Only a Matter of Style: Age Transformation Using a Style-Based Regression Model

Yuval Alaluf
Tel-Aviv University

Or Patashnik
Tel-Aviv University

Daniel Cohen-Or
Tel-Aviv University

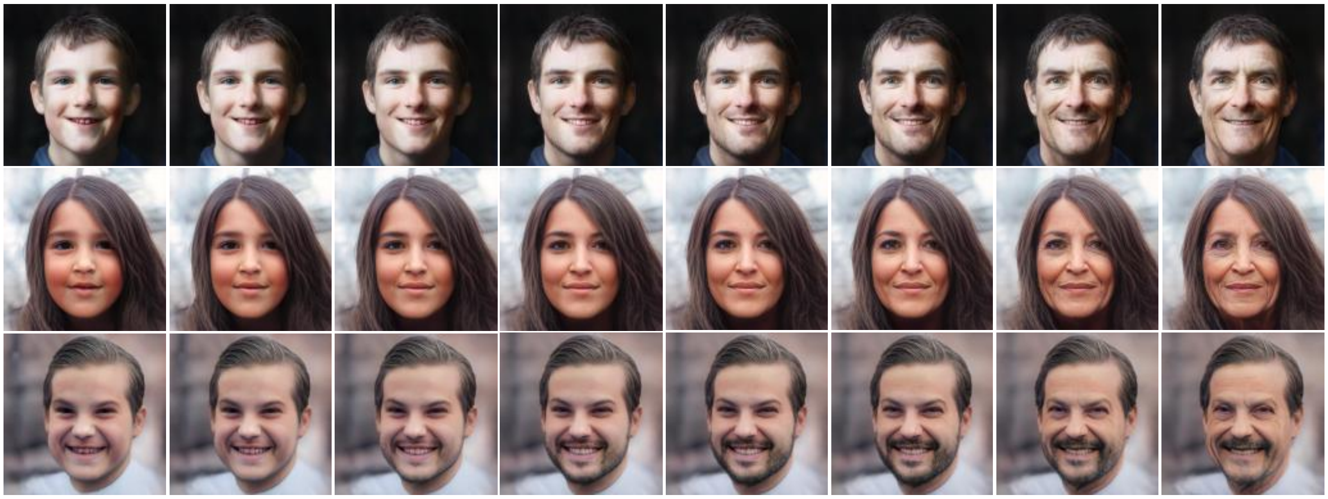


Figure 1: Modeling fine-grained lifelong age transformation using the style-based SAM method.

Abstract

The task of age transformation illustrates the change of an individual's appearance over time. Accurately modeling this complex transformation over an input facial image is extremely challenging as it requires making convincing and possibly large changes to facial features and head shape, while still preserving the input identity. In this work, we present an image-to-image translation method that learns to directly encode real facial images into the latent space of a pre-trained unconditional GAN (e.g., StyleGAN) subject to a given aging shift. We employ a pre-trained age regression network used to explicitly guide the encoder in generating the latent codes corresponding to the desired age. In this formulation, our method approaches the continuous aging process as a regression task between the input age and desired target age, providing fine-grained control over the generated image. Moreover, unlike other approaches that operate solely in the latent space using a prior on the path controlling age, our method learns a more disentangled, non-linear path. Finally, we demonstrate that the end-to-end nature of our approach, coupled with the rich semantic latent space of StyleGAN, allows for further editing of the generated images. Qualitative and quantitative evaluations show the advantages of our method compared to state-of-the-art approaches.

1. Introduction

Age transformation is the process of representing the change in a person's appearance across different ages while preserving their identity. Recently, this task has received increased attention with the rise of applications allowing users to perform facial editing, and age transformation in particular. To model the aging process over a single input facial image one must capture both the change in head shape and texture while faithfully preserving the identity and other key facial attributes of the input face. This becomes increasingly challenging when modeling *lifelong* aging where the desired change in age is significant (e.g. from ages 5 to 85).

To bypass explicitly modeling age transformation, data-driven techniques have been explored. Due to their phenomenal realism, Generative Adversarial Networks (GANs) have been heavily used for synthesizing images in a data-driven fashion, particularly on facial images. These works can typically be classified into two methodologies: image-to-image translation and latent space manipulation.

To model age transformation as an image-to-image problem, most works [14, 30, 54, 64, 67, 71, 76] use age-annotated data and learn a mapping between pre-defined age groups. As the age groups are highly correlated, these methods struggle to model meaningful changes between different ages. Moreover, collecting age-annotated data is often tedious at large scale.

Other works [13, 29, 48, 60, 68] have approached the age transformation task by exploring the semantics of the latent space of a well-trained GAN, such as StyleGAN [38, 39], and perform a latent space traversal to obtain the desired transformed image. These methods often assume the existence of a corresponding linear path in the latent space controlling the attribute of interest. This, however, relies on the existence of a well-behaved, fully disentangled and linear latent space which is difficult to obtain. In addition, while such methods have shown promising editing results on synthetic images generated by StyleGAN, they often struggle to make realistic transformations on *real* images.

In this work, we present a novel method for learning a conditional image generation function capable of capturing the desired change in age while faithfully preserving identity. We approach the age transformation task as an image-to-image translation problem by pairing the expressiveness of a *pre-trained, fixed* StyleGAN generator with an encoder architecture. The encoder is tasked with directly encoding an input facial image into a series of style vectors subject to the desired age change. These style vectors are then fed into StyleGAN to generate the output image representing the desired age transformation. This allows us to easily leverage the state-of-the-art image quality achieved by StyleGAN. To explicitly guide the encoder in generating the corresponding latent codes, we utilize a pre-trained, fixed age regression network to serve as an additional constraint during training. We name our method SAM — *Style-based Age Manipulation* — as our age transformation is controlled via the learned intermediate style representation.

Rather than using labeled data directly, our method attains supervision only through the use of readily available pre-trained networks: (i) a StyleGAN2 [39] generator network trained on facial images, (ii) a pre-trained encoder network from Richardson *et al.* [57] trained to encode real face images into the $\mathcal{W}+$ latent space, (iii) an ArcFace [22] facial recognition network for identity regression, and (iv) a VGG [62] network for age regression. Note that our approach does not assume the existence of multiple age classes or domains. Instead, given a single input image and desired target age, we show that our approach can successfully generate the corresponding image, see Figure 1. Compared to multi-domain approaches that rely on pre-defined age groups [18, 47] or anchor classes [54], viewing human aging as a continuous *regression* process allows for more fine-grained control over the desired transformation.

Furthermore, we analyze the latent path learned by SAM and show it results in a more precise, *non-linear* path that is less entangled with other attributes and conforms well to StyleGAN's latent space manifold. Finally, we demonstrate how the end-to-end nature of SAM, together with the StyleGAN latent space, allows for additional editing on the generated images (e.g. hair color, expression).

Qualitative and quantitative evaluations show that our style-based regression method outperforms current state-of-the-art methods. The main contributions of this paper are:

- A novel style-based regression approach for fine-grained modeling of the age transformation process.
- An analysis of the non-linear latent path learned by SAM showing the benefits of an end-to-end method for modeling age transformation on real images.

We have released our source code and pre-trained model[1]. We also invite the reader to view the accompanying video demonstrating lifespan results generated by SAM.

2. Related Work

Age transformation has been an extensively studied topic in computer graphics. Early works either explicitly modeled the face transformation over time or sought a prototype face for each age group. We refer the reader to [24, 26, 56] for a comprehensive survey of such approaches. More recent works [31, 40, 42, 54, 72] take a data-driven approach for modeling face aging by using deep neural networks.

2.1. Image-to-Image Translation

Image-to-image translation techniques aim at translating a given image of a source domain to a corresponding image of a target domain. Isola *et al.* [34] first introduced the use of conditional GANs [52] for solving various image-to-image translation tasks. As corresponding pairs of source domain and target domain images are not always available or are overly tedious to collect, unpaired approaches have recently been developed for solving such tasks [46, 73, 78]. Some methods [33, 41] translate a given image to a diverse set of corresponding output images, often referred to as multi-modal image synthesis. Furthermore, [17, 18, 33, 47, 79] generalize image-to-image translation into a *multi-domain* translation and train a single network to translate between multiple domains.

Motivated by the early successes of these works, many methods [25, 27, 43, 54, 65] have approached the task of face aging as an image-to-image translation between multiple age groups. These works associate each image with an age label and perform translation between pre-defined age groups. However, as the age groups are highly correlated, these methods often struggle to disentangle age and other attributes. In contrast to these works, we view age transformation as a *regression problem* by using a pre-trained age classifier to estimate age during training without the need for direct supervision. While Yao *et al.* [72] use a similar approach, they are limited to translating images within the $20-70$ age range and model only slight changes in texture.

[1] https://github.com/yuval-alaluf/SAM

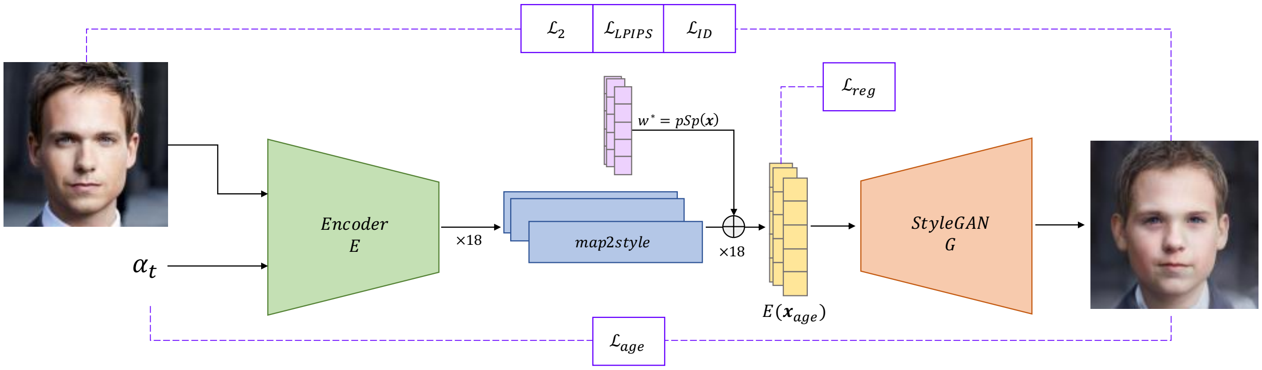


Figure 2: Our SAM architecture. The network receives an input face image and a desired target age α_t. First, the encoder network E is tasked with extracting 18 feature maps corresponding to the 18 style inputs of StyleGAN. The *map2style* layers, introduced in [57], are then used to gradually down-sample each feature map to a 512-dimensional style vector, thereby encoding the input image into the $\mathcal{W}+$ StyleGAN latent space. We additionally employ a fixed, pre-trained pSp [57] encoder to extract the $\mathcal{W}+$ latent code of $\mathbf{x}$, which is then added to the obtained age-transformed latent code. A pre-trained StyleGAN is then used to generate the desired age-transformed image using the resulting latent code $E(\mathbf{x}_{age})$. During training $\mathcal{L}_2, \mathcal{L}_{LPIPS}$ and $\mathcal{L}_{ID}$ ensure visual similarity and identity preservation while $\mathcal{L}_{reg}$ encourages the learned latent codes to be closer to the average latent code. Finally, $\mathcal{L}_{age}$ guides the encoder in generating the desired age-transformed latent code.

2.2. Latent Space of GANs

Recently, many works have explored performing semantically meaningful manipulations in the latent space of a well-trained GAN generator, either by directly learning a disentangled mapping of an attribute of interest [53] or by a latent space traversal [13, 23, 28, 29, 60, 61, 66, 68]. Most notably, given its state-of-the-art image quality and disentangled latent space, StyleGAN [38, 39] has been widely used for this task. Specifically, latent space methods follow an *"invert then edit"* methodology in which real images are first inverted into the latent space of a pre-trained GAN [11, 12, 21, 55, 57, 69, 77]. Then, the resulting latent codes are edited in a semantically meaningful manner by traversing the latent space to obtain a new latent code that is used to generate the edited image. While these approaches allow for extensive editing on real images, they suffer from several drawbacks that should be considered. First, most works rely on the existence of a disentangled linear latent space path controlling age, which is hard to achieve in practice. Second, they are unable to directly generate an image at a desired age. Instead, one must manually traverse the fixed path in search of the desired age-transformed image.

This differs from our approach which can directly encode a real face image, conditioned on a desired target age, into its corresponding latent representation in the StyleGAN domain. Moreover, while previous works assume a prior on the latent path that can be traversed to control age, we train our network to learn this path with no prior assumptions. By doing so, our method learns a non-linear traversal path that is less sensitive to the entanglement of other attributes in the latent space.

3. Method

3.1. Overview

In this section, we present our approach for modeling the age transformation process. Given a source facial image $\mathbf{x}$ at age α_s and a desired target age α_t, our goal is to transform $\mathbf{x}$ to an image $\mathbf{x}' = SAM(\mathbf{x}, \alpha_t)$ representing the source identity at age α_t.

To model the age transformation process, we introduce a complete image-to-image translation architecture by pairing an encoder network and a *fixed*, pre-trained unconditional image generator. The encoder network directly encodes a given image and desired target age to a set of style vectors that capture the desired transformation. Given these style vectors, the generator network is then used to generate the desired output image.

Collecting a series of images of the same person over many years is extremely challenging and therefore we cannot directly rely on pairs of corresponding images. To address this challenge, we apply a cycle consistency loss during training (see Figure 3), which has been shown to be effective in unpaired image-to-image translation tasks [46, 63, 73, 78]. To guide the encoder in generating the appropriate style vectors, we utilize a pre-trained age regression network that serves as an additional loss constraint during the training process. By doing so, the encoder is encouraged to learn a more precise latent path that better approximates the aging process.

3.2. Training

As the age of a given facial image can be estimated using a well-trained age classifier, only the desired target age must be specified. During training, before feeding an image $\mathbf{x}$ through our encoder, the desired target age is randomly generated as

$$\alpha_t \sim \mathcal{U}(5, 100). \tag{1}$$

That is, a target age between 5 and 100 is sampled uniformly at random. The sampled age is then added as a constant-valued channel to the input image, $\mathbf{x}$, resulting in a 4-channel input tensor, which we denote by

$$\mathbf{x}_{age} := \mathbf{x} \parallel \alpha_t. \tag{2}$$

We additionally use a pre-trained Pixel2Style2Pixel [57] encoder trained to encode real face images into the $\mathcal{W}+$ latent space of a pre-trained StyleGAN2 generator. In particular, given an input image $\mathbf{x}$, we first compute,

$$\mathbf{w}^* := pSp(\mathbf{x}) \in \mathbb{R}^{18\times 512}. \tag{3}$$

Using Equations (2)-(3) we then define the output of our model as

$$SAM(\mathbf{x}_{age}) := G(E(\mathbf{x}_{age}) + \mathbf{w}^*)), \tag{4}$$

where $E(\cdot)$ and $G(\cdot)$ denote our encoder and the StyleGAN generator, respectively. Here, our encoder is trained to learn the residual between the latent code representing the age-transformed image and the latent code representing the original input image.

During training, we perform two passes:

$$\mathbf{y}_{out} = SAM(\mathbf{x}_{age}), \tag{5}$$

$$\mathbf{y}_{cycle} = SAM(\mathbf{y}_{out} \parallel \alpha_s), \tag{6}$$

where $\mathbf{y}_{out}$ is the generated image representing the age transformation. We then apply a cycle consistency pass to recover the original image and set the target age equal to the source age α_s (see Figure 3).

3.3. Architecture

Our encoder architecture is based on the encoder presented in Richardson *et al*. [57]. Here, we extend it to an unsupervised setting for modeling face aging. To form a complete image-to-image translation architecture we combine the encoder with the representative power of a pre-trained StyleGAN [38, 39] generator. Given a 4-channel input described in Section 3.2, the encoder, based on a Feature Pyramid Network [44] architecture, is tasked with generating 18 unique style vectors corresponding to the 18 inputs of StyleGAN. To do so, the encoder first extracts 18 feature maps at three different spatial scales. The *map2style* blocks, introduced in [57], are then used to gradually down-sample each feature map to obtain a 512-dimensional style vector. Finally, the 18 style vectors are fed into the fixed StyleGAN generator to obtain the output image representing the age transformation.

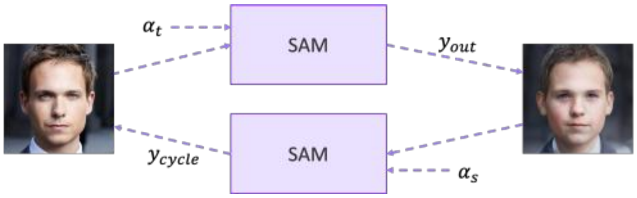


Figure 3: To address the challenge of the unsupervised setting, a cycle consistency pass is performed to recover the input at the source age α_s.

3.4. Losses

Our encoder is trained using a weighted combination of several loss objectives. The same set of losses introduced below is used both for the forward pass (Eq. 5) and cycle pass (Eq. 6). Specifically, the input image $\mathbf{x}$ and transformed image $\mathbf{y}_{out}$ are compared on the forward pass, while $\mathbf{x}$ is compared to the recovered image $\mathbf{y}_{cycle}$ in the cycle pass. For conciseness, we describe only the loss objectives for the forward pass.

First, we use the $\mathcal{L}_2$ loss to learn pixel-wise similarities and the LPIPS [75] loss to learn perceptual similarities:

$$\mathcal{L}_2\left(\mathbf{x}_{age}\right) = ||\mathbf{x} - SAM(\mathbf{x}_{age})||_2 \tag{7}$$

$$\mathcal{L}_{\text{LPIPS}}\left(\mathbf{x}_{age}\right) = ||F(\mathbf{x}) - F(SAM(\mathbf{x}_{age}))||_2, \tag{8}$$

where $F(\cdot)$ denotes the perceptual feature extractor. As humans age, their head shape naturally changes over time. Motivated by this, we apply a higher weight on the $\mathcal{L}_2$ and $\mathcal{L}_{\text{LPIPS}}$ losses in the center region of the image and a lower weight in the outer region.

We additionally adopt a regularization loss [15, 57] that encourages the encoder to output style vectors closer to the average latent vector. We find that this regularization improves image quality without harming the fidelity of the desired age transformation.

Identity Loss. The authors in [57] show the importance of the identity loss in obtaining accurate reconstructions of real facial images. As identity-preservation is a key challenge in modeling the age transformation process, we incorporate this identity loss into our training process by measuring the cosine similarity between the output image and its source image. While distinctive facial features are preserved as we age, our perceived identity may change over time (e.g., a person looks different at age 5 and age 55) [16, 51]. To capture this observation, we compute an identity loss weighted by the change in age such that a larger age change corresponds to a smaller weight. That is, we compute:

$$\mathcal{L}_{\text{ID}}\left(\mathbf{x}_{age}\right) = w(\Delta_{age})\cdot(1 - \langle R(\mathbf{x}), R(SAM(\mathbf{x}_{age}))\rangle) \tag{9}$$

where R is a pre-trained ArcFace [22] network. Estimating the age α_s of the input image using the pre-trained age

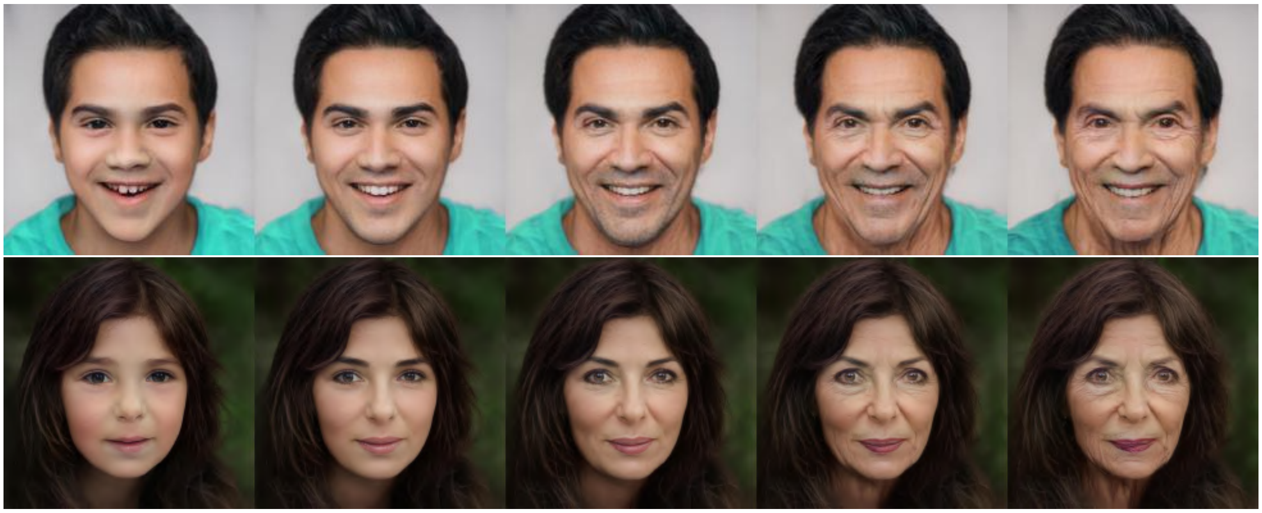


Figure 4: Age transformation results generated using SAM. Observe the state-of-the-art image quality achieved by leveraging a fixed, pretrained StyleGAN generator.

regression network, we define $\Delta_{age} = \frac{1}{100}|\alpha_s - \alpha_t|$. The weight function $w(\cdot)$ is then defined by,

$$w(\Delta_{age}) = 0.25 \cdot \cos(\pi \cdot \Delta_{age}) + 0.75. \tag{10}$$

Here, a weight of 1 corresponds to no change in age, while an increase in Δ_{age} corresponds to a monotonic decrease in $w(\cdot)$ with a minimum weight of 0.5. The motivation behind the weighted identity loss is further discussed in Section 4.1.

Aging Loss. To measure the accuracy of the age transformation, we use a pre-trained age predictor network [58, 59], denoted A. Given an input image $\mathbf{x}$ and desired target age α_t the aging loss is computed as the $\mathcal{L}_2$ loss between α_t and the age of the generated image,

$$\mathcal{L}_{\text{age}}\left(\mathbf{x}_{age}\right) = ||\alpha_t - A(SAM(\mathbf{x}_{age}))||_2. \tag{11}$$

In summary, the objective of the forward pass is given by:

$$\begin{aligned}\mathcal{L}_{for}(\mathbf{x}_{age}) = {} & \lambda_{l2}\mathcal{L}_2(\mathbf{x}_{age}) + \lambda_{lpips}\mathcal{L}_{\text{LPIPS}}(\mathbf{x}_{age}) + \\ & \lambda_{reg}\mathcal{L}_{\text{reg}}(\mathbf{x}_{age}) + \lambda_{id}\mathcal{L}_{\text{ID}}(\mathbf{x}_{age}) + \\ & \lambda_{age}\mathcal{L}_{\text{age}}(\mathbf{x}_{age})\end{aligned} \tag{12}$$

where λ_{l2}, λ_{lpips}, λ_{reg}, λ_{id}, λ_{age} are constants defining the loss weights. Combining both forward and cycle passes, the full objective is then given by,

$$\begin{aligned}\mathcal{L}(\mathbf{x}_{age}, \mathbf{y}_{out} \parallel \alpha_s) = {} & \mathcal{L}_{for}(\mathbf{x}_{age}) + \\ & \lambda_{cycle}\mathcal{L}_{cycle}(\mathbf{y}_{out} \parallel \alpha_s),\end{aligned} \tag{13}$$

where λ_{cycle} defines the weight of the cycle loss. Recall that in the cycle pass the recovered image, $SAM(\mathbf{y}_{out} \parallel \alpha_s)$, is compared to the original input image $\mathbf{x}$.

Note that during training, only the encoder network E is trained. The StyleGAN generator, facial recognition network, age regression network, and Pixel2Style2Pixel encoder for computing the $\mathcal{W}+$ latent codes of the input image all remain fixed. This simplifies the training process, as no discriminator needs to be trained, and allows for faster convergence. Additional implementation details and loss weights can be found in Appendix A.

4. Experiments

In this section, we perform extensive experimentation to explore the effectiveness of our method. In particular, we compare our approach to state-of-the-art age transformation methods and latent space methods. We then show the benefits of using our end-to-end approach for learning the latent path controlling age.

For training, we use the FFHQ [38] dataset with all evaluations performed on the CelebA-HQ [37] test set. A comparison to multi-domain image-to-image translation approaches as well as an ablation study of SAM can be found in Appendices B and C. An accompanying video demonstrating full lifespan results can be found in the project page.

Effectively evaluating an age transformation method is extremely challenging. For a given input image, we wish to generate an image that faithfully resembles the same individual. At the same time, however, it is necessary to perform meaningful changes to the input image that accurately reflect the complex human aging process. Therefore, before evaluating the various aging methods, we explore the effectiveness of state-of-the-art facial recognition networks in recognizing the same individual at different ages. In doing so, we show that measuring the identity preservation capability of aging methods should be done with great care.

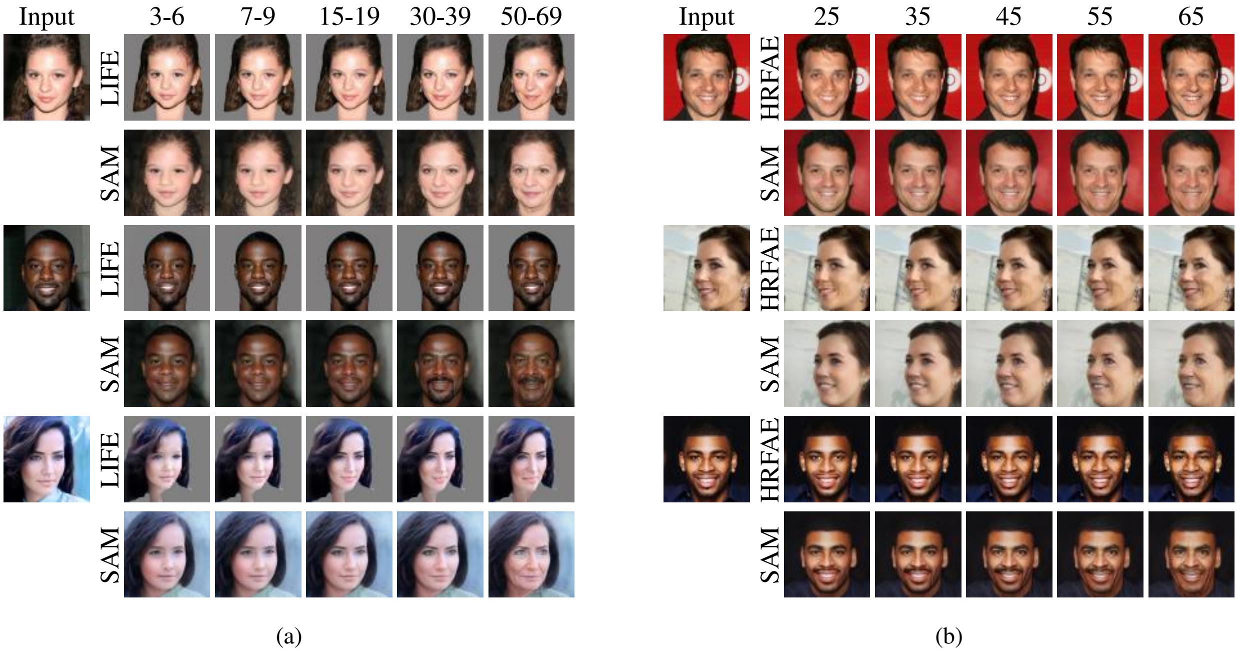


Figure 5: Qualitative comparison of age transformation results with (a) LIFE [54] and (b) HRFAE [72] on the CelebA-HQ [37] test set. For translating our images to the age groups in LIFE, we set the target age equal to the median age of each group. Additional results can be found in the Appendix. Best viewed zoomed-in.

4.1. Age Invariance of Facial Recognition Networks

Numerous experimental psychological studies [16, 35, 36, 51] have shown that although people typically recognize individuals they know across their lifetime, current facial recognition systems struggle in doing so. We perform two experiments to evaluate the age-invariance of the current state-of-the-art facial recognition network. First, we collect 50 pairs of individuals at different ages. We then use the ArcFace [22] recognition network to compute the cosine similarity of each pair of images. We obtain an average similarity score of 0.45 with a standard deviation of ± 0.10. Next, we collect multiple images of the same individual at different ages. Setting aside one of the images as a query image, we measure the cosine similarity of the remaining images to the query. As shown in Figure 6, as we move away from the age of the query image, the similarity monotonically decreases, indicating that identity, as measured by facial recognition systems, does not remain fixed with age.

We conclude that relying on facial recognition systems may be ineffective when faced with large variations in age. Following the above experiments and the conclusions from [16, 35, 51], we find that currently, the most effective approach for quantitatively measuring identity across one's lifespan is through human perceptual evaluation. Moreover, such an evaluation should be performed on well-known individuals as viewers are significantly less accurate in identifying unfamiliar faces [19, 35, 49, 50].

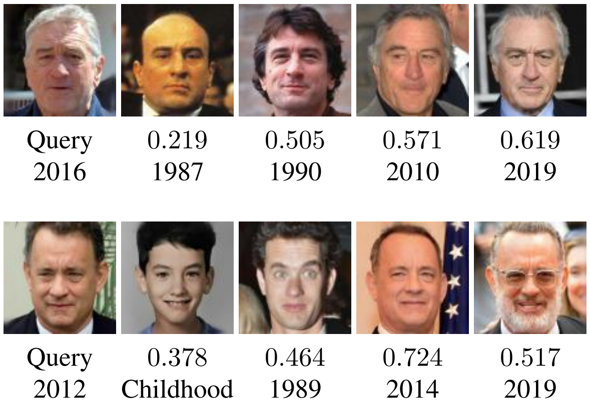


Figure 6: Identity similarity results using the ArcFace [22] recognition network. For each row, we compute the cosine similarity between the query and the remaining images. Image credits in order: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

In addition, motivated by these findings, we employ a weighted identity loss to model this observation (see Section 3.4).

4.2. Comparison with Age Transformation Methods

We begin our evaluation by comparing our proposed method to two state-of-the-art age transformation methods from Or-El *et al.* [54] and Yao *et al.* [72], which we refer to as *LIFE* and *HRFAE*, respectively.

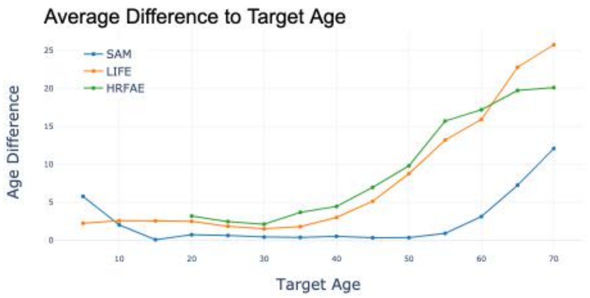


Figure 7: Here, we examine each method's ability to generate a full lifespan of images. Note that the lower the better.

Since LIFE is originally trained on a subset of the FFHQ dataset, we retrain LIFE as a single model using all $70,000$ images. We employ the same pre-processing steps and training procedure as used in the official implementation. As HRFAE is trained on the entire FFHQ dataset, we use their official implementation and pre-trained model for evaluation. Note that each method supports a different age range (0-70 for LIFE, 20-70 for HRFAE, $5-100$ for SAM). Therefore, when comparing two methods, we consider only ages supported by both.

Qualitative Evaluation. We provide a visual comparison with LIFE in Figure 5a. One can see that although LIFE is able to successfully capture the change in head shape across the different ages (see row 3), the resulting images contain unwanted artifacts. We observe the ability of SAM to more naturally alter facial features that change as we age, e.g., facial hair and wrinkles (see row 2). In Figure 5b we compare our method with HRFAE. While HRFAE is able to generate high-resolution images, they are limited to generating subtle changes in texture between the different ages. In contrast, SAM is able to achieve visually pleasing results while better modeling the change in head shape and texture. For example, observe the change in the jaw line of SAM's outputs.

We note that while SAM preserves various features (e.g., hair color) across the different target ages, this may not always be desirable, since often an individual's hair color change over time. To address this, we can apply simple editing techniques to generate multiple outputs for a single input image illustrating possible changes over time. These techniques are explored in Section 4.4.

Quantitative Evaluation. Since LIFE operates using predefined age groups, we are unable to compare with it by directly translating images to specific target ages. However, LIFE is able to interpolate between the different age groups to illustrate a continuous age progression. Although this interpolation can generate a full lifespan of images, the exact age of each image can not be guaranteed. Therefore, we choose to generate a full set of 80 images for each source image and select the generated image whose predicted age is closest to a desired target age. This process is repeated for multiple target ages. We note that although HRFAE and SAM generate images using a specified target age, for a fair comparison of the three methods, we use the same selection protocol for all three. Moreover, for each method we select images only within the method's supported age range.

Human Evaluation I

Target	5		30		65	
Method	Age	Quality	Age	Quality	Age	Quality
HRFAE	–	–	11.16	17.90	9.75	16.30
LIFE	47.9	26.15	5.14	11.60	4.10	13.80
SAM	**52.1**	**73.85**	**83.70**	**70.50**	**86.15**	**69.90**

Table 1: Each cell indicates the percent of respondents who preferred the corresponding method for the evaluated metric. When compared to LIFE [54] and HRFAE [72], SAM achieves superior aging accuracy and image quality across the three different target ages.

Human Evaluation II

Target	5	30	65
Method	ID Recall	ID Recall	ID Recall
HRFAE	–	89.7	87.6
LIFE	79.0	86.7	81.3
SAM	89.2	91.0	86.6

Table 2: We asked respondents to identity images of celebrities transformed to the different target ages using each of the three methods. Each cell indicates the percent of queries correctly identified at the corresponding target age.

Having performed the above selection process, we now evaluate each method's ability to accurately generate images for a wide-range of target ages. To do so, for each target age, we measure the average difference between the target age and the predicted ages of the images chosen by the above selection process. To ensure the predicted ages are independent of our loss function, we use the Microsoft Azure Face API for age estimation. All metrics are computed on $1,000$ randomly-drawn samples from the CelebA-HQ [37] test set.

Note that approximately 80% of the CelebA-HQ test set falls within the $20-40$ age range. We therefore expect the age difference to increase as we move away from this range. Figure 7 presents the comparison of the aging accuracy of the three methods. One may notice LIFE's ability to more accurately generate images with a target age of 5. We note this to be a limitation of SAM, which is discussed further in Section 5. Nevertheless, for the remaining target ages we find that SAM out-performs both LIFE and HRFAE. This

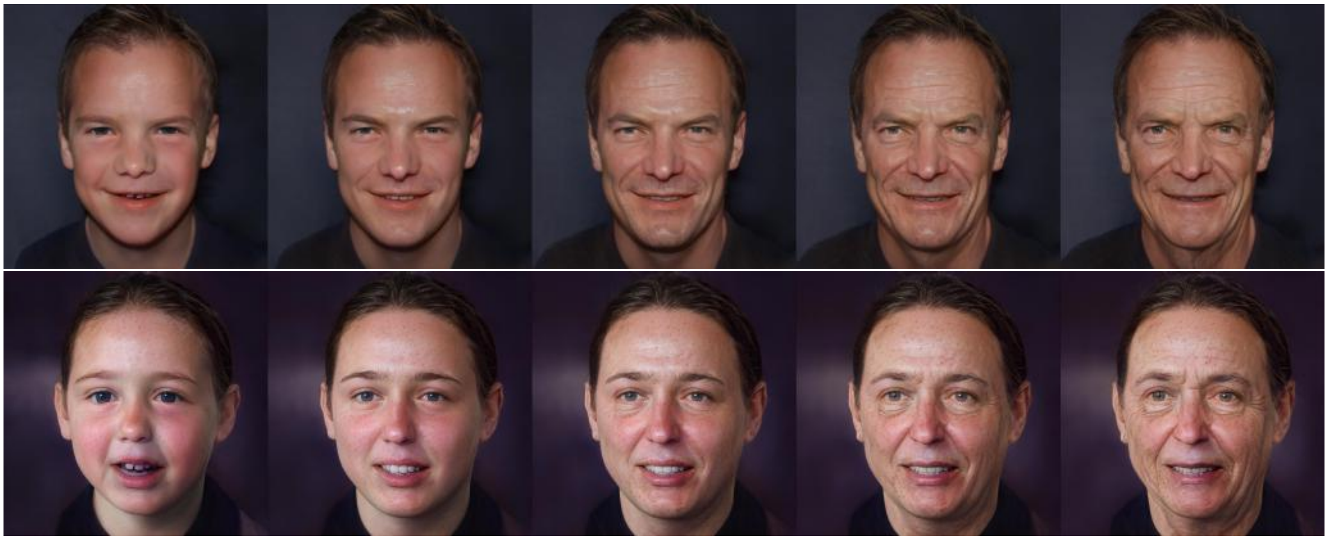


Figure 8: Additional age transformation results generated using SAM.

difference is most notable when translating to the older age range (60+). These results show the significantly wider range of ages supported by the regression-based approach of SAM when modeling full lifespan age progression.

Human Evaluation. Following the observations in Section 4.1, to more reliably quantify the performance of each method evaluated above, we also perform a human evaluation. For measuring the aging accuracy, we show the outputs of the three methods side-by-side and ask which method best portrays an individual at the desired target age. Similarly, to measure overall image quality, we ask the respondent to select the output that is most visually appealing. For a complete analysis, we repeat the above scenario three times and translate the test images to target ages of 5, 30, and 65. The results are shown in Table 1. For each of the three target ages, a total of 150 responses were recorded for each of the two aspects (300 responses for each target age). Note, all images were randomly selected for this study.

To evaluate whether the identity of the images generated by the three methods are well-preserved and recognizable, we take a different approach. Here, we collect images of well-known celebrities and perform age transformation to multiple target ages using the three methods. We then ask each respondent to identify the individual shown in the transformed image. To increase the difficulty, the question is asked as an open question (i.e. respondents are not given any choices to select from). In each cell of Table 2 we show the percent of queries that were correctly identified out of the 150 responses for the corresponding method and target age. Taking into account that some respondents simply do not know the queried individual, we find that the methods are similar in their ability to preserve the input identity.

4.3. Comparison with Latent Space Methods

With the recent emergence of strong image synthesis models such as StyleGAN, latent space traversal has become a popular approach for editing on real images. While these works have shown potential, they typically assume some prior, such as linearity, on the latent path controlling an attribute of interest. Previous works [20, 38, 60, 70] have demonstrated the disentanglement of StyleGAN's latent space. However, since this latent space is a manifold, a path that controls a specific attribute is not necessarily linear. It is also not guaranteed that such a linear path remains within the manifold. Such approaches may therefore fail to fully capture the disentanglement property of the latent space and may generate samples outside the true data distribution.

In contrast, we train an encoder to directly embed real images into the latent space of a pre-trained StyleGAN. As a result, one may view our method as an approach to *learn* a latent space path without explicitly needing to model it. Here, we show the advantages of using an end-to-end approach for learning the latent path without requiring any prior assumptions.

We compare our results with two state-of-the-art latent space approaches: InterFaceGAN [60] and StyleFlow [13]. InterFaceGAN assumes that the latent path controlling a given attribute (e.g., age) is linear. StyleFlow learns non-linear paths in the latent space by using normalizing flows conditioned on the attribute of interest. We show that compared to both of these methods that operate specifically in the latent path, our end-to-end approach obtains improved disentanglement and superior visual quality on real facial images.

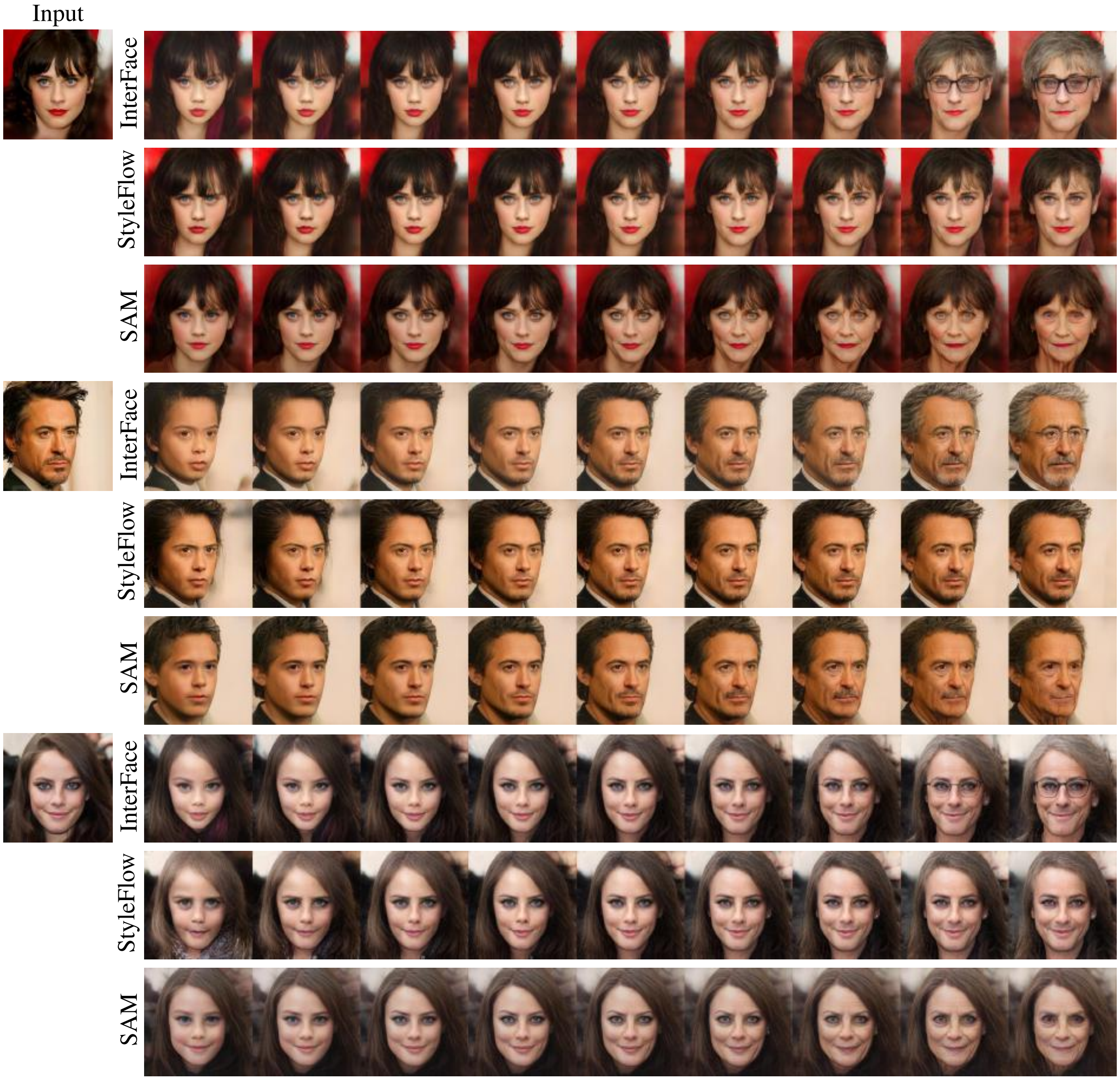
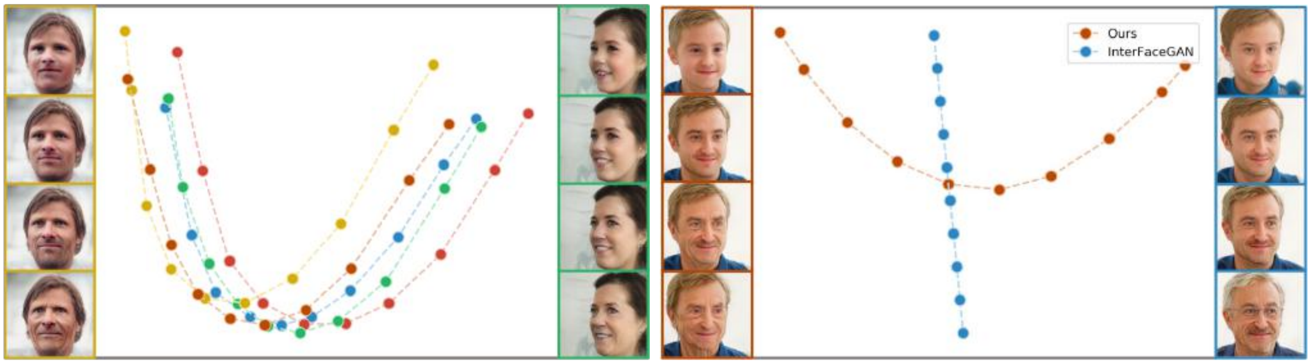


Figure 9: A visual comparison with InterFaceGAN [60] and StyleFlow [13] on real face images. Results for InterFaceGAN and StyleFlow are obtained by inverting the input into StyleGAN's latent space using pSp [57] and traversing along the learned age path.

Qualitative Evaluation. As InterFaceGAN is originally trained using StyleGAN1 generators, we retrain their method using the same StyleGAN2 generator used by our method and follow the same training procedure as described in their paper to obtain the linear aging direction. For StyleFlow, we use their official implementation for generating the transformed images. To edit a given *real* image, we first encode the image into the $\mathcal{W}+$ latent space using pSp [57], the state-of-the-art StyleGAN encoder, and then interpolate the obtained latent code using the latent direction of each method. A visual comparison is provided in Figure 9.

For completeness, we provide additional results for InterFaceGAN in the Appendix using their official implementation and age boundary. As their original paper and official implementation uses a StyleGAN1 [38] generator, we use IDInvert [77] for inverting the real input image.

Although InterFaceGAN is able to generate high-quality images, it can be seen that they fail to disentangle age and other facial attributes. For example, one can observe that eye glasses and hair color are heavily entangled with the progression in age. This may be explained by the bias of the training data as older people are more likely to wear



(a)

(b)

Figure 10: In (a) we project the learned latent paths of five different images using PCA. Illustrated on the sides are intermediate outputs of two of the five images obtained by traversing along the corresponding colored path. As can be seen, the non-linearity of the learned paths are better suited to the complex nature of StyleGAN's latent space manifold. In (b) we compare the linear nature of InterFaceGAN (shown in blue) with the non-linearity of SAM (shown in red).

glasses. This potential data bias, coupled with the linearity of their approach, results in poor lifelong age transformation results. Notice that while StyleFlow obtains better disentanglement of age compared to InterFaceGAN, they struggle to faithfully transform the input images into the older age range (60+) and obtain lower-quality results. In contrast, our approach achieves superior disentanglement between age and other facial attributes while faithfully preserving identity across various ages.

Exploring the Latent Path. To better understand the latent path learned by our approach, we use PCA to project the $\mathcal{W}+$ age-transformed latent codes obtained for various real face images. To emphasize the non-linearity of the learned paths, we compute the projection plane using the age-transformed latent codes of a single input image. As can be seen in Figure 10a, the paths of different images are similar, strengthening the claim that the latent space of StyleGAN is a well-behaved manifold that can be traversed to edit a particular attribute of interest (e.g. age) in a disentangled manner. Note that while these paths are similar, they are each different in nature, illustrating the non-trivial solution learned by SAM.

Although the StyleGAN latent space is well-behaved, we observe that a fully disentangled path is hard to achieve when trying to explicitly model it using some prior on the latent path. To show this, we project the age transformation latent paths learned by InterFaceGAN and SAM for the same real input face image. As can be seen in Figure 10b, the path learned by InterFaceGAN is indeed linear and results in a strong entanglement between age and other facial attributes. Also, notice that explicitly modeling the path learned by SAM is not trivial as the path's behavior changes at different parts of the age progression.

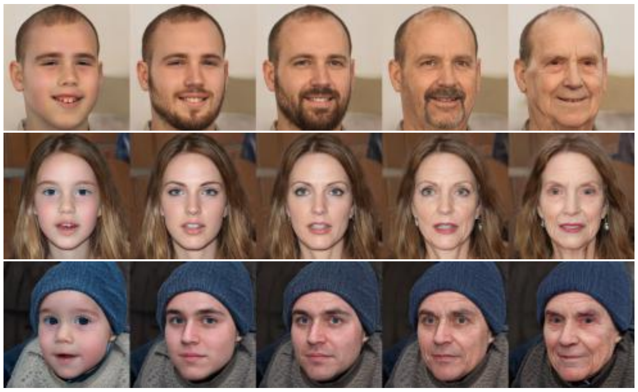


Figure 11: Images generated from randomly sampled w vectors that are traversed along the age manifold direction learned by SAM.

To validate the learned manifold direction that controls age, we randomly sample an image from the $\mathcal{W}$ latent space of StyleGAN and perform age transformation by traversing along a path lying on the learned manifold. We illustrate several examples in Figure 11. As can be seen, we get a smooth and disentangled age progression, which shows the generalization of our method in learning a latent path corresponding to an age shift. This, however, may raise a question as to the benefits of our end-to-end approach. Although it is possible to operate solely in the latent space and apply a "general" learned path to a given image, the end-to-end nature of SAM allows one to achieve more fine-grained control over the resulting age. For example, with SAM one can directly specify the desired target age, which cannot be done when operating solely in the latent space.

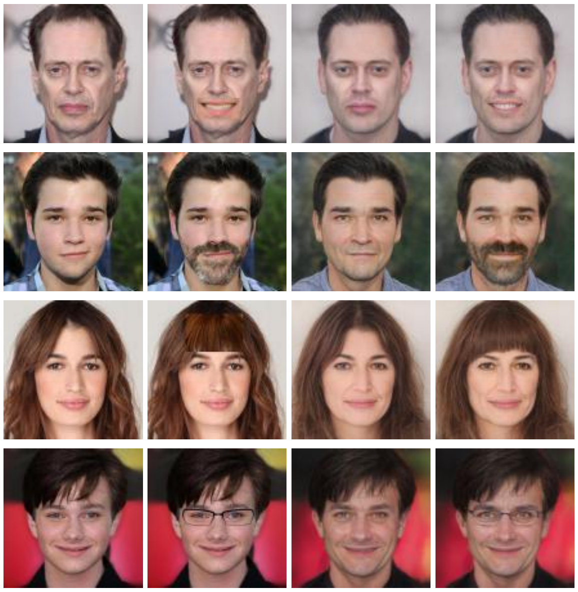


Figure 12: SAM with patch editing allows for more fine-grained control over specific features such as expression, facial hair, hair style, and glasses.

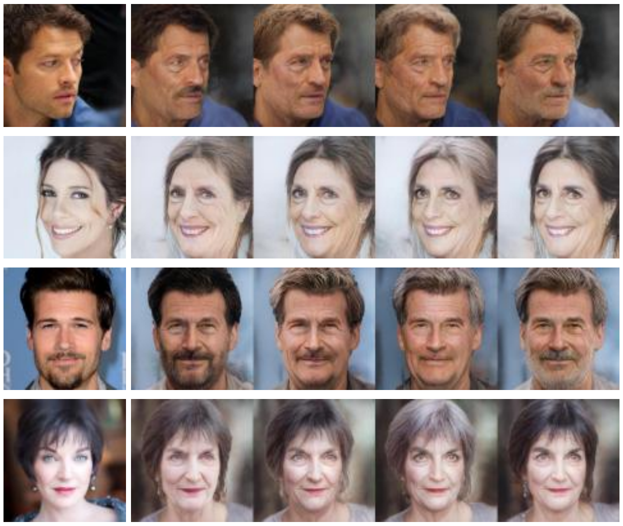


Figure 13: Performing style-mixing on the age-transformed outputs of SAM allows additional editing control over features such as hair color and facial hair. Here, for each input image we perform style-mixing with four references images on layers $8-9$ to obtain multi-modal results.

4.4. Additional Editing

As demonstrated in the previous section, our method is able to successfully disentangle age from other facial attributes. That is, all attributes except for age are faithfully preserved from the input image. As we age, however, attributes such as hair style and hair color naturally change. It is therefore desirable to be able to control such attributes when modeling age progression. Here, we show that by using simple editing techniques, our method provides additional control over the resulting image, while still capturing the desired age.

Patch Editing. Patch editing provides a simple and intuitive editing approach where one can "paste" a particular attribute such as glasses, facial hair, and bangs onto the input source image. The edited image is then passed through SAM to obtain the age-transformed edited image. As shown in Figure 12, SAM is able to seamlessly fuse the patch into the source image, while preserving the aging accuracy of our approach. The ability of SAM to successfully perform this task lies in the end-to-end nature of our approach. In particular, given some source image, edited or not, our encoder is tasked with encoding the input into a latent code of a realistic face image. Therefore, even when the patch is imprecisely pasted, SAM is able to merge the crop into its surrounding context while transforming the input image to the desired age.

Style Mixing. While patch editing allows for fine-grained control over specific attributes, it is typically limited to local edits of the input image. For example, controlling lighting and hair color is harder to achieve through patch editing due to the more global nature of the required edit. Since our approach maps a given input image into StyleGAN's $\mathcal{W}+$ latent space, we can leverage the inherent editing capabilities it offers. As shown in [38], the fine input styles mostly control the lighting and color of the generated image. Motivated by this, after transforming a given source image to the desired target age, we can additionally perform style-mixing on layers $8-9$ with a given reference image. Multi-modal synthesis is then naturally supported by performing style-mixing on multiple reference images. As shown in Figure 13, doing so enables fine-grained control in generating multiple plausible age transformation results.

5. Limitations

Although our suggested approach is effective in representing the age transformation process, there are several limitations that should be considered. Although using a fixed, pre-trained StyleGAN generator simplifies the training process and allows for generating high-quality images, doing so may make it challenging to effectively model extreme poses, challenging expressions, and accessories. Furthermore, as our results are governed by the style represen-

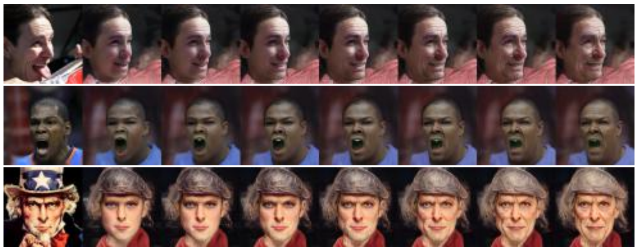


Figure 14: Limitations of SAM. Our approach may struggle when faced with extreme poses or expressions not seen during the training of StyleGAN.

tation, our method is limited to images that can be accurately embedded into StyleGAN's latent space. Thus, modeling faces that reside outside the StyleGAN domain, such as that of Uncle Sam, may be challenging. Another important assumption of our method is the existence of an age predictor that is able to generalize well to all age groups. As most age-annotated datasets are heavily biased toward images of adults, achieving high age prediction accuracy for young children may be challenging. As a result, our method may struggle to generate images of children under the age of 5. This limitation may be resolved by using a more accurate age predictor for computing the aging loss. We present several challenging cases in Figure 14.

6. Conclusion

This work presented a novel approach for modeling the age transformation task using a single input facial image. Our end-to-end approach maps a given input image and desired target age to the latent space of a fixed StyleGAN generator. To model the aging process, we employ an age predictor network that guides the encoder in generating latent codes corresponding to the desired transformation, resulting in a *learned* latent path representing age progression. We showed that treating age progression as a regression problem via the age predictor provides more fine-grained control over the output age. Further, unlike previous latent space approaches that control age using a prior assumption on the latent path, our approach encourages the model to learn a *non-linear* path more suitable for disentangling age from other facial attributes. We then demonstrated how this improved disentanglement allows for further fine-grained editing on the generated aging results. We believe that the key insights provided in this work can be extended for additional editing applications. We leave this exploration as a potential direction for future work.

References

[1] Image taken by Georges Biard and can be found here. License: Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0). 6

[2] Image can be found here by Jon Rubin. License: Attribution 2.0 Generic (CC BY 2.0). 6

[3] Image taken by Gorup de Besanez and can be found here. License: Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). 6

[4] Image taken by David Shankbone and can be found here. License: Attribution 3.0 Unported (CC BY 3.0). 6

[5] Image taken by Jaqueline de Souza and can be found here. License: Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). 6

[6] Image taken by Angela George and can be found here. License: Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0). 6

[7] Image taken by Korush and Millie and can be found here. License: Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). 6

[8] Image taken by Alan Light and can be found here. License: Attribution 2.0 Generic (CC BY 2.0). 6

[9] Image taken from here. Licensed under the Public Domain as a work of the U.S. federal government. 6

[10] Image taken by John Bauld and can be found here. License: Attribution 2.0 Generic (CC BY 2.0). 6

[11] R. Abdal, Y. Qin, and P. Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019. 3

[12] R. Abdal, Y. Qin, and P. Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020. 3

[13] R. Abdal, P. Zhu, N. Mitra, and P. Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows, 2020. 2, 3, 8, 9, 19

[14] G. Antipov, M. Baccouche, and J.-L. Dugelay. Face aging with conditional generative adversarial networks, 2017. 1

[15] Baylies. stylegan-encoder, 2019. Accessed: January 2021. 4

[16] A. M. Burton, R. S. Kramer, K. L. Ritchie, and R. Jenkins. Identity From Variation: Representations of Faces Derived From Multiple Instances. *Cogn Sci*, 40(1):202–223, Jan 2016. 4, 6

[17] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 2

[18] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 2, 15

[19] R. Clutterbuck and R. A. Johnston. Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, 31(8):985–994, 2002. PMID: 12269591. 6

[20] E. Collins, R. Bala, B. Price, and S. Süsstrunk. Editing in style: Uncovering the local semantics of gans, 2020. 8

[21] A. Creswell and A. A. Bharath. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7):1967–1974, 2018. 3

[22] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 2, 4, 6, 15

[23] E. Denton, B. Hutchinson, M. Mitchell, and T. Gebru. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439*, 2019. 3

[24] C. N. Duong, K. Luu, K. G. Quach, and T. D. Bui. Longitudinal face aging in the wild - recent deep learning approaches, 2018. 2

[25] H. Fang, W. Deng, Y. Zhong, and J. Hu. Triple-gan: Progressive face aging with triple translation loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 804–805, 2020. 2

[26] Y. Fu, G. Guo, and T. Huang. Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 32:1955–76, 11 2010. 2

[27] M. Georgopoulos, J. Oldfield, M. A. Nicolaou, Y. Panagakis, and M. Pantic. Enhancing facial data diversity with style-based face aging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[28] L. Goetschalckx, A. Andonian, A. Oliva, and P. Isola. Ganalyze: Toward visual definitions of cognitive image properties, 2019. 3

[29] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. 2, 3

[30] Z. He, M. Kan, S. Shan, and X. Chen. S2gan: Share aging factors across ages and share aging trends among individuals. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9439–9448, 2019. 1

[31] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Attgan: Facial attribute editing by only changing what you want, 2018. 2

[32] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization, 2017. 15

[33] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2

[34] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks, 2018. 2

[35] R. Jenkins, D. White, X. Van Montfort, and A. M. Burton. Variability in photos of the same face. *Cognition*, 121(3):313–323, 2011. 6

[36] R. A. Johnston, M. Kanazawa, T. Kato, and M. Oda. Exploring the structure of multidimensional face-space: The effects of age and gender. *Visual Cognition*, 4(1):39–57, 1997. 6

[37] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5, 6, 7, 17, 18

[38] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 3, 4, 5, 8, 9, 11, 15

[39] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2, 3, 4

[40] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato. Fader networks: Manipulating images by sliding attributes, 2018. 2

[41] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. K. Singh, and M.-H. Yang. Drit++: Diverse image-to-image translation viadisentangled representations. *International Journal of Computer Vision*, pages 1–16, 2020. 2

[42] P. Li, H. Huang, Y. Hu, X. Wu, R. He, and Z. Sun. Uva: A universal variational framework for continuous age analysis, 2019. 2

[43] P. Li, H. Huang, Y. Hu, X. Wu, R. He, and Z. Sun. Hierarchical face aging through disentangled latent characteristics. In *European Conference on Computer Vision*, pages 86–101. Springer, 2020. 2

[44] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[45] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019. 15

[46] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 700–708. Curran Associates, Inc., 2017. 2, 3

[47] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz. Few-shot unsupervised image-to-image translation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 15

[48] Y. Liu, Q. Li, Z. Sun, and T. Tan. Style intervention: How to achieve spatial disentanglement with style-based generators?, 2020. 2

[49] A. M. Megreya and A. M. Burton. Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, 34(4):865–876, Jun 2006. 6

[50] A. M. Megreya and A. M. Burton. Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied, 14(4)*, page 364–372, 2008. 6

[51] M. Mileva, A. W. Young, R. Jenkins, and A. M. Burton. Facial identity across the lifespan. *Cognitive psychology*, 116:101260, 2020. 4, 6

[52] M. Mirza and S. Osindero. Conditional generative adversarial nets, 2014. 2

[53] Y. Nitzan, A. Bermano, Y. Li, and D. Cohen-Or. Face identity disentanglement via latent space mapping. *ACM Trans. Graph.*, 39(6), Nov. 2020. 3

[54] R. Or-El, S. Sengupta, O. Fried, E. Shechtman, and I. Kemelmacher-Shlizerman. Lifespan age transformation synthesis, 2020. 1, 2, 6, 7, 15, 17

[55] S. Pidhorskyi, D. A. Adjeroh, and G. Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14104–14113, 2020. 3

[56] N. Ramanathan, R. Chellappa, and S. Biswas. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages & Computing*, 20(3):131–144, 2009. 2

[57] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation, 2020. 2, 3, 4, 9, 15, 16, 19

[58] R. Rothe, R. Timofte, and L. V. Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015. 5, 15

[59] R. Rothe, R. Timofte, and L. V. Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018. 5, 15

[60] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 2, 3, 8, 9, 19

[61] Y. Shen and B. Zhou. Closed-form factorization of latent semantics in gans. *arXiv preprint arXiv:2007.06600*, 2020. 3

[62] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 2

[63] H. Tang, H. Liu, D. Xu, P. H. S. Torr, and N. Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks, 2020. 3

[64] X. Tang, Z. Wang, W. Luo, and S. Gao. Face aging with identity-preserved conditional generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7939–7947, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society. 1

[65] Y. Viazovetskyi, V. Ivashkin, and E. Kashin. Stylegan2 distillation for feed-forward image manipulation. *arXiv preprint arXiv:2003.03581*, 2020. 2

[66] A. Voynov and A. Babenko. Unsupervised discovery of interpretable directions in the gan latent space. *arXiv preprint arXiv:2002.03754*, 2020. 3

[67] W. Wang, Z. Cui, Y. Yan, J. Feng, S. Yan, X. Shu, and N. Sebe. Recurrent face aging. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2378–2386, 2016. 1

[68] Z. Wu, D. Lischinski, and E. Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation, 2020. 2, 3

[69] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang. Gan inversion: A survey, 2021. 3

[70] C. Yang, Y. Shen, and B. Zhou. Semantic hierarchy emerges in deep generative representations for scene synthesis, 2020. 8

[71] H. Yang, D. Huang, Y. Wang, and A. K. Jain. Learning face age progression: A pyramid architecture of gans, 2019. 1

[72] X. Yao, G. Puy, A. Newson, Y. Gousseau, and P. Hellier. High resolution face age editing. *CoRR*, abs/2005.04410, 2020. 2, 6, 7, 15, 18

[73] Z. Yi, H. Zhang, P. Tan, and M. Gong. Dualgan: Unsupervised dual learning for image-to-image translation. pages 2868–2876, 10 2017. 2, 3

[74] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, pages 9597–9608, 2019. 15

[75] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4

[76] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder, 2017. 1

[77] J. Zhu, Y. Shen, D. Zhao, and B. Zhou. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049*, 2020. 3, 9, 19

[78] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 2, 3

[79] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017. 2

Appendix

A. Implementation Details

Architectures. As in [57], we use a pre-trained ResNet-IR architecture from [22] for our encoder backbone. We use a *fixed* StyleGAN2 generator pre-trained on the FFHQ [38] dataset. Identity embeddings are extracted using a pre-trained ArcFace network from [22]. As in [72], we use the age classifier from [58, 59] pre-trained on the IMDB-WIKI dataset. As the IMDB-WIKI dataset has very few images below the age of 10 or above the age of 60, we fine-tune the age classifier using the FFHQ-Aging dataset [54].

Training. Training is performed on the FFHQ [38] dataset. For each batch, we randomly set the target ages equal to the estimated source ages with probability 0.33, focusing the network on reconstructing the input images. For training the encoder, we use the Ranger optimizer, a combination of Rectified Adam [45] with Lookahead [74], with a constant learning rate of 0.001. The input image resolution is 256×256. Only horizontal flips are used for augmentations. All experiments are performed using a single NVIDIA Tesla P40 GPU with a batch size of 6.

Losses. Before computing the loss functions, the generated 1024×1024 images are resized to 256×256. For $\mathcal{L}_{\text{ID}}$, the images are cropped around the face region and resized to 112×112 before being fed into the facial recognition network. Similarly, for $\mathcal{L}_{\text{age}}$, the images are resized to 224×224 before being fed into the age classifier. Further, we set the loss lambdas as follows: λ_{l2} and λ_{lpips} are set to $\lambda_{l2} = 1$ and $\lambda_{lpips} = 0.6$ in the center region of the image and $\lambda_{l2} = 0.25$ and $\lambda_{lpips} = 0.1$ in the outer region. We additionally set $\lambda_{reg} = 0.005$, $\lambda_{id} = 0.1$, and $\lambda_{age} = 5$. Finally, the cycle loss coefficient is set to $\lambda_{cycle} = 1$.

B. Multi-Domain Translation Methods

Here, we compare our method with two state-of-the-art multi-domain image-to-image methods: FUNIT [47] and StarGANv2 [18]. These methods work by translating a source image based on a reference image taken from the target domain. In the case of aging, the reference image should guide the network in translating only the age of the source to resemble an image from the target domain. We retrain each of the alternative methods using the same age domains used in LIFE [54] with the default training settings.

Qualitative Evaluation. We qualitatively compare the three methods in Figure 15. For performing inference on FUNIT and StarGANv2, we randomly sample a reference image from the target age group and translate the source image using the selected reference image. When performing inference using our approach, we set the target age equal to the median age of each group. As can be observed, in both alternative methods the reference image has a large effect on the translated image and alters attributes besides the age. For example, one can see that the texture of the translated images is mostly taken from the reference image rather than from the source, resulting in unrealistic changes in skin-tone. As a result, since a different reference image is used for each age group, the generated images may vary significantly. This is undesirable when modeling a continuous process such as human aging. This behavior may be explained by the design of these two methods. Both are built upon the AdaIN [32] layer, which mostly alters the texture of the source image based on the chosen reference image. Although this is desirable in tasks such as style transfer, the opposite is desired for age transformation (where the texture, pose, lighting, etc. should be taken from the source).

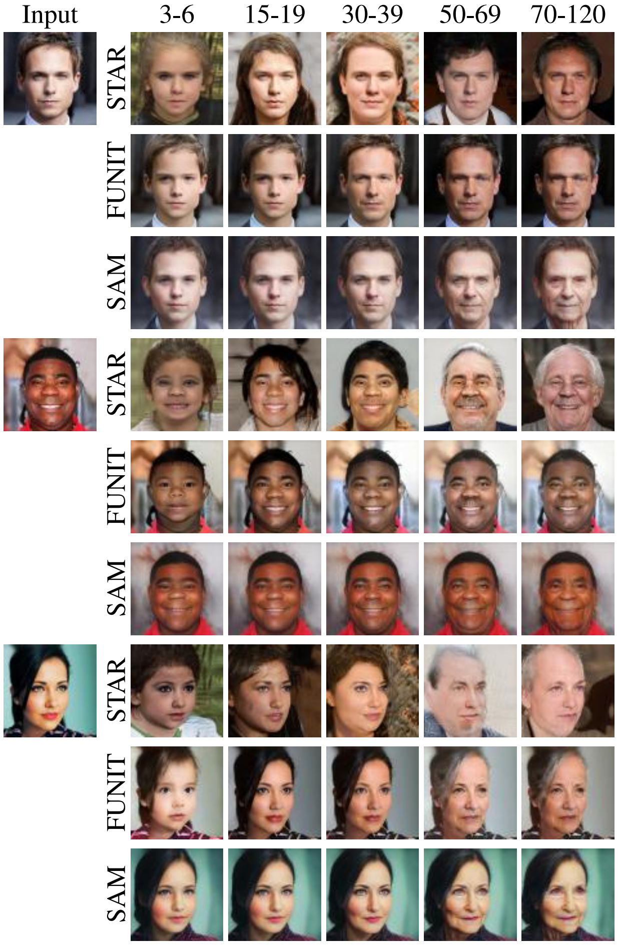


Figure 15: Qualitative Comparison with FUNIT [47] and StarGANv2 [18] (STAR). Note that FUNIT and StarGANv2 translate each source image using the same reference image. For translating images using SAM we set the target age equal to the the median age of each group.

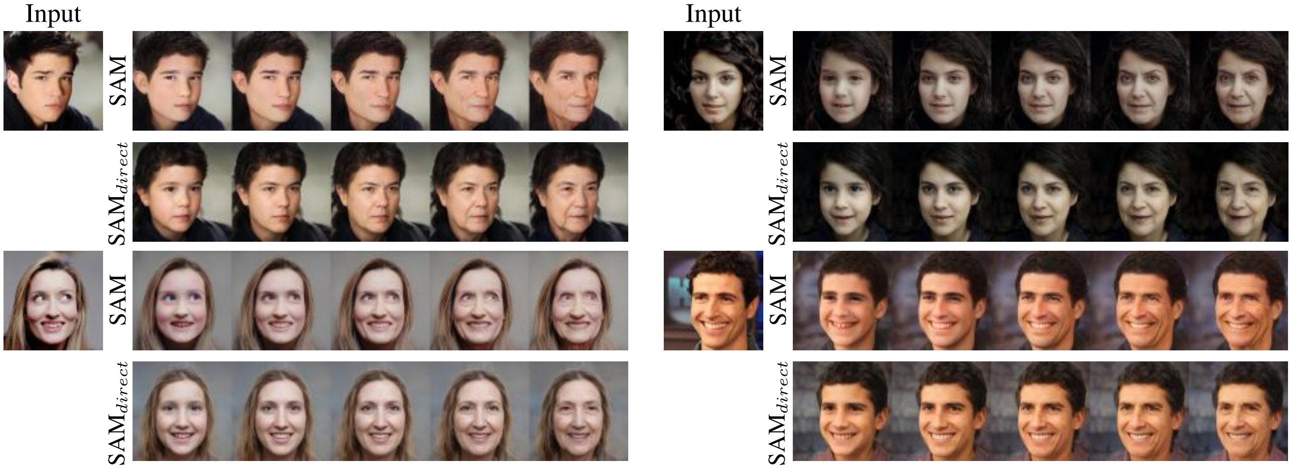


Figure 16: Qualitative comparison between our SAM method and the SAM$_{direct}$ variant. As can be seen, our SAM method is able to more accurately preserve identity and key facial features while faithfully modeling lifelong age transformation.

In contrast to these works, our method is able to better preserve the input identity and is able to better handle non-frontal images. Overall, our results have a much higher visual quality with substantially fewer artifacts.

C. Ablation Study

In this section we perform an ablation study on the training formulation of SAM. Recall that the output of our model is defined as

$$SAM(\mathbf{x}_{age}) := G(E(\mathbf{x}_{age}) + \mathbf{w}^*)). \tag{14}$$

Here, $\mathbf{x}_{age} = \mathbf{x} \parallel \alpha_t$ is a 4-channel input defining the desired age transformation on input image $\mathbf{x}$ and $\mathbf{w}^*$ is the inversion of $\mathbf{x}$ obtained using a fixed, pre-trained Pixel2Style2Pixel [57] encoder. That is,

$$\mathbf{w}^* := pSp(\mathbf{x}) \in \mathbb{R}^{18\times 512}. \tag{15}$$

Observe that in this formulation, we train our SAM encoder to learn the *residual* vector between the latent code of the original input image and the latent code of the age-transformed image. In a sense, SAM is tasked with learning a *shift* in the latent space with respect to $\mathbf{w}^*$.

Another possible approach for modeling age transformation is to directly learn the age-transformed image without relying on the inversion of the input image. That is, one can directly output

$$SAM_{direct}(\mathbf{x}_{age}) := G(E(\mathbf{x}_{age}))). \tag{16}$$

Note that for simplicity, we denote the two variants by SAM and SAM$_{direct}$, respectively.

Here, we show that learning an "age shift" in the latent space results in significant improvement in the identity preservation of the input image across the various target ages without harming the aging accuracy of our method.

Ablation Study: Average Age Difference

Target Age	5	20	35	50	65
SAM	**5.79**	0.74	0.39	**0.36**	7.26
SAM$_{direct}$	6.23	**0.69**	**0.37**	0.41	**6.83**

Table 3: Quantitative evaluation of the two SAM variants. As can be seen, our residual-based SAM approach compares favorably with its counterpart SAM$_{direct}$.

In Figure 16 we provide a visual comparison of lifelong aging results generated using both of the approaches. Observe SAM's ability to better capture non-frontal poses, complex hair styles, and facial expressions. In particular, notice SAM's ability to more accurately retain the eye gaze direction of the bottom left input image and model the complex hair of the top right image.

To verify that our residual-based learning approach does not harm the aging accuracy of the transformations we repeat the quantitative evaluation performed in Section 4.2 and compare the aging accuracy of SAM and SAM$_{direct}$ with respect to five target ages: 5,20,35,50, and 65. For each target age, we compute the average difference between the desired age and the predicted age of each generated image. The full quantitative results are summarized in Table 3. As can be seen, our residual-based approach does not restrict the network's ability to generate images across a wide range of ages.

D. Additional Results

We provide additional visual comparisons and results generated by SAM in the following Figures. It is recommended to zoom-in when viewing all results.

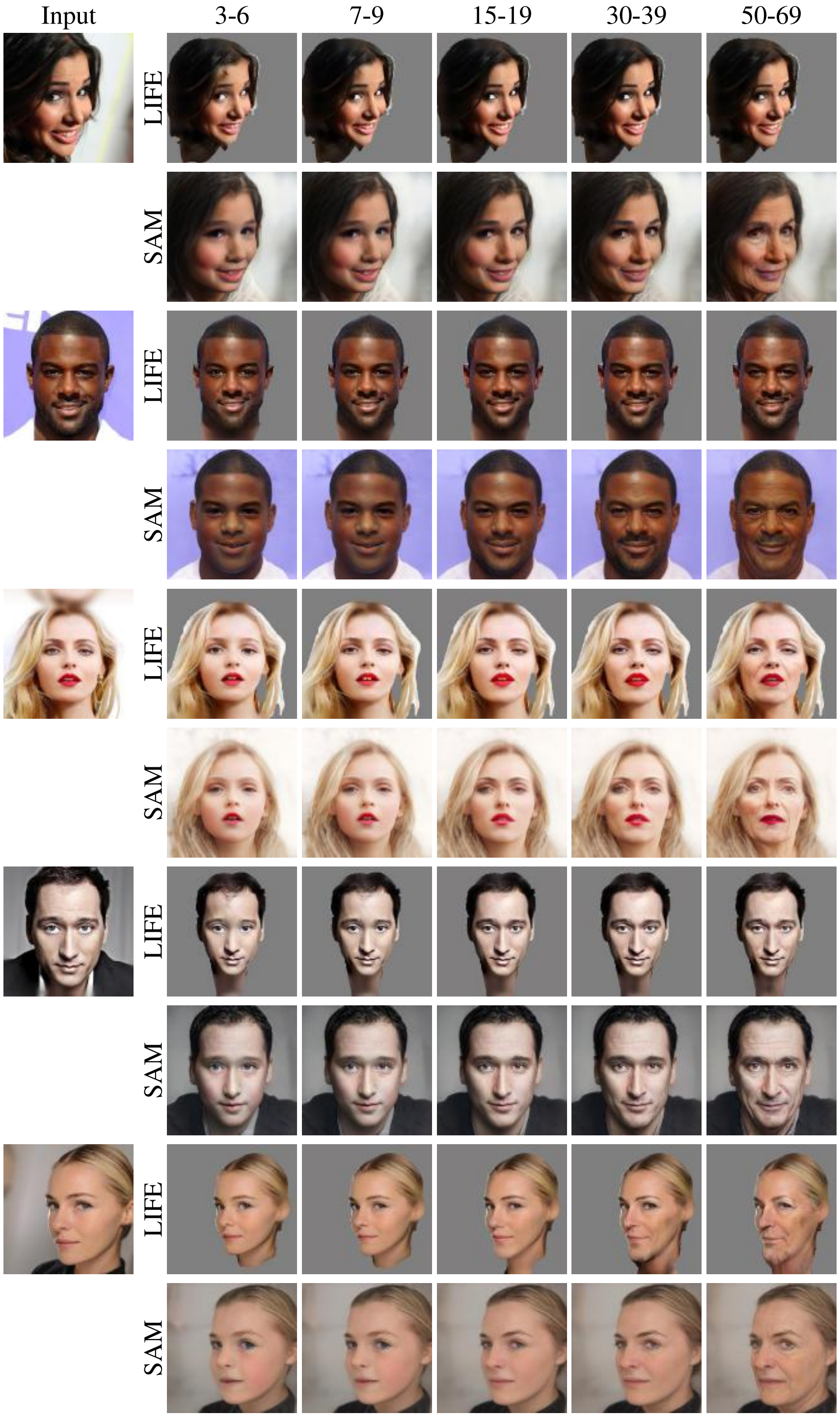


Figure 17: Additional qualitative comparisons of age transformation results with LIFE [54] on the CelebA-HQ [37] test set. For translating our images to the age groups in [54], we set the target age equal to the the median age of each group.

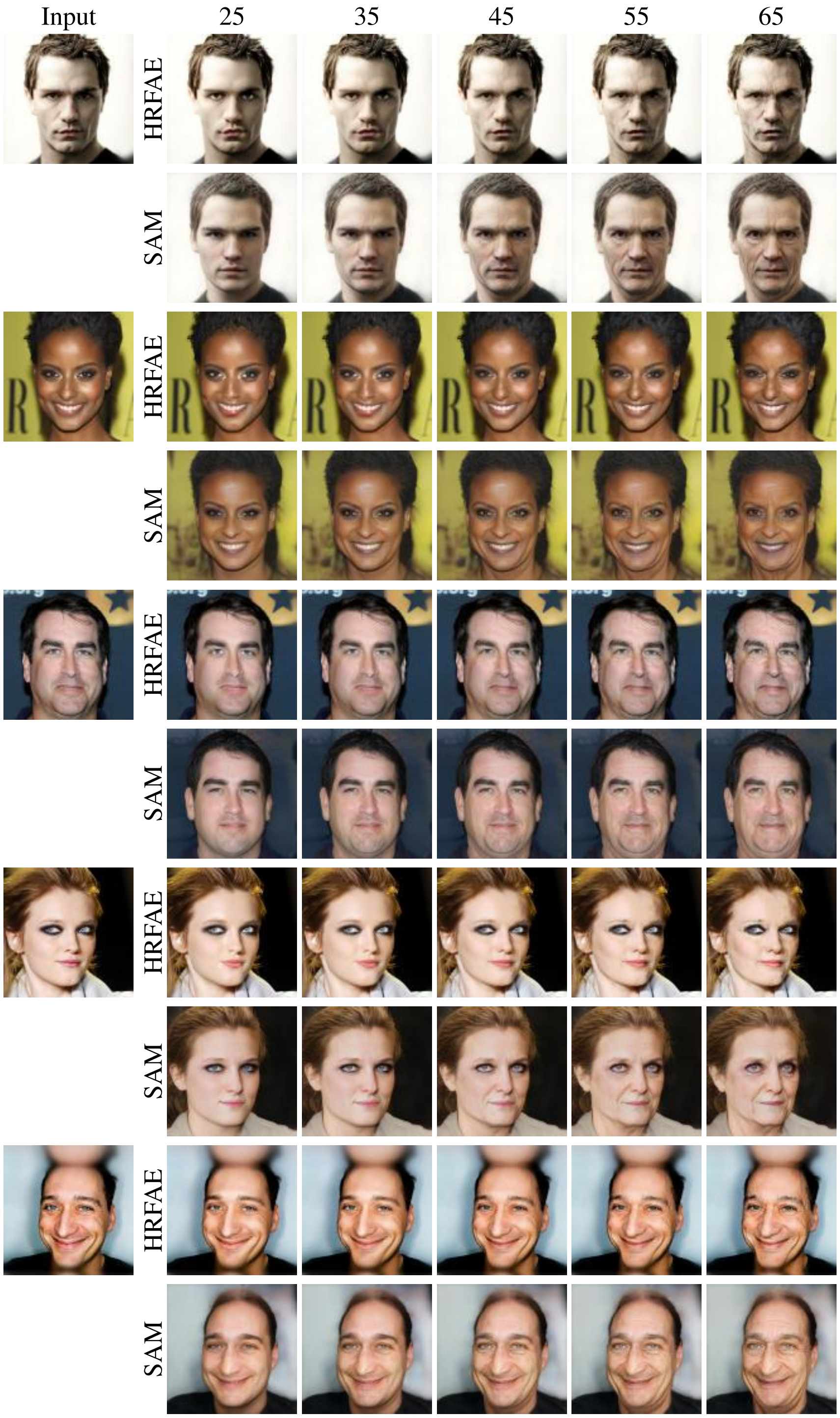


Figure 18: Additional qualitative comparisons of age transformation results with HRFAE [72] on the CelebA-HQ [37] test set.

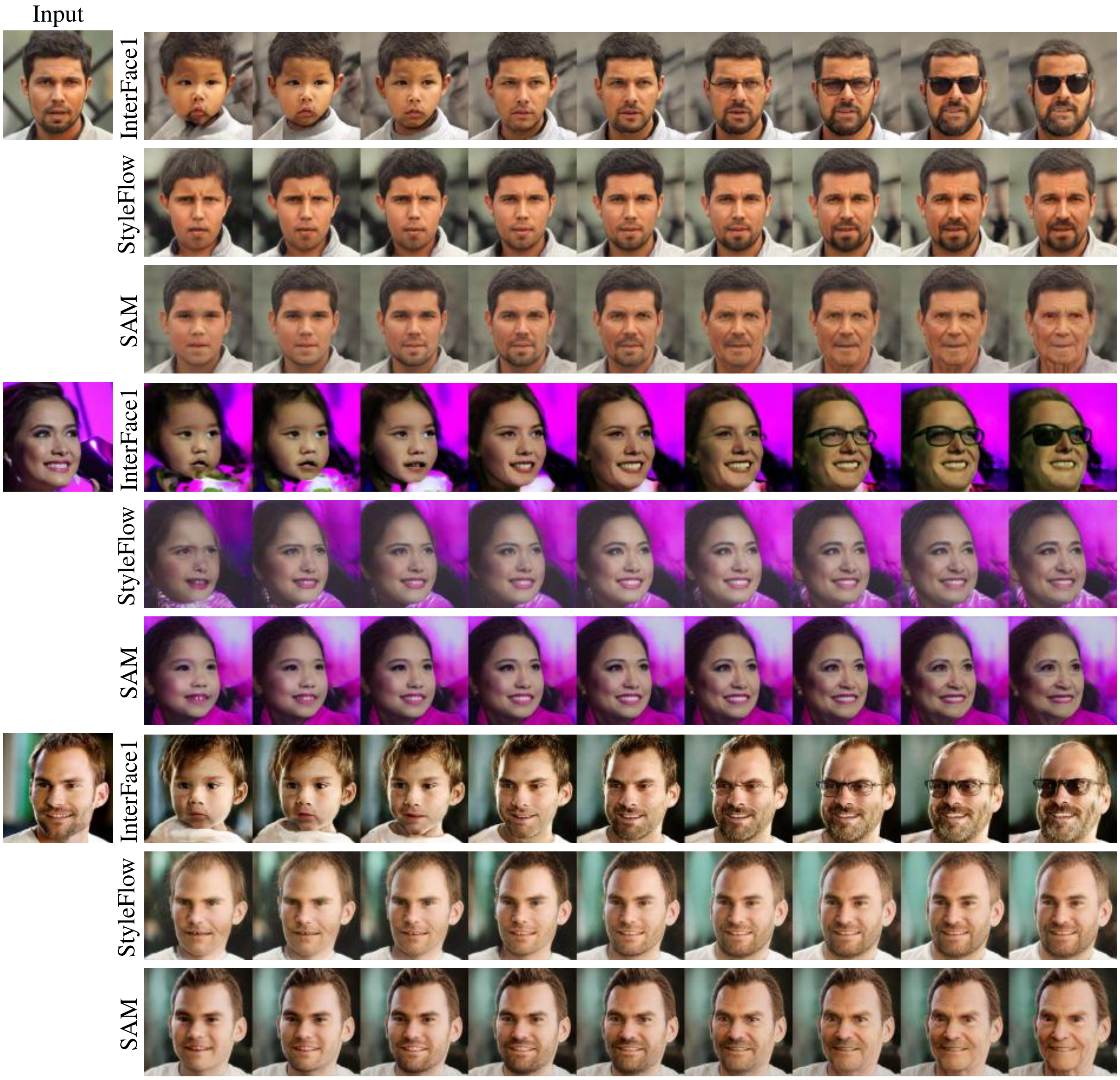


Figure 19: Additional visual comparisons with InterFaceGAN [60] (InterFace1) and StyleFlow [13] on real face images. Results for InterFaceGAN are obtained using IDInvert [77] for inversion into the StyleGAN1 latent space. StyleFlow is performed on latents obtained by inverting the input using a pSp [57] encoder into the StyleGAN2 latent space.

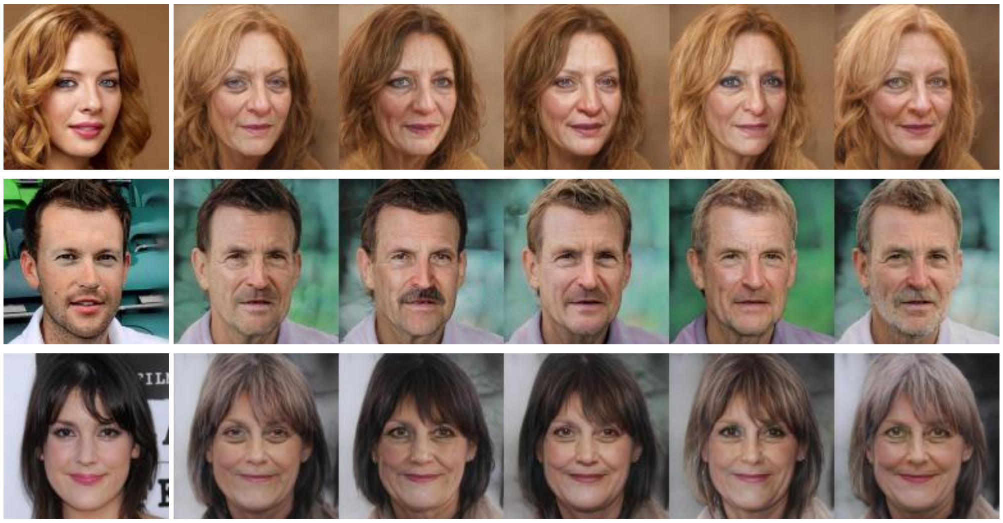


Figure 20: Additional multi-modal results generated by SAM by performing style-mixing on the age-transformed outputs. Here, for each input image we perform style-mixing with five references images on layers $8-9$ to obtain multiple transformation results.

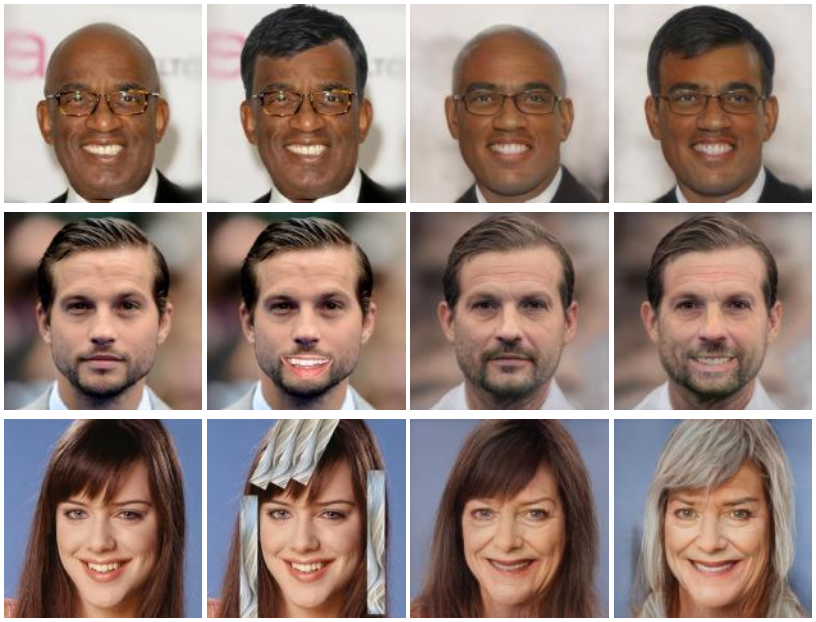


Figure 21: Additional patch editing results generated with SAM.

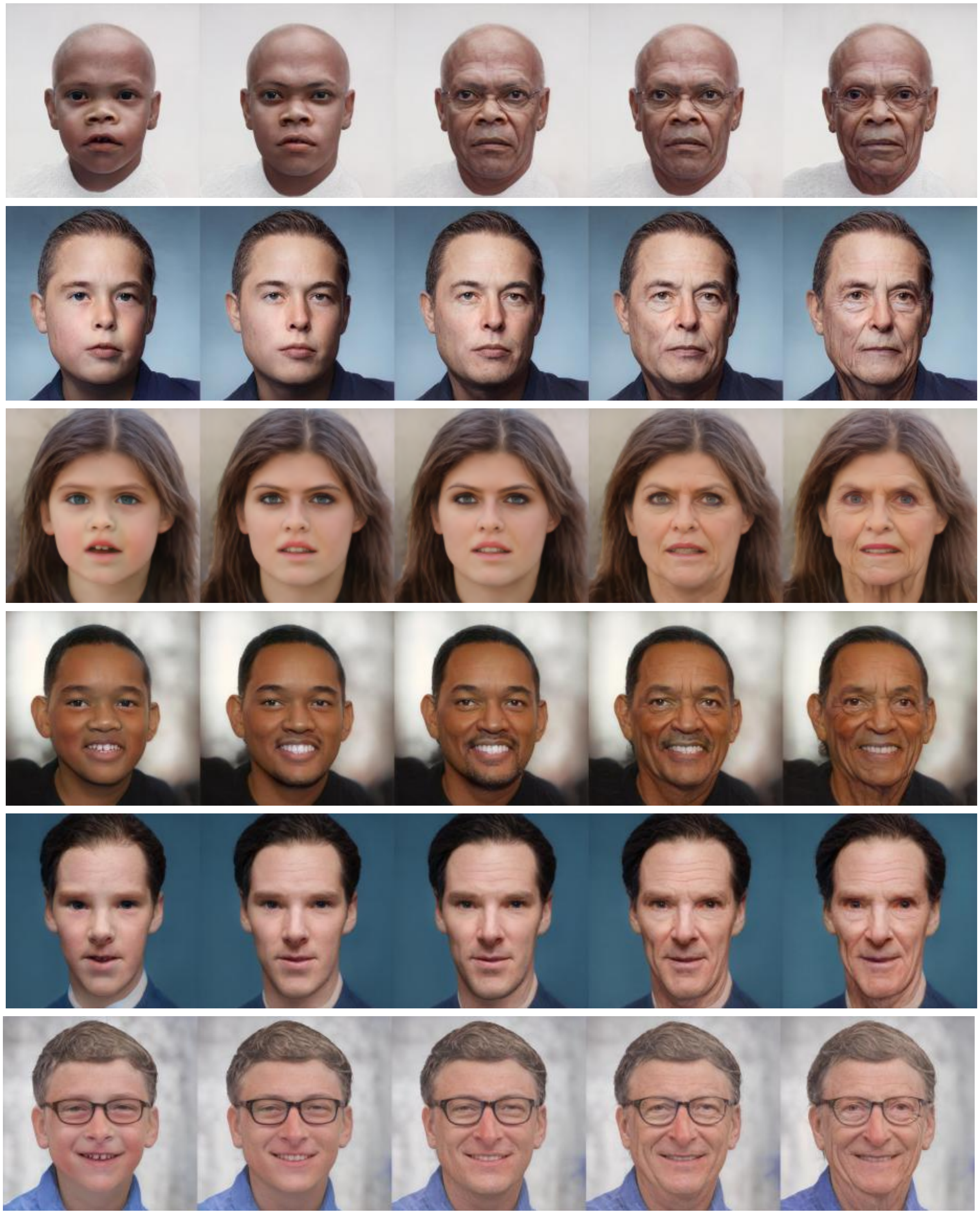


Figure 22: Additional age transformation results generated using SAM.

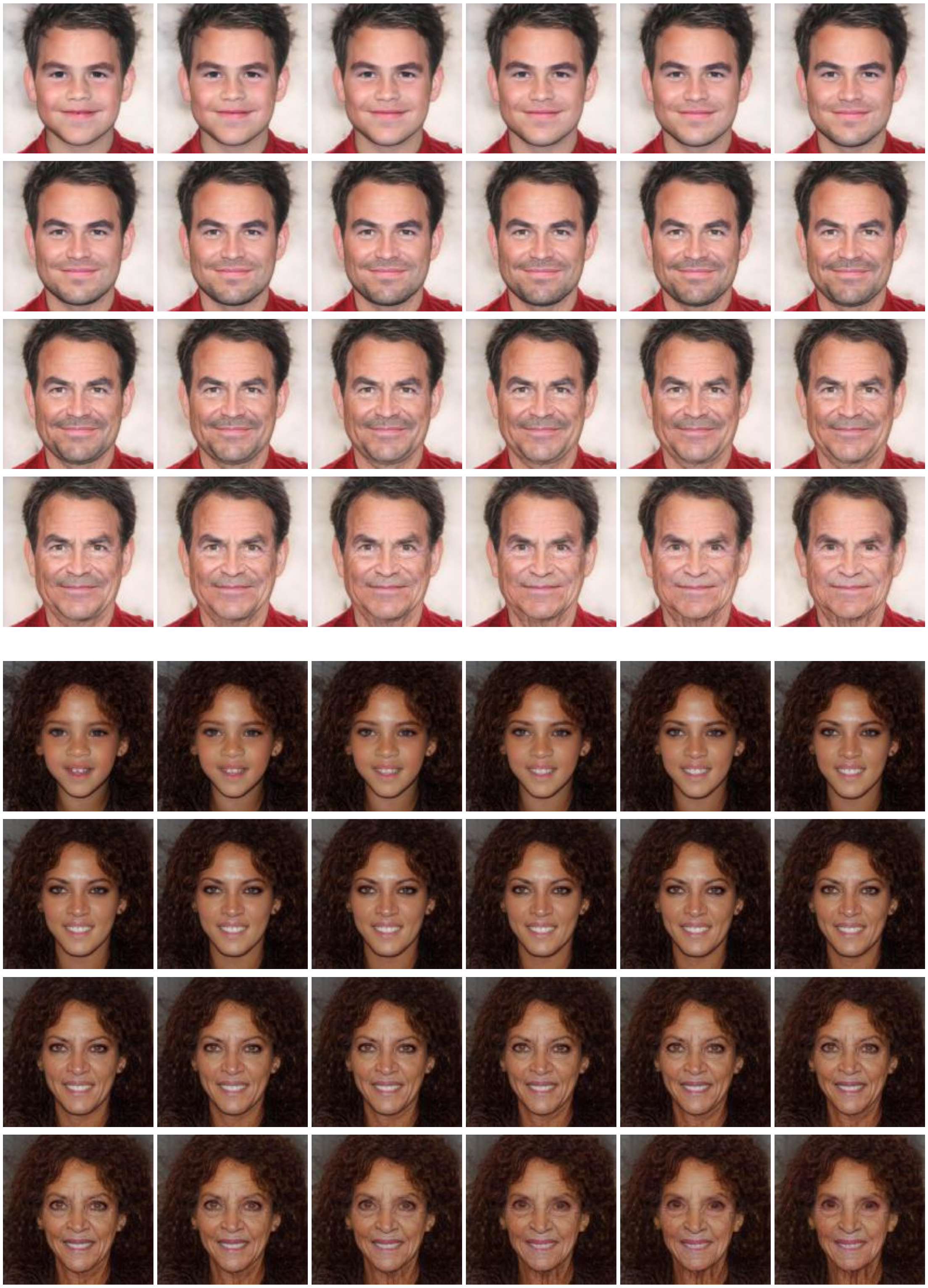


Figure 23: Full lifespan aging results generated using SAM.