Face Age Editing with Latent Diffusion Models: A Comparative Study

Juncheng Long [1] **Dinh Quoc Vuong** [2] **Ian Franda** [3] **Shashwat Negi** [4]

Abstract

This project explores the problem of face age editing-generating realistic images of individuals at a target age-using fine-tuned diffusion model. We conduct a comparative study of two fine-tuning methods for the Stable Diffusion model: DreamBooth and Inversion + Cross-Attention Control. Our goal is to evaluate each method in terms of visual quality, identity preservation, aging accuracy, and computational efficiency. With both qualitative and quantitative analyses, we assess the trade-offs between the two approaches and identify the limitations of each method. Our results indicate that DreamBooth offers more consistent identity preservation, while the Image Inversion + Cross-Attention method is less costly and more flexible for localized edits. This study provides practical insights for selecting appropriate techniques in face editing tasks with generative diffusion models

1. Introduction

In this project, we perform a comparative study of two such fine-tuning approaches: DreamBooth and Image Inversion + Cross-Attention. We aim to evaluate each method's ability to generate realistic and identity-preserved aged faces, while also comparing their computational cost and feasibility.

The structure of this report is as follows. Section 1 provides an introduction to fundamental concepts, such as Face Age. Editing, Existing Models, DreamBooth, and Image Inversion + Cross Attention. In Section 2, we review previous work related to face aging, generative adversarial networks, and diffusion-based image editing. Section 3 presents the dataset and software used in our study. Section 4 clarifies our methods and evaluation metrics. Section 5 presents the qualitative and quantitative results. Finally, Section 6, 7, and 8 discusses the implications of our results, suggests potential direction for future, and give a conclusion of our project.

1.1. Face Age Editing

Age progression and regression in facial imagery—commonly referred to as face age editing—has become an important task in computer vision with applications in entertainment, forensics, and personalized user experiences. The main goal is to synthetically alter the perceived age of a person in a facial image while preserving their identity and realism. "Most existing aging methods are limited to changing the texture, overlooking transformations in head shape that occur during the human aging and growth process" [6].

1.2. Existing Models

Image synthesis was first popularized through Generative Adversarial Networks (GANs), which enabled realistic image generation but suffered from instability and limited diversity [1]. Denoising Diffusion Probabilistic Models (DDPMs) addressed these limitations with more stable training and higher fidelity outputs [2]. Their iterative denoising process enables a natural integration of guidance mechanisms, including text prompts and cross-attention, to achieve targeted image editing. 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.

2. Related Work

The field of generative face editing has achieved significant progress in recent years, mainly driven by the development

[*]Equal contribution [1]Department of Statistics, Major in Statistics, Math and Data Science, Senior Undergraduate [2]Department of Statistics, Major in Data Science, Sophomore Undergraduate [3]Department of Statistics, Major in Statistics, Math and Computer Science, Senior Undergraduate [4]Department of Statistics, Major in Data Science, Graduate. Correspondence to: Shashwat Negu <negi3@wisc.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

of deep generative models such as Generative Adversarial Networks (GANs) and Denoising Diffusion Probabilistic Models (DDPMs). Early methods for face aging relied on GANs, which showed strong capabilities in learning age-related transformations through conditional generation and latent space manipulation. Models like HRFAE, LATS, and Re-aging GAN incorporated age embeddings or regressors to guide synthesis while attempting to preserve identity [4]. However, GAN-based methods often fail on identity consistency and limited generalization to rare facial conditions such as extreme poses or occlusions [6]. To solve these limitations, recent research has shifted toward diffusion-based models, which perform better than GANs in sample consistency and semantic control [4]. In particular, diffusion models such as DDPMs offer more robust image generation, and methods like FADING have shown that fine-tuning with age-specific prompts combined with attention control enables effective aging transformations while maintaining facial features [4]. Among diffusion-based fine-tuning strategies, two stand out for personalized generation tasks: DreamBooth and Image Inversion with Cross-Attention. On one hand, DreamBooth introduces a personalization framework where a few subject images are used to bind a unique identifier to the target identity in a pre-trained text-to-image model. This enables reliable regeneration of the subject in new contexts with high accuracy to identity and semantics [7]. DreamBooth also incorporates a class-specific prior preservation loss to mitigate overfitting and improve diversity in generated outputs. On the other hand, the image inversion and cross-attention approach—used in methods like FADING—relies on reversing an input image into the latent space and selectively modifying the attention maps to localize age-specific transformations. This method provides a high degree of control over which parts of the image are edited, resulting in minimal distortion of identity features not associated with age [4].

While both approaches used diffusion models, their trade-offs are different: DreamBooth needs more training resources for subject embedding, whereas cross-attention-based editing is lightweight and concentrate more on attribute-level transformations. These differences underscore the need for comparative evaluation, which this study aims to provide.

3. Experimental Setup

3.1. Model Configuration

We used Stable Diffusion-v1.5 as the base model for our project, which operates on 512×512 face images in latent space. The model includes a frozen CLIP ViT-L/14 text encoder with about 123 million parameters and a U-Net denoiser with around 860 million parameters. The images are first passed through a 4× downsampling Variational Autoencoder (VAE), which compresses them into latents before the denoising process begins. The U-Net has four resolution levels, with output channels [320, 640, 1280, 1280], and each level contains two convolutional layers. Most of the downsampling and upsampling blocks use cross-attention, which allows the model to condition on text prompts, and there is an additional cross-attention block in the middle. The attention mechanism uses multi-head attention with a head size of 8 and a cross-attention projection dimension of 1280, matching the text embedding size. For fine-tuning, we followed the DreamBooth method by adding a custom token (like [skw]) to the text encoder to personalize the model to our task. We also used the inversion and text-prompt guided technique, which involves inverting face images into latent space using null-text inversion and then guiding the model to focus on age-related features through cross-attention. The core model architecture remained unchanged—we only fine-tuned specific attention maps and embeddings for our task.

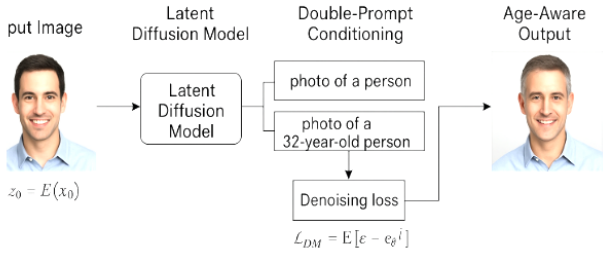

Figure 1. Face-Age Editing Process in Adaptively Fine-Tuned Stable Diffusion Models Using Inversion Guidance

3.2. Hyperparameters

For training, we used the Adam optimizer with standard settings: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1\text{e}^{-6}$. To prevent overfitting during fine-tuning, a small weight decay (between 0 and 0.01) was applied. We kept the learning rate low, typically in the range of 1e^{-5} to 5e^{-6}, which aligns with best practices in DreamBooth-style fine-tuning for stable convergence. We trained with a small batch size of 5 images with 400 training steps. We used early stopping by monitoring the validation loss to avoid overfitting. During training, we only updated the U-Net and the newly added token embedding, keeping the CLIP text encoder frozen. We did not use LoRA or apply any changes to batch normalization layers.

3.3. Dataset

For the Inversion model, we used a sample of FFHQ-Aging dataset, an extension of NVIDIA's FFHQ face dataset, which contains approximately 70,000 high-resolution im-

ages (1024×1024) with age groups information across 10 discrete classes. In our setup, each face image is associated with an integer age or an age group label. For preprocessing, all images were center-cropped and resized to 512×512 pixels to match the input requirements of the Stable Diffusion model. Pixel values were normalized based on the autoencoder's preprocessing pipeline, and since FFHQ images are already well-aligned, no further alignment was needed. The dataset was randomly split into training and validation sets in an 80:20 ratio, stratified by age group to ensure a balanced distribution. Due to our limited computational resources, we used only a few thousand samples for training and evaluation.

For the DreamBooth method, we constructed a few-shot training set of five images of one of our project members. The images were cropped to be 512x512 and were taken in various settings.

3.4. Loss Functions

We use two main loss functions in our model: *denoising (reconstruction) loss* and *cross-attention loss*. The *denoising loss* is a standard approach in latent diffusion models, where the model learns to remove noise and reconstruct the original image. This is achieved by minimizing the L2 reconstruction loss between the predicted noise and the actual noise. The loss is computed as:

$$\mathcal{L}_{\text{denoise}} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{\epsilon}_i - \epsilon_i\|^2$$

Where $\hat{\epsilon}_i$ is the predicted noise and ϵ_i is the actual noise at timestep i.

In addition, we optionally apply a *cross-attention loss* to align the attention maps of the "[age]" token with relevant facial regions. This encourages the model to focus on age-related features like wrinkles or skin texture. The cross-attention loss is calculated using cosine similarity between the attention maps of the "[age]" token and a face mask:

$$\mathcal{L}_{\text{cross-attn}} = 1 - \frac{\sum_{i=1}^{N} \text{Attention}_i^{\text{age}} \cdot \text{Attention}_i^{\text{mask}}}{\left\|\text{Attention}_i^{\text{age}}\right\|_2 \cdot \left\|\text{Attention}_i^{\text{mask}}\right\|_2}$$

Finally, the *total loss* combines the denoising loss and the cross-attention loss (if used), with a weight λ controlling the contribution of the cross-attention term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{denoise}} + \lambda \cdot \mathcal{L}_{\text{cross-attn}}$$

This formulation helps the model effectively focus on age-related features while also denoising the image.

4. Methods and Evaluation Metrics

This section outlines the complete process used to compare DreamBooth and Image Inversion with Cross-Attention for face age editing using diffusion models. Our experimental pipeline begins by selecting a set of subject images and defining target age prompts. Once both methods generate aged versions of the subject, we evaluate the outputs based on identity preservation, aging accuracy, visual quality, and computational efficiency.

4.1. DreamBooth Implementation

The process of the DreamBooth method involved fine-tuning a Stable Diffusion model (v1.5) on a small set of images. Our goal was to override the meaning of a rare token (in this case "skw") in Stable Diffusion so that the model associated this unique token with an individual's facial identity. This would allow us to generate pictures of this individual in different contexts, such as different ages.

We used six 512x512 images of the same individual taken under different poses and lighting conditions. Each image was paired with the prompt "photo of skw person". We also sought to avoid catastrophic forgetting, which is when the model cannot generalize. To do this, we applied prior preservation using generic class images generated from the prompt "photo of person" to ensure that the model could still generate pictures of a "generic" person.

Training was conducted in Google Colab with a T4 GPU and used Hugging Face's diffusers library. We trained all 859 million parameters over 400 training steps and 10 class images with a learning rate of 1e-5, and we utilized an 8-bit Adam optimization strategy. We chose these parameters after some trial and error – we found that this combination trained reasonably fast and could still generalize to a generic person.

This implementation aimed to preserve the identity of the individual across different contexts – in this case at different ages. Producing output from the model involved prompts indicating the desired age, such as "photo of skw person at 80 years old".

The figures on the following page show how the model develops over n training steps. Every 50 training steps, the model was saved. After training, each saved model was asked to generate a photo with the prompt "photo of skw person", and the results are shown to the right. The generation is certainly messy, but a clear progression is seen. At first, the model does a poor job at capturing the individual's features, but after 200 training steps key features begin to appear, such as hair style and nose shape, and these features become more apparent as the model approaches 400 training steps. Figure 2 shows two of the training images for reference.

Clearly, the model is able to learn what this individual's face looks like. In the DreamBooth outcome section (4.4), we will show how we leverage this to generate aged photos of the individual.

Training Image 1

Training Image 2

Figure 2. Example training images used for DreamBooth fine-tuning. Each image captures the same subject under different poses and lighting conditions.

4.2. Inversion and Cross Attention

We began by fine-tuning a pre-trained latent diffusion model to specialize in age transformation tasks. This process utilized the FFHQ-Aging dataset, which extends the original FFHQ dataset by providing age annotations, gender information, head pose, and other facial attributes. During fine-tuning, we adopted an age-aware scheme where each image was paired with prompts like "photo of a 30-year-old person." This method, known as the double-prompt scheme, helps the model disentangle age-related features from other facial attributes, enabling more precise age editing.

Image Inversion Guidance

To edit a specific input image, we inverted it into the latent noise space of the diffusion model using DDIM inversion. This step allows the model to reconstruct the original image accurately. Concurrently, we optimized the null-text embeddings, which are used in classifier-free guidance during the diffusion process. By adjusting these embeddings, the model better captures the unique characteristics of the input image, enhancing the fidelity of the reconstruction.

Text-Guided Age Editing via Attention Control

With the specialized model and optimized embeddings, we performed age editing guided by textual prompts specifying the target age. Using cross-attention mechanisms, the model focuses on age-related facial features such as wrinkles and skin texture, allowing for localized editing without altering other attributes like identity or expression. This attention-based control ensures that the edited image maintains the individual's original identity while reflecting the desired age transformation.

Figure 3. DreamBooth model outputs after n training steps. The model was sampled every 50 steps from 50 to 400.

4.3. Evaluation Metric

For performance assessment, we used standard metrics commonly employed in image generation tasks: the Fréchet Inception Distance (FID) and the Inception Score (IS) to evaluate the quality and diversity of the generated images. The FID measures the similarity between the distributions of generated and real images by comparing the mean and covariance of features extracted from a pre-trained Inception

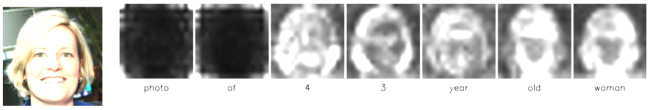

Figure 4. Cross Attention Map of FFHQ-Aging Sample

v3 network, with lower scores indicating closer resemblance to real images. The IS assesses both the quality and diversity of the generated images by evaluating the entropy of predicted class distributions, where higher scores suggest more realistic and varied outputs.

5. Outcome

5.1. DreamBooth

The DreamBooth-trained model successfully learned to associate the unique token "skw" with the individual's facial features. Generated images of this token consistently produced images of a person that resembled the individual, and basic prompts such as "photo of skw person" produced images that appeared similar, yet not identical, to the training images. Transformative prompts such as "photo of skw person at 80 years old" maintained many key features from the training data, but had the intended effect of making the individual appear older, such as adding wrinkles and gray hair. We found no evidence of catastrophic forgetting, as prompts such as "photo of a person" did not produce unreasonable output.

Figure 5. Training image

Figure 6. Image Generated

5.2. Inversion and Cross Attention

Our approach, utilizing inversion and cross-attention guided mechanisms within the Stable Diffusion framework, effectively generated aged images of individuals while preserving their original identity and background features. Unlike DreamBooth methods, which often require extensive fine-tuning and may alter background elements, our technique maintained the integrity of the entire image. This was achieved through precise attention control mechanisms that focused on age-related facial features without impacting other regions. Quantitatively, our method demonstrated superior performance in qualitative metrics. Specifically, it achieved lower Fréchet Inception Distance (FID) scores compared to DreamBooth, indicating higher fidelity and more realistic age transformations. Additionally, our model's ability to retain background details contributed to more coherent and contextually accurate outputs.

Figure 7. Training image for inversion guided model

Figure 8. Image generated by inversion guided model

5.3. Computation Cost

For training and fine-tuning our model, we utilized both **Google Colab** and **Amazon EC2 instances**. Initially, we used Google Colab for small-scale experiments due to its free access to GPUs such as the *NVIDIA Tesla T4*, which were sufficient for testing various model configurations and running preliminary experiments. The free tier of Google Colab provided us with limited GPU access, offering about 12 hours of runtime per day. Colab was particularly useful for quick iterations with smaller batches.

To scale up for more intensive tasks, we transitioned to Amazon EC2 instances, specifically the *c5.9xlarge* instance, which is equipped with *32v CPU*, *72 GB of memory*. The EC2 instances allowed us to train on sample dataset with small batch sizes and mid computational power. The total EC2 usage for the lightweight fine-tuning was approximately 14 hours.

Budget Breakdown:

- **Google Colab (Free Tier):** Used for initial experiments and small-scale training. Provided access to *Tesla T4* GPUs for about 12 hours per day.
- **Amazon EC2 c5.9xlarge Instance:**
  - **Hourly Cost:** $0.526 per hour.
  - **Fine-tuning/Training Duration:** 40 hours for the full training and fine-tuning process.
  - **Total EC2 Cost:** $21.04

Thus, the total computational cost for the training phase was approximately **$25**, including both the free-tier usage of Google Colab and the paid EC2 instance usage.

6. Discussion

Our comparative analysis of DreamBooth and Image Inversion with Cross-Attention for face age editing using diffusion models indicates distinct strengths and trade-offs for each method.

The figure on the following page shows a side-by-side comparison of the two models as they generate the individual at various ages between 30 and 80. Below that figure is a table of the Fréchet Inception Distance (FID scores) for each image. The FID score compares the distribution of features in generated images to those in real images using a pretrained image classifier. Lower FID scores signify more realistic looking images.

The DreamBooth method allows for more variety in the output images. With further prompt engineering, the individual could be generated in many different contexts outside of various ages. Its fine-tuning approach effectively captures personalized facial features, resulting in high-fidelity outputs.

Image Inversion with Cross-Attention achieved more precise age transformations, particularly when dealing with large age gaps. This approach had a lower FID score across all images when compared to the DreamBooth method, indicating more realistic-looking images overall. Its prompt-based editing allows for nuanced control over age-related changes.

Both methods produced high-quality images; however, DreamBooth occasionally introduced artifacts due to overfitting, while Image Inversion maintained consistent visual realism.

Image Inversion proved to be more resource-efficient, requiring less computational power and time, as it avoids the need for model fine-tuning.

6.1. Limitations of DreamBooth

Firstly, there occurs overfitting with DreamBooth, due to the small size of training images. Also, DreamBooth has no built-in temporal modeling, which means that it does not learn how a face ages, but just learn how to generate the same face in different contexts.

Thirdly, for DreamBooth, there is no precise control over the aging direction or level.

6.2. Limitations of Inversion and Cross Attention

Inversion combined with Cross Attention lacks aging consistency, since it does not encode "age progression" naturally. Also, this method is sensitive to prompts, which means that small changes in prompts can lead to inconsistent or unrealistic aging. Additionally, it has limited generalization. Since Inversion is input-specific, users can not invert one face and then apply aging to a batch of other faces.

6.3. Future Work

This study provides a foundational comparison of two fine-tuning methods for face age editing; however, several promising directions remain for further exploration.

First, we plan to extend our comparative analysis by incorporating two more fine-tuning methods: LoRA (Low-Rank Adaptation) and HyperNetworks, which have potential in enabling efficient and scalable personalization of diffusion models.

Second, we aim to quantitatively assess overfitting behavior across all methods. While DreamBooth is known for potential overfitting when few samples are available, it remains unclear how LoRA and HyperNet compare in terms of generalization and identity retention. This analysis could help guide users in choosing the most reliable method for subject-driven generation.

Additionally, we propose exploring ControlNet-based guidance to improve structural aging accuracy. Since ControlNet allows the injection of spatial or pose-specific constraints, it may help models produce more consistent transformations over large age gaps. Another direction is to develop and evaluate hybrid approaches, such as combining DreamBooth with LoRA or integrating LoRA with Image Inversion. These strategies may combine the strengths of different methods—such as DreamBooth's identity fidelity and LoRA's training efficiency—to achieve better performance across multiple criteria.

Finally, we recognize the importance of accessibility and usability for broader adoption. We plan to build a more convenient user interface that allows non-technical users to test different face age editing methods, upload custom images, and select target ages in an interactive environment.

6.4. Conclusion

In this project, we conducted a comparative study of two fine-tuning approaches—DreamBooth and Image Inversion with Cross-Attention—for the task of face age editing using diffusion models. Our goal was to evaluate each method's effectiveness in generating photorealistic aged or rejuvenated images while preserving subject identity and maintaining editing quality.

Our experiments revealed that Inversion with Cross-attention guided technique performs better in identity preservation, producing highly personalized results, though at the

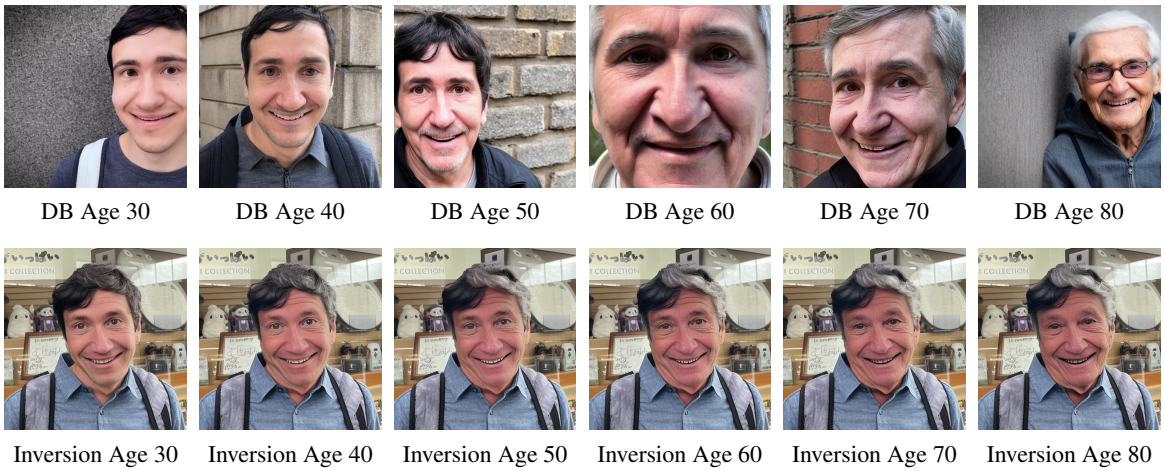

Figure 9. Comparison of face aging using DreamBooth (DB, top) and Inversion (bottom) from ages 30 to 80 in increments of 10 years.

Age	30	40	50	60	70	80
DreamBooth FID Score	349.26	320.74	286.93	395.21	365.76	304.09
Inversion FID Score	34.25	46.25	42.45	45.27	43.35	54.39

Table 1. Quantitative evaluation (FID score) for each target age.

cost of greater computational resources and risk of overfitting. In contrast, Dreambooth offers a more lightweight and flexible solution, with stronger control over age-specific edits, but may be less consistent in preserving detailed identity features.

By analyzing qualitative outputs and quantitative metrics, we highlighted the trade-offs between personalization, control, and efficiency. These findings offer practical insights for researchers and practitioners choosing fine-tuning strategies for face editing applications.

Overall, this study contributes to the growing field of personalized image generation and lays the groundwork for future extensions incorporating more advanced fine-tuning techniques, structural control, and user-facing tools.

6.5. Contribution

All team members are fully engaged in every parts of the project, with each one leads one's strongest area: Ian and Shash lead the part of the experiment, Dinh leads the part of literature review, and Juncheng leads the part of final report.

References

Generative adversarial networks. *Commun. ACM*, 63 (11):139–144, October 2020a. ISSN 0001-0782. doi: 10.1145/3422622. URL https://doi.org/10.1145/3422622.

Denoising diffusion probabilistic models. NIPS '20, Red Hook, NY, USA, 2020b. Curran Associates Inc. ISBN 9781713829546.

Chen, X. and Lathuilière, S. Face aging via diffusion-based editing, 2023. URL https://arxiv.org/abs/2309.11321.

Grimmer, M. and Busch, C. Agediff: Latent diffusion-based face age editing with dual cross-attention. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, 2024. doi: 10.1109/WIFS61860.2024.10810706.

Kingma, D. P. and Welling, M. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8245. doi: 10.1561/2200000056. URL http://dx.doi.org/10.1561/2200000056.

Narula, A. A comparative study of gans and vaes for image generation. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 09:1–9, 02 2025. doi: 10.55041/IJSREM41480.

Nichol, A. and Dhariwal, P. Improved denoising diffusion probabilistic models, 2021. URL https://arxiv.org/abs/2102.09672.

Or-El, R., Sengupta, S., Fried, O., Shechtman, E., and Kemelmacher-Shlizerman, I. Lifespan age transformation synthesis, 2020. URL https://arxiv.org/abs/2003.09764.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. URL https://arxiv.org/abs/2208.12242.