Project Title: Face Age Editing with Fine-Tuned Diffusion Models: A Comparative Study

Name: Dinh Quoc Vuong, Sophomore Undergraduate
Major in Data Science,
Contact: dqvuong@wisc.edu

Name: Shashwat Negi, Graduate student
Major in Data Science,
Contact: negi3@wisc.edu

Name: Ian Franda, Senior Undergraduate
Major in Statistics, Math, and Computer Science,
Contact: ifranda@wisc.edu

Name: Juncheng Long, Senior Undergraduate
Major in Statistics, Math, and Data Science ,
Contact: jlong32@wisc.edu

Course Number: Stat 453, Spring 2025
Team number: 20

1. **Motivation.** In recent years, advancements in diffusion models have placed them at the forefront of image generation research. One interesting application of these models is face aging – the task of simulating how a person's appearance changes over time. Given an image of someone's face, these models aim to generate realistic and identity-preserving representations of that individual at different ages. Our goal is to contribute to the development of diffusion models trained for this particular task, and to compare the effectiveness of various diffusion models with various datasets. We will fine tune common diffusion models by injecting new labeled face datasets and study how architectural differences between models influence the effectiveness of generating realistic face aging transformations.
2. **Existing literature.** Image synthesis was first popularized through Generative Adversarial Networks (GANs) [1], which enabled realistic image generation but suffered from instability and limited diversity. Denoising Diffusion Probabilistic Models (DDPMs) [2] addressed these limitations with more stable training and higher fidelity outputs. Recent literature has applied diffusion models to the task of face aging. In 2023, the model 'Face Aging via DIffusion-based editiNG' (FADING) [4] took a pre-trained diffusion model and specialized it via age-aware training for the task of face aging. The results far exceeded that of any previous model built. Even more recently in 2024, AgeDiff [3] introduces a novel dual cross-attention mechanism for face age editing, supporting both context-fixed and context-free transformations with fine-grained control. The application of diffusion models in the realm of face-aging is a new and exciting subject that this project hopes to add to.

3. **Limitations of existing models or methods.** Vanilla DDPMs, despite producing high-quality samples, suffer from three major limitations: poor log-likelihood performance, inefficient noise scheduling, and slow sampling. The fixed variance used in the reverse process and the simplified training objective lead to weak mode coverage, reflected in high negative log-likelihood (NLL) scores. Additionally, the linear noise schedule destroys useful information too quickly—especially in early steps—reducing both efficiency and sample quality. Finally, generating samples requires hundreds to thousands of denoising steps, making DDPMs computationally expensive and slow. Given these limitations, we aim to compare various existing diffusion models for face age editing (FAE) and explore ways to improve their performance for more efficient and accurate age progression.

Models. We will fine-tune and compare multiple diffusion models for this project. For our experiments, we intend to work with Denoising Diffusion Probabilistic Models (DDPMs), a class of generative models that have recently demonstrated state-of-the-art performance in image synthesis tasks. DDPMs operate by gradually corrupting an image with Gaussian noise over a series of time steps and then learning to reverse this process to generate realistic samples from pure noise. In the context of face aging, we aim to fine-tune a pretrained DDPM on age-annotated face datasets to conditionally generate aged versions of input faces. A number of models we are considering include the following. Improved Denoising Diffusion Probabilistic Model[5], which uses the base DDPM but with 3 distinct improved features: (1) Learned variance - it uses learned vector to interpolate the variance of the reverse process; (2) Hybrid Loss: it combines the simplified noise prediction loss with a small-weighted ELBO term; (3) Cosine Noise Schedule - it replaces the linear noise schedule with a cosine-shaped one to prevent over-noising early or late.

Datasets. For this project, we will utilize publicly available, high-resolution face datasets that are commonly used for facial analysis and face generation tasks. One of the primary datasets considered is CelebA-HQ, a high-quality version of the original CelebA dataset, released by

NVIDIA. It consists of approximately 30,000 high-resolution face images with resolutions up to 1024×1024.The dataset is open-source at https://github.com/switchablenorms/CelebAMask-HQ.
Another key data set is FFHQ (Flickr-Faces-HQ), a high-quality face dataset also released by NVIDIA, originally intended for training StyleGAN and other generative models. It contains 70,000 high-resolution face images (1024×1024) collected from Flickr and curated to cover a wide demographic range, including variations in age, ethnicity, lighting, and backgrounds.
Building upon FFHQ, will use the FFHQ-Aging dataset, a curated and age-annotated subset specifically tailored for facial aging tasks. The dataset organizes images into discrete age brackets (e.g., 0–2, 3–6, 30–39, 60+), either through automated age estimation or manual annotation. This dataset is also publicly available and can be found at https://github.com/royorel/FFHQ-Aging-Dataset.

Measurements. To assess the realism and diversity of the generated aged faces, we will use FID and IS. To measure how accurately the model transforms faces to the intended age group we will use MAE. Details are as follows:

1. Fréchet Inception Distance (FID): Measures the distributional difference between real and generated images using features extracted by an Inception network. Lower FID indicates more realistic and closer-to-real distributions.

2. Inception Score (IS): Evaluates both the quality and diversity of generated images using classification confidence and entropy. Higher IS reflects sharper images and better diversity across age groups.

3. Mean Absolute Error (MAE): Compares the predicted age (using a pretrained age estimator from HuggingFace) with the target age label. Lower MAE indicates better aging accuracy.

Compute budget. Describe how many models you want to experiment with, and the size of each model. For example, you many want to test different CNNs, Vision Transformers on an image task. Estimate the time you need for each model. Some suggestions include the following.

- Count the number of parameters in the model : The total number of parameters depends on the image resolution. For a pre-trained model like AgingDiff (DDIM-based), it has approximately 100M–150M parameters for 512×512 images.

- Conduct preliminary experiment to find how long the model spends on one batch of training data : Fine-tuning a pre-trained model with images at a resolution of 512×512 and a batch size of 16 takes approximately 2–3 seconds per batch on a GPU. Future model training will be conducted on the CHTC Condor compute resource for approximately 80–100 epochs.

Collaboration plan. Each team member is equally responsible, committed, and will actively contribute to the project. Tasks will be divided across model development, data preparation, and evaluation metric creation to ensure balanced and efficient collaboration.

References

[1] Goodfellow, Ian, *et al.* "Generative adversarial networks." Communications of the ACM 63.11 (2020): 139-144.

[2] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.

[3] Grimmer, Marcel, and Christoph Busch. "AgeDiff: Latent Diffusion-based Face Age Editing with Dual Cross-Attention." 2024 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2024.

[4] Chen, Xiangyi, and Stéphane Lathuilière. "Face aging via diffusion-based editing." arXiv preprint arXiv:2309.11321 (2023).

[5] Nichol, Alexander Quinn, and Prafulla Dhariwal. "Improved denoising diffusion probabilistic models." International conference on machine learning. PMLR, 2021.

[6] Or-El, Roy, *et al.* "Lifespan age transformation synthesis." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. Springer International Publishing, 2020.