Report about "Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise"

Diego Martí Monsó (03716066)

Abstract

In their paper "Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise", which is currently under review for submission at the Eleventh International Conference on Learning Representations (ICLR 2023), Bansal et al. demonstrate that Gaussian noise is not necessary for generative diffusion models to work. Instead of sequentially adding Gaussian noise to a clean image and then reversing the noising process, as it is done in standard diffusion models, the authors explore image generation using deterministic degradations. These deterministic degradations can be inverted with the help of a new sampling algorithm that is proposed. While the quality of the produced samples is not comparable to that of the state-of-the-art, the paper serves as a proof of concept that diffusion models can reverse arbitrary image transforms. In this report, we review the studied methods and their implications.

1 Introduction

In this work, we reproduce and analyze the methods for cold diffusion proposed by Bansal et al. in their paper "Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise" [BBC+22a]. Additionally, we comment on both the strengths and weaknesses of the paper, as well as the reception and criticism the paper has received by the community [Ano22].

Standard diffusion models consist of an image degradation process, which usually involves gradually adding small amounts of Gaussian noise, and a reversal process, where a trained neural network tries to decontaminate the image and output a clean, noiseless image. Under the right circumstances, diffusion models can generate new images from pure noise. These generative models have been inspired by Langevin Dynamics, which model the transition between states of heavy noise (at a high temperature) and little or no noise (at a low temperature). Thus, the models explored in [BBC+22a] are called "Cold Diffusion" models, as they rely on the absence of noise, both during training and sampling. Cold diffusion allows for generalized diffusion models that can revert arbitrary degradations. Currently, there is no solid theoretical framework to understand cold diffusion. The authors do not provide much theoretical support for their findings either, but they perform an extensive qualitative and quantitative analysis of their results.

2 Generalized Diffusion

2.1 General Pipeline

The general pipeline to achieve a complex generative behavior with diffusion models consists of three main steps:

First, we apply the degradation operator D to the clean image $x_0 \in \mathbb{R}^N$, sampled from a distribution $\mathcal{X}$, with severity t. Thus, the operation $x_t = D(x_0, t)$ gradually removes information from x_0 with increasing t. By definition, $D(x_0, 0) = x_0$. In standard diffusion models, the degradation operator contaminates the image with noise with a variance schedule that is dependant on t. For cold diffusion, we consider mostly deterministic degradations.

Second, we train the restoration neural network R_θ – or short, R – parameterized with θ, to behave as the inverse of D. It should therefore approximate the operation $R_\theta(x_t, t) \approx x_0$. The network is trained by minimizing

$$\min_\theta \mathbb{E}_{x\sim\mathcal{X}} \|R_\theta(D(x,t),t) - x\|_1$$

with the random image x that belongs to the distribution $\mathcal{X}$ and the ℓ_1 norm $\|\cdot\|_1$.
Third, we use a sampling algorithm to invert the (deterministic) diffusion process and reconstruct the image $\hat{x}_0$. Standard diffusion approaches, such as DDIM/DDPM [SME21], use a naive sampling algorithm (Algorithm 1 in [BBC+22a]). However, a different sampling algorithm (Algorithm 2 in [BBC+22a]) is used in cold diffusion[1]. The two sampling algorithms and their differences are explored in the following subsection.
All in all, the key differences between cold diffusion and standard diffusion models are the deterministic (vs. random) degradation and the sampling algorithm utilized for generation. On the one hand, cold diffusion can be applied to conditional generation (or inverse) tasks, where there is some information available from the original image after the degradation process. Thus, the aim is to reverse the degradation and reconstruct the original image. On the other hand, one can also perform unconditional generation, where there is no information available from the original data distribution after the degradation.

2.2 Sampling Algorithms

Figure 1: Qualitative comparison of the sampling quality of Algoithm 1 (top row) and Algorithm 2 (bottom row) applied to cold diffusion on the CelebA dataset.

When the degradation operator D is smooth or differentiable, Algorithm 1 fails to produce meaningful results, as can be seen Figure 1. Thus, Bansal et al. propose a new sampling methodology, Algorithm 2, which produces higher quality images when sampling from cold degradations. Both algorithms can be found here in Appendix B. In essence, Algorithm 1 – which is the equivalent to sampling method in DDIM/DDPM [SME21] – alternates between applying the restoration operator and the degradation operator, with slightly less severity (i.e., one time step less) until the image is fully reconstructed. Thus, the degradation applied at each step s of the reverse process reads as

$$x_{s-1} = D(\hat{x}_0, s-1).$$

In contrast, Algorithm 2 introduces a new way of applying the degradation at each step s of the reverse process, given by

$$x_{s-1} = x_s - D(\hat{x}_0, s) + D(\hat{x}_0, s-1).$$

One can easily see that both update rules are perfect (i.e., $x_s = D(x_0, s)$ for $s < t$) if the restoration operator R is the exact inverse of D. In the case of Algorithm 2, the term $x_s - D(\hat{x}_0, s)$ cancels out if $\hat{x}_0 = R(x_s, s) = x_0$.
Bansal et al. do not provide a clear theoretical reasoning to explain the fact that Algorithm 1 fails to produce qualitatively good samples for cold diffusions with smooth degradations. However, they demonstrate the local stability of Algorithm 2 with respect to errors in the restoration operator R when using linear degradations, which is an attribute that Algorithm 1 does not have (or, at least, has not been proven to have). Nevertheless, it must be insisted that this stability is local (around $x = x_0$ and $s = 0$) and only true for linear degradations of the form $D \approx x + s \cdot e$, as the authors tend to generalize the statement that Algorithm 2 is immune to errors in R and each iteration behaves as if R was a perfect inverse of D. In the mathematical proof, which is an inductive proof, Bansal et al. argue that it is valid to assume a linear degradation function, because the Taylor expansion of a smooth

[1] In other versions of the paper, Algorithm 2 is also called TACoS, which is an acronym that stands for "Transformation Agnostic Cold Sampling". For more information on the versions of the paper, please see Appendix A.

degradation function $D(x, s)$ is $D(x, s) \approx x + s \cdot e + \text{HOT}$ (considering that $D(x, 0) = 0$ by definition), where HOT are higher order terms. The authors, however, ignore the fact that e is not necessarily constant and in the general case is dependent on x, ergo $e(x)$ should be regarded. A constant e can be assumed for some truly linear degradations, such as the "animorphism" presented later, but could not be true for other degradation functions, like the blurring operation that is also reviewed later.

3 Generalized Diffusions with Various Transformations

Figure 2: Conditional sampling models, trained on the CelebA dataset and each with a different deterministic degradation. The degradations correspond to the deblurring, inpainting, super-resolution, and snowification tasks. **Left column to right column:** degraded image $D(x_0, T)$, direct restoration $R(D(x_0, T), T)$, sample produced with Algorithm 2, and original image x_0.

The authors study the effects of different deterministic degradations on diffusion processes for unconditional generation. They present four degradation functions and provide empirical results for each of the degradations. For the sake of conciseness, we only show one result per degradation, which is qualitatively representative for the CelebA dataset. Furthermore, we only comment on the quantitative analysis derived from the provided metrics. Screenshots of the extended material, including more sampled images and tables with the Fréchet Inception Distance (FID) scores can be found here in Appendix C. A different restoration network has to be trained separately for each degradation and each dataset, where the chosen hyperparameters may vary. Further implementation details can be found in the official GitHub repository [BBC+22b].

3.1 Deblurring

The clean image x_0 is contaminated with a Gaussian blur degradation that grows with each time step t. Hence, the image is convoluted with the Gaussian kernels $\{G_s\}$ of size 11×11. The standard deviation either remains constant (in the case of MNIST), grows at a rate proportional to t (in the case of CIFAR-10), or exponentially (in the case of CelebA). Then, the degradation operator $D(x_0, t)$ can be defined as

$$x_t = G_t * x_{t-1} = G_t * \cdots * G_1 * x_0 = \overline{G}_t * x_0 = D(x_0, t),$$

where $*$ is the convolution operator. Interestingly, from an image processing point of view, these operations correspond to removing high frequencies from the original image in each time step. During sampling, Algorithm 2 sequentially adds a difference of Gaussians to the degraded image, which is equivalent to adding back the frequencies that have been removed. In Figure 2, we see that the sampling method presented in Algorithm 2 produces sharper samples than the direct reconstruction of the image through a single application of the restoration operator R. However, the image is still not photo-realistic, the produced face lacks detail, and features are too smooth. The FID scores of the sampled images are slightly better compared to the direct reconstruction, but the SSIM and RMSE scores are marginally worse.

3.2 Inpainting

The image is now degraded via the sequential multiplication with the Gaussian masks $\{z_{\beta_i}\}$, where β_i denotes the schedule of the variance with $\beta_1 = 1$ and $\beta_{i+1} = \beta_i + 0.1$. The degradation $D(x_0, t)$, which

effectively grays out pixels in an area of Gaussian shape, can be written as

$$x_t = x_0 \cdot \prod_{i=1}^{t} z_{\beta_i} = D(x_0, t)$$

with the operator $\cdot$ denoting the element-wise multiplication. Both the qualitative and the quantitative analysis of the results are very similar to those of the deblurring case in the previous subsection. Additionally, we see in Figure 2 that Algorithm 2 fails to fully remove the gray stain from the image, while the direct method manages to do so. This fact is not mentioned by the authors.

3.3 Super-Resolution

In the forward process, we downsample (by halving the resolution at each step) the input image to a resolution of 4×4 in the case of MNIST and CIFAR-10, and 2×2 for CelebA. Then, we upscale to the original image size with nearest-neighbor interpolation. As seen in Figure 2, the sample quality of Algorithm 2 is qualitatively superior to the direct reconstruction, but still far from photo-realistic. The quantitative analysis shows a poor distributional similarity to the original datasets, especially for CIFAR-10 and CelebA. Algorithm 2 tends to yield a better FID score, but is still outperformed in the other metrics (SSIM and RMSE).

3.4 Snowification

Finally, Bansal et al. explore the degradation of the data point by applying the ImageNet-C "snowification" transform [HD19] that adds a synthetic snow effect on top of the image. When reversing the degradation operation, the direct method clearly produces better quality samples than Algorithm 2, as illustrated in Figure 2. We see that the image generated with Algorithm 2 has still some of the synthetic snow contaminating the image. Furthermore, some areas of the image are colored erroneously. The authors seem to ignore this finding in their analysis and do not provide metrics to compare the direct reconstruction with Algorithm 2, as they do for the other degradations studied before.

4 Cold Generation

Here, Bansal et al. examine unconditional generation for cold diffusion. The idea is that the cold diffusion models produce new data samples that belong to the original image distribution, starting the reverse process from a degradaded image x_T that is completely devoid of information. As we see in the later discussion, the authors put their focus not only on optimizing the quality of the samples, but more importantly, on establishing ways to promote diversity during generation.

4.1 Generation using deterministic noise degradation

	Hot Diffusion		**Cold Diffusion**	
Dataset	Fixed noise	Estimated noise	Perfect symmetry	Broken symmetry
CelebA	59.91	23.11	97.00	49.45
AFHQ	25.62	20.59	93.05	54.68

Table 1: FID scores using hot (noise) and cold diffusion (with blur) for the CelebA and AFHQ datasets.

The degradation operator $D(x, t) = \sqrt{\alpha_t}x + \sqrt{1 - \alpha_t}z$ is defined as an interpolation between the image x and a noise pattern $z \sim \mathcal{N}(0, 1)$, sampled from a normal distribution. As the degradation D is only applied once during training, the noise pattern is only sampled once. However, in the generation process with Algorithm 2, the degradation D is applied sequentially. Sampling a new noise pattern z in each application of D during image generation would make the generation process non-deterministic. Therefore, there are two options for deterministic generation using a deterministic noise degradation. The quantitative results of the two presented approaches are available in Table 1.

Fixed noise: the first option is to sample the noise vector z once during each individual image generation process and reuse the same z in each application of D.
Estimated noise: the second option, which is equivalent to the deterministic sampling in DDIM[2] [SME21], is to estimate the noise pattern in step t of the reverse process as

$$\hat{z}(x_t, t) = \frac{x_t - \sqrt{\alpha_t} R(x_t, t)}{\sqrt{1 - \alpha_t}}.$$

4.2 Image generation using blur

In standard diffusion models, the distribution of the final degraded image x_T at step T is an isotropic Gaussian. In contrast, the image at the end of a blurring process, as described in Section 3.1, with large T is a constant image x_T that can be represented with a 3-dimensional vector holding the RGB value, which is the channel-wise mean of the input image x_0. The distribution of the degraded image can then be modeled with a single-channel Gaussian Mixture Model (GMM).
Perfect symmetry: if we sample a new RGB value of a fully blurred image from the GMM and use it to create a monochromatic image x_T of the original image size, there is a perfect correlation between the pixels of x_T. Hence, there is very low diversity during the deterministic image generation process.
Broken symmetry: to increase the variability of the produced samples, the perfect symmetry in x_T can be broken by adding very low-variance Gaussian noise to the image.
As seen in Table 1, the estimated noise approach is quantitatively superior to the fixed noise method for deterministic noise degradation. While still having a worse performance than diffusion with deterministic noise degradation, breaking the symmetry improves both the quality and the variability of the generated samples for cold diffusion with blur degradation.

4.3 Generation using other transformations

Lastly, the authors study generalized diffusion with arbitrary transformations from one data distribution to a different one in the forward process. The goal is to have predictable final distributions of the degraded image, which allow for variability in the generation process.
Random masking: to use the masking degradation from Section 3.2 with large T, until there is no information left in the final image x_T, we sample a random color for the mask. This way, x_T is an image c of a randomly sampled color. Conversely, always using the same color, e.g., black, would not allow for diversity in generation. We therefore redefine the degradation to

$$x_t = G_t \cdot x_0 + (1 - G_t) \cdot c = D(x_0, t)$$

with the Gaussian mask $G_t = \prod_{i=1}^{t} z_{\beta_i}$.
Super-resolution: similarly as already described for the blurring case, we can fit a GMM to the images x_T, degraded as in Section 3.3, and sample a new data point to initialize the generation process.
Animorphosis: as a proof of concept, the authors interpolate images from the CelebA dataset with images from the AFHQ dataset, from which a new image is generated sampling with Algorithm 2. The interpolation can also be applied to arbitrary initial data manifolds. Conceptually, the described transformation is equivalent to the deterministic noise degradation in Section 4.1.

5 Conclusion

Overall, the paper serves as an empirical proof that Gaussian noise is not necessary for diffusion models to show complex generative behavior. Despite the arguments of the authors, it is still unclear that Algorithm 2 is superior to Algorithm 1. The quality of the sampled images is worse than the samples generated with standard diffusion models. A potential future line of research could be in the area of conditional generation, where one could guide the generation process via the configuration of the degraded image. Also, the paper opens the door to the study of general diffusion models as a whole.

[2] In DDIM, the noise estimate $\hat{z}$ is used to predict the image $\hat{x}_0$. Here, $\hat{x}_0$ is found first and used then to estimate the noise $\hat{z}$.

References

[Ano22] Anonymous. Official review of "Cold diffusion: Inverting arbitrary image transforms without noise". *https://openreview.net/forum?id=slHNW9yRie0*, 2022.

[BBC+22a] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *Version available at https://arxiv.org/abs/2208.09392*, 2022.

[BBC+22b] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold-diffusion-models. https://github.com/arpitbansal297/Cold-Diffusion-Models, 2022.

[HD19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[SME21] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

A Disclaimer About the Paper

Bansal et al. have submitted their paper "Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise" [BBC+22a] for blind review, to be published at ICLR 2023 [Ano22]. Even though the identity of the authors should be anonymous, versions of the work are available with the revealed names of the authors. In this report, we always refer to the version referenced in the bibliography (i.e., [BBC+22a]) and not to other versions that are also available. Although the notation used by Bansal et al. is inconsistent, we carry on with the same notation in order to allow the identification of the formulas referenced in this report.

B Pseudocode of Sampling Algorithms

The two sampling algorithms presented in [BBC+22a] can be found here.

Algorithm 1 Naive Sampling

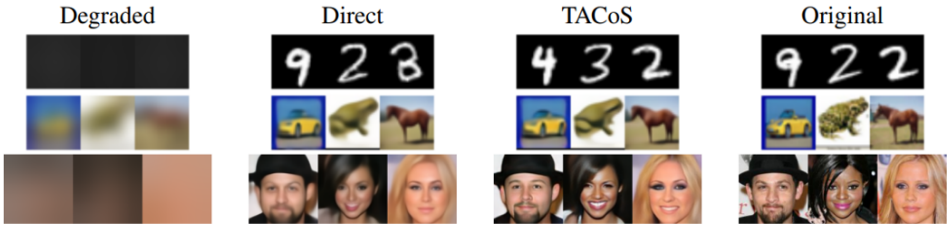
Input: A degraded sample x_t
for $s = t, t-1, \dots, 1$ **do**
  $\hat{x}_0 \leftarrow R(x_s, s)$
  $x_{s-1} = D(\hat{x}_0, s-1)$
end for
Return: x_0

Algorithm 2 Improved Sampling for Cold Diffusion (or Transformation Agnostic Cold Sampling)

Input: A degraded sample x_t
for $s = t, t-1, \dots, 1$ **do**
  $\hat{x}_0 \leftarrow R(x_s, s)$
  $x_{s-1} = x_s - D(\hat{x}_0, s) + D(\hat{x}_0, s-1)$ ▷ New update rule
end for
Return: x_0

C Extended Results

Here, we can find screenshots of the extended qualitative and quantitative analysis provided in [BBC+22a] for the described generative models with deterministic degradation functions.

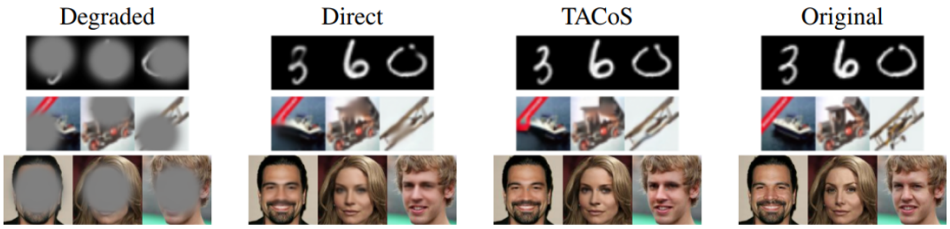


(a) Deblurring models trained on MNIST, CIFAR-10, and CelebA datasets. **Left to right:** degraded inputs $D(x_0, T)$, direct reconstruction $R(D(x_0, T), T)$, sampled reconstruction with Algorithm 2, and original image.

	Degraded			Sampled			Direct		
Dataset	FID	SSIM	RMSE	FID	SSIM	RMSE	FID	SSIM	RMSE
MNIST	438.59	0.287	0.287	**4.69**	0.718	0.154	5.10	**0.757**	0.142
CIFAR-10	298.60	0.315	0.136	**80.08**	0.773	0.075	83.69	**0.775**	0.071
CelebA	382.81	0.254	0.193	**26.14**	0.568	0.093	36.37	**0.607**	0.083

(b) Screenshot of the table with quantitative results for the sample quality using deblurring models.

Figure 3: Screenshots of the empirical results for deblurring.

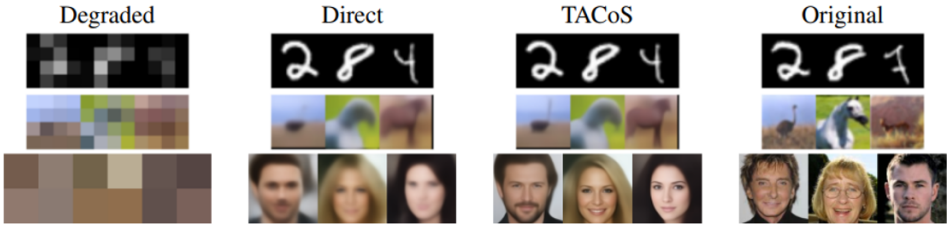


(a) Inpainting models trained on MNIST, CIFAR-10, and CelebA datasets. **Left to right:** degraded inputs $D(x_0, T)$, direct reconstruction $R(D(x_0, T), T)$, sampled reconstruction with Algorithm 2, and original image.

	Degraded			Sampled			Direct		
Dataset	FID	SSIM	RMSE	FID	SSIM	RMSE	FID	SSIM	RMSE
MNIST	108.48	0.490	0.262	**1.61**	0.941	0.068	2.24	**0.948**	0.060
CIFAR-10	40.83	0.615	0.143	**8.92**	0.859	0.068	9.97	**0.869**	0.063
CelebA	127.85	0.663	0.155	**5.73**	0.917	0.043	7.74	**0.922**	0.039

(b) Screenshot of the table with quantitative results for the sample quality using inpainting models.

Figure 4: Screenshots of the empirical results for inpainting.



(a) Superresolution models trained on MNIST, CIFAR-10, and CelebA datasets. **Left to right:** degraded inputs $D(x_0, T)$, direct reconstruction $R(D(x_0, T), T)$, sampled reconstruction with Algorithm 2, and original image.

	Degraded			Sampled			Direct		
Dataset	FID	SSIM	RMSE	FID	SSIM	RMSE	FID	SSIM	RMSE
MNIST	368.56	0.178	0.231	4.33	0.820	0.115	**4.05**	**0.823**	0.114
CIFAR-10	358.99	0.279	0.146	**152.76**	0.411	0.155	169.94	**0.420**	0.152
CelebA	349.85	0.335	0.225	**96.92**	0.381	0.201	112.84	**0.400**	0.196

(b) Screenshot of the table with quantitative results for the sample quality using superresolution models.

Figure 5: Screenshots of the empirical results for superresolution.

(a) Desnowification models trained on MNIST, CIFAR-10, and CelebA datasets. **Left to right:** degraded inputs $D(x_0, T)$, direct reconstruction $R(D(x_0, T), T)$, sampled reconstruction with Algorithm 2, and original image.

	Degraded Image			Reconstruction		
Dataset	FID	SSIM	RMSE	FID	SSIM	RMSE
CIFAR-10	125.63	0.419	0.327	31.10	0.074	0.838
CelebA	398.31	0.338	0.283	27.09	0.033	0.907

(b) Screenshot of the table with quantitative results for the sample quality using desnowification models.

Figure 6: Screenshots of the empirical results for desnowification.

Figure 7: Examples of generated samples from 128×128 CelebA and AFHQ datasets using cold diffusion with blur tranformation.