Diffusion Models for MetFaces-styled Image Generation

Singaravel, Siddharthan
singaravel.s@northeastern.edu

Rosen, Joshua
rosen.jos@northeastern.edu

Jebaraj, Ainsley
jebaraj.a@northeastern.edu

***Abstract*—This project aims to build a model that will generate novel artwork influenced by the style of well-known paintings. To that end, in the style of New York Metropolitan Museum's publicly available MetFaces dataset which features a select set of curated face images. Employing state-of-the-art diffusion models within a modified U-Net architecture, this research focuses on synthesizing high-quality facial images by gradually transforming a random noise distribution into structured artwork. The approach involves a comprehensive evaluation of different noise schedules—linear and cosine—and varying numbers of iterations to refine the model's output. Results from the experiments indicate that the cosine noise schedule, especially at higher iteration counts, significantly enhances the quality of the generated images, as evidenced by lower Fréchet Inception Distance (FID) scores. This study not only underscores the capabilities of diffusion models in artistic image generation but also opens avenues for future research in creative AI applications, offering a promising method for both preserving and reinterpreting cultural heritage through technology.**

I. Introduction

A. Overview of the Project

The interplay between art and technology has long captured the imagination, leading to innovations that redefine how we perceive and create art. This project stands at the confluence of these realms, aiming to generate original and novel images of faces in styles inspired by artworks curated at the Metropolitan Museum of Art. By leveraging cutting-edge deep learning techniques, specifically diffusion models, this initiative seeks not only to synthesize art but also to deepen our understanding of generative model capabilities.

B. Motivation

Artistic content generation using deep learning models is a burgeoning field that promises to democratize art creation, enhance creative processes, and provide insights into the cognitive aspects of art perception. By automating the synthesis of art in styles reflective of human creativity, such models can assist artists in exploring new creative avenues and help preserve artistic heritage in digital forms. Additionally, this project contributes to the academic and practical discourse on the potentials and ethics of AI in creative industries.

C. Approach

Our approach utilizes state-of-the-art diffusion models, a class of generative models that have shown significant promise over traditional Generative Adversarial Networks (GANs) in generating high-quality images. Diffusion models operate by gradually transforming a random noise distribution into a structured image, mimicking the process of developing a photograph. The choice of diffusion models is based on their recent successes in producing images that closely resemble natural photographs, their flexibility in handling diverse datasets, and their robustness against mode collapse—a common issue in traditional GANs. In pursuit of our objectives, we will engage in a detailed examination of the open-source codebases available for diffusion models, ensuring a thorough understanding of their mechanisms and optimizing them for our specific task of generating art-inspired images.

D. Dataset for Experiments and Evaluation

The practical aspect of our research will utilize the "MetFaces" dataset, consisting of 1336 PNG images. This dataset is curated by NVIDIA Corporation and made available under the Creative Commons BY-NC 2.0 license. These images, which capture a variety of facial expressions and features drawn from the Metropolitan Museum of Art's extensive collection, provide a rich basis for training our models. This selection not only aligns with our goal of generating art-inspired faces but also ensures compliance with licensing and ethical standards for data usage.

II. Background

A. Evolution of Image Generation with Deep Learning

The quest for automated image generation has been a significant area of focus within the field of artificial intelligence, particularly within the domain of deep learning. The initial strides were made with the advent of Generative Adversarial Networks (GANs), introduced by Goodfellow et al. in 2014. GANs set a new standard for image quality and diversity by effectively learning to mimic various data distributions. The foundational concept of GANs involves two neural networks—generative and discriminative models—engaged in a zero-sum game to respectively generate new images and evaluate their authenticity.

Over time, various enhancements and iterations such as Deep Convolutional GANs (DCGANs), Wasserstein GANs (WGANs), and StyleGAN have emerged. Each of these variants brought improvements in stability, image quality, and the ability to handle higher resolutions. DCGANs, for example, introduced convolutional layers into GAN architectures, significantly boosting their performance on image tasks. WGANs modified the loss function used in training GANs to improve

model training and address the issue of mode collapse. StyleGAN, later on, provided a means to control specific features in generated images, allowing for unprecedented customization of generated outputs. [1]–[4].

B. Shift to Diffusion Models

Despite these advancements, a transformative shift occurred with the development of likelihood-based diffusion models. These models represent a different approach where images are generated through a process of adding and then iteratively removing noise. Research indicates that diffusion models tend to achieve better fidelity and diversity in generated samples compared to state-of-the-art GANs. [5], [6] They operate on the principle of starting with a distribution of noise and gradually converting it into a structured image across numerous steps, akin to developing a photograph.

Diffusion models have shown remarkable success in various domains, including image synthesis and text-to-image generation, demonstrating their versatility and robustness. The capability of these models to produce highly detailed and realistic images stems from their unique training dynamics, which learn an optimal path for reversing the noise addition process.

C. Physics-Inspired Generative Models

Adding another layer to the generative landscape are physics-inspired models like Poisson Flow Generative Models (PFGMs) and PFGM++. [7] These models integrate concepts from physical processes to guide the generation of new data points in the model's latent space. By emulating the flow of a physical system, such as the diffusion of heat or the distribution of particles, these models can introduce an additional level of naturalism and coherence to the generation process.

D. The Role of Open Source and Community Collaboration

A pivotal aspect of advancing diffusion model technology has been the role of open-source contributions. By examining and utilizing publicly available code, researchers can build upon pre-existing work, accelerating innovation and ensuring a broad testing ground for new theories and techniques. Open-source projects also foster a collaborative environment where ideas can be quickly disseminated and iteratively improved upon, which is crucial for tackling complex problems such as high-fidelity image generation.

III. Approach

In this project, we employ a diffusion-based generative model tailored to create art-styled facial images, using a U-Net architecture optimized for our specific application. This section discusses the diffusion process, the architectural specifics of our model, and the dataset preparation integral to training our model.

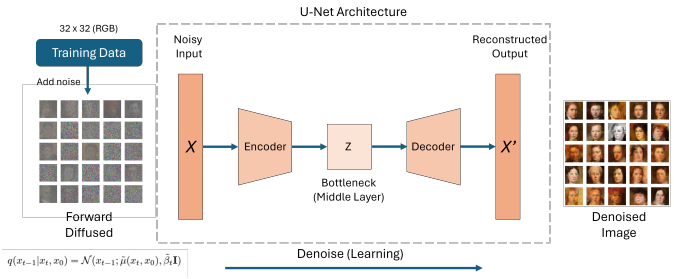

Fig. 1: U-Net architecture for encoding-decoding operations including a bottleneck multi-layer perceptron network; noisy input passed as input to be denoised is the learning objective

Fig. 2: Addition of gaussian noise to the training set

A. Network Architecture

Our model is built upon a modified U-Net architecture, which is fundamentally composed of a series of downsampling and upsampling blocks connected by skip connections. This architecture is well-suited for tasks that involve image generation from noisy data, as it effectively captures and synthesizes details at multiple scales.

B. Encoder-Decoder Structure with Skip Connections

The encoder part of the network progressively reduces the spatial dimensions of the input image while increasing the depth of feature maps. This downsampling process captures the essential features at different resolutions. The decoder part then progressively reconstructs the image from the condensed feature representation using upsampling layers. Skip connections between corresponding downsampling and upsampling layers help in recovering fine details by directly propagating features from the encoder to the decoder.

C. Positional Embeddings

To incorporate the sequential nature of the diffusion process, positional embeddings are integrated into the architecture. These embeddings provide the model with temporal context necessary to guide the generation process through various stages of noise reduction.

D. Diffusion Process

The core idea behind diffusion models is to start with a distribution of noise and gradually refine this distribution through a series of learnable reverse steps until it converges to the distribution of the target data.

E. Forward Process

The forward process involves adding noise to the original images in a controlled manner over several steps. At each step, the image becomes noisier, but crucially, the process is designed so that the noise can be reversed.

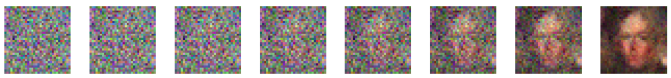

Fig. 3: Separating the forward diffused noise is the objective of learning

F. Reverse Process

The reverse process is where our model learns to generate images. Starting from pure noise, the model applies a series of transformations that gradually denoise the input to form an image. Each step of this process is guided by learned parameters that effectively invert the forward noise addition.

G. Training Strategy

The model is trained by first executing the forward process to generate noisy images at various stages. These noisy images are then used as input to the reverse process. The goal during training is to minimize the difference between the original images and the reconstructed images from the reverse process, allowing the model to learn how to effectively denoise inputs.

H. Loss Function

The primary metric for training efficacy is the difference in pixel values between the original clean images and the images output by the reverse process. We utilize an L1 loss, which promotes less blurring, a desirable attribute for maintaining the artistic integrity of generated images.

I. Dataset and Preprocessing

For training and validation, we use the MetFaces dataset , which consists of 1,336 images derived from artworks in the Metropolitan Museum of Art. [**?**] These images are preprocessed to a uniform size to suit the input requirements of our network and normalized to facilitate faster and more stable convergence.

To enhance the robustness of our model and to prevent overfitting, we apply data augmentation techniques such as random cropping, flipping, and rotation. This ensures that the model is not just memorizing specific artworks but learning to generalize from the artistic styles embodied in the dataset.

RESULTS

The primary metric for assessing the quality of the generated images in this project is the Fréchet Inception Distance (FID) score. The FID score is a widely recognized measure in the domain of generative models, providing a quantitative estimate of the similarity between the distributions of generated images and real images. Lower FID scores indicate that the generated images are more similar to the real images, suggesting better model performance.

We conducted experiments to evaluate the effectiveness of our diffusion model under different training configurations, particularly focusing on the number of training iterations and the type of noise schedule used during the diffusion process.

Iterations	Noise Schedule	FID
1000	Linear	222.26
500	Linear	325.45
1000	Cosine	189.77
500	Cosine	190.02

Table 1: FID scores for different iterations and noise schedules

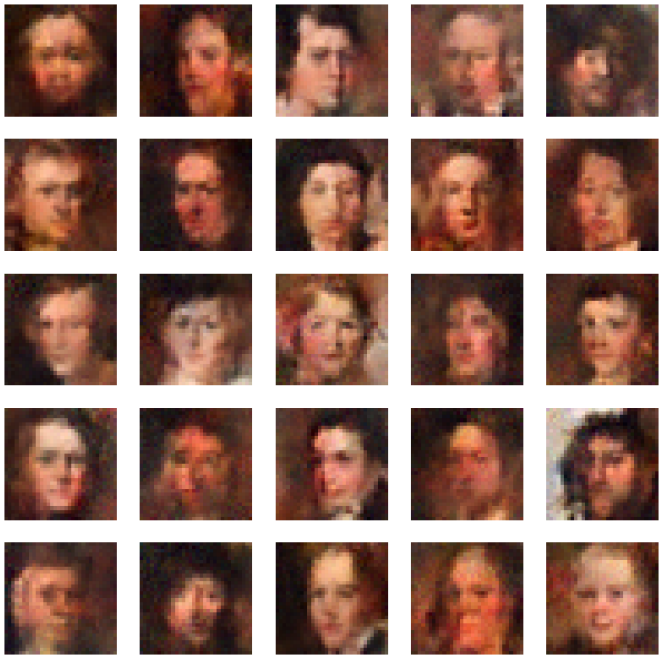

Fig. 4: Images synthesized by the Diffusion model

The experiments were structured to compare the impact of using linear and cosine noise schedules across different iteration counts.

The FID scores obtained from the experiments clearly demonstrate the influence of both the number of iterations and the type of noise schedule on the quality of images generated by the diffusion model. Increasing the number of iterations from 500 to 1000 consistently leads to lower FID scores, indicating better synthesis of images. This improvement is evident in both noise schedules, but it is particularly significant in the linear noise schedule, where the FID score decreases from 325.45 to 222.26. This suggests that more iterations allow the model to refine its denoising process more effectively, which is critical in achieving higher fidelity in the generated images.

The cosine noise schedule outperforms the linear noise schedule across both iteration counts, with notably lower FID scores in each case. For example, at 1000 iterations, the cosine schedule achieves an FID of 189.77 compared to 222.26 for the linear schedule. This indicates that the cosine method, which adjusts noise levels in a non-linear fashion, is more adept at modeling the complexities involved in generating high-quality images. The cosine schedule's superior performance is also evident at the lower iteration count of 500, where it almost matches the performance of the higher iteration count with

an FID score of 190.02. This near equivalence suggests that the cosine schedule is not only more effective but also more efficient, achieving similar quality outputs in fewer iterations.

This efficiency is particularly significant as it implies that the model can deliver high-quality results with reduced computational demands. This efficiency makes the cosine noise schedule a preferable choice in scenarios where computational resources are a constraint. Moreover, these findings highlight that optimizing the noise schedule could yield more substantial improvements in model performance than simply increasing the number of iterations. This insight opens up new avenues for research, particularly in exploring other noise modulation strategies that could further enhance the quality of generated images while maintaining or even reducing the computational load.

Discussion

This project highlights the efficacy of diffusion models in generating high-quality, art-inspired images, showcasing significant advancements in the interplay between deep learning and artistic creation. The results emphasize the importance of the noise schedule and iteration count in optimizing model performance.

The superior performance of the cosine noise schedule, particularly at lower iteration counts, suggests that this approach can efficiently manage noise to produce high-quality images. This finding underlines the potential for diffusion models to not only create art but also to serve as tools for preserving and reinterpreting cultural heritage.

For future research, exploring various adaptive noise scheduling techniques could optimize performance further. Additionally, integrating feedback mechanisms that refine noise adjustments dynamically during training may enhance both the quality and efficiency of the generative process. This approach could expand the model's utility in artistic and other creative domains, pushing the boundaries of what generative models can achieve.

Conclusion

This project successfully leveraged diffusion-based generative models, specifically optimized through a U-Net architecture, to create high-quality, art-inspired facial images. By synthesizing images that mimic the styles curated at the Metropolitan Museum of Art, this study demonstrates the potential of deep learning technologies in bridging the gap between art and artificial intelligence. The efficacy of diffusion models, particularly under the influence of different noise schedules and iteration counts, highlights their robustness and versatility in generating complex image distributions. The results clearly show that the cosine noise schedule, combined with sufficient training iterations, optimally enhances the quality of generated images. This research not only pushes the boundaries of generative AI in artistic domains but also suggests a promising avenue for preserving and reinterpreting cultural heritage through technology. The takeaway message is clear: advanced diffusion models hold significant promise for the future of art creation and preservation, marrying the richness of artistic expression with the precision of machine learning.

Acknowledgment

This authors would like to thank Lilian Weng of OpenAI [9] for her blog explaining the math behind diffusion models and Hugging Face's annotated diffusion models repository [10] which were helpful in building equations for the code implementations. The authors would also want to thank NVIDIA for curating the dataset used in this project.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

[2] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

[3] Arjovsky, M., Chintala, S., & Bottou, L. (2017, July). Wasserstein generative adversarial networks. In International conference on Machine Learning (pp. 214-223). PMLR.

[4] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp.4401-4410).

[5] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851.

[6] Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems, 34, 8780-8794.

[7] Xu, Y., Liu, Z., Tegmark, M., & Jaakkola, T. (2022). Poisson flow generative models. Advances in Neural Information Processing Systems, 35, 16782-16795.

[8] https://github.com/NVlabs/metfaces-dataset

[9] https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

[10] https://huggingface.co/blog/annotated-diffusion

All authors contributed equally.