Exploring Denoising Diffusion Models for Realistic Anime Character Generation

Vishakha Kumari, Robin Singh Bhadoria
Department of Computer Science & Engineering
National Institute of Technology Hamirpur, Himachal Pradesh
{23mcs113, robin.bhadoria}@nith.ac.in

2024 IEEE 16th International Conference on Computational Intelligence and Communication Networks (CICN) | 979-8-3315-0526-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/CICN63059.2024.10847556

***Abstract*— In recent years, many brands have started incorporating anime characters into their marketing strategies to boost brand recognition and appeal to a broader audience. These generated characters can also be utilized in online games, further enhancing creativity within the entertainment industry. However, image generation in style transfer tasks presents significant challenges due to the complex variations seen in anime images. It is crucial to retain key features, such as emotions and gestures, during image generation. Generative models like Denoising Diffusion Probabilistic Models (DDPMs) are effective in producing high-quality and detailed images. They work by progressively refining noisy inputs, making them well-suited for capturing intricate details like emotions and gestures. However, the sampling process in DDPMs requires running a Markov chain through numerous steps, making the process computationally expensive. To address these challenges, Denoising Diffusion Implicit Models (DDIMs) have been introduced. DDIMs generate high-quality samples with improved efficiency by using a more implicit, deterministic method for denoising at each stage, significantly speeding up the image generation process.**

***Keywords* – Denoising Diffusion Probabilistic Models (DDPM); Denoising Diffusion Implicit Models (DDIM); Forward Diffusion; U-Net; Reverse Diffusion; KID**

I. INTRODUCTION

Image generation have achieved a remarkable progress over the years, from traditional deterministic methods to generative models. Generative Adversarial Networks (GANs) gained popularity for their capability to produce highly realistic and high-resolution images. Generative models learns to approximate complex data distributions to produce highly realistic data. However, GANs often suffered from many problems such as instability during training mode collapse and the need of fine tuning hyper- parameters.

To overcome these limitations, Diffusion models showed up as a novel approach for image generation, providing an alternative to adversarial training. These models function by gently introducing noise to the data during the training process and reversing the process during generation. This step-by-step denoising process enables diffusion models to create diverse and high-quality outputs with greater stability and efficiency. While GANs operate within a game-theoretic framework, involving a competitive game between a generator and a discriminator, diffusion models adopt a probabilistic framework, making their training more stable and adaptable. Recent developments, including DDPMs (Denoising Diffusion Probabilistic Models) [1], DDIMs (Denoising Diffusion Implicit Models) [2], and latent diffusion models, have further advanced their capabilities, enabling high-resolution image synthesis while ensuring computational efficiency.

This paper consists of using Denoising Diffusion Implicit Models (DDIMs) based framework for image generation. DDIMs marks a remarkable progress in diffusion models by offering a more efficient way to image synthesis. Unlike traditional diffusion models, which involve multiple steps of noise addition and denoising, DDIMs provide a more direct and computationally less expensive method through an implicit sampling process.DDIMs involves a non-Markovian framework which means that DDIMs can achieve high-quality image generation in just a fewer steps. Our study showcases the the effectiveness of DDIMs in producing high-resolution images, highlighting their potential for various creative and practical applications in generative modeling.

In this paper, we utilize the High-Resolution Animeface Dataset (512x512), easily available on Kaggle, which is part of the Danbooru2019 Portraits collection. This dataset, curated by Gwern Branwen and the Danbooru Community, consists of 303,000 high-quality anime face images.

This proposed work combines the elements of deterministic and stochastic processes, incorporates adaptable noise schedules, and emphasizes a clear separation of signal and noise components. These features offer several advantages:

1. Accelerated sampling with a reduced number of reverse steps.
2. Enhanced output quality by improving the reconstruction of the original clean image.
3. Greater versatility, enabling the model to be adaptable for various tasks, datasets, and conditions.

II. UNDERLYING CONCEPTS OF DIFFUSION MODELS

A. Key Ideas Related to Diffusion Models

Diffusion models, a type of generative approach that generate data by iteratively transforming random noise into more meaningful results. These models are built on principles of probabilistic modeling and stochastic processes. The process of diffusion involves two primary phases: forward diffusion or reverse diffusion.

Forward diffusion process involves progressively corrupting a data sample x_0 through a sequence of noise in a step-by-step manner until it completely turn into pure noise. Mathematically, this process is illustrated as a Markov chain[2]:

$$q(x_t \mid x_{t-1}) = N(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I), \quad (1)$$

such that $(1-\beta_t) = \alpha_t$

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{\bar{\alpha}_t} x_{t-1}, (1-\bar{\alpha}_t)I) \quad (2)$$

Equation (2) represents the conditional probability distribution where x_t denotes the noisy data at each timestamp t and x_{t-1} represents the data from the previous timestamp, α_t regulates the amount of noise introduced at each timestamp t. After many steps T, the data is effectively transformed into pure Gaussian noise.

The reverse process aims to reconstructs the original image by gradually recovering from the noise in the corrupted data, undoing the forward diffusion. A trained neural network needed to approximate the conditional probability p_θ $(x_{t-1}|x_t)$, progressively denoising x_t back to x_0.

The denoising process also follows a markov chain as given below:

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)) \quad (3)$$

Equation (3) models to tranform the noisy data x_t into a less noisy data x_{t-1} as step in the reverse diffusion process and right part of the equation illustrates Gaussian distribution, having $\mu\theta(x_t,t)$ as mean and $\Sigma_\theta(x_t,t)$ as variance[2]. This can be illustrated as:

$$\mu_\theta(x_t,t) = 1/(\alpha\sqrt{t}).\ (x_t - \beta_t/(1-\sqrt{\bar{\alpha}_t})\ .\ \epsilon\theta(x_t,t))$$

$$\text{and } \Sigma_\theta(x_t,t) = \beta_t/(1-\bar{\alpha}_t)I$$

The overall training objective is to approximate the forward and reverse diffusion, so for that we will use a neural network usually a U-Net like structure. The objective is to simply minimize the discrepancy between the actual noise ε and the predicted noise ε_θ. This loss is often derived from variational lower bound (ELBO) which ensures that the learned distribution closely aligns with the true data distribution.

$$L = Eq(x_t,x_0)[\|\epsilon - \epsilon_\theta(x_t,t)\|^2] \quad (4)$$

$Eq(x_t,x_0)$ in equation (4) indicates the expectation or average taken over the joint probability distribution of the noisy data x_t and the original data x_0 from the forward diffusion process.

$[\|\epsilon - \epsilon_\theta(x_t,t)\|^2$ represents squared difference between the two, original noise ϵ and the predicted noise ϵ_θ [2].The goal is to minimize this error, enabling the model to accurately estimate the noise which is being introduced during the forward diffusion process.

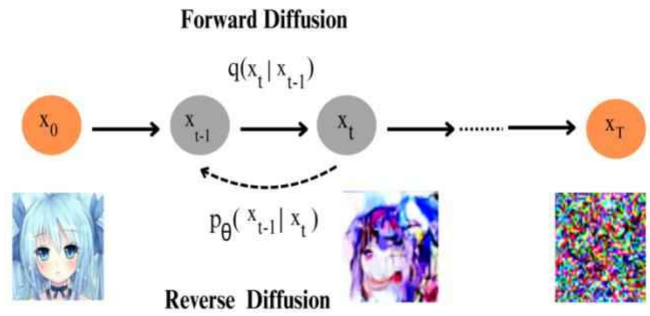


Figure 1. Forward and Reverse Diffusion

The transition from an anime-style clean image x_0 to a completely noisy image x_T in forward diffusion, and its gradual restoration back to x_0 during reverse diffusion, demonstrates the model's ability to synthesize or reconstruct high-quality outputs through this collaborative diffusion process. This figure effectively visualizes the interaction between noise addition and removal in the generative framework.

B. Denoising Diffusion Implicit Models(DDIMs)

Denoising Diffusion Implicit Models, an extention of traditional diffusion models which offers a more efficient approach to generative modeling by emphasizing the denoising process[1]. Unlike the DDPMs which operate through the iterative process of introducing and eliminating noise, implicit model aims to learns a faster reversal of a diffusion process with fewer steps thereby improving generation speed while improving high quality outputs. There is a deterministic mappnig between the timestamps i.e from one timestamp to the next which skips the need for stochastic sampling during the reverse process and enhances efficiency by improving the way the process is carried out.The overall model works by leveraging a non-Markovian approach to transition between noisy states, effectively providing more flexibility in terms of model architecture and training.

Rather than relying on stochastic sampling for x_{t-1} , Implicit models utilizes a reparameterized determination equation to compute x_{t-1} directly from x_t eliminating the need for randomness[1].

$$x_{t-1} = \sqrt{\alpha_t}(x_t - \sqrt{1-\alpha_t}.\epsilon_\theta(x_t,t))/\sqrt{\alpha_t}) + 1 - \sqrt{\alpha_t}.\epsilon_\theta(x_t,t) \quad (5)$$

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)) \quad (6)$$

such that, $\mu_\theta(x_t,t) = \sqrt{\alpha_t}(x_t - \beta_t/(1-\sqrt{\bar{\alpha}_t}).\epsilon\theta(x_t,t))$
and $\Sigma_\theta(\mathbf{x}_t,t) = 0$

III. Research Methdology

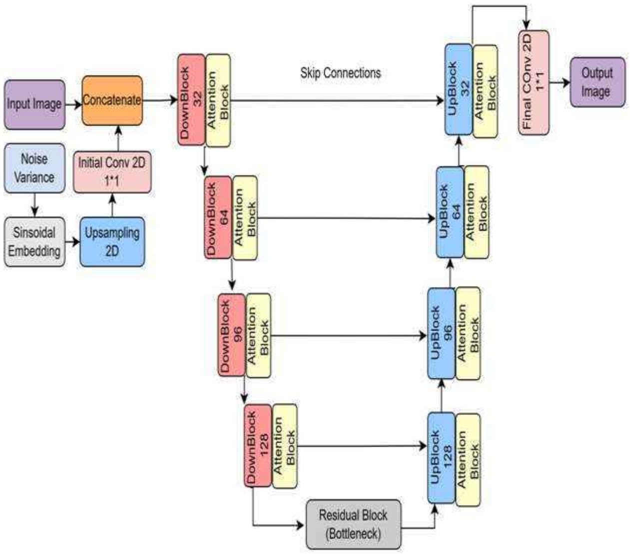


Figure 2. U-Net Model Architecture

Algorithm 1: Diffusion Process

Input: X_0: Clean image, T: Forward timesteps, S: Reverse timesteps, noise_schedule(t): Function providing signal_rate$_t$ and noise_rate$_t$, network: trained model predicting $\epsilon\theta$.

Output: Reconstructed clean image $\bar{x}_0$.

1. **Forward Diffusion (Training Phase):**
   Initialize an empty sequence {xt}
   For t=1 to T:
   1. Compute signal_rate$_t$, noise_rate$_t$ using noise_schedule(t,T).
   2. Sample noise ϵ from a standard normal distribution, $\epsilon \sim N(0,I)$.
   3. Compute x_t= signal_ratet$\cdot x_0$+noise_rate$_t \cdot \epsilon$
   4. Append x_t to the sequence.

   End For
   Set $x_T = x_t$ (final noisy image).

2. **Reverse Diffusion (Generation Phase):**
   Initialize x_T as the input noisy image.
   For s=S to 1:
   1. Set T=s/S.
   2. Compute signal_rate$_t$, noise_rate$_t$ using noise_schedule(T,s).
   3. Compute signal_rate$_{t-1}$, noise_rate$_{t-1}$ for the next timestamp.
   4. Predict noise ϵ_θ = network(x_t,t).
   5. Compute $\bar{x}_0$= (x_t−noise_rate$_t \cdot \epsilon_\theta$)/(signal_rate$_t$)
   6. Update x_{t-1}= signal_rate$_{t-1}$.$\bar{x}_0$ +noise_rate$_{t-1}$. (x_t−signal_rate$_t \cdot \bar{x}_0$)/(noise_rate$_t$)

   End For.
   Return $\bar{x}_0$

During training, the U-Net as in figure (2) learns to estimate the noise ϵ_θ which is added to image at each timestep by reducing the difference between the true noise ϵ and its prediction. During generation, the U-Net uses its learned ability to iteratively predict ϵ_θ for each noisy image x_t, progressively refining it back to the clean image x_0. Thus, the collaborative process between forward and reverse diffusion, powered by the U-Net, enables the synthesis of high-resolution images [11]. The diffusion_schedule function in the model governs the diffusion process, where noise is gradually introduced and later removed from the image. It transforms diffusion times into angles, computes signal and noise rates using trigonometric relationships, and ensures that their squared sum equals 1. This scheduler controls the specific amount of noise applied at each step, allowing the model to progressively refine noisy data and produce high-quality images over multiple iterations. EMA with a decay factor (β=0.999) is employed to create a more reliable model by averaging the weights, making them less affected by short-term fluctuations. This contributes to better generalization and performance, particularly in scenarios involving noisy updates, such as in GANs or diffusion models. In this research, we utilize the Kernel Inception Distance (KID) metric to check the quality of synthesized images. This metric computes the Maximum Mean Discrepancy (MMD) between the original and generated image distributions, using a pretrained network mainly Inception network for feature extraction and lower value of kID is better as it ensures dependable comparisons throughout different stages of model training.

IV. Result & Anaysis

This research utilized diffusion models for generating anime images and evaluated their performance on the Danbooru, CelebA, and Oxford Flowers datasets.The results emphasize the models effectiveness, with Danbooru achieving a KID score of 0.1;CelebA a KID score of 0.2 and Oxford Flowers yielding a KID score of 0.1.These outcomes demonstrate the model's capability for the generation of diverge high-quality images across different datasets.

Danbooru

CelebA

OXFORD FLOWER

TABLE I. KID SCORES ON DIFFERENT DATASETS

Dataset	KID	Noise Loss	Image Loss
Danbooru[15]	0.12	0.13	0.21
Celeba[16]	0.20	0.14	0.13
Oxford Flower [17]	0.10	0.12	0.21

TABLE II. VALIDATION RESULTS ON DANBOORU

Epoch	KID	Noise Loss	Image Loss
1	1.8	0.7	2.2
5	1.1	0.3	0.9
10	0.6	0.1	0.2
15	0.2	0.1	0.2
20	0.2	0.1	0.2
25	0.2	0.1	0.2
30	0.1	0.1	0.2
35	0.1	0.1	0.2

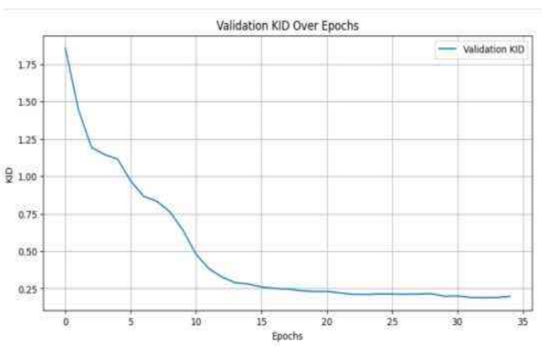
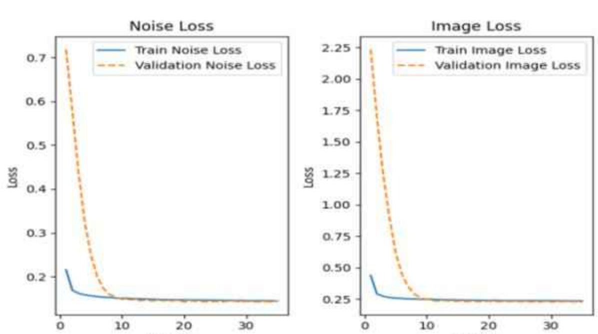


Figure 3. KID versus epochs



(a) (b)

Figure 4. (a) Train and validation noise loss over epochs (b) Train and validation image loss over epochs

Fig (3) and (4) illustrates the KID score and training or validation loss over 35 epochs on danbooru [15]. Both training and validation losses show a consistent decrease, indicating effective model optimization. Image loss indicates that difference between the generated image and the clean image should be low, reflecting the model's ability to generate accurate images. Noise loss identifies the model's ability to predict and remove noise during the diffusion process. Lower noise and image loss indicates better performance in restoring the image to its clean state.

V. CONCLUSIONS

This paper focuses on the design and assessment of a Denoising Diffusion Implicit Model (DDIM) for creating high-quality anime character images, leveraging a U-Net architecture. The proposed approach showcased the capability of DDIMs to generate intricate and visually compelling images while providing a more efficient sampling process compared to conventional diffusion models. Future work could involve adding conditional diffusion, utilizing larger datasets for better generalization and incorporating positional embeddings and thoughtfully crafted noise schedules.

REFERENCES

[1] Jiaming Song, Chenlin Meng & Stefano Ermon Denoising Diffusion Implicit International Conference on Learning Representations 2021.

[2] Jonathan Ho, Aditi Choi, Timothy Salimans Denoising Diffusion Probabilistic Models Proceedings of the 38th International Conference on Machine Learning 2020.

[3] D. Zhang, N. Tang and Y. Qu, "Joint Motion Deblurring and Super-Resolution for Single Image Using Diffusion Model and GAN," in IEEE Signal Processing Letters, vol. 31, pp. 736-740, 2024, doi: 10.1109/LSP.2024.3370491..

[4] L. Papa, L. Faiella, L. Corvitto, L. Maiano and I. Amerini, "On the use of Stable Diffusion for creating realistic faces: from generation to detection," 2023 11th International Workshop on Biometrics and Forensics (IWBF), Barcelona, Spain, 2023, pp. 1-6, doi: 10.1109/IWBF57495.2023.10156981.

[5] Alec Radford & Luke Metz, Soumith Chintala Unsupervised Representation Learning with Deep Convolution Generative Adversial Networks ICLR 2016.

[6] Noor, N.Q., Zabidi, A.B., Jaya, M.I., & Ler, T.J. (2024). Performance Comparison between Generative Adversarial Networks (GAN) Variants in Generating Anime/Comic Character Images - A Preliminary Result. 2024 IEEE Symposium on Industrial Electronics & Applications (ISIEA).

[7] Cao Y, Meng X, Mok PY, Lee TY, Liu X, Li P. AnimeDiffusion: Anime Diffusion Colorization. IEEE Trans Vis Comput Graph. 2024;30(10):6956-6969. doi:10.1109/TVCG.2024.3357568

[8] A. Purwanto, Kusrini, E. Utami and D. Agustriawan, "A Comprehensive Literature Review on Generative Adversarial Networks (GANs) for AI Anime Image Generation," 2024 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), Bandung, Indonesia, 2024, pp. 1-6, doi: 10.1109/AIMS61812.2024.10513308.

[9] S. Ruan, "Anime Characters Generation with Generative Adversarial Networks," 2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China,2022,pp.1332-1335,doi:10.1109/AEECA55500.2022.9918869.

[10] Bing, Li & Zhu, Yuanlue & Wang, Yitong & Lin, Chia-Wen & Ghanem, Bernard & Shen, Linlin. (2021). AniGAN: Style-Guided

Generative Adversarial Networks for Unsupervised Anime Face Generation. 10.48550/arXiv.2102.12593.

[11] Jing, B., Ding, H., Yang, Z. et al. Image generation step by step: animation generation-image translation. Appl Intell 52, 8087–8100 (2022). https://doi.org/10.1007/s10489-021-02835-z.

[12] Zhu, Gaofeng & Qu, Zhiguo & Sun, Le & Liu, Yuming & Yang, Jianfeng. (2024). Realistic real-time processing of anime portraits based on generative adversarial networks. 10.21203/rs.3.rs-4080250/v1.

[13] Z. Hua and L. Jie, "An Improved Deep Convolutional Generative Adversarial Network for Anime-Style Character Image Painting," 2024 9th International Conference on Computer and Communication Systems (ICCCS), Xi'an, China, 2024, pp. 102-106, doi: 10.1109/ICCCS61882.2024.10603201

[14] Z. Huang, K. Wang, Y. Xiao and Z. Xiang, "Research and Application of LSTM Model and Diffusion Model," 2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE), Jinzhou, China, 2024, pp. 875-878, doi: 10.1109/ICSECE61636.2024.10729552.

[15] S. An, "High-Resolution Anime Face Dataset (512x512)," Kaggle, 2020.[Online].Available:https://www.kaggle.com/datasets/subinium/highresolution-anime-face-dataset-512x512. [Accessed: Sept.6, 2024].

[16] M. Odhiambo, "CelebA-HQ resized (256x256)," Kaggle, 2020. [Online].Available:https://www.kaggle.com/datasets/mosisodhiambo/celeba-hq-resized-256x256. [Accessed: Nov. 20, 2024].

[17] C. L. Storkey, "Oxford Flowers dataset," TensorFlow Datasets, 2018. [Online].Available:https://www.tensorflow.org/datasets/community_catalog/huggingface/oxford_flowers102. [Accessed: Nov. 5, 2024].

[18] J. Ho, "DDPM: Denoising Diffusion Probabilistic Models," GitHub repository,2020.[Online].Available:https://github.com/hojonathanho/diffusion. [Accessed: Oct. 24, 2024].

[19] L. X. Moreh, "DDIM Keras Example," GitHub repository, 2023. [Online]. Available: https://github.com/loctxmoreh/ddim-keras-example.