[Re] Denoising Diffusion Restoration Models

Dalim Wahby
dalim@kth.se

Philipp Ahrendt
pcah@kth.se

Iga Pawlak
ipawlak@kth.se

Abstract

In this work we aim to re-implement the Denoising Diffusion Restoration Models (DDRM), on selected tasks, such as 1) denoising, 2) deblurring, and 3) super-resolution. Furthermore, we investigate the performance of a less-informed priors. Additionally, we introduce a novel metric, that assesses the quality of a generated image, called the GIQA score. Our experiments yield similar results to those of the original paper and we found that a less-informed prior does not necessarily impact the performance of a model, thanks to a multitude of features being present in training data. However, for specialized tasks, the prior should be chosen carefully, since the generated images might lack realism. Furthermore, we found that the GIQA score returns realistic values, with the generated image having a score between the degraded and original score, implicating that it is a relevant measure for the quality of generated images. In conclusion, this work contributes to the understanding of DDRM by re-implementing the method and clarifying unclear passages of the original paper.

1 Introduction

Advancements in imaging technologies have led to the acquisition of large amounts of visual data, however, the existing limitations of imaging systems often introduce unwanted artifacts, including, but not limited to, noise and blurriness. Alongside the massive image data collection, technologies to counteract information loss have been developed [1], to obtain a more clean measurement of reality, in a process called *restoration of images*.

In this work, we will focus on re-implementing a more recent technique from 2022, which was published by Kawar et al. [11], called *Denoising Diffusion Restoration Models* (DDRM). It builds upon the work of Ho et al. [9], using their diffusion models trained using denoising score matching, for solving various inverse linear problems. We re-implement the proposed logic, using the model pre-trained on the LSUN bedroom dataset by Ho et al. [9], and modify their denoising function. The performance of this model is evaluated on an Out-of-Distribution (OOD) dataset and pitted against the baseline model, using the PSNR, and SSIM metrics.

Furthermore, we implement an additional metric, called Generated Image Quality Assessment (GIQA). This metric aims to quantitatively evaluate the quality of each generated image [8]. Opposed to conventional scores, including but not limited to the PSNR and SSIM scores, the GIQA score is assessed on a single image and gives you a relative score for the quality of the image in question. According to Gu et al. [8], the GIQA score is to be consistent with human assessment. By integrating the new metric, we aim to compare the qualities of the generated image to the corresponding degraded and original images and aim to shed light on how well the proposed method of Kawar et al. performs concerning generating high quality images. All implementations can be found in our GitHub repository[1].

[1] https://github.com/citrovin/denoising-diffusion-restoration-model

2 Background and Related Work

To properly understand the essence of the DDRM paper, we first need to take a look at definitions and existing literature. In this section we aim to elaborate on the necessary background, to understand what differentiates the proposed method by Kawar et al. [11] compared to existing literature. This mainly refers to the restoration of images and diffusion models, which will be introduced in this section.

2.1 Restoration of Images

Restoration of images from observations is an inverse problem, where the measurements are obtained in a forward degradation process which is often non-invertible. For such problems, restoring a unique solution from the observations is not possible without prior knowledge about the data [14]. The measurements $\mathbf{y}$ (with additive noise) are obtained using the model:

$$\mathbf{y} = \mathcal{H}\mathbf{x} + \mathbf{z}, \tag{1}$$

where $\mathbf{z} \sim \mathcal{N}(0, \sigma_{\mathbf{y}}^2\mathbf{I})$ is the noise level and the degradation matrix $\mathcal{H}$ is in general non-linear. Some tasks, however, such as denoising, deblurring, super resolution or inpainting are ill-posed linear inverse problems [6], where $\mathcal{H}$ is a linear matrix, written as $\mathbf{H}$.

The availability of the forward operator $\mathcal{H}$, which may be entirely known, partially known, only known at test time, or completely unknown, will often determine how this problem is solved. Deep Generative models that may be trained either in an unsupervised or supervised manner are frequently used in recent work [14].

The models that learn from input-output pairs $(\mathbf{x}, \mathbf{y})$ with $\mathcal{H}$ fully known are the Neumann networks [6, 10], and Denoising Auto-Encoders [13]. These often have excellent performance, however they are also susceptible to variations and uncertainties in $\mathcal{H}$. For a wider range of situations, frameworks like AUTOMAP [20], which learn with an unknown $\mathcal{H}$, provide for greater flexibility.

When presented with only the ground truth during training, two approaches seem to have gained popularity. Some works are based on Plug-and-Play (PnP) [16] using denoising algorithms as priors, including Regularisation by Denosing (RED) [15] where neural networks are used as denoisers. Others, using Compressed Sensing using Generative models(CSGM) [3] involve learning a generative prior from data. Moreover, models such as DeblurGAN [12] or AmbientGAN [2] (learning only from the output) achieve good results, however, all these types of models are usually difficult to train and computationally expensive, even at inference time [14].

In this project, we re-implement the paper *Denoising Diffusion Restoration Models* (DDRM) [11], where linearity of the degradation matrix $\mathbf{H}$ is assumed to recover the original signal of the measurements without the need for additional training. The paper builds up on Denoising Diffusion Probabilistic Models (DDPM) [9], using their diffusion models trained based on denoising score matching, to solve various inverse linear problems. In particular, the Singular Value Decomposition (SVD) of the $\mathbf{H}$ matrix allows for modifications of the reverse diffusion process, to perform it in the matrix's spectral space.

2.2 Diffusion Probabilistic Models

Diffusion models are a family of generative models, that progressively inject noise into data (in the forward diffusion) and learn the reverse diffusion process to generate a sample from noise [19]. They have outperformed GANs in terms of quality of generated samples, hence becoming the new state-of-the-art [5].

In classical diffusion models, the forward process is a Markov chain, that gradually adds Gaussian noise to the input, according to a β-schedule (variance schedule), $\beta_1, ..., \beta_T$ [9]. With the original image denoted as $\mathbf{x}_0$ and T time steps we get:

$$q(\mathbf{x}_{1:T}|x_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \qquad\qquad q(\mathbf{x}_t|\mathbf{x}_{t-1}) := (\mathbf{x}_t, \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

We can directly sample $\mathbf{x}_t$ at a given time step, according to:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_t$.

After degrading the images, the new goal is to recreate the original image. To this end, the transitions in the reversed process, are learned Gaussians:

$$p(x_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}; \mathbf{I}) \qquad p_\theta(\mathbf{x}_{t-1}, \mathbf{x}_t) := \mathcal{N}(x_{t-1}; \mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)). \tag{4}$$

Denoising Diffusion Probabilistic Models [9] define $\Sigma_\theta(\mathbf{x}_t, t)) = \sigma_t^2 \mathbf{I}$, with σ_t dependent on β_t (or equivalently α_t). The model aims to predict the mean of $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ from $\mathbf{x}_t$. With the reparametrisation of (3) as:

$$\mathbf{x}_t(\mathbf{x}_0, \epsilon) = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{(1 - \bar{\alpha}_t)}\epsilon. \tag{5}$$

Moreover, the model can predict $\epsilon_\theta(\mathbf{x}_t, t)$ and then calculate the mean based on this value. This approach has provided the best results while using a simplified objective function.

3 Methodology

With DDPM as the starting point we move on to the description of the formulas proposed by Kawar et al. on the diffusion steps as well as the scaling necessary to use pre-trained models with this formulation. This section also contains information about the metrics used for evaluation of the technique, namely PSNR and SSIM as well as a novel metric, the GIQA [8] score.

3.1 Denoising Diffusion Restoration Models

In DDRM [11], instead of static noise, the input of the model is a measurement, degraded according to (1) with a linear $\mathbf{H}$, denoted as $\mathbf{y}$. From SVD decomposition of $\mathbf{H}$ we obtain the $\mathbf{U}, \mathbf{\Sigma}$ and $\mathbf{V}^\mathrm{T}$ matrices, with $\mathbf{\Sigma}$ containing the singular values s_i of $\mathbf{H}$.

The $\epsilon_\theta(\mathbf{x}_t, t)$ predicted by the pre-trained model from DDPM is used to compute the predicted $\mathbf{x}_0$, denoted as $\mathbf{x}_{\theta,t}$. Here, the sampling of $\mathbf{x}_t$ is defined as

$$\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon \tag{6}$$

as opposed to (5). This means that σ_t is calculated as $\frac{\sqrt{1-\bar{\alpha}_t}}{\bar{\alpha}_t}$ and that $\mathbf{x}_t = \mathbf{x}_t / \sqrt{\bar{\alpha}_t} = \mathbf{x}_t \sqrt{1 + \sigma_t^2}$.

With $\bar{\mathbf{y}} = \mathbf{V}^\mathrm{T}\mathbf{y}$ and $\bar{\mathbf{x}}_{\theta,t} = \mathbf{V}^\mathrm{T}\mathbf{x}_{\theta,t}$ as well as $\mathbf{a}^{(i)}$ denoting the i-th element (pixel) of a vector $\mathbf{a}$, we can write the first step of the reverse diffusion as

$$p_\theta^{(T)}(\bar{\mathbf{x}}_T^{(i)}|\mathbf{y}) = \begin{cases} \mathcal{N}(\bar{\mathbf{y}}^{(i)}, \sigma_T^2 - \frac{\sigma_\mathbf{y}^2}{s_i}) & \text{if } s_i > 0 \\ \mathcal{N}(0, \sigma_T^2) & \text{if } s_i = 0 \end{cases} \tag{7}$$

Then we compute $\mathbf{x}_T = \mathbf{x}_T / \sqrt{1 + \sigma_t^2}$. This is then the input to the model, on the basis of which we get $\epsilon_\theta(\mathbf{x}_T, T)$ and $\mathbf{x}_{\theta,T}$ according to (5). Then, the next steps of the reverse diffusion process are as follows

$$p_\theta^{(t)}(\bar{\mathbf{x}}_t^{(i)}|\mathbf{x}_{t+1}, \mathbf{y}) = \begin{cases} \mathcal{N}(\bar{\mathbf{x}}_{\theta,t}^{(i)} + \sqrt{1-\eta^2}\sigma_t \frac{\bar{\mathbf{x}}_{t+1}^{(i)} - \bar{\mathbf{x}}_{\theta,t}^{(i)}}{\sigma_{t+1}}, \eta^2\sigma_t^2) & \text{if } s_i = 0 \\ \mathcal{N}(\bar{\mathbf{x}}_{\theta,t}^{(i)} + \sqrt{1-\eta^2}\sigma_t \frac{\bar{\mathbf{y}}^{(i)} - \bar{\mathbf{x}}_{\theta,t}^{(i)}}{\sigma_\mathbf{y}/s_i}, \eta^2\sigma_t^2) & \text{if } \sigma_t < \frac{\sigma_\mathbf{y}}{s_t} \\ \mathcal{N}((1-\eta_b)\bar{\mathbf{x}}_{\theta,t}^{(i)} + \eta_b\bar{\mathbf{y}}^{(i)}, \sigma_t^2 - \frac{\sigma_\mathbf{y}^2}{s_i}\eta_b) & \text{if } \sigma_t \geq \frac{\sigma_\mathbf{y}}{s_t} \end{cases} \tag{8}$$

Similarly to the first step, we also calculate $\mathbf{x}_t = \mathbf{x}_t \sqrt{\bar{\alpha}_t}$ and get $\mathbf{x}_{\theta,t}$ using this scaled input.

3.2 Metrics and GIQA

In this work, we mainly focus on the metrics used by Kawar et al., namely the PSNR and the SSIM, to compare the performance of the DDRM to the baseline [11]. Since the PSNR and the SSIM were used in the original paper we will not define them in the main body of the text. For a more detailed definition of the first two, refer to Appendix A. In both cases we consider a score to be good, the higher it is, i.e. for PSNR $\to \infty$ and for SSIM $\to 1$.

In addition, to the aforementioned metrics we implement the Generated Image Quality Assessment (GIQA) score to add a novel component to the re-implementation of the original paper and evaluate the performance of the generated images accordingly. The GIQA score was proposed as a method to objectively asses the quality of an image generated by a Neural Network [8]. However, as it is a relative scoring method, the absolute numbers do not provide any insights. Hence, we used the GIQA score to validate if our restored images were of better quality i.e. have a higher score than the degraded images using the original images as a reference with the normalized score of 1.

The proposed GIQA score adopted a Gaussian Mixture Model (GMM) and is calculated according to:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} \mathbf{w}^i \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i) \tag{9}$$

$$S_{GMM}(\mathcal{I}_g) = p(f(\mathcal{I}_g)|\lambda) \tag{10}$$

where $\lambda = \mathbf{w}^i, \boldsymbol{\mu}^i, \boldsymbol{\Sigma}^i$ are the mixture weights, mean vectors and covariance matrices from M component densities. To assess the realistic performance of the generative model the mean quality score is calculated as follows:

$$QS(G) = \frac{1}{N_g} \sum_{i}^{N_g} S(\mathcal{I}_g^i) \tag{11}$$

where we have N_g generated samples $\mathcal{I}_g^i, i = 1, 2, ..., N_g$.

As the DDRM does not generate images from pure noise, but rather restores degraded images and the GIQA score is a relative metric, we normalized the scores and calculated its ratio in comparison to the original image:

$$\overline{QS}(G) = \frac{1}{N_g} \sum_{i}^{N_g} \frac{\bar{S}(\mathcal{I}_g^i)}{\bar{S}(\mathcal{I}_{orig}^i)} \tag{12}$$

where $\bar{S}$ is the normalized GIQA score and $\mathcal{I}_g^i$ is the image sample of the degraded or restored image and $\mathcal{I}_{orig}^i$ is the original sample.

3.3 Data

For the re-implementation of this paper, we did not need to train our model. Hence, we were not required to collect huge amounts of data. We were able to use some test data provided by the authors of the original paper and some data similar to what they used. The provided data consists of a small set of OOD images (41 images that can be found on this GitHub[2]), and the similar data we used ImageNet 1k, which is a small subset of ImageNet, containing 1000 images [4]. While the ImageNet 1k data comes from the same dataset the authors used it does not contain the same exact images.

Additionally, since the GIQA score we used is a Gaussian Mixture Model (GMM) parameterized on LSUN cat dataset, we evaluate the performance of our implementation concerning the GIQA score on a subset of the LSUN cats dataset, which is provided in the GitHub repository of the GIQA score[3]. Furthermore, to compare the choice of prior on the results we evaluated three different LSUN-based DDPMs on the LSUN church validation dataset containing 300 images.

3.4 Experimental Setup

The conducted experiments aim to reproduce the results of the original paper. To this end, we implemented the described processes in the previous sections, and compare our implementation with our baseline, and the computed results of the original paper. For all of our experiments, the baseline is defined as the score between the original image and the degraded image.

The original paper uses a model pre-trained on ImageNet, however, since we encountered problems while loading this model, we decided to opt for the model trained on the LSUN Bedrooms data set. This model was trained on less data, however, it is still one of the mentioned DDPM models and is

[2]https://github.com/jiamings/ddrm-exp-datasets

[3]https://github.com/cientgu/GIQA

trained on a significant amount of data. Additionally, since the GIQA score is trained on the LSUN cats dataset, we also use the model trained on the LSUN cats dataset, to be as consistent as possible with training and testing data.

We ran all of our experiments on the small OOD data set and two experiments on ImageNet 1k, namely super-resolution and deblurring with noise. We decided to depict the same samples, to compare the differences between our experiments with and without noise. Moreover, we run our experiments concerning the GIQA score in the aforementioned subset of the LSUN cats dataset and the cross-comparison of different priors on the LSUN church validation dataset. Finally, for our experiments to be as close as possible to the original implementation, we chose values of the hyperparameters, identical to those in the original paper, as shown in Appendix B.

4 Results

The results of the five experiments, conducted for the tasks of denoising, deblurring, and super-resolution are reported and discussed in this section.

4.1 Denoising

In our first experiment, the restoration of noisy images on out-of-distribution (OOD) test data, we evaluated the performance of the DDRM compared to the baseline score. The experiment was conducted under varying levels of noise, represented by standard deviations $\sigma_{\mathbf{y}} = 0.1$ and $\sigma_{\mathbf{y}} = 0.3$. The results in Table 1, showcase the performance of our DDRM compared to the baseline score. We can see that even under the more challenging conditions of $\sigma_{\mathbf{y}} = 0.3$, the DDRM model maintained a robust performance with a PSNR of 26.7845 and an SSIM of 0.7818. With the help of Figure 3 in Appendix C.1, we can observe that the restored images appear to the naked eye to be very close to the original image, validating its high performance. These findings underscore the effectiveness of the DDRM method in mitigating noise and do not contradict the findings of Kawar et al [11].

Table 1: Denoising with different standard deviations on OOD test data

	$\sigma_{\mathbf{y}} = 0.1$		$\sigma_{\mathbf{y}} = 0.3$	
Method	PSNR	SSIM	PSNR	SSIM
Baseline	20.0	0.3389	10.4581	0.0919
DDRM	29.7769	0.8642	26.7845	0.7818

4.2 Deblurring

The second task is the deblurring of images, both with and without noise, applied to the OOD test data and ImageNet1k. The baseline, representing the original image contrasted with the degraded image, serves as a reference point. Our experiments were conducted under varying noise conditions with standard deviations $\sigma_{\mathbf{y}} = 0.1$ for OOD data and $\sigma_{\mathbf{y}} = 0.05$ for ImageNet1k. We can see in Table 2 that restoration performed using DDRM improves the scores, concerning the baseline and allows us to achieve results very similar to the ones reported in the original paper. Additionally, in Figure 1a, we can see the performance on example images. For the comparison of noisy and noiseless deblurring, refer to Appendix C.2. The values of PSNR and SSIM decrease slightly when adding more noise to the image, which can also be observed in the original paper and is in compliance with the logic, that with a higher noise level, more information about the original image is lost, and that directly impacts the backward diffusion process.

Table 2: Deblurring on OOD test data and ImageNet 1k

	OOD Data ($\sigma_{\mathbf{y}} = 0.1$)				ImageNet 1k ($\sigma_{\mathbf{y}} = 0.05$)			
	Noiseless		Noise		Our experiments		Reported by paper	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Baseline	21.89	0.60	17.68	0.139	19.71	0.26	18.35	0.2
DDRM	29.84	0.85	25.30	0.71	24.60	0.66	25.45	0.66

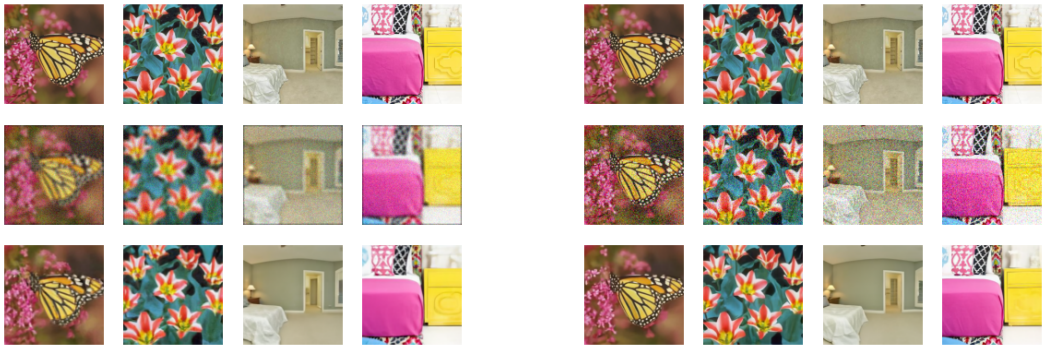
4.3 Super-Resolution

The super-resolution task aims to create an image with a higher resolution. In our case, the original image with 256x256 pixels is degraded to an image with 128x128 pixels. The task is to use the information about the degradation process, to restore the original 256x256 pixels image. Our experiments were conducted on the OOD data as well as ImageNet 1k. In both cases, we can observe that the DDRM scores higher than the baseline, which corresponds to our expectation, as shown in Table 3. For the comparison of noisy and noiseless super-resolution, refer to Appendix C.3.

Table 3: Super-resolution on OOD test data

	OOD Data ($\sigma_{\mathbf{y}} = 0.1$)				ImageNet 1k ($\sigma_{\mathbf{y}} = 0.05$)			
	Noiseless		Noise		Our experiments		Reported by paper	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Baseline	29.1919	0.8673	22.5351	0.4365	24.7875	0.6156	22.55	0.46
DDRM	30.5242	0.8937	26.3954	0.7631	25.9678	0.7348	25.21	0.66

While evaluating the performance on ImageNet 1k, an interesting observation is that our baseline SSIM score (0.6156) deviates from the reported score in the original image (0.46), as shown in Table 3. This divergence was solely observed in the super-resolution task, where the divergence could be due to the resizing method used. To compare the degraded with the generated image, we had to ensure their same dimensionality. Hence, the degraded image had to be resized to 256x256 pixels. This process has likely upscaled the degraded images, yielding higher scores on the baseline compared to the original paper. This is even more apparent when looking at the noiseless example. When there is no noise, a good upsampling method in effect increases the resolution of an image in turn increasing the scores for the baseline.



(a) Deblurring (b) Super-resolution

Figure 1: Deblurring and super-resolution on OOD test data ($(\sigma = 0.1)$)

4.4 Cross-comparison of models

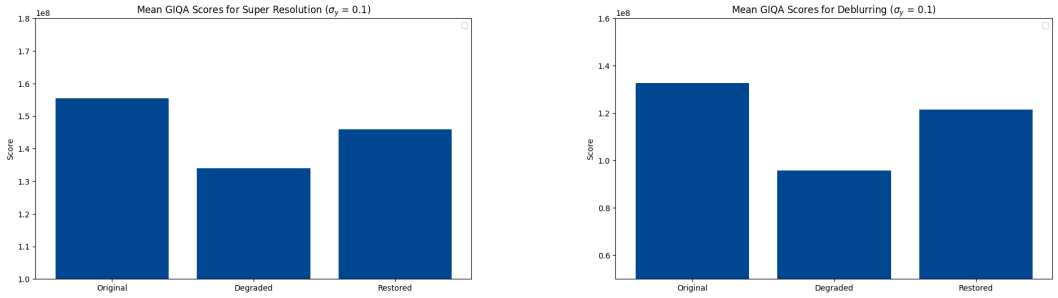
While conducting the experiments we found some interesting results that hint at the fact that for image restoration the prior might not be as relevant as for pure image generation. We tested the three pre-trained diffusion models: LSUN-church, LSUN-Bedroom, and LSUN-cat on the validation set of LSUN-church. The results can be seen in Table 4. As can be noticed all three models have highly similar scores and looking at the restored images 6 in Appendix C.4, they seem comparably good as well. This might be related to the fact that deblurring or super-resolution in DDRM relies on features in the latent space that are similar to various pre-trained models such as edges. Furthermore, the data the models are trained on might contain similarly relevant information for the restoration tasks such as buildings and horizons in LSUN-cat data. However, when running the experiments of these pre-trained models for more specialized tasks the realism might still be lacking if the prior was not chosen carefully. One example would be that tested on human images, the restored faces of the humans might not be as accurate as when using a model trained on that data. The results of this experiment can also serve to explain why despite not using the ImageNet model like the original paper described we still got similar results on the ImageNet 1k dataset.

Table 4: Comparison of different pre-trained models evaluated on LSUN-church ($\sigma_{\mathbf{y}} = 0.1$)

	LSUN-church model		LSUN-bedroom model		LSUN-cat model	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Baseline	16.85	0.13	16.85	0.13	16.85	0.13
DDRM	22.68	0.64	22.74	0.63	22.73	0.64

4.5 GIQA

The GIQA model was trained on the LSUN cat dataset, so we denoised the images with a DDRM pre-trained on that dataset. In Figure 2 we can see that the mean of the GIQA score of the restored images is slightly higher than the degraded images while being slightly lower than the original images. This highlights our findings that the restoration process worked and we have generated a higher quality image than the degraded image. However, it also shows that restored images are not perfectly restored.



(a) Super Resolution

(b) Deblurring

Figure 2: Comparison of the mean GIQA score for the different image categories

Table 5: Super-resolution and Deblurring on pre-trained and evlauated LSUN Cat test data

	Super-Resolution ($\sigma_{\mathbf{y}} = 0.1$)			Deblurring ($\sigma_{\mathbf{y}} = 0.1$)		
Method	PSNR	SSIM	GIQA	PSNR	SSIM	GIQA
Baseline	23.09	0.42	0.76	18.27	0.14	0.59
DDRM	28.30	0.78	0.89	27.52	0.73	0.89

5 Conclusion

In this work, we successfully replicated the main components of the *Denoising Diffusion Restoration Models* by Kawar et al. [11], undertaking a comprehensive evaluation through the reproduction of five experiments on both the same and different datasets. Notably, our experimentation included two trials on the full ImageNet 1k dataset and three on Out-of-Distribution (OOD) data, which was provided by the authors. It is important to highlight that, while we tried as much as possible to align our experimental design with the original one, we utilized a different model, namely the model trained on LSUN bedrooms instead of ImageNet. Despite using a different model, our results yielded a reasonable alignment with those of the original paper.

Moreover, we conducted a cross-comparison analysis of models with different priors on different datasets. We found that for image restoration, the relevance of prior information might be less critical than in pure image generation. Testing three pre-trained diffusion models on the LSUN-church validation set revealed similar performance, suggesting shared latent space features. However, for specialized tasks, careful prior selection remains crucial. For example, restoring human faces may lack accuracy without a model trained on human-specific data.

As an additional novel component, we introduced a new metric that evaluates the quality of the generated image, called the GIQA score. We found that the mean GIQA score for the restored images was slightly higher than the degraded images, yet slightly lower than the original images.

This highlights our findings that the restoration process worked and we have generated a higher quality image than the degraded image. However, it also shows that restored images are not perfectly restored.

During the implementation process, we noticed that the original publication sometimes lacked motivation and explanations of the mathematical formulations. We specifically aimed to reformulate these passages more precisely, and hence contribute to the understanding of DDRM.

Finally, we would like to note that this work could be extended in a couple of directions. For example, the DDRM paper restricts itself to linear degradation of images. Further work could investigate how to use diffusion models to restore images for non-linear degradation processes. Another interesting direction could be to analyze to which extent a less informed prior affects the performance of the restoration.

6 Self Assessment

We re-implemented the method as described in the paper for selected tasks, namely denoising, deblurring, and super-resolution as well as running multiple experiments. The performance was evaluated both on OOD data coming from the corpus of data the model was pre-trained on and data from a different dataset. Furthermore, we conducted additional experiments by investigating different noise levels.

The performance was evaluated on two out of the three metrics as used in the original paper, namely PSNR and SSIM. The results are similar to the ones reported in the original paper. Additionally, we computed the GIQA score for the original, degraded, and restored images, to assess whether the restoration produces realistic images from a human point of view.

Our work contains an explanation of the procedure, that we aimed to describe step-by-step with all the information necessary to reproduce the method. The original paper lacks this structure, making it difficult to re-implement their proposed approach. To this end, we provided an extensive reflection on the challenges and issues.

Given all of the above and the high complexity of the paper, we believe that our contribution, including validating the paper's results, applying a novel metric to assess the performance, cross-comparing models with different priors, and identifying issues with the clarity of the reputability of the original paper, which makes us eligible for several bonus points and should justifiy the grade *Excellent* (A).

7 Challenges

The challenges faced in the project arose primarily from the original paper's lack of implementation details, leading to difficulties in replicating some steps. For example, the memory-efficient SVD was missing practical guidance and some operations and their results did not have a clear interpretation.

The appendix provided information on the scaling necessary for using pre-trained DDPM models, however some details on how the described process aligned with DDPM remained unclear. Given also the absence of explicit formulas for important computations, such as σ_t and its relationship with β_t, the team was required to make informed assumptions for successful interpretation. These observations highlight the importance of detailed implementation guidance and clarity in mathematical formulations in the original paper for effective reproduction. For a more detailed description of the issues mentioned, see Appendix D.

8 Ethical considerations and alignment with UN SGD

The proposed methods by Kawar et al. could increase the performance of applications in image processing, which bears the potential of having a positive impact on everything that has to do with images. However, this comes with the responsibility of ensuring that it is not abused in a way that it negatively impacts the conservation of human rights. Furthermore, this model does not rely on training its own generative model, it relies on using a pre-trained one. Hence, the proposed approach helps in not increasing the amount of CO2 emissions through training new models, which adheres to the development of the UN SDGs.

References

[1] M.R. Banham and A.K. Katsaggelos. "Digital image restoration". In: *IEEE Signal Processing Magazine* 14.2 (1997), pp. 24–41. DOI: 10.1109/79.581363.

[2] Ashish Bora, Eric Price, and Alexandros G. Dimakis. "AmbientGAN: Generative models from lossy measurements". In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=Hy7fDog0b.

[3] Ashish Bora et al. "Compressed Sensing using Generative Models". In: (Mar. 2017).

[4] Jia Deng et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[5] Prafulla Dhariwal and Alexander Nichol. "Diffusion models beat gans on image synthesis". In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794.

[6] Davis Gilton, Greg Ongie, and Rebecca Willett. "Neumann Networks for Linear Inverse Problems in Imaging". In: *IEEE Transactions on Computational Imaging* 6 (2020), pp. 328–343. DOI: 10.1109/TCI.2019.2948732.

[7] B. Girod. "Psychovisual Aspects Of Image Processing: What's Wrong With Mean Squared Error?" In: *Proceedings of the Seventh Workshop on Multidimensional Signal Processing*. 1991, P.2–P.2. DOI: 10.1109/MDSP.1991.639240.

[8] Shuyang Gu et al. *GIQA: Generated Image Quality Assessment*. 2020. arXiv: 2003.08932 [eess.IV].

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *CoRR* abs/2006.11239 (2020). arXiv: 2006.11239. URL: https://arxiv.org/abs/2006.11239.

[10] Kyong Hwan Jin et al. "Deep Convolutional Neural Network for Inverse Problems in Imaging". In: *IEEE Transactions on Image Processing* 26.9 (2017), pp. 4509–4522. DOI: 10.1109/TIP.2017.2713099.

[11] Bahjat Kawar et al. *Denoising Diffusion Restoration Models*. 2022. arXiv: 2201.11793 [eess.IV].

[12] Orest Kupyn et al. "DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks". In: *CoRR* abs/1711.07064 (2017). arXiv: 1711.07064. URL: http://arxiv.org/abs/1711.07064.

[13] Ali Mousavi, Ankit B. Patel, and Richard G. Baraniuk. "A deep learning approach to structured signal recovery". In: *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. 2015, pp. 1336–1343. DOI: 10.1109/ALLERTON.2015.7447163.

[14] Gregory Ongie et al. "Deep learning techniques for inverse problems in imaging". In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 39–56.

[15] Yaniv Romano, Michael Elad, and Peyman Milanfar. "The Little Engine that Could: Regularization by Denoising (RED)". In: *SIAM Journal on Imaging Sciences* 10 (Nov. 2016). DOI: 10.1137/16M1102884.

[16] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. "Plug-and-play priors for model based reconstruction". In: *2013 IEEE global conference on signal and information processing*. IEEE. 2013, pp. 945–948.

[17] Zhou Wang, Alan C. Bovik, and Ligang Lu. "Why is image quality assessment so difficult?" In: *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 4. 2002, pp. IV-3313-IV–3316. DOI: 10.1109/ICASSP.2002.5745362.

[18] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861.

[19] Ling Yang et al. *Diffusion Models: A Comprehensive Survey of Methods and Applications*. 2023. arXiv: 2209.00796 [cs.LG].

[20] Bo Zhu et al. "Image reconstruction by domain-transform manifold learning". In: *Nature* 555.7697 (2018), pp. 487–492.

Appendix A Metrics

A.0.1 Peak-Signal-to-Noise Ratio

The Peak Signal-to-Noise Ratio is a widely used metric for quantifying the quality of restored or processed images by measuring the ratio between the maximum possible signal strength and the noise introduced during the processing. It is expressed in decibels (dB) and is calculated using the mean squared error (MSE). The PSNR formula is given by:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right) \tag{13}$$

where MAX is the maximum possible pixel value of the image (1 in our case, since we use pixel values between 0 and 1), and MSE is the mean squared error between the original and processed images, defined as:

$$\text{MSE} = \frac{1}{HW} \sum_{i=1}^{W} \sum_{j=1}^{H} (I(i,j) - K(i,j))^2 \tag{14}$$

Here, $I(i,j)$ represents the pixel value at position (i,j) in the original image, $K(i,j)$ represents the corresponding pixel value in the processed image, and H and W are the height and width of the images. We consider a higher score of PSNR to be better than a lower one.

The PSNR struggles with capturing the image quality [7, 17], which is why we additionally use the SSIM, as introduced in the upcoming section.

A.0.2 Structural Similarity Index Measure

The Structural Similarity Index is a metric used for evaluating the similarity between two images, considering not only pixel-wise differences but also incorporating information about the structures present in the images [18]. SSIM is a decimal value between -1 and 1, with 1 indicating perfect similarity. The SSIM index is calculated using three components: luminance (l), contrast (c), and structure (s). The overall SSIM ($SSIM_{\text{total}}$) is the product of these three components and is expressed as:

$$SSIM_{\text{total}}(I,K) = l(I,K) \cdot c(I,K) \cdot s(I,K) \tag{15}$$

where I and K are the original and processed images, respectively. The luminance component (l) is given by:

$$l(I,K) = \frac{2\mu_I \mu_K + C_1}{\mu_I^2 + \mu_K^2 + C_1} \tag{16}$$

The contrast component (c) is given by:

$$c(I,K) = \frac{2\sigma_I \sigma_K + C_2}{\sigma_I^2 + \sigma_K^2 + C_2} \tag{17}$$

And the structure component (s) is given by:

$$s(I,K) = \frac{\sigma_{IK} + C_3}{\sigma_I \sigma_K + C_3} \tag{18}$$

Here, μ_I, μ_K, σ_I, σ_K, and σ_{IK} represent the mean, standard deviation, and cross-covariance of pixel values in the corresponding windows of images I and K. The constants C_1, C_2, and C_3 are used for numerical stability. We consider a higher score of SSIM to be better than a lower one.

Appendix B Hyperparameters in the experimental setup

The hyperparameters used in the experiments were the same as the ones in the original paper. The complete list of values for each of them is reported in Table 6.

Table 6: Hyperparameters

Hyperparameters	Value
η	0.85
η_b	1
Time steps	20
Time steps (model)	1000
Kernel size	(9,9)
Type of super-resolution	4x
Image size	(256,256)
Batch size	4

Appendix C Additional results

C.1 Denoising

Since denoising is an implicit part of the DDRM paper and the results are not specifically reported, we only report the scores of the metrics in the main body of the text. In this section, we want to show a plot of the results, for the reader to see that the denoising process works on its own.

As aforementioned, we consider different noise levels, 1) $\sigma = 0.1$ and 2) $\sigma = 0.3$. In both cases we can observe that the denoising process works well in both cases, as shown in Figure 3. This indicates that the proposed method of Kawar et al. can successfully denoise images that were degraded in terms of adding noise.

(a) $\sigma = 0.1$ (b) $\sigma = 0.3$

Figure 3: Denoising on OOD test data

C.2 Deblurring

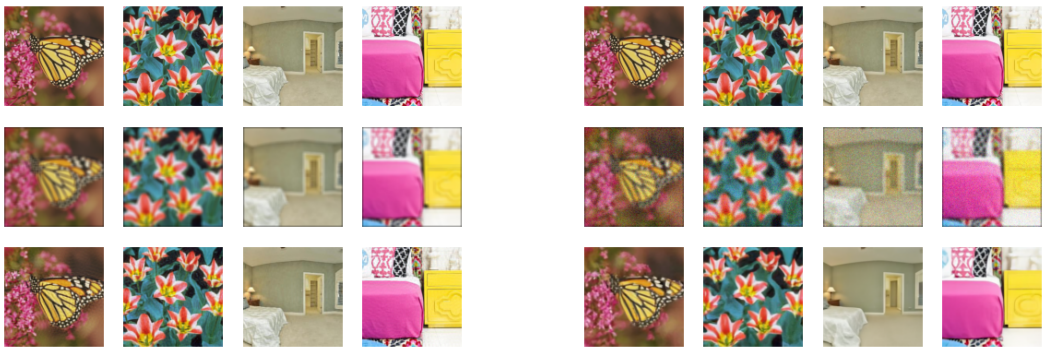
In the main body of this work, we only show the results of noisy deblurring tasks. For completeness purposes, we include Figure 4, to show the noiseless and noisy deblurring directly next to one another.

C.3 Super-Rosultion

In the main body of this work, we only show the results of noisy super-resolution tasks. For completeness purposes, we include Figure 5, to show the noiseless and noisy super-resolution directly next to one another.

C.4 Cross-comparison of models

In Section 4.4 in the main body we described the performance of different priors on the same data set. In Figure 6 we can see a sample of restored images to compare the quality of the restoration.



(a) Noiseless (b) Noise ($\sigma = 0.1$)

Figure 4: Deblurring on OOD test data

(a) Noiseless (b) Noise ($\sigma = 0.1$)

Figure 5: Super-resolution on OOD test data

Figure 6: Comparison of restored images of three models on LSUN-church validation data

Appendix D Challenges and general comments

During the project we encountered multiple challenges, in most cases resulting from the original paper lacking details on implementation and description of several steps of the process, as well as their interpretation.

As one of the first tasks, we looked into the implementation of the memory-efficient SVD for our selected degradation models. The mathematical concepts here were quite clear, however, the description could have used some tips on the actual implementation, such as the fact that the permutation does not necessarily need to be a matrix multiplied by the other matrices, but simply performed using available methods. It has also been difficult to understand, whether the implementation was correct, given that the obtained values could not easily be compared with anything.

This leads us to the observation, that some values and steps in the paper lack a proper interpretation, and sometimes any interpretation. For example, for the $\bar{\mathbf{y}}$ vector obtained as $\bar{\mathbf{y}} = \mathbf{\Sigma}^{\dagger}\mathbf{U}^{\mathrm{T}}\mathbf{y}$ we do not find any explanation, for why it is calculated that way and how it corresponds to the original $\mathbf{y}$ vector. Therefore, it is hard to see if the result of some manipulation was right, as we do not know what to expect.

Similarly, it is said, that $\mathbf{x}_{\theta,\mathbf{t}}$ represents the model's prediction and that in DDPM the authors predict the noise values to subtract to recover $\mathbf{x}_{\theta,\mathbf{t}}$. To our understanding, however, in DDPM the authors use the value of ϵ_θ to predict the $\mathbf{x_0}$ as well as μ_θ, the mean from equation (4). We have deduced that since we modified the steps, it is not the mean that we are supposed to compute using ϵ_θ but the $\mathbf{x_0}$ and proceeded that way. This detail could have been stated explicitly to improve clarity.

Furthermore, we had an issue when adding either more noise to the image with no other degradation, or adding noise to a transformation such as deblurring. Initially, our implementation used the values of σ_t computed based on the β schedule, according to the DDPM implementation. Also, the $\mathbf{x}_0$ has been extracted from noise accordingly. However, with more noisy transformations, this meant that the variance in the initial step $\sigma_T^2 - \frac{\sigma_\mathbf{y}}{s_i}$ was lower than zero, and hence, sampling was not possible. Subsequently, in the appendix, we found the description of scaling of $\mathbf{x_t}$ to use Variance Exploding hyperparameters with a pre-trained DDPM model. Moreover, the $\mathbf{x}_{\theta,\mathbf{t}}$ that we understood was obtained using $\mathbf{x_t}$, which is not mentioned in this section. This made it unclear whether this value should still be extracted using the same method as DDPM or in a different way and whether it should also be scaled.

It seems that for reproducing the paper, it would also be very useful to see an explicit statement on how σ_t is computed, based on the β schedule. Here we find that $\alpha_t = \frac{1}{1+\sigma_t^2}$ which allows us to compute the σ_t, but also the connection between α_t and β_t is not directly reported. In [9] we find that $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_t$ and from the equations for the sampling of $\mathbf{x}_t$ in both papers we concluded that the α_t used in the DDRM paper is, in fact, $\bar{\alpha}_t$ from the DDPM paper. Only that observation allowed us to compute the σ_t from the β schedule used in DDPM and it could have rather been described in the DDRM paper.