# Generative AI Theory & Applications
# Report 4: Denoising Diffusion Probabilistic Model

Syed Abraham Ahmed

April 9, 2025

## Contents

# 1 DDPM Introduction vs. WGAN-GP

## 1.1 DDPM & WGAN-GP

Diffusion models are a class of generative models. In a Denoising Diffusion Probabilistic Model (DDPM), a forward Markov chain gradually adds Gaussian noise to an image until it becomes nearly pure noise, while a reverse process is learned to recover the data from the noise. For our scenario, the CelebALike dataset was used to train our diffusion model.

The forward process is defined where $\beta_t$ is a variance schedule that increases over $t = 1, \ldots, T$:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\Big(x_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1},\ \beta_t\mathbf{I}\Big),$$

The reverse process is parameterized with learnable parameters $\theta$. Training minimizes a variational upper bound on the negative log-likelihood, which, under proper choices of $\Sigma_\theta$, reduces to a simplified loss function measuring the discrepancy between the true and predicted noise.

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\Big(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\Big)$$

### 1.1.1 DDPM Parameters

The DDPM hyperparameters were chosen based on GPU limitations. The batch size was set to 64 to fit within memory constraints, and the number of epochs was limited to 5. A total of 1000 time steps was used for denoising. To balance performance and memory usage, the channel multiplier was set to [1, 2, 4]. Only 1 residual block was used to reduce memory load. [1]

These choices resulted in an epoch time of approximately 7 minutes which led a total train time of around 35 to 40 minutes for a total of 5 epochs. This is from an initial state of approximately 45 minutes per epoch.

### 1.1.2 WGAN-GP Parameters

The hyperparameters were selected for performance. The batch size was set to 16 due to GPU constraints. The image size was fixed at 64x64 with 3 channels. [2]

The discriminator architecture consists of LeakyReLU activations and dropout to prevent overfitting. The generator uses dense and upsampling layers with LeakyReLU activations. The model was trained for 3 extra discriminator steps per generator update to stabilize training. [2]

## 1.2 Evaluation Metrics

### 1.2.1 Qualitative

The DDPM images generally exhibit a more coherent and appealing style in comparison to the WGAN-GP generated images. Certain artifacts are well generalized with regard to eye, face, nose and mouth structure. However there are details that lack in quality with regard to the background of the faces, and transition from one face feature to another.

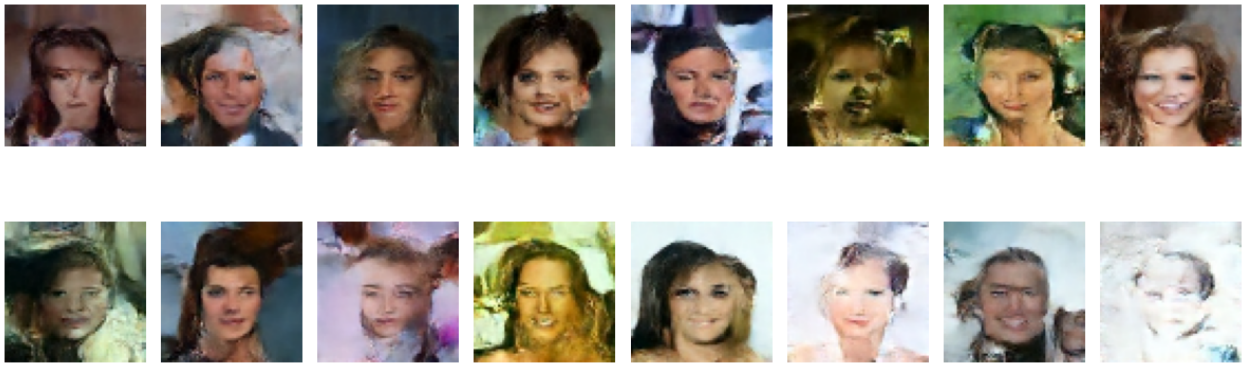

Figure 1: DDPM Generated Samples, Epoch 5

The WGAN-GP images shows blocky artifacts and lacks fine detail, indicating potential training instability. Unlike DDPMs' smooth denoising process yielding coherent results, WGAN-GP may have struggled with mode collapse, leading to a lower quality output. The "muddy" appearance contrasts sharply with the arguably more appealing style of DDPM generated images.

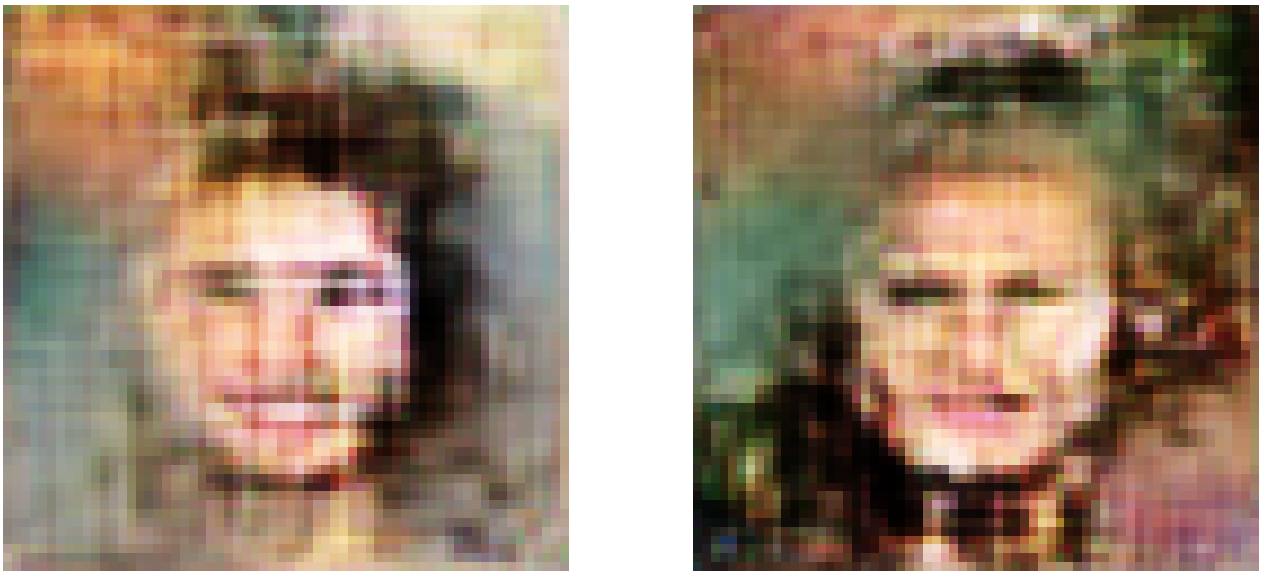

Figure 2: WGAN-GP Generated Samples, Epoch 5

#### 1.2.2 Quantitative

To properly analyze our generated images in a quantitative manner, we are able to use the Fréchet Inception Distance (FID) and the Inception Score (IS) for both model examples. Using a provided template of the FID and IS score [3], we are able to generate images, sample from our respective CelebALike dataset, and evaluate our scores respectively.

Table 1: Generative Model Performance Metrics

| Metric | DDPM | WGAN-GP |
|---|---|---|
| FID Score | 278.75 | 245.15 |
| IS Score | 1.0001 (std dev: 0.0003) | 1.8351 (std dev: 0.2511) |

It is worth reminding that the DDPM model was ran for 5 epochs in comparison to the 30 epochs ran on the WGAN-GP model.

Based on these metrics, WGAN-GP appears to generate more realistic (lower FID) and potentially higher quality (higher mean IS) images compared to DDPM. However, DDPM produces more consistent results (lower IS standard deviation), even if the overall quality as perceived by the IS is lower. The qualitative assessment mentioning DDPMs might not be fully captured by these specific metrics alone, which shows the importance of evaluating generated images both quantitatively and qualitatively.

### 1.3 Pros and Cons

#### 1.3.1 DDPM

**Pros:** DDPMs demonstrated training stability due to its gradual denoising process. Qualitatively, the generated samples exhibited a more coherent and appealing style with good generalization of facial structures. The reverse diffusion process encouraged learning a smooth mapping, leading to globally consistent outputs. The diversity potential is high by reversing a complex noise distribution.

**Cons:** The primary drawbacks were the slow sampling speed and high computational cost of the iterative reverse process (1000 steps in our case). Achieving high quality requires extensive training, which is computationally intensive (7 minutes per epoch with incredibly limited parameters). Our limited 5-epoch training may have resulted in lower quantitative metrics (higher FID, lower IS) and a lack of fine details and smooth transitions.

#### 1.3.2 WGAN-GP

**Pros:** WGAN-GP had faster evaluations per epoch (which was noticed during training) and achieved better quantitative metrics (lower FID, higher mean IS) with more extensive training (30 epochs). The model showed potential for good diversity quantitatively. A more detailed explanation is located under section 1.2.2.

**Cons:** The main challenge was training instability, leading to blocky artifacts and a less coherent visual style. The "muddy" appearance suggested incomplete learning of fine details.

## 2 References

[2] A.K.Nain, DDPM, Keras, 2022.DIRECT LINK

[2] F.Chollet, A.K.Nain, WGAN-GP, Keras, 2023.DIRECT LINK

[3] J. Brownlee, How to Implement the Inception Score (IS) for Evaluating GANs, Machine Learning Mastery, Oct, 2019. DIRECT LINK