# Denoising Diffusion GANs

**Parth Pujari, Anish Kulkarni, Cheshta Damor, Ashwin Abraham**

## 1 Introduction

Stable Diffusion is a deep learning, text-to-image model released in 2022. It is primarily used to generate detailed images conditioned on text descriptions, though it can also be applied to other tasks such as inpainting, outpainting, and generating image-to-image translations guided by a text prompt.

Generative Adversarial Networks (GANs) are an approach to generative modeling (for generating realistic and high-quality synthetic data) using deep learning methods. GANs consist of two neural networks, the generator, and the discriminator, engaged in a competitive training process. The generator aims to produce data that is indistinguishable from real data, while the discriminator's role is to distinguish between real and generated data. Through this adversarial training, both networks improve iteratively.

### 1.1 Tackling a trilemma

We plan to reproduce the paper on ”TACKLING THE GENERATIVE LEARNING TRILEMMA WITH DENOISING DIFFUSION GANs” by Zhisheng Xiao, Karsten Kreis and Arash Vahdat. The gist of the paper is as follows.

**Abstract**

A wide variety of deep generative models have been developed in the past decade. Yet, these models often struggle with simultaneously addressing three key requirements including: high sample quality, mode coverage, and fast sampling. We call the challenge imposed by these requirements the generative learning trilemma, as the existing models often trade some of them for others. Particularly, denoising diffusion models have shown impressive sample quality and diversity, but their expensive sampling does not yet allow them to be applied in many real-world applications. In this paper, the authors argue that slow sampling in these models is fundamentally attributed to the Gaussian assumption in the denoising step which is justified only for small step sizes. To enable denoising with large steps, and hence, to reduce the total number of denoising steps, they propose to model the denoising distribution using a complex multimodal distribution. They introduce denoising diffusion generative adversarial networks (denoising diffusion GANs) that model each denoising step using a multimodal conditional GAN.

The datasets used in the paper are the CIFAR-10, the stacked MNIST and/or CelebA-HQ all of which are found on Kaggle.

# 2 Multimodal denoising distributions for large denoising steps

A common assumption in the diffusion model literature is to approximate $q(x_{t-1}|x_t)$ with a **Gaussian** distribution. Here, we question when such an approximation is accurate.

The true denoising distribution $q(x_{t-1}|x_t)$ can be written as $q(x_{t-1}|x_t) \propto q(x_t|x_{t-1})q(x_{t-1})$ using Bayes' rule where $q(x_t|x_{t-1})$ is the forward Gaussian diffusion shown berore and $q(x_{t-1})$ is the marginal data distribution at step $t$. It can be shown that in two situations the true denoising distribution takes a Gaussian form.

First, in the limit of infinitesimal step size $\beta_t$, the product in the Bayes' rule is dominated by $q(x_t|x_{t-1})$ and the reversal of the diffusion process takes an identical functional form as the forward process (Feller, 1949). Thus, when $\beta_t$ is sufficiently small, since $q(x_t|x_{t-1})$ is a Gaussian, the denoising distribution $q(x_{t-1}|x_t)$ is also Gaussian, and the approximation used by current diffusion models can be accurate. To satisfy this, diffusion models often have thousands of steps with small $\beta_t$.

Second, if data marginal $q(x_t)$ is Gaussian, the denoising distribution $q(x_{t-1}|x_t)$ is also a Gaussian distribution. The idea of bringing data distribution $q(x_0)$ and consequently $q(x_t)$ closer to Gaussian using a VAE encoder was recently explored in LSGM (Vahdat et al., 2021). However, the problem of transforming the data to Gaussian itself is challenging and VAE encoders cannot solve it perfectly. That is why LSGM still requires tens to hundreds of steps on complex datasets.

When neither of the conditions are met, i.e., when the denoising step is large and the data distribution is non-Gaussian, there are no guarantees that the Gaussian assumption on the denoising distribution holds. To illustrate this, in Fig. 2, we visualize the true denoising distribution for different denoising step sizes for a multimodal data distribution. We see that as the denoising step gets larger, the true denoising distribution becomes more complex and multimodal.

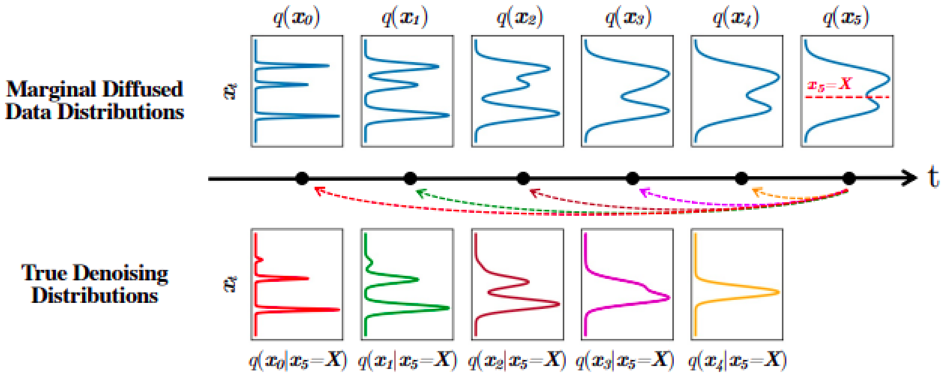


Figure 1: **Top**: The evolution of 1D data distribution $q(x_0)$ through the diffusion process. **Bottom**:, The visualization of the true denoising distribution for varying step sizes conditioned on a fixed $x_5$. The true denoising distribution for a small step size (i.e., $q(x_4|x_5 = X)$) is close to a Gaussian distribution. However, it becomes more complex and multimodal as the step size increases.

## 2.1 Probabilistic Modelling of the DDGAN

## 2.2 Background

We define the Diffusion model as a *latent variable* model of the form

$$p_\theta(x_0) = \int p_\theta(x_{0:T})\, dx_{1:T}$$

where $x_1 ... x_T$ are the latents as defined earlier. Note that they have the same dimensionality as the data $x_0$. The joint distribution $p_\theta(x_0)$ is called the *reverse process* and is defined as a Markov chain with learned Gaussian transitions starting from $p(x_T) = \mathcal{N}(x_T; \mathbf{0}, \mathbf{I})$ and

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \qquad p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \boldsymbol{\mu}_\theta(x_t, t), \boldsymbol{\Sigma}_\theta(x_t, t)) \tag{1}$$

The *forward process* is fixed to a Markov chain that gradually adds Gaussian noise to data as per a variance schedule $\beta_1 .... \beta_T$:

$$q(x_{1:T}) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \qquad q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1},\ \beta_t \mathbf{I}) \tag{2}$$

Training is performed (similar to variational Bayes) by minimizing the negative log likelihood;

$$\mathbb{E}[-\log p_\theta(x_0)] \leq \mathbb{E}_q \left[ -\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} \right] = \mathbb{E}_q \left[ -\log p(x_T) - \sum_{t \geq 1} \log \frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right] =: L \tag{3}$$

The forward process parameters are held constant (as hyperparameters). The forward process admits sampling $x_t$ at an arbitrary time step in closed form, the notation used is : $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. Then,

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0,\ (1 - \bar{\alpha}_t)\mathbf{I}) \tag{4}$$

We note that the forward process is derived from the conditional probability distribution $q(x_t|x_{t-1})$, our sampling is to be estimated by the probability distribution $p_\theta(x_{t-1}|x_t)$.

Our goal is to reduce the number of denoising diffusion steps T required in the reverse process of diffusion models. Inspired by the observation above, we propose to model the denoising distribution. with an expressive multimodal distribution. Since conditional GANs have been shown to model complex conditional distributions in the image domain , we adopt them to approximate the true denoising distribution $q(x_{t-1}|x_t)$.

To set up the **adversarial training** we denote the time dependent discriminator as $D_\phi(x_{t-1}, x_t, t) : \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R} \to [0, 1]$, with parameters $\phi$. It takes the $N$ dimensional images $x_{t-1}$ and $x_t$ and decides whether $x_{t-1}$ is a plausible denoised version of $x_t$. The generator takes as input $x_t$ and tries to generate a plausible $x_0$ (called $x'_0$) from $p_\theta(x_0|x_t)$.

## 2.3 Neural Network model

The idea is that now our approximation to the posterior $p_\theta(z|x)$ is going to be a neural network and the parameters $\theta$ (and consquently $\phi$ for the discriminator) are jointly optimized using minibatch gradient descent. We model the Discriminator as a Convolutional Neural Network with heavy downsampling and the Generator as a UNet, both with timestep embeddings. We also train a DDPM model on the Diffuser's UNet. Finally we train the models used in the paper and test them on CIFAR-10 and stacked MNIST.

**Algorithm 2:** Denoising Diffusion GANs Training Algorithm

```
Input: Training dataset $\mathcal{D}$, Generator $G$, Discriminator $D$, Diffusion process parameters,
       Learning rate $\alpha$
Output: Trained Generator $G$
for each training iteration do
    Sample a mini-batch $\mathcal{B}$ from $\mathcal{D}$;
    Sample a timestep $t$ uniformly;
    for each data point $\mathbf{x}$ in $\mathcal{B}$ do
        Sample $\mathbf{y_t} = (x_{t-1}, x_t)$ from the forward process;
        Sample $x'_0$ from $p_\theta(x_0|x_t)$ and $x'_{t-1}$ from $q(x'_{t-1}|x'_0, x_t)$;
        Let $\tilde{\mathbf{y}}_\mathbf{t} \coloneqq (x'_{t-1}, x_t)$, $\mathbf{x} \in \mathcal{B}$;
    Update Discriminator $D$ using $\mathcal{B}$ and $\tilde{\mathcal{B}}$ according to the GAN objective:
        $\nabla_{\theta_D} \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} -\log D(\mathbf{y_t}, t) + \nabla_{\theta_D} \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} -\log(1 - D(\tilde{\mathbf{y}}_\mathbf{t}), t)$
     Sample another $t$ uniformly and another $\tilde{\mathbf{y}}_\mathbf{t}$;
    Update Generator $G$ using $\mathcal{B}$ to minimize the denoising diffusion GAN objective:
        $\nabla_{\theta_G} \frac{1}{|\mathcal{B}|} \sum_{\mathbf{x} \in \mathcal{B}} -\log D(\tilde{\mathbf{y}}_\mathbf{t})$
```

# 3 Results

## 3.1 Diffusion on MNIST and Fashion MNIST

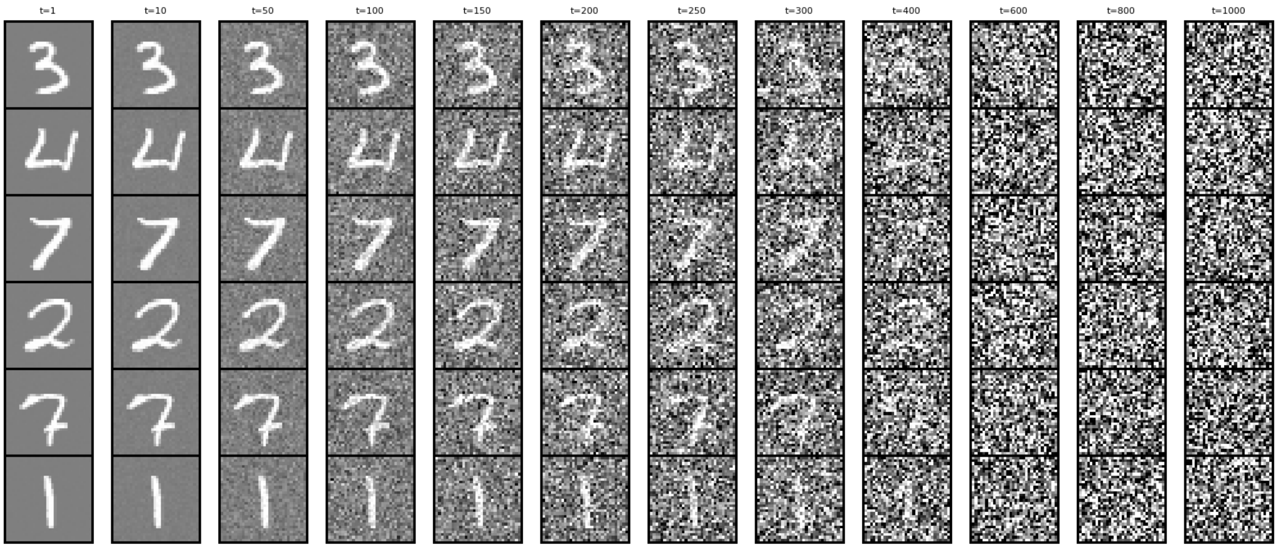
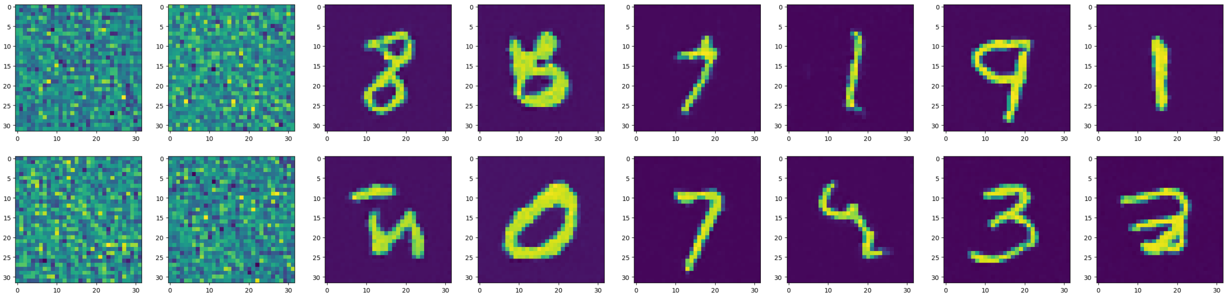
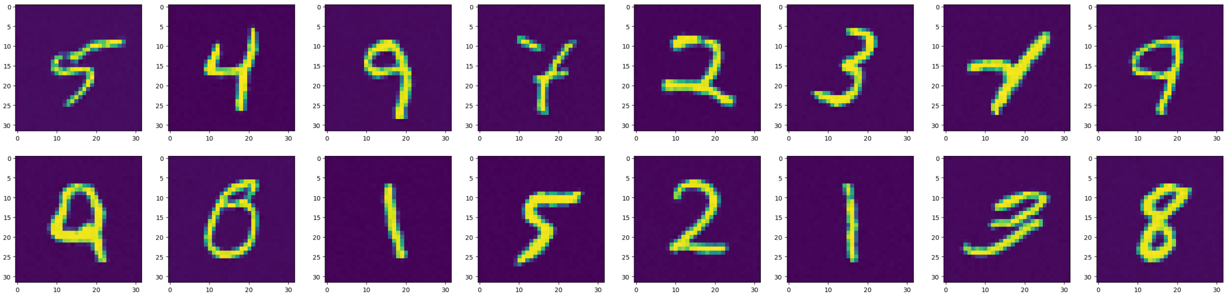


Figure 2: Forward Diffusion on MNIST



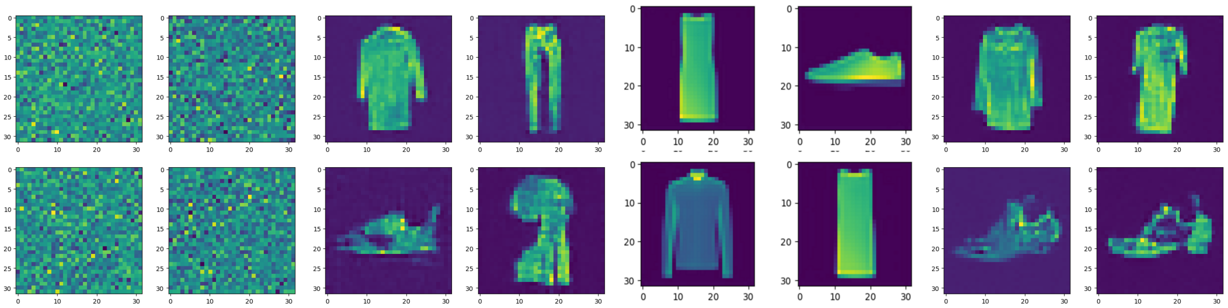
(a) (b) (c) (d)

Figure 3: Left to right, epochs 1 through 4



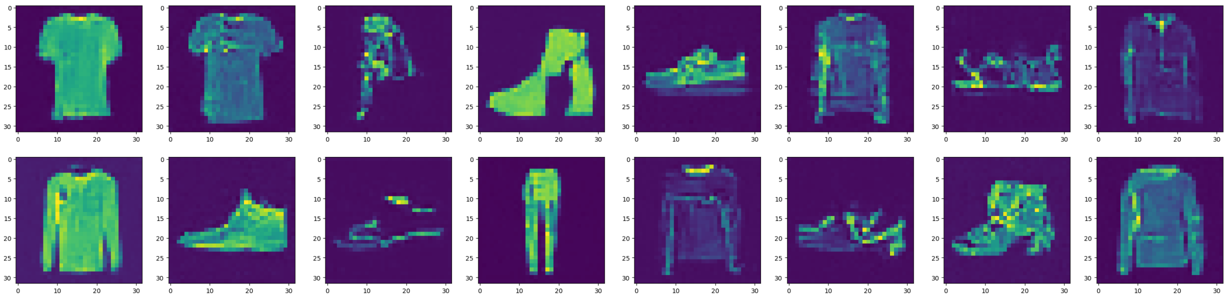
(a) (b) (c) (d)

Figure 4: Epoch 4



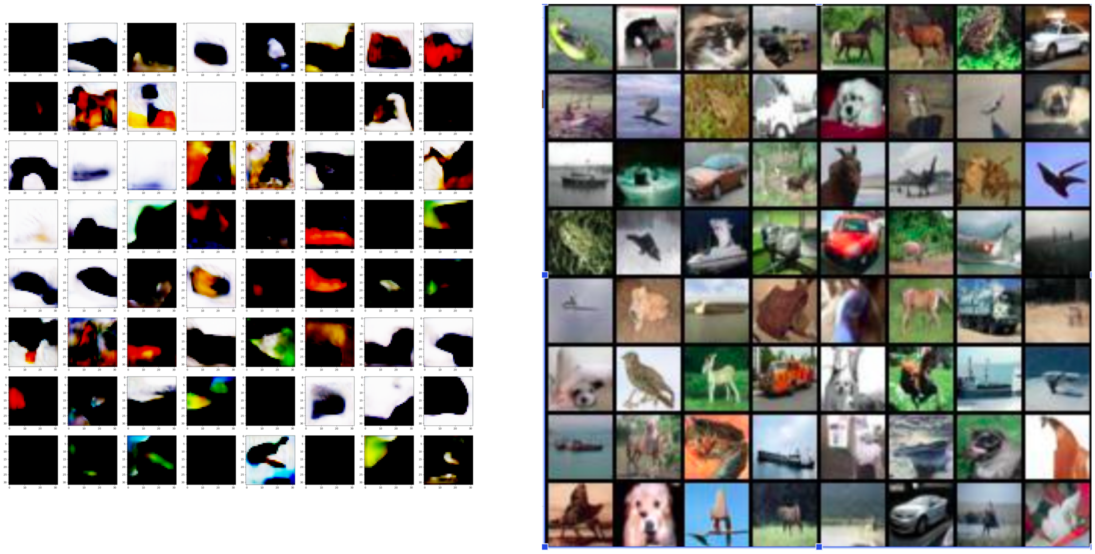
(a) (b) (c) (d)

Figure 5: Left to right, epochs 1 through 4



(a) (b) (c) (d)

Figure 6: Epoch 4



(a) (b)

Figure 7: Epoch 20 and 1000