM	23.7	0.686	**29.9**	**1.6**
DDRM	**27.4**	**0.799**	**29.6**	0.1
DDRM-SRLDM	14.68	0.141	135	0.1

TABLE 2
Results of gaussain deblurring ($\sigma = 1$).

Method	PSNR ↑	SSIM ↑	FID ↓	Throughput ↑
Baseline	25.5	0.778	27.8	-
SRLDM	**22.8**	**0.664**	35.6	**1.6**
DDRM	20.1	0.384	**34.5**	0.1
DDRM-SRLDM	18.2	0.309	45.7	0.1

TABLE 3
Results of denoising ($\sigma = 0.1$).

Method	PSNR ↑	SSIM ↑	FID ↓	Throughput ↑
Baseline	26.3	0.634	17.47	-
SRLDM	23.7	0.657	29.9	**1.6**
DDRM	**34.8**	**0.932**	**8.94**	0.1
DDRM-SRLDM	22.1	0.588	36.5	0.1

TABLE 4
Results of noisy image super resolution ($64 \rightarrow 256$, $\sigma = 0.05$).

Method	PSNR ↑	SSIM ↑	FID ↓	Throughput ↑
Baseline	24.9	0.6603	97.12	-
SRLDM	22.7	0.553	38.3	**1.6**
DDRM	**26.6**	**0.759**	36.66	0.1
DDRM-SRLDM	13.4	0.109	168	0.1

6 Discussion

6.1 Analysis of DDRM-SRLDM pipeline

The main challenge of using SRLDM in the DDRM pipeline is the transformation between pixel space and latent space. We have experimented with two strategies for the pixel-latent space transformation as depicted in Figure 5. The first one is exactly the same as what we described in Section 3.3, while the second one obtains ϵ_t by decoding the predicted noise latent vector n_t. We found that the latter method generates complete noise, and adopt the first method. From this observation, we infer that noise isn't transformed properly forth and back by the encoder and the decoder. Since DDRM performs diffusion in pixel space, we have to transform between latent and pixel space at every time step. Therefore, the transformation may result in some artifacts in the intermediate noisy variable x_t.

To further understand the diffusion mechanism, we look into the intermediate results x_t of the SRLDM, DDRM and DDRM-SRLDM pipelines as depicted in Figure 6. The noise

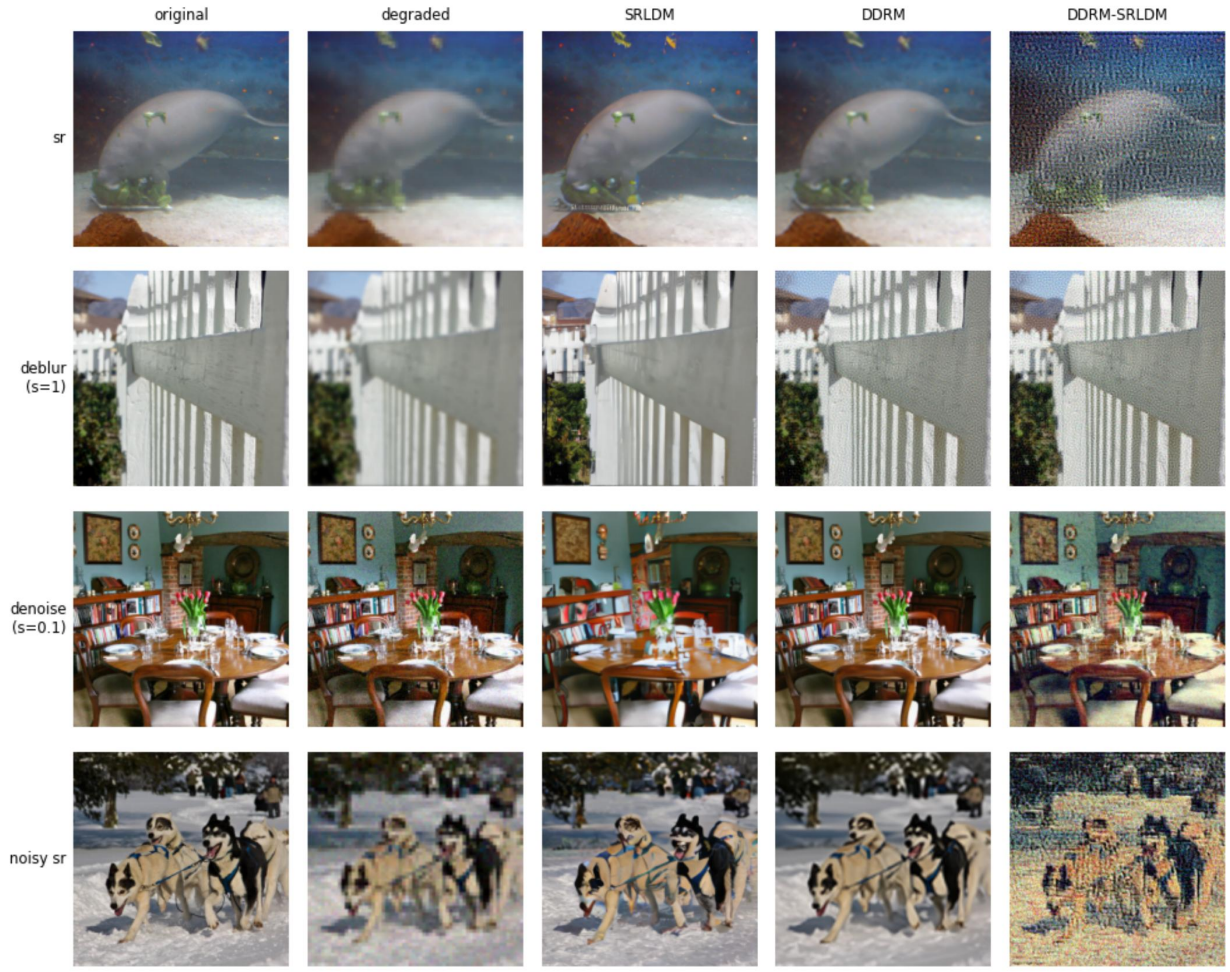

Fig. 4. Generation result. Include original imgaes, degraded images, outputs of SRLDM, DDRM, and SRLDM over 4 different tasks: super resolution ($64 \rightarrow 256$), deblurring ($\sigma = 1$), denosing ($\sigma = 0.1$), and noisy super resolution ($64 \rightarrow 256$, $\sigma = 0.05$)

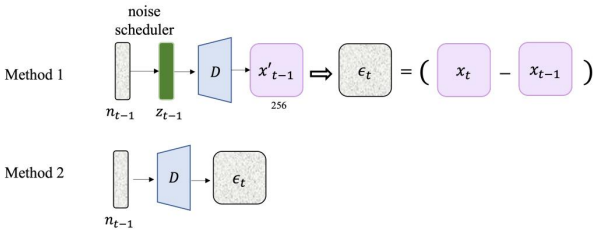

Fig. 5. Two different methods for transforming between pixel and latent space for the DDRM-SRLDM pipeline. n_t is the output of denoising UNet, and ϵ_t is required by the DDRM pipeline.

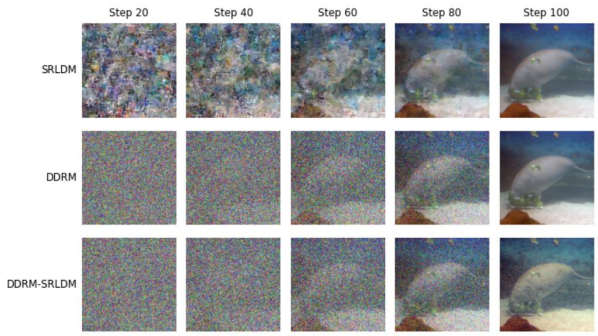

Fig. 6. Intermediate results of SRLDM, DDRM, and DDRM-SRLDM. All experiments are run with 100 time steps. The initial time step is zero, and the final time step is 100.

of the SRLDM looks very different from the noise of the DDRM. Therefore, when using the SRLDM model in the DDRM pipeline, it may fail to recognize the noise, resulting in low performance.

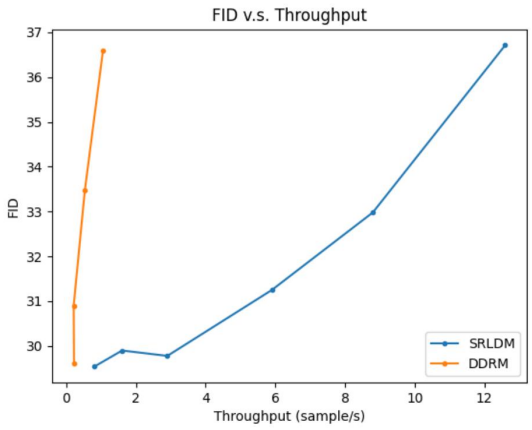

Fig. 7. FID v.s throughput for SRLDM and DDRM pieplines.

6.2 FID v.s. Throughput

We can tradeoff between resulting image quality and inference time by adjusting the number of inference steps. We ran SRLDM with timesteps {5, 10, 20, 50, 100, 200} and DDRM with timesteps {10, 20, 50, 100}, and plot the FID score versus throughput (sample/second) in Figure 7. The result shows that we get higher generation quality with lower throughput, which follows our intuition about the tradeoff. Also, we observe that both methods can generate high-quality images with a small number of inference steps ($\sim$10). Last but not least, SRLDM is much faster than DDRM, since it performs the diffusion process in the latent space.

7 Conclusion

We have experimented with three diffusion based methods: SRLDM, DDRM, and DDRM-SRLDM. We found that DDRM achieved the highest score in most tasks, but it is 16 times slower than SRLDM. On the other hand, SRLDM generates good results much faster. SRLDM can be applied to tasks other than super resolution, and surprisingly achieves the best result in the deblurring task. Lastly, DDRM-SRLDM generates huge artifacts, which we infer that the failure is due to the transformation between pixel and latent space. We leave exploring a better combination of DDRM of latent diffusion models as future work.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022.
[2] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," 2022.
[3] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," 2015.
[4] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," 2017.
[5] A. Vahdat and J. Kautz, "Nvae: A deep hierarchical variational autoencoder," 2021.
[6] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," 2018.
[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," 2020.
[8] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," 2021.
[9] S. Gao, X. Liu, B. Zeng, S. Xu, Y. Li, X. Luo, J. Liu, X. Zhen, and B. Zhang, "Implicit diffusion models for continuous super-resolution," 2023.
[10] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.
[11] B. Kawar, G. Vaksman, and M. Elad, "Snips: Solving noisy inverse problems stochastically," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. Curran Associates, Inc., 2021.
[12] E. T. Reehorst and P. Schniter, "Regularization by denoising: Clarifications and new interpretations," 2018.
[13] X. Pan, X. Zhan, B. Dai, D. Lin, C. C. Loy, and P. Luo, "Exploiting deep generative prior for versatile image restoration and manipulation," 2020.