Identity-Preserving Aging of Face Images via Latent Diffusion Models

Sudipta Banerjee* Govind Mittal* Ameya Joshi Chinmay Hegde
Nasir Memon
New York University
{sb9084, mittal, ameya.joshi, chinmay.h, memon}@nyu.edu

Abstract

The performance of automated face recognition systems is inevitably impacted by the facial aging process. However, high quality datasets of individuals collected over several years are typically small in scale. In this work, we propose, train, and validate the use of latent text-to-image diffusion models for synthetically aging and de-aging face images. Our models succeed with few-shot training, and have the added benefit of being controllable via intuitive textual prompting. We observe high degrees of visual realism in the generated images while maintaining biometric fidelity measured by commonly used metrics. We evaluate our method on two benchmark datasets (CelebA and AgeDB) and observe significant reduction ($\sim 44\%$) in the False Non-Match Rate compared to existing state-of the-art baselines.

1. Introduction

Motivation. It is well known that facial aging can significantly degrade the performance of modern automated face recognition systems [16, 19, 29]. Improving the robustness of such systems to aging variations is therefore critical for their lasting practical use. However, building systems that are robust to aging variations requires high quality longitudinal datasets: images of a large number of individuals collected over several years. Collection of such data constitutes a major challenge in practice. Datasets such as MORPH [5] contains longitudinal samples of only 317 subjects from a total of $\sim$13K subjects over a period of five years [8]. Other datasets like AgeDB [28] and CACD [10] contains unconstrained images with significant variations in pose, illumination, background, and expression.

An alternative approach to gathering longitudinal data is to digitally simulate face age progression [21]. Approaches include manual age-editing tools, such as YouCam Makeup, FaceApp, and AgingBooth [1, 13]; more recently, GAN-based generative models, such as AttGAN, Cafe-GAN, Talk-to-Edit [17, 20, 23, 24, 37] have also been used to simulate age progression in face images. However, we find that generative models struggle to correctly model biological aging, which is a complex process affected by genetic, demographic, and environmental factors. Moreover, training high quality GANs for adjusting facial attributes themselves require a large amount of training data.

Our Approach. Existing generative models often struggle to manipulate the age attribute and preserve facial identity. They also require auxiliary age classifiers and/or extensive training data with longitudinal age variations. To address both of the above issues, we propose a new latent generative model for simulating high-quality facial aging, while simultaneously preserving biometric identity. The high level algorithmic idea is to finetune latent text-to-image diffusion models (such as Stable Diffusion [32]) with a novel combination of contrastive and biometric losses that help preserve facial identity. See Fig. 1 for an overview of our method.

The proposed method requires: (i) a pre-trained latent diffusion model (see Sec. 2), (ii) a small set (numbering $\approx$ 20) of training face images of an individual, and (iii) a small auxiliary set (numbering $\approx$ 600) of image-caption pairs. The pairs contain facial images of individuals and captions indicating their corresponding age. This auxiliary set of image-caption pairs serve as the regularization set. The individuals in the training set and the regularization set are disjoint. We use the training images during fine-tuning to learn the identity-specific information of the individual, and the regularization images with captions to learn the association between an image (face) and its caption (age). Finally, we simulate age regression and progression of the trained individual using a text prompt specifying the target age. See the details of our method in Sec. 3.

Summary. Our main contributions are as follows.

- We adapt latent diffusion models to perform age regression and progression in face images. We introduce two key ideas: an identity-preserving loss (in addition to perceptual loss), and a small regularization set of image-caption pairs to resolve the limitations posed by existing GAN-based methods.

*Both authors contributed equally.

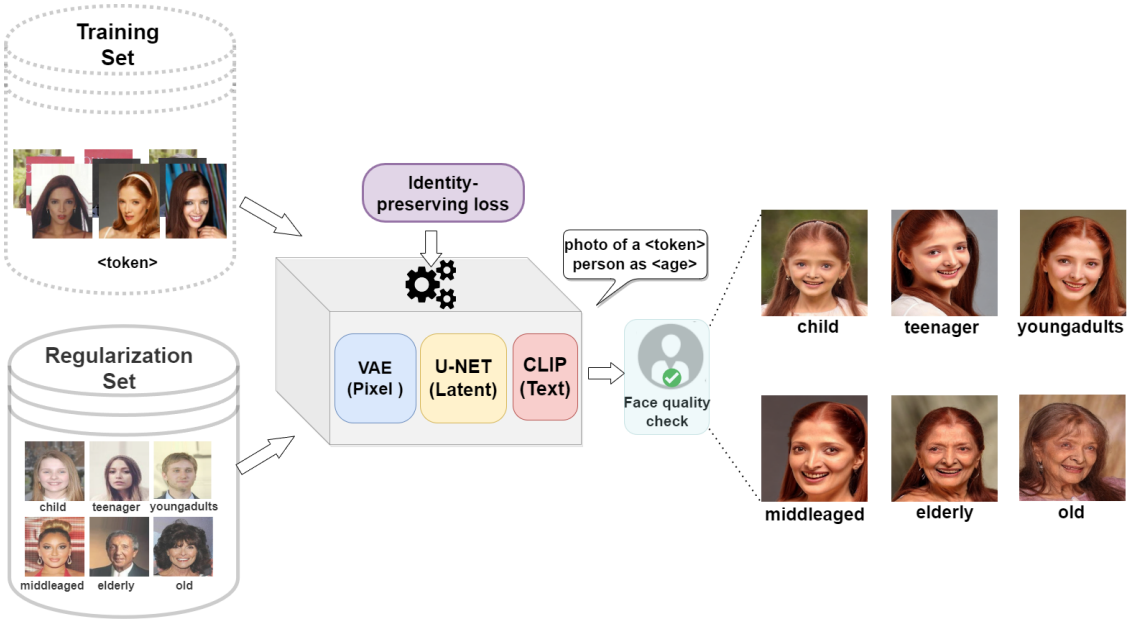


Figure 1. Overview of the proposed method. The proposed method needs a fixed *Regularization Set* comprising facial images with age variations and a variable *Training Set* comprising facial images of a target individual. The latent diffusion module (comprising a VAE, U-Net and CLIP-text encoder) learns the concept of age progression from the regularization images and the identity-specific information from the training images. We integrate biometric and contrastive losses in the network for identity preservation. At inference, the user prompts the trained model using a rare token associated with the trained target subject and the desired age to perform age editing.

- As a secondary finding, we show that face recognition classifiers may benefit by fine-tuning on generated images with significant age variations as indicated in [31].
- We conduct experiments on CelebA and AgeDB datasets and perform evaluations to demonstrate that the synthesized images i) appear visually compelling in terms of aging and de-aging through qualitative analysis and automated age predictor, and ii) match with the original subject with respect to human evaluators and automated face matcher. We demonstrate that our method outperforms SOTA image editing methods, namely, IPCGAN [34], AttGAN [17] and Talk-to-Edit [20].

The rest of the paper is organized as follows. Sec. 2 outlines existing work. Sec. 3 describes the proposed method for simulating facial aging and de-aging. Sec. 4 describes the experimental settings. Sec. 5 presents our findings and analysis. Sec. 6 concludes the paper.

2. Related Work

Previous automated age progression models have used a variety of architectures, including recurrent ones [36] and GANs. [37] uses a hierarchy of discriminators to preserve the reconstruction details, age and identity. STGAN [24] utilizes selective transfer units that accepts the difference between the target and source attribute vector as input, resulting in more controlled manipulation of the attribute. Cafe-GAN [23] utilizes complementary attention features to focus on the regions pertinent to the target attribute while preserving the remaining details. HRFAE [38] encodes an input image to a set of age-invariant features and an age-specific modulation vector. The age-specific modulation vector re-weights the encoded features depending on the target age and then passes it to a decoder unit that edits the image. CUSP [15] uses a custom structure preserving module that masks the irrelevant regions for better facial structure preservation in the generated images. The method performs style and content disentanglement while conditioning the generated image on the target age. ChildGAN [9] is inspired from the self-attention GAN and uses one-hot encoding of age labels and gender labels appended to the noise vector to perform age translation in images of young children.

We focus on three methods in our comparisons. IPCGAN [34] uses a conditional GAN with an identity preserving module and an age classifier to perform image-to-image style transfer for age-editing. AttGAN [17] performs binary facial attribute manipulation by modeling the relationship between the attributes and the latent representation of the face. The network enables high quality facial attribute editing while controlling the attribute intensity and style. Talk-to-Edit [20] provides fine-grained facial attribute editing via dialog interaction, similar to our approach. The method uses a language encoder to convert the user's request into an 'editing encoding' that encapsulates information about the

degree and direction of change of the target attribute, and seeks user feedback to iteratively edit the desired attribute.

We also highlight two recent methods that also use diffusion models for face generation. In DCFace [21], the authors propose a dual condition synthetic face generator to allow control over simulating intra-class (within same individual) and inter-class (across different individuals) variations. In [30], the authors explore suitable prompts for generating realistic faces using stable diffusion and investigate their quality. Neither method focus on identity-preserving text guided facial aging and de-aging, which is our goal.

3. Our Proposed Method

Although a suite of age editing methods exist in the literature as discussed above, the majority of them focuses on perceptual quality instead of biometric quality. A subset of latent space manipulation methods struggle with 'real' face images and generate unrealistic outputs. Existing works reiterate that age progression is a smooth but non-deterministic process that requires incremental evolution to effectively transition between ages. This motivates the use of diffusion models, which naturally model the underlying data distribution by incrementally adding and removing noise. We start with a brief mathematical overview.

3.1. Preliminaries

Denoising diffusion probabilistic models (DDPMs) [18] perform the following steps: 1) a forward diffusion process $\boldsymbol{x}_0 \xrightarrow{>>\eta_t} \boldsymbol{x}_t$ [1] that incrementally adds Gaussian noise, η sampled from a normal distribution, $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, to the clean data, $\boldsymbol{x}_0$ sampled from a real distribution, $p(\boldsymbol{x})$ over t time steps. 2) a backward denoising process $\boldsymbol{x}_0 \xleftarrow{<<\eta_t} \boldsymbol{x}_t$ [1] that attempts to recover the clean data from the corrupted or noisy data $\boldsymbol{x}_t$ by approximating the conditional probability distribution, $p(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ using a neural network that serves as a noise estimator. The forward and backward processes can be considered analogous to VAEs [22].

Note that DDPMs are computationally expensive as the estimated noise has the same dimension as the input. Alternatively, stable diffusion [32] is a class of latent diffusion models that performs diffusion on a relatively lower dimensional latent representation. Latent diffusion generates high quality images conditioned on text prompts. It comprises three modules: an autoencoder (VAE), a U-Net and a text-encoder. The encoder in the VAE converts the image into a low dimensional latent representation fed as the input to the U-Net model. The U-Net model estimates the noise needed to recover the high resolution output from the decoder of the VAE. [32] further added cross-attention layers in the U-Net backbone to use text embedding as a conditional input, thereby enhancing the model's generative capability.

[1] $>>$ denotes noise addition while $<<$ denotes noise removal.

In this work, we focus on DreamBooth [33], a latent diffusion model that fine-tunes a text-to-image diffusion framework for re-contextualization of a single subject. To accomplish this, it requires (i) a few images of the subject, and (ii) text prompts containing a unique identifier and the class label of the subject. The class label denotes a collective representation of multiple instances while the subject will correspond to a specific example belonging to the class. The objective is to associate a unique token or a rare identifier to each subject (a specific instance of a class) and then recreate images of the same subject in different contexts as guided by the text prompts. The class label harnesses the prior knowledge of the trained diffusion framework for that specific class. Incorrect class labels or missing class labels may result in inferior outputs [33]. The unique token acts as a reference to the particular subject, and needs to be rare enough to avoid conflict with other concepts. The authors use a set of rare tokens corresponding to a sequence of 3 or fewer Unicode characters and the T5-XXL tokenizer. See [33] for more details. DreamBooth uses a class-specific prior preservation loss to increase the variability of generated images while ensuring minimal deviation between the target subject and the output images. The original training loss can be written as follows.

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x},\boldsymbol{c},t}[w_t \| f_\theta(g_t(\boldsymbol{x}), c) - \boldsymbol{x}\|_2^2 + \\ \lambda w_{t'} \| f_\theta(g_{t'}(\boldsymbol{x'}), c_{class}) - \boldsymbol{x'}\|_2^2].\end{aligned} \tag{1}$$

The first term in Eqn. 1 denotes the squared error between the ground-truth images, $\boldsymbol{x}$, (training set) and the generated images, $f_\theta(g_t(\boldsymbol{x}), c)$. Here, $f_\theta(\cdot, \cdot)$ denotes the pre-trained diffusion model (parameterized by θ) that generates images for a noise map and a conditioning vector. The noise map is obtained as $g_t(\boldsymbol{x}) = \alpha_t \boldsymbol{x} + \sigma_t \eta$, where $\eta \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and α_t, σ_t, w_t are diffusion control parameters at time step $t \sim \mathbb{U}[0, 1]$. The conditioning vector $\boldsymbol{c}$ is generated using a text encoder for a user-defined prompt. The second term refers to the prior-preservation component using generated images that represents the prior knowledge of the trained model for the specific class. The term is weighted by a scalar value, $\lambda = 1$. The conditioning vector in the second term, $\boldsymbol{c}_{class}$, corresponds to the class label.

3.2. Methodology

DreamBooth works effectively with the aid of prior preservation for synthesizing images of dogs, cats, cartoons, etc. But in this work, we are focusing on human face images that contain intricate structural and textural details. Although the class label 'person' can capture human-like features, this may not be adequate to capture identity-specific features that vary across individuals. Therefore, we include an identity-preserving term in the loss function. The identity-preserving component minimizes the distance between the biometric features from the original and gener-

ated images as follows.

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{c},t}[w_t \| f_\theta(g_t(\boldsymbol{x}), c) - \boldsymbol{x}\|_2^2 + \lambda w_{t'} \| f_\theta(g_{t'}(\boldsymbol{x'}), c_{class}) - \boldsymbol{x'}\|_2^2 + \lambda_b \mathcal{B}(f_\theta(g_t(\boldsymbol{x}), c_{class}), \boldsymbol{x})]. \quad (2)$$

We use this new loss to fine-tune the VAE. The third term in Eqn. 2 refers to the biometric loss computed between the ground-truth image of the subject, $\boldsymbol{x}$, and the generated image weighted by $\lambda_b = 0.1$. Note that $f_\theta(g_{t'}(\boldsymbol{x}), c_{class})$ uses the training set (*i.e.*, images of an individual subject), whereas $f_\theta(g_{t'}(\boldsymbol{x'}), c_{class})$ uses the regularization set that contains representative images of a class. Here, $\mathcal{B}(\cdot, \cdot)$ computes the L_1 distance between the biometric features extracted from a pair of images (close to zero for same subjects, higher values correspond to different subjects). We use a pre-trained VGGFace [4] feature extractor, such that,

$$\mathcal{B}(i, j) = \|VGGFace(i) - VGGFace(j)\|_1 .$$

Now, we turn to target-specific fine-tuning. The implementation used in our work [3, 14] uses a frozen VAE and a text-encoder while keeping the U-Net model unfrozen. U-Net denoises the latent representation produced by the encoder of VAE, $g_t(\boldsymbol{x}) = \boldsymbol{z}_t = \alpha_t \boldsymbol{x} + \sigma_t \eta$. Therefore, we use identity-preserving contrastive loss using the latent representation. We adopted the SimCLR [11] framework that uses a normalized temperature-scaled cross-entropy loss between positive and negative pairs of augmented latent representations, denoted by $\mathcal{S}(\cdot, \cdot)$ in Eqn. 3. We compute the contrastive loss between the latent representation of the noise-free inputs ($\boldsymbol{z}_0$) and the de-noised outputs ($\boldsymbol{z}_t$) with a weight term $\lambda_s = 0.1$ and a temperature value = 0.5. Refer to [11] for more details. The contrastive loss between the latent representation in the U-Net architecture enables us to fine-tune the diffusion model for each subject as follows.

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{c},t}[w_t \| f_\theta(g_t(\boldsymbol{x}), c) - \boldsymbol{x}\|_2^2 + \lambda w_{t'} \| f_\theta(g_{t'}(\boldsymbol{x'}), c_{class}) - \boldsymbol{x'}\|_2^2 + \lambda_s \mathcal{S}(\boldsymbol{z}_t, \boldsymbol{z}_0)]. \quad (3)$$

In addition to customizing the losses, we use the regularization set to impart the concept of facial age progression and regression to the latent diffusion model. The regularization set contains representative images of a class, in our case, 'person'. A regularization set comprising face images selected from the internet would have sufficed if our goal was to generate realistic faces as done in [30]. However, our task involves learning the concept of aging and de-aging, and then apply it to any individual. To accomplish this task, we use face images from different age groups and then pair it with one-word captions that indicate the age group of the person depicted in the image. The captions correspond to one of the six age groups: 'child', 'teenager', 'youngadults', 'middleaged', 'elderly', and 'old'. We could have used numbers as age groups, for example, twenties, forties or sixties, but we found that a language description is more suitable than a numeric identifier. Another reason for pairing these age descriptions with the images is that we can use these same age identifiers while prompting the diffusion model during inference (photo of a ⟨ token ⟩ ⟨ class label ⟩ as ⟨ age group ⟩). We use the following six prompts during inference. 1) photo of a sks person as child, 2) photo of a sks person as teenager, 3) photo of a sks person as youngadults, 4) photo of a sks person as middleaged, 5) photo of a sks person as elderly, and 6) photo of a sks person as old. We have explored other tokens (see Sec. 5.4).

4. Experiments

Setup and implementation details. We conduct experiments using DreamBooth implemented using Stable Diffusion v1.4 [3]. The model uses CLIP's [2] text encoder trained on laion-aesthetics v2 5+ and a vector quantized VAE [35] to accomplish the task of age progression. The text encoder stays frozen while training the diffusion model. We use two datasets, namely, **CelebA** [27] and **AgeDB** [28]. We use 2,258 face images belonging to 100 subjects from the CelebA [27] dataset, and 659 images belonging to 100 subjects from the AgeDB dataset to form the 'training set'. CelebA does not contain age information, except a binary 'Young' attribute annotation. We do not have ground-truth for evaluating the generated images synthesized from the CelebA dataset. On the other hand, AgeDB dataset comprises images with exact age values. We then select the age group that has the highest number of images and use them as the training set, while the remaining images contribute towards the testing set. Therefore, 2,369 images serve as ground-truth for evaluation in AgeDB dataset.

We use a regularization set comprising image-caption pairs where each face image is associated with a caption indicating its corresponding age label. We use 612 images belonging to 375 subjects from the CelebA-Dialog [20] dataset,where the authors provide fine-grained annotations of age distributions. We convert the distribution to categorical labels to use as captions in the regularization images. We refer to them as {Child: <15 years, Teenager: 15-30 years, Youngadults: 30-40 years, Middleaged: 40-50 years, Elderly: 50-65 years and Old: >65 years}. We use 612 (102×6) images in the subject disjoint regularization set.

The success of generating high quality images often depend on effectively prompting the diffusion model during inference. The text prompt at the time of inference needs a rare token/identifier that is associated with the concept learnt during fine-tuning. We use four different rare tokens {*wzx*, *sks*, *ams*, *ukj*} [6] in this work for brevity.

We use the implementation of DreamBooth using stable diffusion in [3] and used the following hyperparameters.

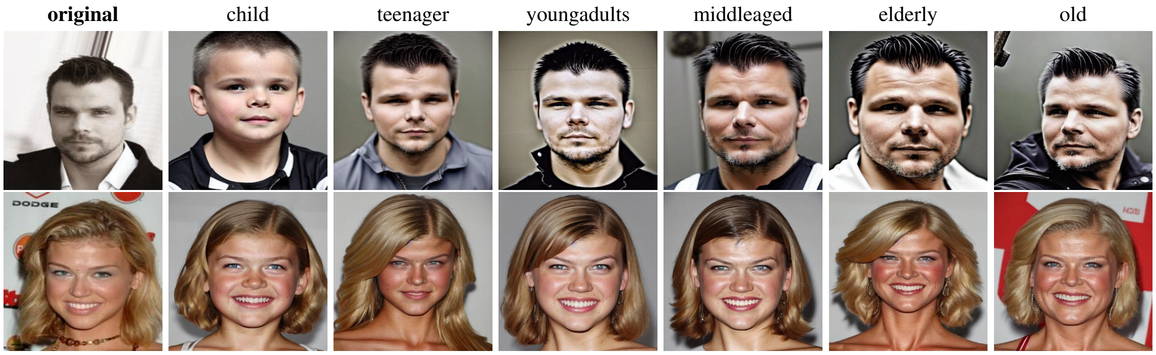


Figure 2. Illustration of age edited images generated from the CelebA dataset.

We adopt a learning rate = 1e-6, number of training steps $= 800$, embedding dimensionality in autoencoder $= 4$, and batch size $= 8$. The generated images are of size 512×512. We use $\lambda = 1, \lambda_b = 0.1$ and $\lambda_s = 0.1$ (refer to Eqns. 2 and 3). We generate 8 samples at inference. However, we perform a facial quality assessment using EQFace [26] to limit the number of generated face images to 4, such that, each generated image contains a single face with frontal pose. We adopt a threshold of 0.4, and retain the generated images if quality exceeds the threshold, else, discard them. Training each subject requires ∼5-8 mins. on a A100 GPU.

We perform **qualitative evaluation** of the generated images by conducting a user study involving 26 volunteers. The volunteers are shown a set of 10 face images (original) and then 10 generated sets; each set contains five images belonging to five age groups (excluding old), resulting in a total of 60 images. They are assigned two tasks: 1) identify the individual from the original set who appears most similar to the subject in the generated set; 2) assign each of the five generated images to the five age groups they are most likely to belong to. We compute the proportion of correct face recognition and age group assessment.

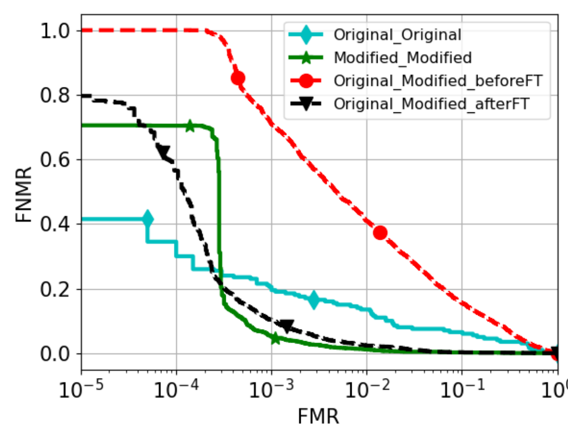
Further, we perform **quantitative evaluation** of the generated outputs using the ArcFace [12] matcher (different from VGGFace used in identity-preserving biometric loss). We utilize the genuine (intra-class) and imposter (inter-class) scores to compute Detection Error Trade-off (DET) curves and report the False Non-Match Rate (FNMR) at a False Match Rate (FMR) of 0.01% and 0.1%.

5. Results

We report the biometric matching performance using the ArcFace matcher between **original and modified** images in Table 1 for the CelebA dataset. See examples of generated images in Fig. 2. In CelebA, we do not have access to ground-truths, so we perform biometric matching with disjoint samples of the subject not used in the training set. We refer this as the 'simulation' result. We achieve the best biometric matching using the initial loss settings of latent diffusion (Eqn. 1). The biometric matching impacts the sim-

Table 1. CelebA simulation results for biometric matching between **Original-Modified** images. The metrics are False Non-Match Rate (FNMR) at False Match Rate (FMR) = 0.01/0.1%.

Age group	Initial loss		With contrastive loss	
	sks	wzx	sks	wzx
child	0.49/0.21	0.58/0.27	0.56/0.26	0.60/0.29
teenager	0.23/0.07	0.32/0.12	0.29/0.10	0.34/0.12
youngadults	0.25/0.08	0.30/0.10	0.28/0.08	0.31/0.10
middleaged	0.20/0.07	0.28/0.09	0.27/0.09	0.30/0.10
elderly	0.22/0.07	0.29/0.10	0.25/0.09	0.29/0.11
old	0.24/0.10	0.31/0.12	0.29/0.11	0.32/0.12



Matching scenarios	FNMR@FMR=0.01/0.1%
Ori-Ori	0.14/0.07
Mod-Mod	0.02/0.01
Ori-Mod (w/o fine-tune)	0.41/0.16
Ori-Mod (w/ fine-tune)	**0.03/0.01**

Figure 3. (Top:) DET curves of face matching using generated images from the CelebA dataset. (Bottom:) Recognition performance in the table indicating FNMR @ FMR=0.01/0.1%. The age-edited images are generated using the *wzx* token with contrastive loss.

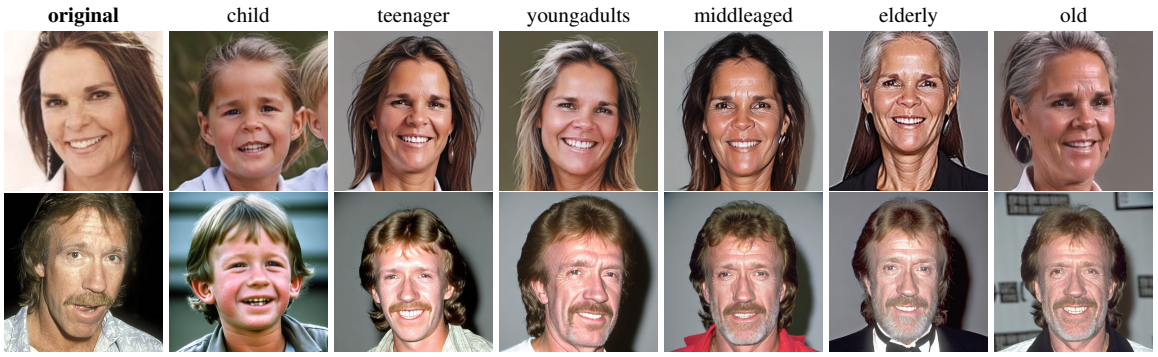
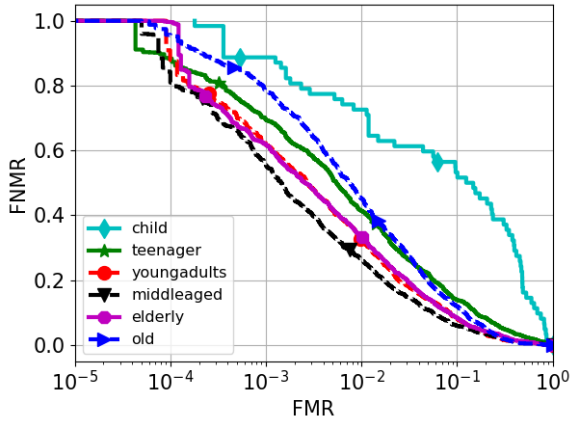


Figure 4. Illustration of age edited images generated from the AgeDB dataset.



Age group	FNMR@FMR=0.01/0.1%
child	0.73/0.54
teenager	0.42/0.14
youngadults	0.33/0.09
middleaged	0.27/0.06
elderly	0.34/0.09
old	0.46/0.12

Figure 5. (Top:) DET curves of face matching using generated images from the AgeDB dataset for the six age groups. (Bottom:) Recognition performance in the table indicating FNMR @ FMR=0.01/0.1%. The age-edited images are generated using the *wzx* token with contrastive loss.

ilarity between generated and gallery images and does not quantify the success of age editing. On the other hand, the generated images using contrastive loss[2] (Eqn. 3) successfully accomplish aging/de-aging but achieve low matching as the ArcFace model is not trained on generated images.

Therefore, we conduct an additional experiment of fine-tuning the ArcFace model on subject disjoint age-edited images (∼3,400) and then repeat the matching experiments for the CelebA dataset. We report the **original-original**, **modified-modified**, **original-modified** (before fine-tuning ArcFace) and **modified-modified** (after fine-tuning ArcFace) face matching performance and the corresponding DET curves in Fig. 3 for the contrastive loss and *wzx* token combination. Note that there is a significant improvement in face matching performance between the modified-modified images and original-modified images after fine-tuning. We achieve **FNMR=3% at FMR=0.01%** and **FNMR=1% at FMR = 0.1%** with the fine-tuned face matcher on the age-edited images. We down-sample the modified images to the same resolution as the original images, and observe similar results. Additionally, the fine-tuned face matcher drastically improves when comparing original-modified images, indicating that synthetic images can improve the robustness of existing face matchers as suggested in [31].

We report the biometric matching performance using the ArcFace matcher between **original and modified** images for the Age DB dataset. In AgeDB, we have a separate gallery set consisting of images across age groups different than the images used during training. We use them as ground-truth for evaluation and refer this as the 'imputation' result. As anticipated, we observe modest performance across a majority of the age groups barring 'child'. We had only 28 images from 18 subjects (out of 100) corresponding to child group, and some of the images were of extremely poor quality, thereby resulting in an abnormal high value of FNMR. See examples of generated images in Fig. 4. We present the the DET curves and the corresponding FNMR values @FMR=0.01/0.1% in Fig. 5.

5.1. Comparison of auxiliary loss functions

We compare the proposed loss functions: 1) VGGFace-based Biometric loss and 2) Contrastive loss and observe a reduction in FNMR up to 46% @FMR=0.01% averaged across all age groups when using contrastive loss with respect to biometric loss. Genuine match scores (scores between original and age-edited images of the same individual) that indicate intra-class fidelity are much better preserved when using contrastive loss (see Fig. 6). We ex-

[2]We also compare with VGGFace-based biometric loss (Eqn. 2), and observed contrastive loss outperforms biometric loss. See Sec. 5.1.

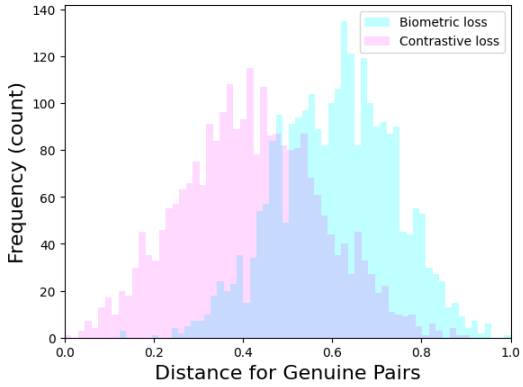


Figure 6. Comparison of auxiliary loss functions (VGGFace-based biometric loss vs. Contrastive loss) in terms of cosine distance scores computed for genuine pairs using the ArcFace matcher. Contrastive loss produces desirable lower distance between genuine pairs.

plored different values of λ_b and λ_s, $= \{0.01, 0.1, 1, 10\}$, and observe 0.1 produces the best results for both variables.

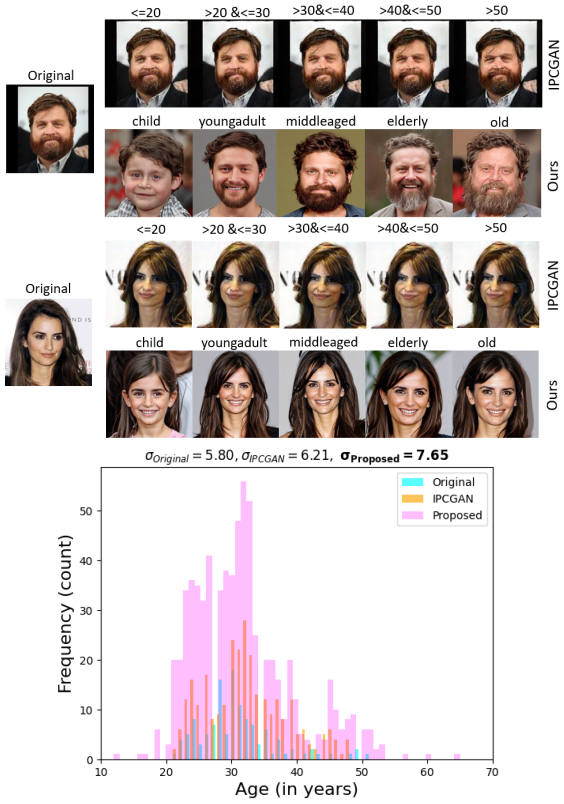
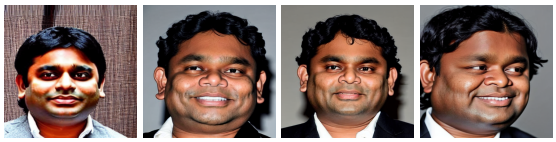


Figure 7. (Top and middle): Comparison of outputs produced by IPCGAN and the proposed method. (Bottom): Age predictions by automated age predictor shows that our method generates images with a wider age dispersion compared to original CACD images and IPCGAN-generated images.

Age group	Methods		
	AttGAN	Talk-to-Edit	Proposed
child	-	0.99/0.40	**0.56/0.26**
teenager	-	1.0/0.50	**0.29/0.10**
youngadults	0.47/0.20	0.70/0.21	**0.28/0.08**
middleaged	-	0.51/0.13	**0.27/0.09**
elderly	-	0.83/0.39	**0.25/0.09**
old	0.31/0.11	0.56/0.22	**0.29/0.11**
Average	0.39/0.15	0.76/0.31	**0.32/0.12**

Figure 8. (Top): Comparison of 'young' outputs (columns 2-4) and 'old' outputs (columns 5-7) generated by the proposed method with baselines: AttGAN and Talk-to-Edit. The original images are in the first column. (Bottom): False Non-Match Rate (FNMR) at False Match Rate (FMR) = 0.01/0.1%



(a) *sks* (b) *ukj* (c) *wzx* (d) *ams*

Figure 9. Comparison of the images generated using the four tokens in this work.

5.2. Comparison with existing methods

We use IPCGAN [34], AttGAN [17] and Talk-to-Edit [20] as baselines for comparison. We evaluate using the pre-trained models of the baselines provided by the authors. As IPCGAN was trained on the CACD dataset [10], we fine-tune our method on 62 subjects from the CACD dataset. We observe an FNMR=2% (IPCGAN), compared to FNMR=11% (Ours) @ FMR=0.01. IPCGAN defaults to the original when it fails to perform aging or de-aging resulting in spuriously low FNMR. We perform automated age prediction using the DeepFace [7] age predictor. We observe the images synthesized by our method result in wider dispersion of age predictions compared to the original images and the IPCGAN-generated images, indicating successful age editing. See Fig. 7. We apply AttGAN and Talk-to-Edit on the CelebA dataset. See comparison between generated images of proposed and baseline methods, and biometric matching performance in Fig. 8. We observe that the proposed method (contrastive loss, *sks*) outperforms AttGAN by 19% on 'young' images and by 7% on 'old' images at FMR=0.01. AttGAN can only edit to young or

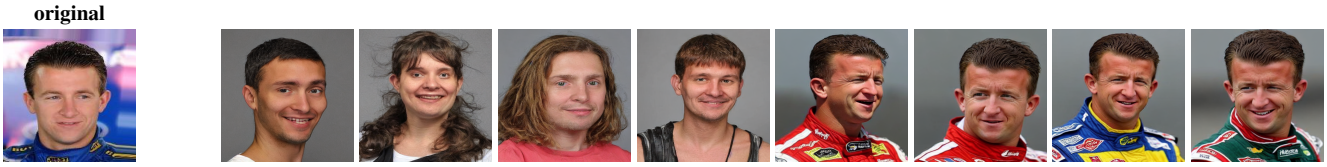


Figure 10. Impact of token (*wzx*) and class label (*person*) on generated images: "photo of a person" (left) vs. "photo of a ***wzx*** person" (right). Note the token is strongly associated with a specific identity belonging to that class.

old ages. Further, we observe that the method outperforms Talk-to-Edit by an average FNMR =44% at FMR=0.01. The different age groups are simulated using a target value parameter in Talk-to-Edit that varies from 0 to 5, each value representing an age group. However, we observe several cases of distorted or absence of outputs in Talk-to-Edit.

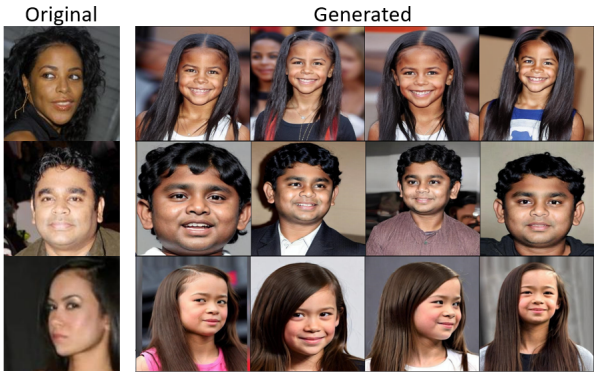


Figure 11. Examples of generated images pertaining to diverse sex and ethnicity for 'child' group.

Figure 12. Failure cases corresponding to 'child' age group.

5.3. User study

We collected 26 responses from the user study. Rank-1 biometric identification accuracy (averaged across the total number of responses) is equal to 78.8%. The correct identification accuracy of the age groups are: child = 99.6%, teenager = 72.7%, youngadults = 68.1%, middleaged = 70.7% and elderly = 93.8%. The users were able to successfully distinguish between generated images from different age groups with reasonably high accuracy.

5.4. Effect of rare tokens

We use four tokens in this work, namely, {*sks*, *ukj*, *ams*, *wzx*}, for the sake of brevity. We observe *sks* and *wzx* tokens result in visually compelling results compared to the remaining two tokens, and have been used for further evaluation. Note these tokens are condensed representations provided by the tokenizer that are determined by identifying rare phrases in the vocabulary (see Fig. 9). Additionally, we evaluate the effect of the token and the class label in the prompt in Fig. 10; removing the token results in lapse in identity-specific features.

5.5. Effect of demographics

We also observed the following effects. **Age:** The generated images can capture different age groups well if the training set contains images in the middle-aged category. We observe that if training set images comprise mostly elderly images, then the method struggles to render images in the other end of the spectrum, *i.e.*, the child category, and vice-versa. We also observe that we obtain visually compelling results of advanced aging when we use 'elderly' in the prompt instead of 'old'. **Sex:** The generated images can effectively translate the training images into older age groups for men compared to women. This can be due to the use of makeup in the training images. **Ethnicity:** We do not observe any strong effects of ethnicity/race variations in the outputs. See Fig. 11. Although in some cases, the proposed method struggles with generating 'child' images if most of the training images belong to elderly people or contain facial hair. See Fig. 12.

6. Conclusion

Existing facial age editing methods typically struggle with identity-preserved age translation. In this work, we harness latent diffusion coupled with biometric and contrastive losses for enforcing identity preservation while performing facial aging and de-aging. We use a regularization image set to impart the understanding of age progression and regression to the diffusion model, that in turn, transfers the effects onto an unseen individual while preserving their identity. The generation process is guided by intuitive text prompts indicating the desired age. Our method demonstrates significantly better results in terms of both qualitative and quantitative evaluation, and outperforms existing methods with a reduction in FNMR up to 44% at FMR=0.01%.

Future work will focus on designing zero-shot age editing without fine-tuning, and utilizing composable diffusion models [25] for fine-grained age editing.

References

[1] Age editing apps. https://www.perfectcorp.com/consumer/blog/selfie-editing/best-age-progression-apps. [Online accessed: April 13, 2023]. 1

[2] CLIP. https://openai.com/research/clip. [Online accessed: April 14, 2023]. 4

[3] DreamBooth Using Stable Diffusion. https://github.com/XavierXiao/Dreambooth-Stable-Diffusion. [Online accessed: April 14, 2023]. 4

[4] Facenet pytorch. https://github.com/timesler/facenet-pytorch. [Online accessed: April 14, 2023]. 4

[5] MORPH Facial Recognition Database. https://uncw.edu/oic/tech/feeding_flock.html. [Online accessed: April 13, 2023]. 1

[6] Rare tokens for DreamBooth Stable Diffusion. https://www.reddit.com/r/StableDiffusion/comments/zc65l4/rare_tokens_for_dreambooth_training_stable/. [Online accessed: April 13, 2023]. 4

[7] ArcFace and VGGFace implementation. https://pypi.org/project/deepface/. [Online accessed: 18th May, 2022]. 7

[8] L. Best-Rowden and A. K. Jain. Longitudinal study of automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):148–162, 2018. 1

[9] P. K. Chandaliya and N. Nain. ChildGAN: Face aging and rejuvenation to find missing children. *Pattern Recognition*, 129:108761, 2022. 2

[10] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Proceedings of the European Conference on Computer Vision*, 2014. 1, 7

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 4

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4685–4694, 2019. 5

[13] FaceApp. https://www.faceapp.com/. [Online accessed: 17th May, 2022]. 1

[14] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 4

[15] G. Gomez-Trenado, S. Lathuilière, P. Mesejo, and Ó. Cordón. Custom structure preservation in face aging. In *European Conference on Computer Vision*, pages 565–580. Springer, 2022. 2

[16] P. Grother, M. Ngan, and K. Hanaoka. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. *NIST IR 8280* https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf, 2019. 1

[17] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. AttGAN: Facial Attribute Editing by Only Changing What You Want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019. 1, 2, 7

[18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020. 3

[19] A. K. Jain, A. A. Ross, and K. Nandakumar. *Introduction to Biometrics*. Springer Publishing Company, Incorporated, 2011. 1

[20] Y. Jiang, Z. Huang, X. Pan, C. C. Loy, and Z. Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13799–13808, 2021. 1, 2, 4, 7

[21] M. Kim, F. Liu, A. Jain, and X. Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, 2023. 1, 3

[22] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations*, 2014. 3

[23] J. Kwak, D. K. Han, and H. Ko. CAFE-GAN: Arbitrary Face Attribute Editing with Complementary Attention Feature. In *European Conference on Computer Vision*, 2020. 1, 2

[24] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen. STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3668–3677, 2019. 1, 2

[25] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum. Compositional visual generation with composable diffusion models. In *17th European Conference on Computer Vision*, 2022. 8

[26] R. Liu and W. Tan. Eqface: A simple explicit quality network for face recognition. In *In Proceeding of IEEE Computer Vision and Pattern Recognition Workshop*, 2021. 5

[27] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision*, December 2015. 4

[28] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 2, page 5, 2017. 1, 4

[29] NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software. https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software. [Online accessed: 4th May, 2022]. 1

[30] 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. In *11th International Workshop on Biometrics and Forensics*, 2023. 3, 4

[31] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao. SynFace: Face Recognition with Synthetic Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2, 6

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, June 2022. 1, 3

[33] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[34] X. Tang, Z. Wang, W. Luo, and S. Gao. Face aging with identity-preserved conditional generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7939–7947, 2018. 2, 7

[35] A. van den Oord, O. Vinyals, and k. kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 4

[36] W. Wang, Z. Cui, Y. Yan, J. Feng, S. Yan, X. Shu, and N. Sebe. Recurrent face aging. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2378–2386, 2016. 2

[37] H. Yang, D. Huang, Y. Wang, and A. K. Jain. Learning Face Age Progression: A Pyramid Architecture of GANs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31–39, 2018. 1, 2

[38] X. Yao, G. Puy, A. Newson, Y. Gousseau, and P. Hellier. High resolution face age editing. In *25th International Conference on Pattern Recognition*, pages 8624–8631, 2021. 2