Identity-Aware Facial Age Editing Using Latent Diffusion

Sudipta Banerjee,[§] *Member, IEEE,* Govind Mittal[§], Ameya Joshi, Sai Pranaswi Mullangi, Chinmay Hegde, *Senior Member, IEEE,* and Nasir Memon, *Fellow, IEEE*

(Invited Paper)

Abstract—Aging in face images is a type of intra-class variation that has a stronger impact on the performance of biometric recognition systems than other modalities (such as iris scans and fingerprints). Improving the robustness of automated face recognition systems with respect to aging requires high quality longitudinal datasets that should contain images belonging to a large number of individuals collected across a long time span, ideally decades apart. Unfortunately, there is a dearth of such good operational quality longitudinal datasets. Synthesizing longitudinal data that meet these requirements can be achieved using modern generative models. However, these tools may produce unrealistic artifacts or compromise the biometric quality of the age-edited images. In this work, we simulate facial aging and de-aging by leveraging text-to-image diffusion models with the aid of few-shot fine-tuning and intuitive textual prompting. Our method is supervised using identity-preserving loss functions that ensure biometric utility preservation while imparting a high degree of visual realism. We ablate our method using different datasets, state-of-the art face matchers and age classification networks. Our empirical analysis validates the success of the proposed method compared to existing schemes. Our code is available at https://github.com/sudban3089/ID-Preserving-Facial-Aging.git

Index Terms—Face recognition, Age editing, Diffusion models.

1 Introduction

Face recognition is affected by intra-class variations (variations within the same individual) due to pose, illumination and expression, commonly known as PIE. Biological aging, on the other hand, is an intrinsic form of intra-class variation affected by genetic, demographic and environmental factors. Although several methods exist to compensate for the PIE variations, facial aging is a major factor that affects automated face recognition (FR) systems [22], [28], [44]. Developing FR systems that are robust to aging variations would essentially require high quality longitudinal datasets: images of a large number of individuals spanning several years, ideally decades. This will help the automated ML algorithm to effectively learn and model the variations in face images with time. However, there are practical limitations to collecting such data for a representative population over an extended duration. Some longitudinal datasets such as MORPH [6], AgeDB [43] and CACD [13] exist but are often constrained by either a short duration for which the images were collected or were collected in-the-wild. For example, MORPH (academic licensed) dataset contains longitudinal samples of only 317 subjects from a total of ∼13K subjects over a period of five years [11].

Collecting longitudinal data can be a tedious process; alternatively, digital simulation can aid in generating age edited images seamlessly [31]. Numerous software-based age progression approaches exist such as, manual age-editing tools, *e.g.*, YouCam Makeup, FaceApp, and AgingBooth [1], [19], and more recently, GAN-based generative models, *e.g.*, AttGAN, Cafe-GAN, Talk-to-Edit [23], [29], [33], [37], [58]. However, we find that generative models often struggle to correctly model biological aging, which is a complex process, resulting in inconsistent synthetic images that may contain unnatural artifacts. Moreover, training high quality GANs for adjusting facial attributes themselves require a large amount of training data (with or without age labels). We observe that some datasets may provide ground truth age labels, while in other cases, they might need to be inferred using age prediction tools. Age labels corresponding to web-scraped images may not be correct and can result in incorrect age modeling. Therefore, we propose to use a text-to-image generator that relies on a small set of images with assigned age groups (instead of exact age values) to *learn* the change in facial features as an individual transitions through different phases in their life beginning with childhood and adolescence, advancing to middle age, and culminating in old age. Our objective is to first learn the mapping between text description that indicates a specific age category, and the visual cues that appear in a face image of individuals belonging to that age, and subsequently apply the learned mapping to perform facial aging/de-aging for *any* individual.

When successfully implemented, identity-preserving facial aging simulation can be used in several settings:

1. Curate large-scale longitudinal datasets: Digital age progression can easily produce age simulated faces from an existing dataset at a large-scale, thus, alleviating cumbersome data collection. Longitudinal datasets can help improve the robustness of the face matchers by effectively modeling the intra-class variations.

2. Social media application: Digital editing filters are widely used for creative pursuits to improve the visual aesthetcis of facial images, *e.g.*, change hair color, whiten teeth, add make up, etc. Facial aging, is also a type of attribute editing operation that allows an individual to envision how their appearance changes with age [53].

3. Forensic analysis: Facial age progression is a critical tool

This work was supported by the National Science Foundation under SaTC grant no. 2154119.

The authors are with New York University, Email (sb9084@nyu.edu, mittal@nyu.edu, ameya.joshi@nyu.edu, sm11006@nyu.edu, chinmay.h@nyu.edu, memon@nyu.edu)

[§]Equal contribution.

used in investigating cases of missing individuals and help relocating them even after significant time has transpired [51]. Photosketch [34] was developed in 1991 and used by National Center for Missing and Exploited Children and FBI.

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 fine-tune latent text-to-image diffusion models (such as Stable Diffusion [48]) with novel losses (cosine, 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 $\approx$ 10-20) 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.

Main contributions.

- 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.
- 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 [47].
- We conduct experiments on CelebA, LFW 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 matchers, namely ArcFace and AdaFace. We demonstrate that our method outperforms GAN-based age editing methods, namely, IPCGAN [52], AttGAN [23] and Talk-to-Edit [29] as well as a diffusion model-based framework, ProFusion [62], and improves upon the work done in [10] in terms of methodology and empirical analysis.

Summary. We extend our previous work [10] as follows:

- **Methodology**
  (i) We upgrade the stable diffusion model from SDv1.4 to SDv1.5. Both models were trained on v1.2 as base model but v1.5 was trained for larger number of steps compared to v1.4 (595K steps vs. 225K steps) resulting in improved photorealism of images[1]. We select SDv1.5 instead of SDv2.0 as existing reports and our findings indicate the former generates higher quality face images than the latter.
  (ii) We utilize cosine embedding loss function in the diffusion model and compare with contrastive and biometric losses. We use the cosine embedding loss to supplement the prior preservation loss. Our rationale for using cosine embedding loss stems from the paper CosFace [55] that learns better identity mapping using the cosine angular margin between facial embedding.
- **Experimental validation**
  (i) We use a new dataset, LFW (Labelled Faces in the Wild) [27], a popular benchmark for face recognition algorithms. We use 100 subjects, each subject having 10 images from the LFW dataset to perform age simulation.
  (ii) We further supplement biometric evaluation using a second matcher, AdaFace [30], in addition to ArcFace. The new face matcher aids in studying the variations in performance across face matchers.
  (iii) Finally, we introduce a new baseline that relies on diffusion models for image customization, namely, ProFusion [62]. [10] outperformed GAN-based age editing methods. In this work, (i) we assess how the proposed method fares in comparison to another diffusion model, and (ii) we examine whether to opt for regularization-free (ProFusion) or regularized (DreamBooth) framework for generating identity-preserving realistic human faces.

2 Related Work

Existing work on automated age progression explored a variety of architectures, including recurrent ones [56] and GANs. Recurrent face network uses a gated recurrent unit to model the intermediate stages of age progression that result in smoother optical flow and higher quality age-edited images. [58] uses a hierarchy of discriminators to preserve the reconstruction details, age and identity. STGAN [37] 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 [33] utilizes complementary attention features to focus on the regions pertinent to the target attribute while preserving the remaining details. HRFAE [59] 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 [21] 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 [12] 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. Guidance via Masking-Based Attention (GMBA) [42] is a GAN-based framework that incorporates an age-aware guidance module to modulate between age-specific and age-irrelevant attributes to perform facial age editing. Other GAN-based age editing methods include A^3GAN [40], Age Progression and Regression with spatial attention [36], Attribute-aware aging [39], Wavelet age synthesis [35], Lifespan synthesis [45] and CAE-aging [61]. We also highlight three recent methods that also use diffusion models for face generation. In DCFace [31], 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 [46], the authors explore suitable prompts for generating realistic faces using stable diffusion and investigate their quality. Finally, a con-

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5

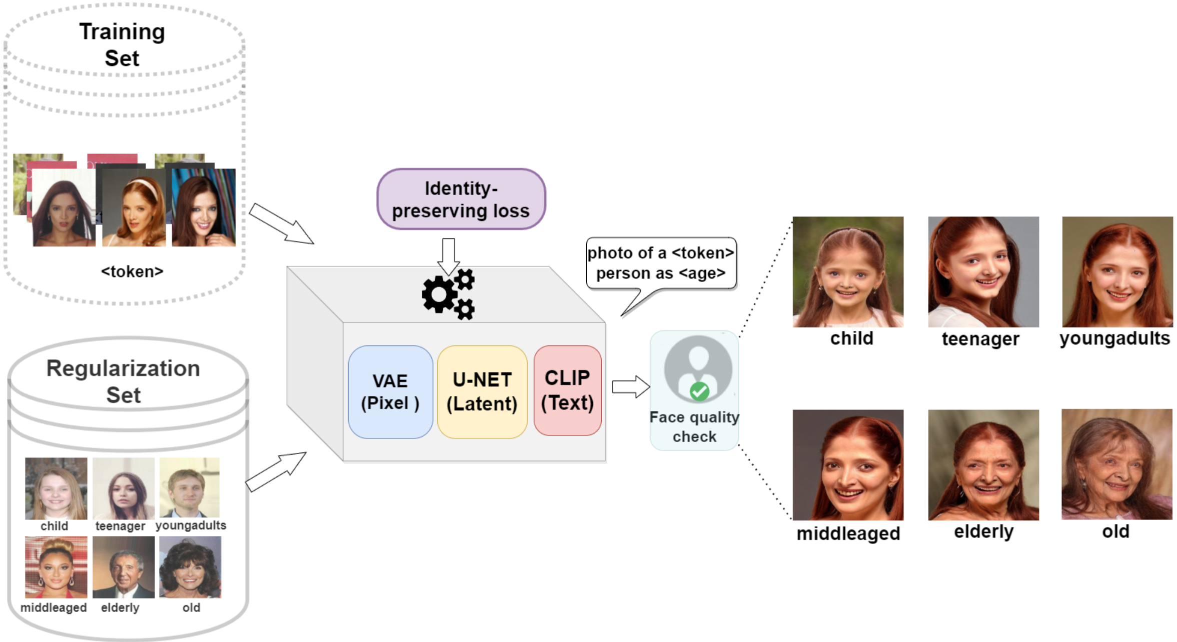


Fig. 1: Overview of the proposed method. The proposed method needs a fixed *Regularization Set* comprising facial images with age variations and a variable *Training Set* comprising facial images of a target individual. The latent diffusion module (comprising a VAE, U-Net and CLIP-text encoder) learns the concept of age progression from the regularization images and the identity-specific information from the training images. We use biometric, cosine 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.

current work that uses an age-aware fine tuning of latent diffusion model by using prompt engineering and a localized age editing using attention control [16]. Existing methods either require input age labels or do not explicitly enforce identity preservation [16] or focus on synthetic identity generation [31]. Photoverse [14] uses explicit facial identity loss for retaining details while performing editing using prompts.

We focus on four methods in our comparisons. IPCGAN [52] uses a conditional GAN with an identity preserving module and an age classifier to perform image-to-image style transfer for age-editing. AttGAN [23] 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 [29] 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. The authors use the semantic field to preserve attribute localization and an identity-keeping loss for maintaining identities. All the above methods are GAN-based age progression schemes. Additionally, we use a diffusion model-specific age editing framework for comparison. We use ProFusion [62], a regularization-free text-to-image customization method that combines an encoder, known as PromptNet, and a novel sampling technique called Fusion sampling. It avoids the problem of over-fitting that is typically handled by fine-tuning using multiple images belonging to the same entity in existing diffusion-based image re-contextualization by leveraging Fusion sampling at the time of inference. The authors claim that elimination of regularization helps in enhanced retention of fine-grained details in the image while significantly reducing the training time.

Recently, work has been done in [10] that uses a special class of latent diffusion model, known as DreamBooth [49] that has gained attention for realistic image re-contextualization for facial age editing. The authors utilize text-to-image generative models for learning the task of age-specific image generation. They integrate identity-specific loss with prior preservation loss (refer to Sec. 3) and evaluated on CelebA and AgeDB datasets in terms of biometric matching. They further compared with GAN-based age editing methods and showed that their method significantly outperformed the SoTA by reducing the False Non-Match Rate by 44%. In this paper, we improve upon the work done in [10] by updating to a recent model and modifying the loss function; conducting extensive ablation with new dataset, face matcher and comparing our method with another diffusion-based framework.

3 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.

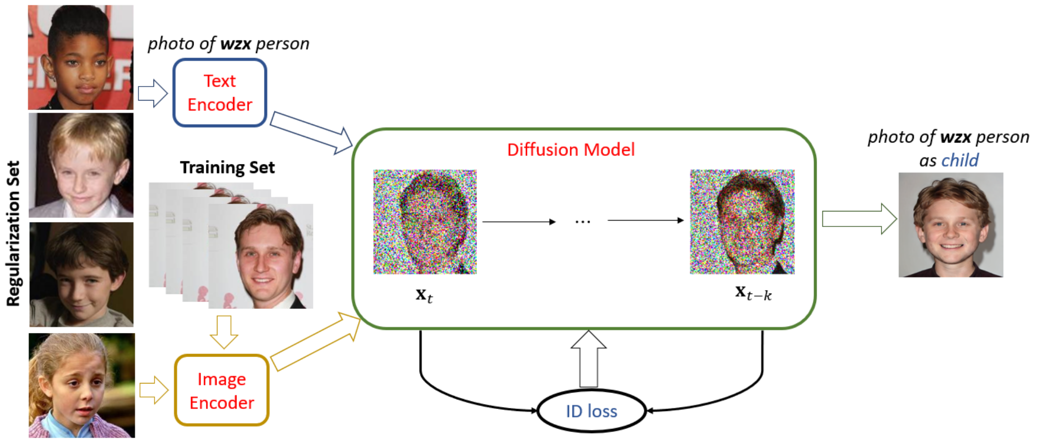


Fig. 2: Our framework learns the association between the visual appearance of an image with respect to a specific age group (e.g., *child*) and transfers it to an individual while preserving their identity.

3.1 Preliminaries

Denoising diffusion probabilistic models (DDPMs) [26] perform the following steps: 1) a forward diffusion process $\boldsymbol{x}_0 \xrightarrow{>>\eta_t} \boldsymbol{x}_t$ [2] 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 $\boldsymbol{x}_0 \xleftarrow{<<\eta_t} \boldsymbol{x}_t$ [2] that attempts to recover the clean data from the corrupted or noisy data $\boldsymbol{x}_t$ by approximating the conditional probability distribution, $p(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$. The conditional probability distribution is parameterized by mean and variance. But for a fixed variance, the neural network needs to only estimate the mean of the conditional probability distribution. To make the process simpler, the forward and backward processes can be considered analogous to variational auto-encoders (VAE), resulting in adoption of evidence lower bound (ELBO) to estimate the mean of the denoising process. See [32] for further details about VAE and ELBO. This further simplifies the objective function to minimizing the mean squared error between the actual and predicted noise.

Latent Diffusion Models. Denoising Probabilistic Models (DPMs) belong to the class of likelihood-based models that typically operate directly in pixel space, and therefore optimizing a high-resolution image generating DPM is computationally expensive during training as well as during inference due to sequential evaluations. This paved the way for Latent Diffusion Models (LDMs) [48] that apply the diffusion process on the latent representations that are considerably lower dimensional than the original data. Latent diffusion generates high quality images while reducing the computational complexity. It comprises three modules, namely, an autoenocder (VAE), U-Net and a text-encoder. LDMs perform two stages of training. In the first stage, the encoder in the VAE converts the image into a low dimensional latent representation, *i.e.,* the encoder downsamples the input $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$ in the RGB space to a latent representation, $\boldsymbol{z} \in \mathbb{R}^{h \times w \times c}$ by a factor $f = \frac{H}{h} = \frac{W}{w}$. The latent representation is fed as the input to the U-Net model. The U-Net model estimates the noise to recover the high resolution de-noised output from the decoder of the VAE. The authors in [48] further added cross-attention layers in the U-Net backbone allowing it to use text embedding produced by the text encoder as a conditional input. The text input is provided as a user-defined prompt that is transformed into text embedding via the text encoder. The cross-attention mechanism enhances the overall generative capability of the diffusion models resulting in high-resolution image synthesis while combining the input (image) and condition (text) in the latent space. DreamBooth [49] uses frozen text and image encoder and fine-tunes the diffusion model with a training set of images to learn a specific entity and a regularization set to learn a specific concept. The identity loss between the latent representations is backpropagated through the diffusion model to learn entity-specific cues, which is transferred during inference to the output image guided by the text prompt that indicates the subject token, class word and desired attribute. Refer to Fig. 2 that provides a detailed outline of the proposed framework involving each component of the LDM for fine-tuning.

In this work, we focus on DreamBooth [49], 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 [49]. 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 [49] 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.

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

The first term in Eqn. 1 denotes the squared error between the ground-truth images, $\boldsymbol{x}$, (training set) and the generated im-

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

ages, $f_\theta(g_t(\boldsymbol{x}), c)$. 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 $\eta \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, 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 generated images as follows.

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

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 λ_b. 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 [5] 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], [20] 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 [15] 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. Temperature-scaled cross entropy loss used in contrastive learning is a popular choice for pulling similar embeddings closer while pushing away dissimilar embeddings. 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 λ_s and a temperature value = 0.5. Refer to [15] 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.

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

Cosine embedding loss function computes the cosine distance between embeddings if they belong to the same class as $1 - cos(x_1, x_2)$ or $\max(cos(x_1, x_2), 0) - m$, if they belong to different classes, where x_1 and x_2 are a pair of embeddings and m is a margin value varying between $[-1, 1]$. During fine-tuning, the outputs from the U-Net module after reverse de-noising should be similar the inputs prior to forward noise addition which is analogous to the working principle of CosFace [55]. Their framework focuses on minimizing intra-class variance and maximizing inter-class variance via L_2 normalization of the facial embedding and cosine decision margin maximization. Inspired by this approach, we incorporated the cosine embedding loss function between the de-noised and noisy latent representations, $\mathcal{C}(\cdot,\cdot)$ with a regularization parameter, λ_c in fine-tuning as follows.

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x},\boldsymbol{c},t}[w_t \| f_\theta(g_t(\boldsymbol{x}), c) - \boldsymbol{x}\|_2^2 + \\ &\lambda w_{t'} \| f_\theta(g_{t'}(\boldsymbol{x'}), c_{class}) - \boldsymbol{x'}\|_2^2 + \lambda_c \mathcal{C}(\boldsymbol{z}_t, \boldsymbol{z}_0)].\end{aligned} \tag{4}$$

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 [46]. 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 and v1.5 [3]. The model uses CLIP's [2] text encoder trained on laion-aesthetics v2 5+ and a vector quantized VAE [54] to accomplish the task of age progression. The text encoder stays frozen while training the diffusion model. We use three datasets, namely, **CelebA** [41], **AgeDB** [43] and **LFW** [27]. We use 2,258 face images belonging to 100 subjects from the CelebA dataset, 659 images belonging to 100 subjects from the AgeDB dataset and 1,000 images belonging to 100 subjects from the LFW dataset to form the 'training set'. CelebA and LFW 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

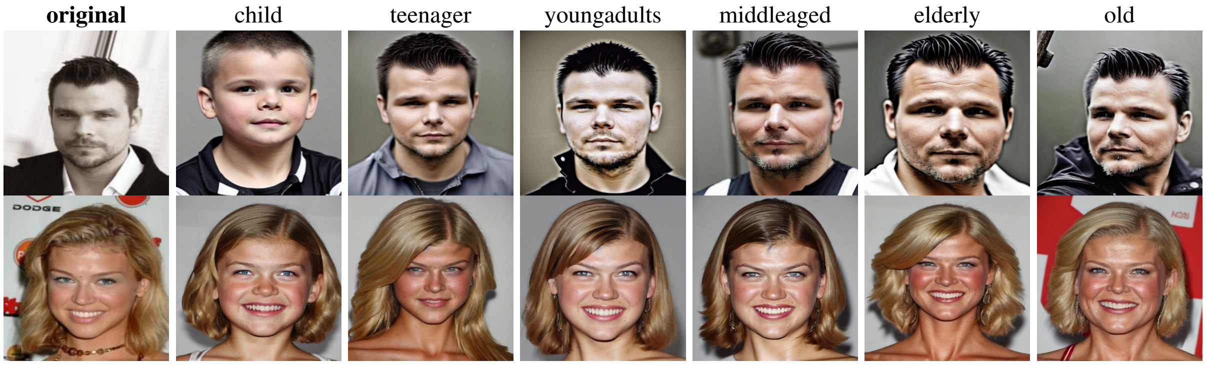


Fig. 3: Illustration of age edited images generated from the CelebA dataset.

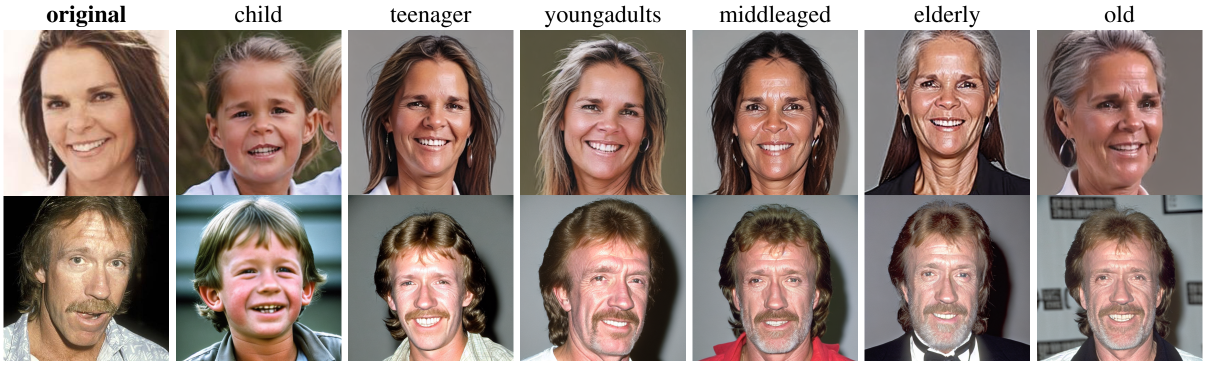


Fig. 4: Illustration of age edited images generated from the AgeDB dataset.

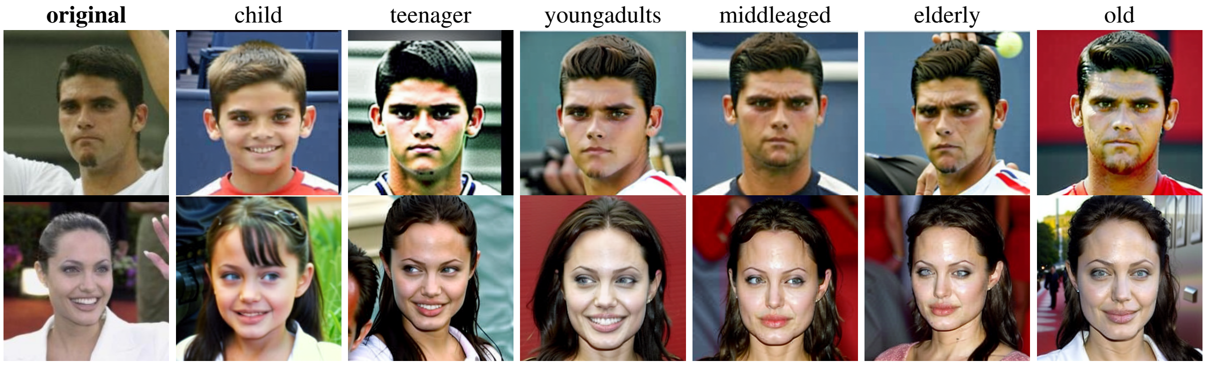


Fig. 5: Illustration of age edited images generated from the LFW dataset.

age group that has the highest number of images as the training set, and the remaining images as the testing set.

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 [29] dataset,where the authors provide fine-grained annotations of age distributions. We convert the distribution to categorical labels to be uses as captions. 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. We use the fixed regularization set across all three datasets — CelebA, LFW and AgeDB.

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*} [8] in this work.

We use the implementation of DreamBooth using stable diffusion in [3] and used the following hyperparameters. 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 $\lambda = 1, \lambda_b = \lambda_c = \lambda_s = 0.1$ (refer to Eqns. 2, 3 and 4). We generate 8 samples at inference. However, we perform a facial quality assessment using EQFace [38] 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.

TABLE 1: Biometric matching on CelebA and AgeDB datasets for aging and de-aging using SDv1.4 and wzx token between Original-Modified images. We report FNMR @FMR = 0.01/0.1%.

Age group	CelebA	AgeDB
child	0.60/0.29	0.73/0.54
teenager	0.34/0.12	0.42/0.14
youngadults	0.31/0.10	0.33/0.09
middleaged	0.30/0.10	0.27/0.06
elderly	0.29/0.11	0.34/0.09
old	0.32/0.12	0.46/0.12

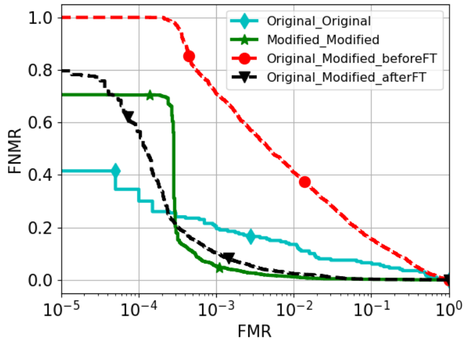
Further, we perform **quantitative evaluation** of the generated outputs using the ArcFace [18] matcher (with RetinaFace [17] detector) and AdaFace [30] matcher (different from VGGFace used in identity-preserving biometric loss). We utilize the genuine (intra-class) and imposter (inter-class) scores 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

5.1 Biometric performance evaluation

We report the biometric matching performance using the ArcFace matcher between **original and modified** images in Table 1 for the CelebA and AgeDB datasets. Our observations indicate that the method performs better on the CelebA dataset than the AgeDB dataset. See examples of generated images in Figs. 3 and 4. In AgeDB, we have a gallery set separate from the training set. 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. 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 observe that the generated images using contrastive loss (Eqn. 3) successfully accomplish aging/de-aging but achieve modest matching results with an average FNMR=0.36 @FMR=0.01 and FNMR= 0.14 @FMR=0.1%. We believe that the ArcFace matcher is typically not trained on generated images, and therefore, struggles to perform matching between the original gallery and generated age-edited probe images.

To test our hypothesis, 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. 6 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% @FMR=0.01%** and **FNMR=1% @FMR = 0.1%** with the fine-tuned face matcher on the age-edited images. The fine-tuned face matcher drastically improves when comparing original-modified images, thus, demonstrating the utility of synthetic images in improving robustness of face matchers as suggested in [47].



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**

Fig. 6: (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.

We illustrate examples of generated images belonging to the LFW dataset in Fig. 5. See the biometric matching results of SDv1.5 model on CelebA and LFW for *sks* and *wzx* tokens in Fig. 7. In this set of experiments, we randomly selected a subset of three images that were used for fine-tuning the diffusion model (SDv1.5) as the gallery set and then computed the biometric matching performance for the six age groups. Additionally, we use a reconstruction prompt, *i.e.*, "photo of a *sks* person" and "photo of a *wzx* person" that ideally outputs images with subtle variations as the input training set. The objective of this prompt is to measure the fidelity of identity preservation in absence of age variations. Our findings indicate that the SDv1.5 model performs better on the CelebA than LFW dataset.

5.2 Comparison of auxiliary loss functions

We compare three loss functions: 1) VGGFace-based Biometric loss, 2) Contrastive loss and 3) Cosine embedding loss. We observe a reduction in FNMR up to 46% @FMR=0.01% when using contrastive loss with respect to biometric loss, and a reduction in FNMR up to 3% @FMR=0.1% with respect to cosine loss averaged across all age groups computed on the CelebA dataset. See Fig. 8. We explored different values of λ_b, λ_c and λ_s, $= \{0.01, 0.1, 1, 10\}$, and observe 0.1 produces the best results for all three hyper-parameters (higher values of hyper-parameters were resulting in over-smooth generated images).

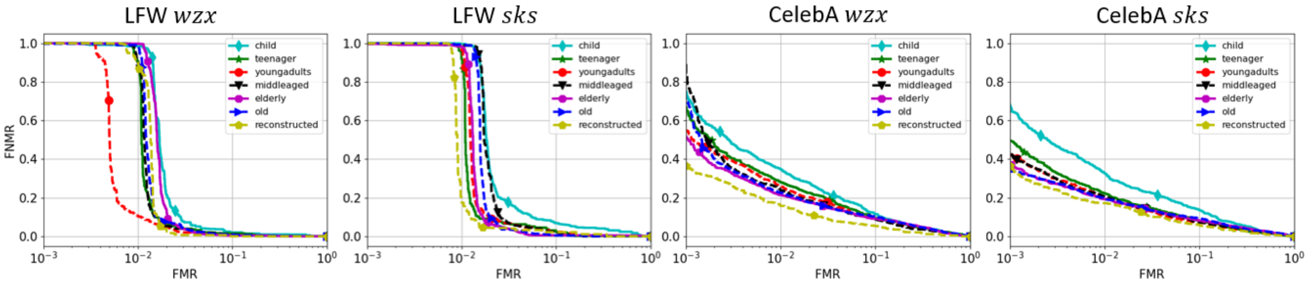
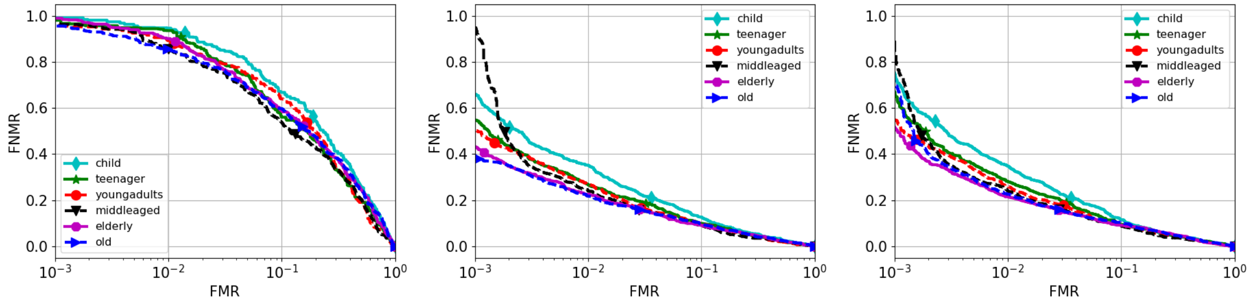


Fig. 7: ArcFace matcher performance on CelebA and LFW for sks and wzx tokens using SDv1.5.

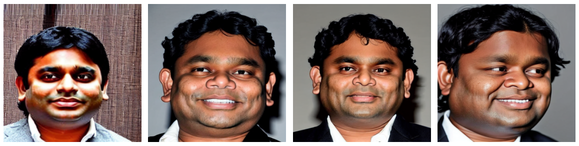


(a) Biometric (b) Cosine (c) Contrastive

Fig. 8: Comparison of auxiliary loss functions (VGGFace-based biometric loss vs. Cosine loss vs. Contrastive loss) in terms of DET curves that indicate that contrastive loss marginally outperforms cosine embedding loss, while both contrastive and cosine embedding losses outperform VGGFace-based biometric loss.

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.



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

Fig. 9: Comparison of images generated using different tokens.

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

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. See Fig. 11. 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. 12.

TABLE 2: AdaFace performance on CelebA and LFW datasets using SDv1.5. We report FNMR @FMR = 0.01/0.1%.

Age group	CelebA		LFW	
	sks	*wzx*	*sks*	*wzx*
child	0.03/0.02	0.04/0.02	0.10/0.04	0.07/0.06
teenager	0.04/0.03	0.05/0.04	0.04/0.04	0.07/0.06
youngadults	0.04/0.03	0.06/0.04	0.05/0.04	0.07/0.06
middleaged	0.04/0.03	0.05/0.04	0.06/0.06	0.06/0.05
elderly	0.05/0.04	0.05/0.04	0.05/0.05	0.07/0.07
old	0.06/0.04	0.06/0.05	0.05/0.04	0.07/0.06

5.6 Effect of variations in face matchers

To study the variations in the performance of face matchers, we utilize another face matcher, namely, AdaFace [30] to compare and contrast withe ArcFace. The reason for selecting AdaFace is because it integrates the quality of the face images while learning

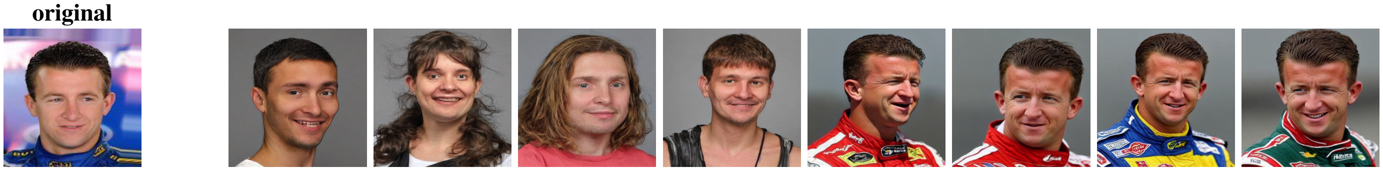


Fig. 10: Impact of token (*wzx*) and class label (*person*) on generated images: "photo of a person" (left) vs. "photo of a ***wzx*** person" (right). Note the token is strongly associated with a specific identity belonging to that class.

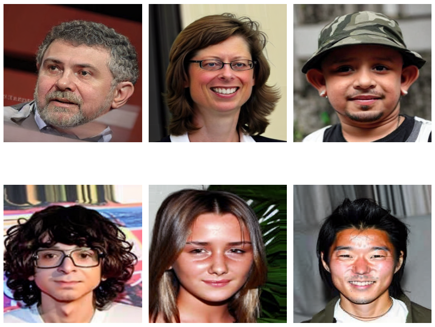


Fig. 11: Failure cases for simulating ages in the child (top row) and old (bottom row) age groups.

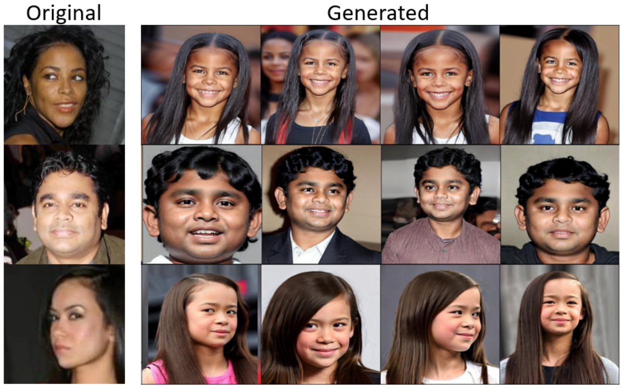


Fig. 12: Examples of generated images pertaining to diverse sex and ethnicity for 'child' group.

the facial embedding for recognition, and is shown to outperform ArcFace on challenging data. We conduct this experiment on CelebA and LFW datasets using sks and wzx tokens. We first present the results of AdaFace on both LFW and CelebA in Table 2. Next, we compare the performance between ArcFace and AdaFace, specifically, on the CelebA dataset in Table 3. We observe that AdaFace outperforms ArcFace by 16% @FMR=0.01% and by 4% @FMR=0.1%. AdaFace is supervised by image quality which is influenced by several factors such as, resolution, illumination, expression and pose variations. Additionally, introducing age variations in the face images may produce some visual artifacts such as wrinkles that may affect texture of the face and can impair genuine matches. AdaFace therefore performs better than ArcFace. We further observe AdaFace performs worse on LFW compared to CelebA because we use aligned and cropped images from the CelebA that are of higher quality than the LFW images resulting in an increase in FNMR by 12% @FMR=0.01% and by 9% @FMR=0.1% when computed on the Reconstructed images, *i.e.*, without any age editing.

TABLE 3: Performance variations in face matchers — ArcFace vs. AdaFace in terms of FNMR@FMR=0.01/0.1% on CelebA dataset using SDv1.5 with *wzx* token.

Age group	ArcFace @FMR(%)=0.01/0.1	AdaFace @FMR(%)=0.01/0.1
Reconstructed	0.16/0.06	**0.02/0.02**
child	0.35/0.12	**0.02/0.04**
teenager	0.29/0.10	**0.04/0.05**
youngadults	0.26/0.09	**0.04/0.06**
middleaged	0.25/0.09	**0.04/0.05**
elderly	0.22/0.10	**0.04/0.06**
old	0.23/0.11	**0.05/0.06**
Average	0.20/0.09	**0.04/0.05**

5.7 Effect of variations in diffusion models

We compare the two diffusion models used in this work in Fig. 13. We compute the biometric scores between genuine pairs, *i.e.*, images belonging to the same individual and we averaged it across all the six age groups, and then plot the histogram of the scores. We observe that SDv1.5 results in lower genuine distance/dissimilarity scores compared to SDv1.4 model indicating better generative capability for fine-tuning while retaining original identity characteristics.

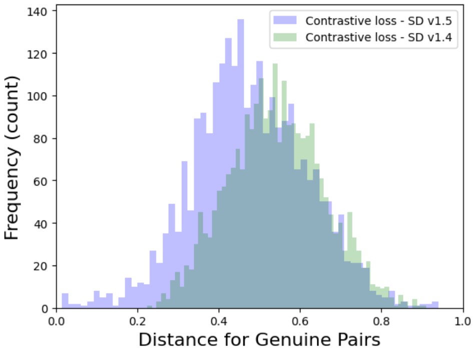


Fig. 13: Variations in genuine scores between SDv1.4 and SDv1.5 with contrastive loss on the CelebA dataset. We observe SDv1.5 outperforms SDv1.4 resulting in lower distance.

5.8 Comparison with existing methods

We use IPCGAN [52], AttGAN [23] and Talk-to-Edit [29]— three GAN-based methods for comparison. We use the pre-trained models provided by the authors. As IPCGAN was trained on the CACD dataset [13], 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 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. 15. We observe that the proposed method (contrastive loss, *sks*) outperforms AttGAN by 19% on 'young' images and by 7% on 'old' images @FMR=0.01. AttGAN can only edit to young or old ages. Further, we observe that the method outperforms Talk-to-Edit (ToE) 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. We observe several cases of inconsistent outputs in ToE. We compare with ProFusion [62] a regularization-free diffusion framework that uses a single test image and creates augmentations of it at inference time for fine-tuning. It is pre-trained on FFHQ dataset. We randomly sample one of the training images for each subject from the CelebA dataset and use it on ProFusion with the following hyperparameters: cfg=7.0 (guidance from image), ref_cfg=5.0 (guidance from prompt), refine_cfg =7.0 (guidance for fusion step sampling) and number of sampling steps=100 to follow as close as possible as the hyperparameters used during Dreambooth inference. See Fig. 16 to compare between images generated by ProFusion and our method. Although ProFusion is much faster than DreamBooth, but lags behind our method both in terms of perceptual and biometric utility by 18%FNMR @FMR=0.01%. Our findings indicate reliable age editing while maintaining biometric fidelity requires regularization provided by DreamBooth as evidenced by the superior biometric matching performance.

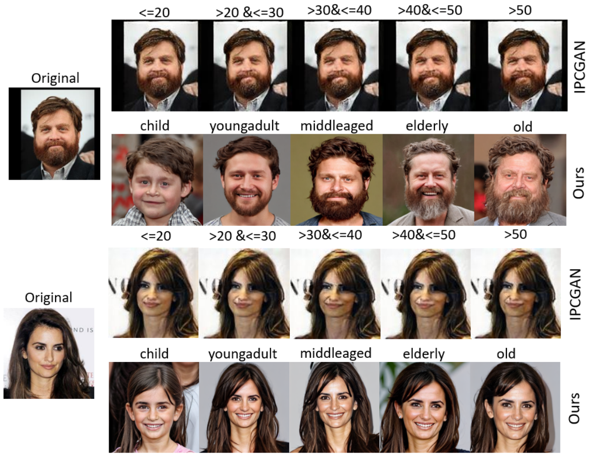


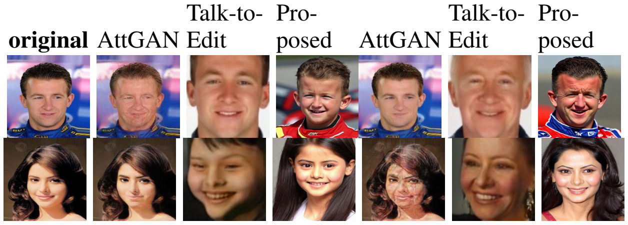
Fig. 14: Comparison of outputs produced by IPCGAN and our method for a male-presenting image (top) and a female-presenting image (bottom).

5.9 Additional evaluation

We perform additional evaluation in terms of age prediction and automated image quality assessment.

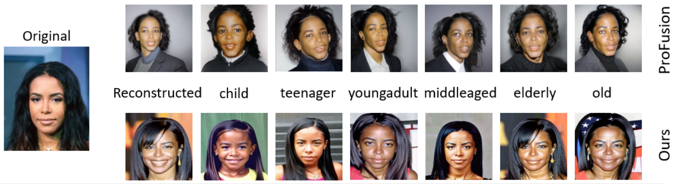
5.9.1 Age prediction

We evaluate the outputs of our method by performing age prediction. First, we present statistical analysis of the age predicted using deepface [9] library for images generated using SDv1.5 in Fig. 17. We observe that the reconstructed images conform mostly to the middleaged group which can be considered as the age group of the training images. We observe a gradual increase in the predicted median age values across the age groups. Second, we compare the performance of our method with IPCGAN [52] in terms of age prediction. 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. 18.



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**

Fig. 15: (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): Biometric matching in terms of FNMR @FMR = 0.01/0.1%.



Age group	Methods		
	ProFusion	Ours (SDv1.4)	Ours (SDv1.5)
child	0.63/0.32	0.56/0.26	0.33/0.14
teenager	0.38/0.18	0.29/0.10	0.23/0.08
youngadults	0.36/0.13	0.28/0.08	0.20/0.08
middleaged	0.39/0.14	0.27/0.09	0.21/0.08
elderly	0.35/0.12	0.25/0.09	0.20/0.09
old	0.37/0.13	0.29/0.11	0.20/0.10
Average	0.41/0.17	0.32/0.12	**0.23/0.09**

Fig. 16: (Top): Comparison of outputs produced by ProFusion and the proposed method (SDv1.4 and SDv1.5) using *sks*. The original images are in the first column. (Bottom): Biometric matching in terms of FNMR @FMR = 0.01/0.1%. 'Reconstructed' corresponds to outputs without age-editing.

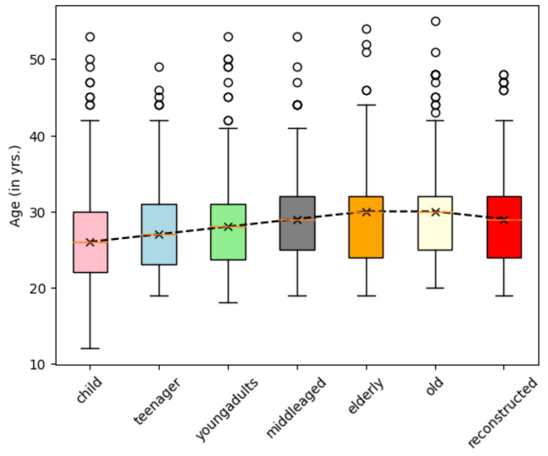


Fig. 17: Box plot of ages computed for generated images using SDv1.5. ‘Reconstructed’ corresponds to generating the images of the subject to be fine-tuned without subjecting it to age variations. The reconstructed images have the median age value consistent with the middleaged group (indicated by the dashed line).

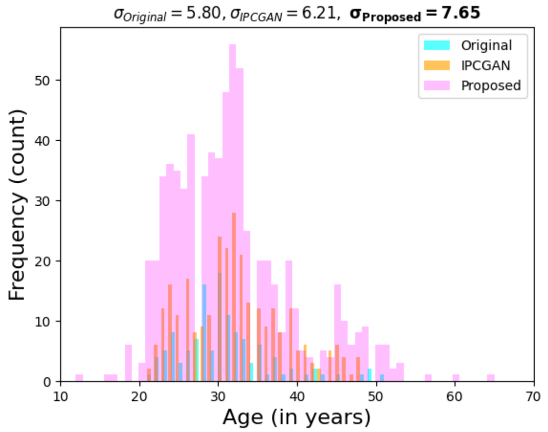


Fig. 18: Age prediction shows that our method generates images with a wider age dispersion compared to original CACD images and IPCGAN-generated images.

5.9.2 Image quality evaluation

We have conducted evaluations to compute image (perceptual) quality in terms of FID (lower is better) and IS (higher is better) metrics using the implementation in [7]. On CelebA dataset, using the *sks* token, we achieve a **FID = 39.54 and an IS = 3.2±0.15**; using the *wzx* token, we achieve a **FID = 42.14 and an IS = 3.2±0.14**. We provide the IS metrics for each age group and reconstructed images in Table 4.

As FID/IS scores are known to produce biased evaluations against Diffusion model [50], we augment our evaluation with a CLIP-based score. CLIPScore [25] measures the cosine similarity between image and text embedding and weighs it by a scalar term, and denotes the semantic similarity between images and their corresponding captions. We used the six captions corresponding to each age groups and the corresponding age-edited images. We used the implementation outlined in Hugging Face [4] with stable diffusion v1.4 and openai/clip-vit-large-patch14 (ViT-L/14) text-encoder. We observed an average **CLIPScore = 18.34**. This score can vary with the CLIP text encoder, diffusion model architecture and the detail of the prompt.

TABLE 4: Inception score (with standard deviation) for different age-edited and un-edited (reconstructed) images produced by the proposed method on the CelebA dataset for *sks* and *wzx* tokens.

Age group	*sks*	*wzx*
reconstructed	2.9 ± 0.31	2.8 ± 0.29
child	3.1 ± 0.4	3.0 ± 0.5
teenager	2.7 ± 0.26	2.7 ± 0.23
youngadults	2.8 ± 0.22	2.9 ± 0.27
middleaged	2.9 ± 0.36	2.9 ± 0.25
elderly	2.9 ± 0.18	3.0 ± 0.31
old	2.9 ± 0.22	2.9 ± 0.23

5.10 Ablation Study

We conduct ablation by evaluating the generated outputs by (i) varying the number of training images for fine-tuning diffusion, and (ii) varying the number of sampling steps during inference.

5.10.1 Effect of varying number of training images

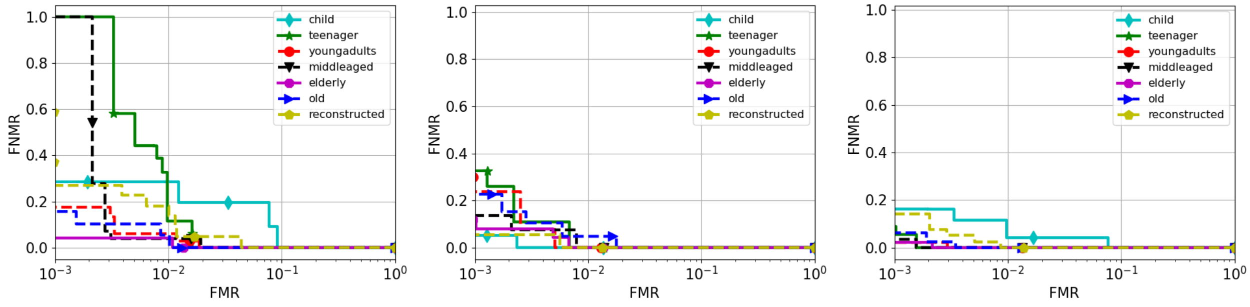
To assess the effect of varying the size of the training set, we conduct the following experiment. We vary the number of training images as $\{10, 20, 30\}$ and fine-tuned DreamBooth [49] for the same 10 subjects in each case, with identity-preserving loss and regularization set for age editing. We then performed biometric evaluation using the ArcFace matcher and presented the respective DET curves in Fig. 19. Although higher number of training images seem to benefit matching performance in some of the age groups, we observe that with ∼20 training images per subject we can obtain successful matching across all age groups. Refer to *child* age group that has higher error rate when 30 images are used for training than 20 images.

5.10.2 Effect of varying number of sampling steps

During inference, we select values such as *ddim_steps* that indicate the number of sampling steps required by the Denoising Diffusion Implicit Model (DDIM)-based sampler, *n_samples* that indicate the number of samples produced for each prompt. In our experiments, we used *ddim_steps*=100 and *n_samples*=8 with deterministic sampling option. The inference speed will be faster if we reduce the number of samples to 1 (same as GANs) and the number of sampling steps without compromising the quality of the age-edited images (see Fig. 20). Alternatively, we can use a different sampler such as DPM (Denoising Probabilistic Model Solver) and DPM++ that are faster than DDIM.

Key Findings. We summarize our observations.

- We observe that the proposed latent diffusion-guided age editing is able to preserve (i) biometric utility (ii) perceptual quality, and (iii) age-specific facial cues. We use ArcFace and AdaFace as automated face matchers; FID, IS and CLIPScore as perceptual quality metrics; open-source library as age estimator. We further augment it with human-based evaluation that performs well on both identification and age estimation. We achieve an FNMR=3% @FMR=0.01% and FNMR=1% @ FMR=0.1% by fine-tuning ArcFace on age edited images.
- We conduct extensive experiments on multiple datasets using different loss functions and model architectures to perform



(a) 10 images

(b) 20 images

(c) 30 images

Fig. 19: Comparison of biometric matching in terms of DET curves as a function of variation of number of training images. We note that increasing number of training images improve matching performance for some of the age groups (teenager and middleaged).

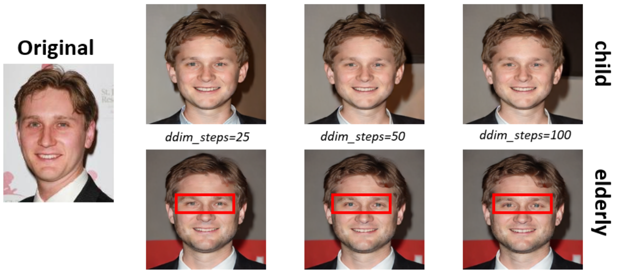


Fig. 20: Age-edited images for different number of sampling steps (*ddim_steps*=$\{25, 50, 100\}$), number of samples (*n_samples*=1) with deterministic sampling. Adequate sampling steps are needed to ensure convergence and preservation of finer details in the face, such as eyes. See the red bounding box for enhanced detail preservation such as corneal reflections.

facial age editing. Overall, SDv1.5 with contrastive loss and cosine embedding loss performs comparably.

- We compare our method with state-of-the-art GAN and diffusion based age editing methods. Our method outperforms them by a significant margin (by 18% on ProFusion and 44% on Talk-to-Edit).
- We analyze the effect of number of training images and sampling steps and observe that $\sim$ 20 training images and $\sim$ 100 steps are adequate for synthesizing high-quality age-edited images. We observe no strong variation in terms of sex and ethnicity on the generated images.
- **Limitations:** First, we observe that if the images in the training set pertain to only a single age, say old-age then the method struggles to generate images on the other side of the spectrum, *i.e.*, child. Second, our method requires multiple images of the same subject with different poses and expressions for diversity. Third, the effect of rare tokens on the generative model needs principled analysis for better understanding and controllability of the synthesis process.

6 Conclusion

We present a novel facial aging and de-aging technique using regularization-based conditional image generation via latent diffusion models. We achieve this by curating a small regularization set of image and caption pairs to teach the model the concept of facial aging. Additionally, we enforce identity-preservation using a novel combination of biometric, cosine and contrastive losses to preserve biometric integrity of the original individual while customizing their appearance to fit the target age in the prompt. We conduct extensive experiments on three datasets (CelebA, LFW and AgeDB), two face matchers (ArcFace and AdaFace) and two diffusion models (SDv1.4 and SDv1.5). We compare our method with three GAN-based age editing methods (IPCGAN, AttGAN and Talk-to-Edit), where we achieved a significant reduction in FNMR upto 44%. We achieved realistic age-edited faces compared to ProFusion, a regularization-free diffusion model based image customization and achieved low FNMR values by 18%. We further boosted the performance of an existing face matcher, ArcFace by 38% reduction in FNMR @FMR=0.01% by fine-tuning it on synthetic age-edited images.

Future work will focus on accomplishing age editing reliably from a single image. Current regularization-free methods struggle between inducing age-specific changes while retaining sufficient identity cues. Another research direction will be integrating neural radiance fields for reliable 3D face reconstruction [60] with diffusion models for 3D facial aging [24], [57].

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]. 5

[3] DreamBooth Using Stable Diffusion. https://github.com/XavierXiao/Dreambooth-Stable-Diffusion. [Online accessed: April 14, 2023]. 5, 6

[4] Evaluating Diffusion Models. https://huggingface.co/docs/diffusers/conceptual/evaluation. [Online accessed: February 02, 2024]. 11

[5] Facenet pytorch. https://github.com/timesler/facenet-pytorch. [Online accessed: April 14, 2023]. 5

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

[7] Pytorch GAN metrics. https://pypi.org/project/pytorch-gan-metrics/. [Online accessed: January 25, 2024]. 11

[8] 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]. 6

[9] ArcFace and VGGFace implementation. https://pypi.org/project/deepface/. [Online accessed: 18th May, 2022]. 10

[10] S. Banerjee, G. Mittal, A. Joshi, C. Hegde, and N. Memon. Identity-preserving aging of face images via latent diffusion models. *International Joint Conference in Biometrics (IJCB)*, 2023. 2, 3

[11] 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

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

[13] 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, 9

[14] L. Chen, M. Zhao, Y. Liu, M. Ding, Y. Song, S. Wang, X. Wang, H. Yang, J. Liu, K. Du, and M. Zheng. Photoverse: Tuning-free image customization with text-to-image diffusion models, 2023. 3

[15] 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. 5

[16] X. Chen and S. Lathuilière. Face aging via diffusion-based editing. *34th British Machine Vision Conference (BMVC)*, 2023. 3

[17] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 7

[18] 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. 7

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

[20] 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. 5

[21] 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

[22] 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

[23] 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, 3, 9

[24] F. M. Z. Heravi and A. Nait-Ali. Adult-child 3D backward face aging model (3D B-FAM). *Journal of Visual Communication and Image Representation*, 72:102803, 2020. 12

[25] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *ArXiv*, abs/2104.08718, 2021. 11

[26] 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. 4

[27] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 2, 5

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

[29] 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, 3, 6, 9

[30] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 7, 8

[31] 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, 2, 3

[32] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations*, 2014. 4

[33] 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

[34] M. Levine and D. Martin. Have you seen me? aging the images of missing children (ncj number 130323). *Law Enforcement Technology*, 18(6):34–40, 1991. 2

[35] P. Li, Y. Hu, R. He, and Z. Sun. Global and local consistent wavelet-domain age synthesis. *IEEE Transactions on Information Forensics and Security*, 14(11):2943–2957, 2019. 2

[36] Q. Li, Y. Liu, and Z. Sun. Age progression and regression with spatial attention modules. In *AAAI*, 2020. 2

[37] 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

[38] 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. 6

[39] Y. Liu, Q. Li, and Z. Sun. Attribute-aware face aging with wavelet-based generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11869–11878, 2019. 2

[40] Y. Liu, Q. Li, Z. Sun, and T. Tan. A3GAN: An Attribute-Aware Attentive Generative Adversarial Network for Face Aging. *IEEE Transactions on Information Forensics and Security*, 16:2776–2790, 2021. 2

[41] 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. 5

[42] J. Maeng, K. Oh, and H.-I. Suk. Age-aware guidance via masking-based attention in face aging. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, page 4165–4169, 2023. 2

[43] 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, 5

[44] 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

[45] R. Or-El, S. Sengupta, O. Fried, E. Shechtman, and I. Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[46] 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 *International Workshop on Biometrics and Forensics*, 2023. 2, 5

[47] 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, 7

[48] 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. 2, 4

[49] 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, 4, 11

[50] G. Stein, J. C. Cresswell, R. Hosseinzadeh, Y. Sui, B. L. Ross, V. Villecroze, Z. Liu, A. L. Caterini, J. E. T. Taylor, and G. Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2023. 11

[51] K. W. Strandberg. Age progression and kidcare (ncj number 147113). *Law Enforcement Technology*, 21(2):46–49, 1994. 2

[52] 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, 3, 9, 10

[53] TikTok's Hyper-Realistic 'Aged' Face Filter is Praised By Users. https://petapixel.com/2023/07/11/tiktoks-hyper-realistic-aged-face-filter-is-praised-by-users/. [Online accessed: 13th October, 2023]. 1

[54] A. van den Oord, O. Vinyals, and k. kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 5

[55] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu. CosFace: Large Margin Cosine Loss for Deep Face Recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 2, 5

[56] 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

[57] Y. Wu, R. Wang, M. Gong, J. Cheng, Z. Yu, and D. Tao. Adversarial uv-transformation texture estimation for 3d face aging. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4338–4350, 2022. 12

[58] 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

[59] 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

[60] J. Zhang, X. Li, Z. Wan, C. Wang, and J. Liao. FDNeRF: Few-shot Dynamic Neural Radiance Fields for Face Reconstruction and Expression Editing. In *Proceedings of the SIGGRAPH Asia*, 2022. 12

[61] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4352–4360, 2017. 2

[62] Y. Zhou, R. Zhang, T. Sun, and J. Xu. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. *arXiv preprint arXiv:2305.13579*, 2023. 2, 3, 10

Sudipta Banerjee is a Research Assistant Professor in the Department of Computer Science at New York University, US. She received her Bachelor in Technology from WBUT, India in 2011, Masters' in Electronics and Telecommunication Engineering from Jadavpur University, India in 2014 and Doctorate in Computer Science from Michigan State University, US in 2020. Her research focuses on biometrics, generative modeling and image forensics.

Govind Mittal earned his Bachelor in Computer Science and Master in Mathematics from BITS Pilani, India. He is currently pursuing a Ph.D. in Computer Science at New York University, Tandon School of Engineering, where he is investigating applications of machine learning to media authentication, making them more usable by enabling human-AI collaborations. His research encompasses detecting real-time video and audio deepfakes and building trustworthy machine-learning models. Govind is driven by a passion for addressing AI Safety challenges, and aspires to contribute to policy-making and developing safe AI practices post-Ph.D.

Ameya Joshi received his Bachelors in Electrical Engineering (Hons.) from BITS Pilani, India in 2014, and his PhD in Electrical Engineering from New York University in 2023, His research interests include robust representation learning, generative models, and multimodal learning. He is currently a Research Scientist at Instadeep, developing representation learning methods for single-cell genomics and multi-omics.

Sai Pranaswi Mullangi is currently working as a Graduate Assistant while pursuing her MS in Computer Engineering at New York University, Tandon School of Engineering. She received her B.Tech (2020) in Electronics and Communications Engineering from the National Institute of Technology, Andhra Pradesh, India. Her research interests include Machine Learning, Deep Learning, Biometrics, and specifically Generative models.

Chinmay Hegde (M '10, S '18) is an Associate Professor at NYU, jointly appointed with the CSE and ECE Departments. His research focuses on foundational aspects of machine learning (such as reliability, robustness, efficiency, and privacy). He also works on applications ranging from computational imaging, materials design, and cybersecurity. He is a recipient of the NSF CAREER and CRII awards, the Black and Veatch Faculty Fellowship, multiple teaching awards, and best paper awards at ICML, SPARS, and MMLS. He is a Senior Member of the IEEE.

Nasir Memon is a professor in the Department of Computer Science and Engineering at NYU Tandon School of Engineering and a co-founders of the Center for Cyber-Security at NYU. His research interests include digital forensics, biometrics, authentication, security and human behav- ior. Memon earned a Bachelor of Engineering in Chemical Engineering and a Master of Science in Mathematics from Birla Institute of Technology and Science (BITS) in Pilani, India. He received a PhD in Computer Science from the University of Nebraska. Professor Memon has published over 350 articles in journals and conference proceedings and holds a dozen patents in image compression and security. He has won several awards including the Jacobs Excellence in Education award and several best paper awards. He has been on the editorial boards of several journals and was the Editor-In-Chief of Transactions on Information Forensics and Security. He is a Fellow of the IEEE, IAPR and SPIE.