Improving Diffusion Models with Discriminator Guidance

reproducibility study

Timo Kleger **Viktoria Sartor**

Abstract

In this reproducibility report we will check the results of the Diffusion Guidance technique from Kim et al. [1] to improve the quality of the generated images from existing pretrained Diffusion Models. Furthermore we try to even further improve the Discriminator Guidance method by applying the ensemble method in the backward diffusion step. We show that the results of the paper are valid. With our own DG Model we achieve slightly worse results than stated in the paper. The ensemble method that we suggested did not improve the prediction as we expected. All our code for generating, training and for the experiments is available under `https://github.com/ChlaegerIO/KTH_DeepLearning-diffusion`.

1 Introduction

This study is about the famous Diffusion Models [2] first introduced in 2015 that generate astonishing images from noise. In our highly digital world it is important to be able to generate images, because sometimes it is economically too expensive to produce image data to train our AI systems. Another important point that we address in this report is that the results should be reproducible. We dive into the the challenge of reproducing scientific results especially in Machine Learning as J. Pineau et. al. [3] stated in a recent paper.

In recent years the Diffusion Models were improved drastically. We focus on one of such an improvement technique called Discriminator Guided Diffusion Models as stated in [1]. It is used as an extension on already fully trained Diffusion Models without requiring to retrain or fine tune the Diffusion Model. Our goal is to replicate the results from Kim et al. [1] on the CIFAR10 data set as we did not have the computational power to train the discriminator on larger data sets such as ImageNet 256x256. Our task consists of training a discriminator, generating images and then comparing the results, thus FID score, to the original paper [1]. Furthermore we show generated images with or without Discriminator Guidance.

In addition to the replication we experiment with an ensemble method to further improve the DG backward Diffusion step and thus to improve the quality of the image generation. We train several ensemble members and then average over all predicted discrimination.

In our studies we show that the FID score improves from 2.09 to 1.88 with the pretrained model of the original paper. Our own trained Discriminator achieves an FID of 2.08 and the ensemble Discriminator has a worse score of 14.35.

Project report for DD2412 or FDD3412

2 Related Work

2.1 Diffusion Models

Diffusion Models in computer vision, particularly Denoising Diffusion Models, have shown promising results in generative modeling. Based on a forward Diffusion stage and a reverse Diffusion stage, these models have been applied in various frameworks [4], including Denoising Diffusion Probabilistic Models [5]. Despite their computational complexity, recent research has focused on making these models more efficient, particularly emphasizing design strategies that improve their computational efficiency [6] while also improving the performance.

2.2 Discriminator Guidance

In our paper Kim et al. [1] proposes a method called Discriminator Guidance that improves sample generation in pretrained Diffusion Models by incorporating a discriminator for realistic sample supervision. This approach achieves state-of-the-art results on ImageNet 256x256 and CIFAR10 without requiring joint training of score and discriminator networks. To understand our study we want to introduce the forward and especially the backward path of Stochastic Differential Equations (SDEs). In the forward path we add step by step noise to the image. In the backward path we learn to revert this process. In most cases the noise that was added to the image in the forward path is learned $\mathbf{s}_\theta$ and then subtracted from the image as seen in equation 2. In or case we further guide the path with a learned Discriminator $\mathbf{d}_\phi$ that distinguishes real from fake images. This is visualized in figure 1. The continuous mathematical formulation of the forward SDE is

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)dt + g(t)d\mathbf{w} \tag{1}$$

and of the backward SDE it is

$$d\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - g^2(t)(\mathbf{s}_\theta + w_t^{(DG)}\mathbf{d}_\phi)(\mathbf{x}_t, t)\right] d\bar{t} + g(t)d\bar{\mathbf{w}}_t. \tag{2}$$

Where $\mathbf{f}(\mathbf{x}_t, t)$ and $g(t)$ are the drift and the volatility coefficients, dt and dw are the infinitesimal small time step and Brownian motion, thus where noise is added, and $\mathbf{s}_\theta$, $w_t^{(DG)}$ and $\mathbf{d}_\phi$ are the learned reverse Diffusion score, a weight and the learned Discriminator value.

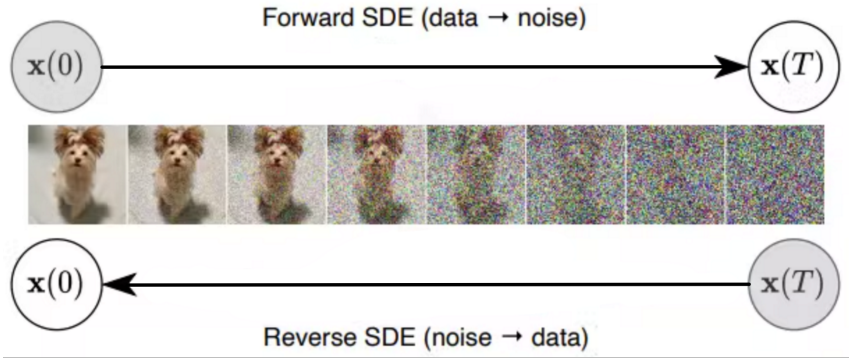


Figure 1: Visualized forward and backward SDE. The image is adapted from Song et al. [7].

Building upon this concept, there are other techniques to improve the performance of Diffusion Models like Classifier-Free Diffusion Guidance introduced by Ho et al. [5]. They demonstrate that guidance can be performed without a classifier by jointly training conditional and unconditional Diffusion Models. Or Lim et al. [8] extend the application of score-based Diffusion Models to function spaces by introducing Denoising Diffusion Operators for training in these spaces. They demonstrate the applicability of these Models in scientific computing and 3D geometric data analysis.

2.3 Ensemble method for Discriminators

There are papers which use ensemble methods for Discriminators. For example Xu et al. [9] developed FairGAN, a fairness-aware generative adversarial network that generates discrimination-free data while preserving data utility. The model uses two discriminators to create fairness. Another model called PATE-GAN was developed by Yoon et al. [10]. It uses several discriminators, each trained discriminator ensure the possibility to generate synthetic data with differential privacy guarantees. It uses the Private Aggregation of Teacher ensembles method.

3 Method

The three main steps of the paper are to generate fake samples 3.2, to train the Discriminator 3.3 and to generate DG samples for evaluation 3.4.

3.1 Data

The paper [1] was tested on CIFAR10, CelebA, FFHQ and ImageNet256. Based on our computation power we decide to use CIFAR10 dataset [11] to reproduce the Discriminator Guidance method and to experiment with ensemble method. We test the unconditional case of the CIFAR10 data set. This is a harder task than the conditional one as we have less information about the image. The CIFAR10 data consists of 50'000 images in a resolution of (32,32,3) pixels. We have a dataloader which takes 50'000 fake images and 50'000 CIFAR10 images, then randomly shuffle the samples before we split the data set into train set, validation set and test set in the proportion of 80%, 10% and 10%.

3.2 Generate fake sample

For the training of the Discriminator we need generated fake samples with the underlying pretrained Diffusion Model $\mathbf{s}_\theta$. For that we randomly sampled noisy images $x_T \sim \mathcal{N}(0, I)$. Via the backward Difussion in equation 2 with zero DG weight $w_t^{(DG)} = 0$ we get fake images.

3.3 Discriminator Training

In the second step we train the Discriminator, which distinguishes real from fake images. The whole Discriminator network consists of a pretrained and fixed classifier, which is a UNet architecture. This classifier takes the CIFAR10 images as an input (B,32,32,3) and outputs a feature representation (B,8,8,512) of it. The second stage consists of another pretrained Discriminator Model, which in our case is also a UNet. This Discriminator is now fine tuned to take the features and output a single number (B, 1). We use a hardware constrained batch size B of 64 instead of the 128 used in the paper [1]. During training we calculate accuracy and loss, both for the training data and for the validation data. The hyperparameters that are used for learning of all different Discriminators are in table 2.

Table 1: Hyperparameter for generating samples with / without DG

Parameter	**Diffusion Samples**	**DG Samples**
Number of Diffusion steps	35	35
Distance	$[10^{-5}, 1 - 10^{-5}]$	$[10^{-5}, 1 - 10^{-5}]$
Image size	32	32
DG weight 1st order	0	2.0
DG weight 2nd order	0	0
Time	[0.01,1]	[0.01,1]
Boosting	False	True
Batch size	64	64
Number of samples	50,000	50,000

3.4 Generate Discriminator Guided samples and evaluation

For the evaluation we need fake samples generated from the trained Discriminator. To generate these samples with DG we use the parameters of table 1. The only difference is that with Discriminator Guidance we use a first order weight of 2 and boosting as it was done in the paper [1].

To evaluate our results we used FID, precision and recall, similar to the original paper. The results are in section 4. The FID score measures the similarity between two groups of images described by Heusel et al. [12]. In our case this are the CIFAR10 images and our generated images.

3.5 Ensembled Discriminator Guidance

To test our idea of using the ensemble method we train six Discriminators. The first ensemble is equal to the default configuration of the paper, as far as we can tell. So we can use it for the reproducibility task. As a first intuition if our idea works we use different hyperparameters as seen in table 2 and simply average the ensembles. However it is recommended to have a bigger variance in the model selection to get better results and thus also use different architectures as ensembles and not only different hyperparameters [13]. One could use Bagging for example to further improve the ensemble method.

Table 2: Hyperparameter for all ensembles used in Discriminator training

Discriminator	Paper (D0)	D1	D2	D3	D4	D5
Number of Epochs	60	40	40	40	40	200
Learning Rate	0.0003	0.001	0.0001	0.0001	0.00005	0.00001
Weight Decay	1×10^{-7}	1×10^{-7}	0	1×10^{-3}	1×10^{-9}	1×10^{-11}
Min Difference Time	1×10^{-5}	1×10^{-5}	0.01	1×10^{-3}	1×10^{-3}	1×10^{-5}

To sample images with ensembled Discriminator Guidance we averaged the output of all six discriminator values

$$\mathbf{d}'_{\phi} = \frac{1}{6} \sum_{i=0}^{5} d_{\phi(i)}. \tag{3}$$

4 Experiments and results

In the this chapter we explain the experiments and discuss the achieved results. We list all our results in table 3 and discuss it in the following chapter.

Table 3: Evaluation results of reproduction and of our ensemble method

	EDM	EDM-G++	Our EDM-G++	Our ensemble EDM-G++
FID	2.094	1.880	2.084	14.348
Recall	-	0.761	0.733	0.680
Precision	-	0.500	0.525	0.417

4.1 Reproduction

The first column of table 3 list the outcome of the reproducibility experiments. We get a FID score of 2.09 for the Diffusion Model without any Discriminator Guidance. This is a bit higher than the 1.97-2.03 which are stated in our paper for the EDM for CIFAR10 [1]. In a second step we evaluated the pretrained EDM-G++ from the paper and got an FID score of 1.88, which is also higher than stated in the paper with 1.77 [1]. The difference of improvement is with 0.21 in our case and 0.2-0.26 as stated in the paper.

In a second step we trained our own EDM-G++ Discriminator and we achieved an FID score of 2.08, which is a bit lower than in the paper. We tried to use the same hyperparameters as in the paper,

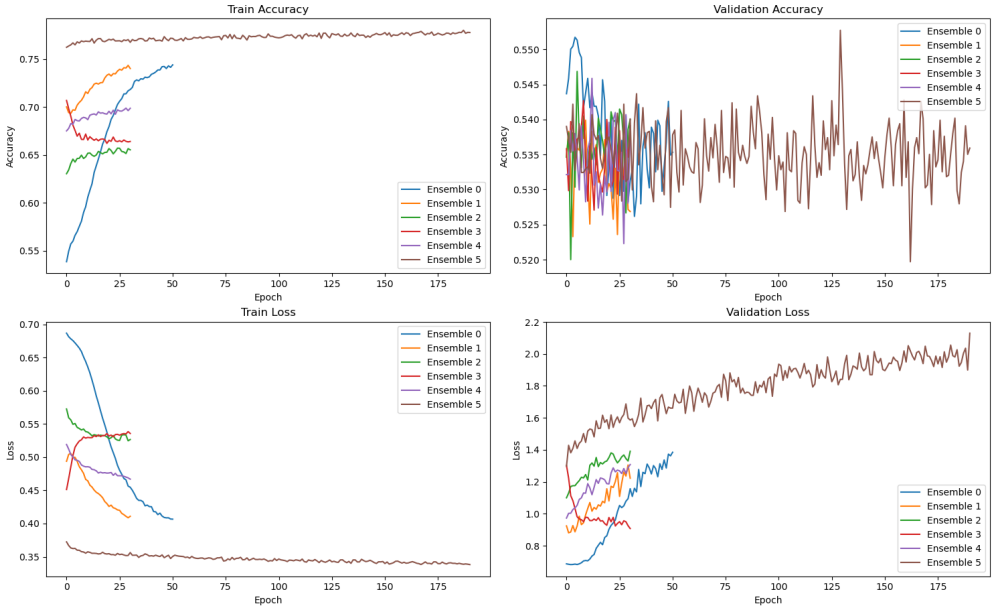
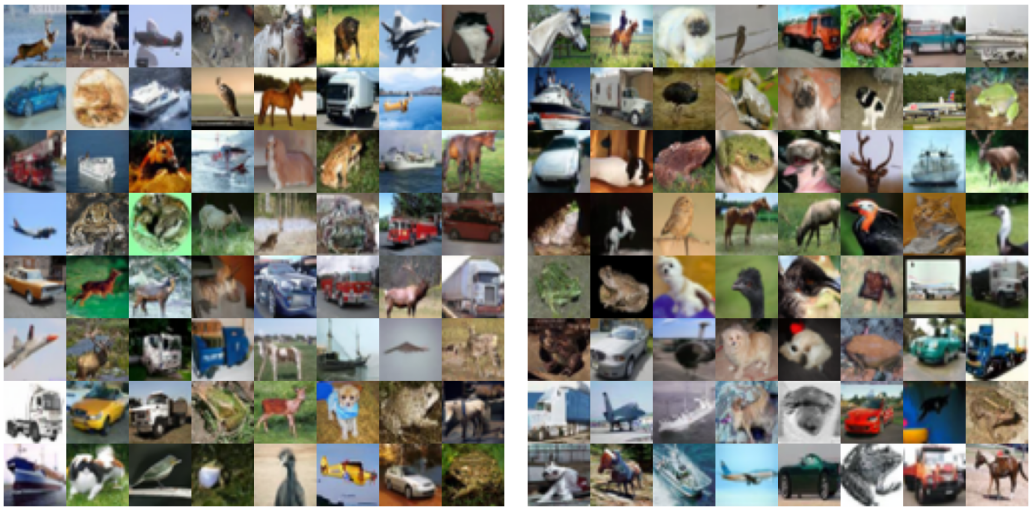


Figure 2: Training and validation statistics of all six ensembles

but not all are clearly visible. For example for the learning rate they just told a default value, but no configuration for the specific results. Different than in the paper we used a validation set during training to gather more information about the training procedure. From figure 2 we conclude that Discriminator 5 should have a better FID score than Discriminator 0 that was used for our EDM-G++ FID score. In the same figure we also noted that we overfit on the training set as the validation loss increases. We think that this is one reason why we perform worse than in the paper as we have not seen the entire data set for training and thus have less samples and a worse estimation of the CIFAR10 data set distribution. Images from our Diffusion model with and without DG can be seen in figure 3.



(a) Samples of Diffusion Model without DG (b) Samples of Diffusion model with our DG

Figure 3: Comparison between samples of Diffusion Model without DG and with our DG

4.2 Ensemble method for Discriminator Guidance

For our ensemble model we get a very bad FID score of 14.35. And also the recall and precision scores are a little bit worse than our EDM-G++ model.

Unluckily we had too less time to train a lot more Discriminators. Otherwise we would have trained each ensemble member individually and then compare it with the ensemble to get more reasoning about why it performs so badly. In our git repository we have generated images for the interessted readers.

5 Challenges

First of all, Diffusion Models was a new field for all of us before this course. Therefore we successfully managed this interesting challenge to understand this field. Then an even bigger challenge was that one of our team mates got sick and gave up after the first three weeks. So we finished the project with only two persons.

In the beginning we had trouble to run the pretrained Diffusion Model, classifier and discriminator as we copied the code from the original paper of the classifier and not from our paper until we realized that in our paper they adapted the original UNet definition slightly.

Some hyperparameters were in table 8 of Kim et al. [1], but for others we had to look in the git repository of the paper. For example the learning rate, the weight decay of the Adam optimizer, the minimum and maximum Diffusion time and churn rate to sample [14]. Generally it was quite hard to find a hyperparameter search that was not already done in the paper and relevant for the method.

6 Conclusion

We conclude that the statement of the paper is quantitatively right. In our own Discriminator we however were not able to get the same performance as stated in the paper. And also with the pretrained versions we got weaker results.

Furthermore we saw that our ensemble method made the generation process a lot worse. This is where we see future work that could rethink about the ensemble method and how to better implement it for example with Bagging or other algorithms [13]. A fist step would be to thoroughly analyse the FID score of each ensemble member individually and then compare it with the combined ensemble model.

7 Ethical consideration, societal impact, alignment with UN SDG targets

The societal impact of generative AI like Diffusion Models is high as this Diffusion Models can generate DeepFakes images and videos. This leads to ethical challenges, as it can be exploited for malicious purposes such as spreading misinformation and identity theft. Beyond ethical concerns, the societal impact involves eroding trust in digital media. It is crucial to inform the public about recent development that the society can adapt to the risks and opportunities. In terms of the UN SDG targests we fulfill the development goal number 17, since we come from two different countries, working together on this project in Sweden [15].

8 Self Assessment

First of all we successfully managed to reproduce the core statement of the paper. But we have gone even further by adapting the learning procedure of the paper by adding a validation set in training and by applying the ensemble method to Diffusion Models. Based on these further significant extensions to the paper we think that a B-grade would be valid even though we did not got the expected results of our ensemble approach.

9 Acronym

Adam	Adaptive moment estimation
AI	Artificial Intelligence
DG	Discriminator Guidance
EDM	Equivariant Diffusion Model
EDM-G++	Equivariant Diffusion Model with DG
FID	Frechet Inception Distance
GAN	Generative Adversarial Network
SDE	Stochastic Differential Equation
UN SDG	United Nations Sustainable Development Goals

References

[1] Dongjun Kim, Yeongmin Kim, Se Jung Kwon, et al. Refining generative process with discriminator guidance in score-based diffusion models, 2023. URL https://icml.cc/virtual/2023/oral/25468. ICML 2023.

[2] Sohl-Dickstein, Jascha, Weiss Eric, et al. Deep unsupervised learning using nonequilibrium thermodynamics. *PMLR. 37: 2256–2265*, 2015.

[3] J. Pineau et al. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). In *arXiv:2003.12206*, 2020.

[4] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:10850–10869, 2022. URL https://api.semanticscholar.org/CorpusID:252199918.

[5] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. URL https://api.semanticscholar.org/CorpusID:219955663.

[6] Anwaar Ulhaq, Naveed Akhtar, and Ganna Pogrebna. Efficient diffusion models for vision: A survey. *ArXiv*, abs/2210.09292, 2022. URL https://api.semanticscholar.org/CorpusID:252918532.

[7] Yang Song, Jascha Sohl-Dickstein, et al. Score-based generative modeling through stochastic differential equations. *ICLR (oral)*, 2021.

[8] Jae Hyun Lim, Nikola B. Kovachki, Ricardo Baptista, Christopher Beckham, Kamyar Azizzadenesheli, Jean Kossaifi, Vikram S. Voleti, Jiaming Song, Karsten Kreis, Jan Kautz, Christopher Joseph Pal, Arash Vahdat, and Anima Anandkumar. Score-based diffusion models in function space. *ArXiv*, abs/2302.07400, 2023. URL https://api.semanticscholar.org/CorpusID:256868496.

[9] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575, 2018. URL https://api.semanticscholar.org/CorpusID:44106659.

[10] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1zk9iRqF7.

[11] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[13] Derrick Mwiti. A comprehensive guide to ensemble learning: What exactly do you need to know, 2023. URL https://neptune.ai/blog/ensemble-learning-guide.

[14] Tero Karras, Miika Aittala, et al. Elucidating the design space of diffusion-based generative models. *NeurIPS*, 2022.

[15] United Nations. Strengthen the means of implementation and revitalize the global partnership for sustainable development, 2023. URL https://sdgs.un.org/goals/goal17.