

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

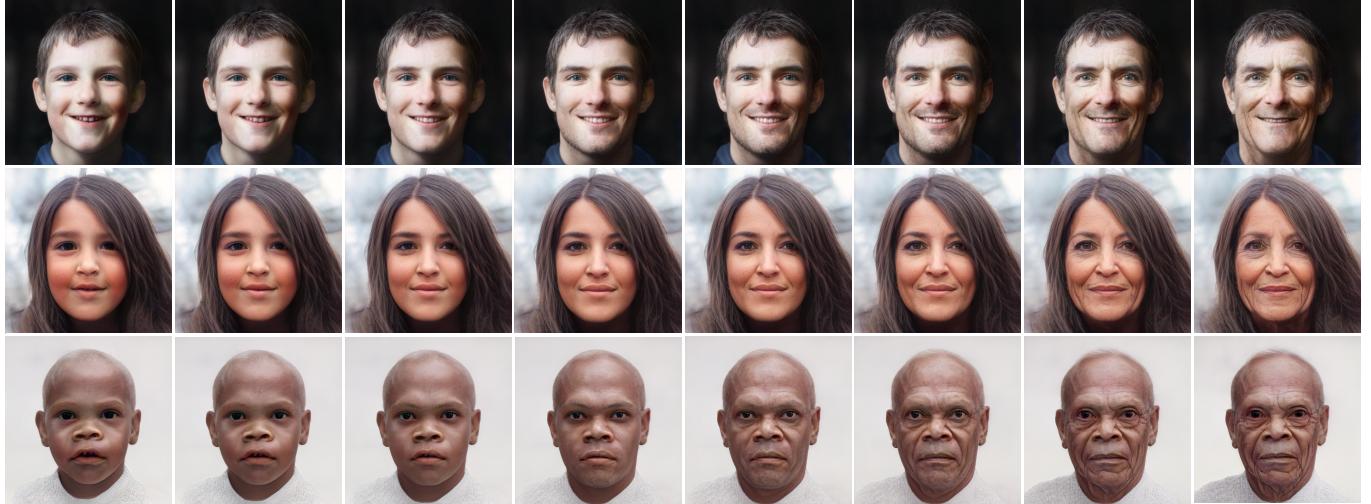


Fig. 1. Modeling fine-grained lifelong age transformation using the style-based SAM method.

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, 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 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 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. Code is available at our project page: <https://yuval-alaluf.github.io/SAM>.

Authors’ addresses: Yuval Alaluf, Tel-Aviv University, yuvalalaluf@gmail.com; Or Patashnik, Tel-Aviv University; Daniel Cohen-Or, Tel-Aviv University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.
 0730-0301/2021/8-ART45 \$15.00
<https://doi.org/10.1145/3450626.3459805>

CCS Concepts: • Computing methodologies → Image manipulation; Computer graphics; Neural networks.

Additional Key Words and Phrases: Generative Adversarial Networks, Age Transformation, Image Editing

ACM Reference Format:

Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. 2021. Only a Matter of Style: Age Transformation Using a Style-Based Regression Model. *ACM Trans. Graph.* 40, 4, Article 45 (August 2021), 21 pages. <https://doi.org/10.1145/3450626.3459805>

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 becomes significant (e.g., from ages 5 to 85).

To bypass explicitly modeling age transformation, data-driven techniques have recently 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 [Antipov et al. 2017; He et al. 2019; Or-El et al. 2020; Tang et al. 2018; Wang et al. 2016; Yang et al. 2019; Zhang et al. 2017] 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 [Abdal et al. 2020b; Härkönen et al. 2020; Liu et al. 2020; Shen et al. 2020; Wu et al. 2020] have approached the age transformation task by exploring the semantics of the latent space of a well-trained GAN, such as StyleGAN [Karras et al. 2019, 2020], 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 fixed, pre-trained age regression network to serve as an additional constraint during the training process. 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 [Karras et al. 2020] generator network trained on facial images, (ii) a pre-trained encoder network from Richardson et al. [2020] trained to encode real face images into the \mathcal{W}^+ latent space, (iii) an ArcFace [Deng et al. 2019] facial recognition network for identity regression, and (iv) a VGG [Simonyan and Zisserman 2015] 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 [Choi et al. 2020; Liu et al. 2019a] or anchor classes [Or-El et al. 2020], 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 learned disentangled path and the StyleGAN latent space, together with the end-to-end nature of SAM, allow for additional fine-grained editing on the generated aging results (e.g. hair color and 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: (i) A novel style-based regression approach for fine-grained modeling of the age transformation process; and (ii) an analysis of the non-linear latent path learned by our approach showing the benefits of an end-to-end method for modeling age transformation on real facial images.

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 [Duong et al. 2018; Fu et al. 2010; Ramanathan et al. 2009] for a comprehensive survey of such approaches. More recent works [He et al. 2018; Lample et al. 2018; Li et al. 2019; Or-El et al. 2020; Yao et al. 2020] take a data-driven approach for modeling face aging by using deep neural networks for image synthesis.

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. [2018] first introduced the use of conditional GANs [Mirza and Osindero 2014] 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 [Liu et al. 2017; Yi et al. 2017; Zhu et al. 2017a]. Some methods [Huang et al. 2018; Lee et al. 2020] translate a given image to a diverse set of corresponding output images, often referred to as multi-modal image synthesis. Furthermore, [Choi et al. 2018, 2020; Huang et al. 2018; Liu et al. 2019a; Zhu et al. 2017b] 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 [Fang et al. 2020; Georgopoulos et al. 2020; Li et al. 2020; Or-El et al. 2020; Viazovetskyi et al. 2020] 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 and may struggle to faithfully preserve identity when modeling lifelong aging. 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. [2020] use a similar approach, they are limited to translating images within the 20 – 70 age range. Moreover, the encoder-decoder architecture of Yao et al. [2020] introduces a wide bottleneck (256×256). As shall be shown, although this assists in reconstructing the input image, they are limited to modeling slight changes in texture and struggle to capture more global changes such as changes in head shape.

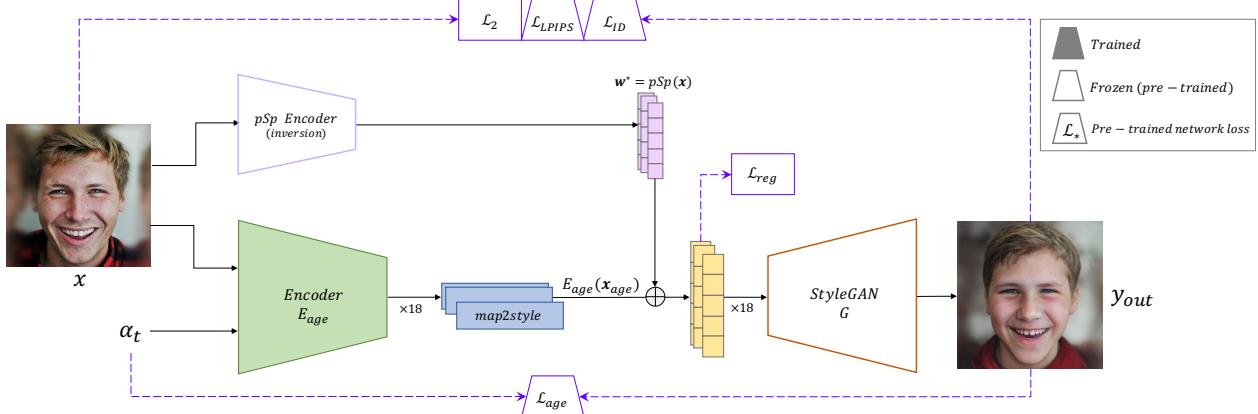


Fig. 2. Our SAM architecture. The network receives an input face image and a desired target age α_t . First, the aging encoder E_{age} is tasked with extracting feature maps at 3 different spatial scales. Then, 18 *map2style* blocks, introduced in [Richardson et al. 2020], are used to gradually down-sample the 3 feature maps into 18 different 512-dimensional style vectors, thereby encoding the input image into the ‘W+’ StyleGAN latent space. We additionally employ a fixed, pre-trained pSp [Richardson et al. 2020] encoder to extract the ‘W+’ latent code of x , denoted w^* , which is then added to the age-transformed latent code, denoted $E_{age}(\mathbf{x}_{age})$. A pre-trained StyleGAN is then used to generate the desired age-transformed image using the aggregated latent code. 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. Observe that during training, only the aging encoder and *map2style* blocks are trained. Moreover, \mathcal{L}_{LPIPS} , \mathcal{L}_{ID} , and \mathcal{L}_{age} are computed via fixed, pre-trained networks as described in Section 3.4.

2.2 The 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 [Nitzan et al. 2020] or by a latent space traversal [Abdal et al. 2020b; Denton et al. 2019; Goetschalckx et al. 2019; Härkönen et al. 2020; Shen et al. 2020; Shen and Zhou 2020; Voynov and Babenko 2020; Wu et al. 2020]. Most notably, given its state-of-the-art image quality and disentangled latent space, StyleGAN has been widely used for this task.

Such latent space methods first perform *GAN inversion*, first introduced by Zhu et al. [2016], in which real images are projected into the latent space of the GAN by finding the latent vector representation that best approximates the input image. Recent years have seen diverse approaches for performing such an inversion [Abdal et al. 2019, 2020a; Bau et al. 2019; Creswell and Bharath 2018; Pidhorskyi et al. 2020; Richardson et al. 2020; Tewari et al. 2020; Tov et al. 2021; Xia et al. 2021; Zhu et al. 2020]. Given the inversion, the resulting latent codes are then 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 manipulations 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 latent 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 aging 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 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 is then used to generate the desired output image. An overview of our architecture is presented in Figure 2.

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, which has been shown to be effective in unpaired image-to-image translation tasks [Liu et al. 2017; Tang et al. 2020; Yi et al. 2017; Zhu et al. 2017a]. 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.

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

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

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

$$\mathbf{x}_{age} = \mathbf{x} \| \alpha_t. \quad (2)$$

We additionally use a fixed, pre-trained pixel2style2pixel [Richardson et al. 2020] 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,

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

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

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

where $E_{age}(\cdot)$ and $G(\cdot)$ denote our aging encoder and the StyleGAN generator, respectively. During training, we perform two passes:

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

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

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).

Observe that we employ two encoders: $pSp(\cdot)$ and $E_{age}(\cdot)$. Here, $pSp(\cdot)$ is used to invert the given image for creating the initial latent embedding of the input. $E_{age}(\cdot)$ is then trained to learn the residual between this initial latent embedding and the latent embedding representing the age-transformed image. Note that the pre-trained pSp encoder remains *fixed* throughout training while $E_{age}(\cdot)$ is trained using the loss objectives described below.

3.3 Architecture

The architecture of our aging encoder E_{age} is based on the pSp encoder presented in Richardson et al. [2020]. 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 [Karras et al. 2019, 2020] generator. Given a 4-channel input described in Section 3.2, the encoder, based on a Feature Pyramid Network [Lin et al. 2017] architecture, is tasked with generating 18 unique style vectors corresponding to the 18 inputs of StyleGAN. To do so, the encoder first extracts feature maps at three different spatial scales. Then, 18 *map2style* blocks, are used to gradually down-sample each feature map to obtain a 512-dimensional style vector. Finally, the 18 style vectors are fed into the pre-trained StyleGAN generator to obtain the output image representing the desired transformation. Note that both $pSp(\cdot)$ and E_{age} share the same FPN-based architecture from Richardson et al. [2020].

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.

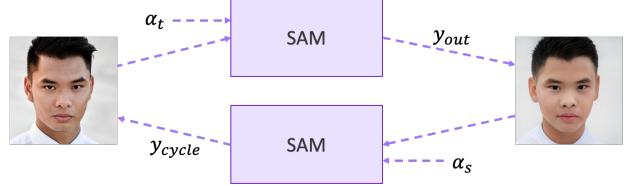


Fig. 3. To address the challenge of the unsupervised setting, a cycle consistency pass is performed to recover the input at the source age α_s .

First, we use the \mathcal{L}_2 loss to learn pixel-wise similarities and the LPIPS [Zhang et al. 2018] loss to learn perceptual similarities:

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

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

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}_{LPIPS} losses in the center region of the image and a lower weight on the outer region.

We additionally adopt a regularization loss [Baylies 2019; Richardson et al. 2020] that encourages the encoder to output style vectors closer to the average latent vector. We find that this regularization improves image quality by removing unwanted image artifacts without harming the fidelity of the desired age transformation.

Identity Loss. The authors in [Richardson et al. 2020] 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) [Burton et al. 2016; Mileva et al. 2020]. 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:

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

where R is a pre-trained ArcFace [Deng et al. 2019] network. Estimating the age α_s of the input image using the pre-trained age regression network, we define $\Delta_{age} = \frac{1}{100} |\alpha_s - \alpha_t|$. The weight function $w(\cdot)$ is then defined by,

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

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 [Rothe et al. 2015, 2018], 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,

$$\mathcal{L}_{Age}(\mathbf{x}_{age}) = \|\alpha_t - A(SAM(\mathbf{x}_{age}))\|_2. \quad (11)$$



Fig. 4. Aging results generated using SAM. Observe the state-of-the-art image quality achieved by leveraging a fixed, pretrained StyleGAN generator.

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

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

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,

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

where λ_{cycle} defines the weight of the cycle loss. Recall that in the cycle pass the recovered image, $SAM(\mathbf{y}_{out} \| \alpha_s)$, is compared to the original input image \mathbf{x} . 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 and latent space methods. We then show the benefits of using our end-to-end approach for learning the latent path controlling age. All evaluations are conducted on the CelebA-HQ [Karras et al. 2017] test dataset.

A comparison to multi-domain image-to-image translation approaches as well as an ablation study of SAM can be found in the Appendix B and Appendix C. Full lifespan animations and an accompanying video are also provided in the project page. Please note that due to the licensing agreement of the CelebA [Liu et al. 2015] dataset, we cannot present the original input images. As an alternative, we display each image's inversion obtained using the pSp encoder [Richardson et al. 2020].

Query 2016	0.505	0.503	0.571	0.619
Query 2012	0.378	0.464	0.724	0.517
	Childhood	1989	2014	2019

Fig. 5. Identity similarity results using the ArcFace [Deng et al. 2019] recognition network. For each row, we compute the cosine similarity between the query and the remaining images. Image credits in order: [Biard 2016],[de Be-sanez 1990],[Shankbone 2008],[Shankbone 2010],[de Souza 2019],[George 2012],[Korush and Millie 2020],[Light 1989],[U.S. 2014],[Bauld 2019]

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 current state-of-the-art facial recognition networks in recognizing the same individual at different ages. In doing so, we show that measuring identity preservation of aging methods should be done with great care.

4.1 The Age-Invariance of Facial Recognition Networks

Numerous experimental psychological studies [Burton et al. 2016; Jenkins et al. 2011; Johnston et al. 1997; Mileva et al. 2020] have shown that although people typically recognize individuals they

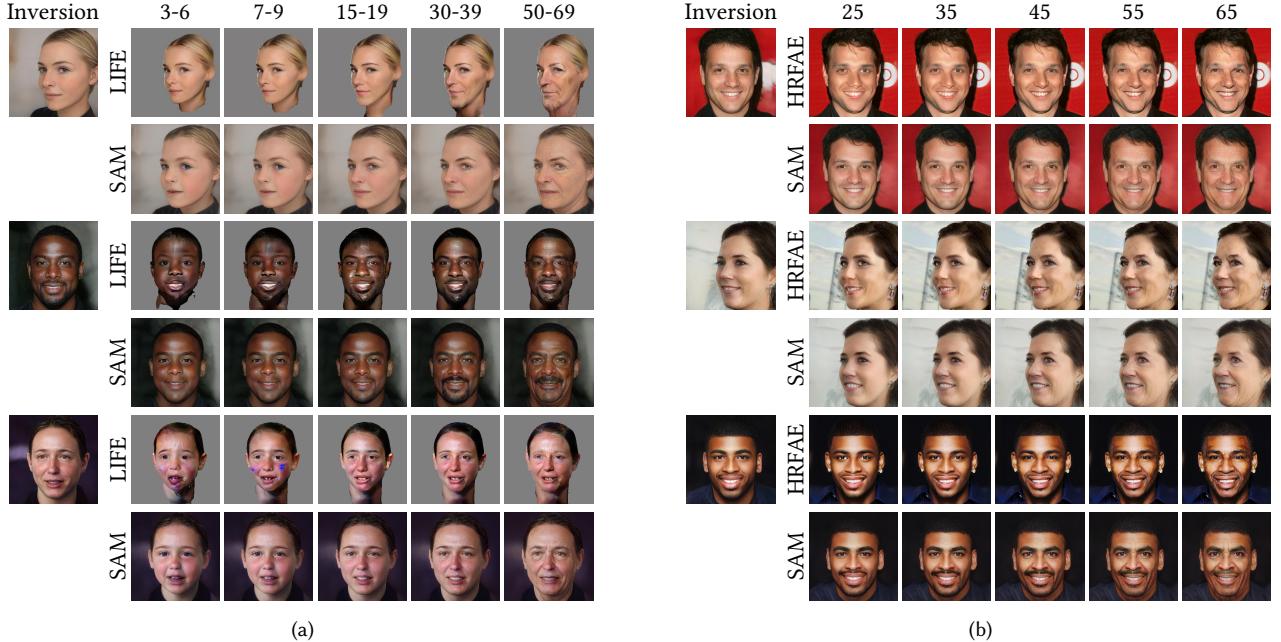


Fig. 6. Qualitative comparison of age transformation results with (a) LIFE [Or-El et al. 2020] and (b) HRFAE [Yao et al. 2020] on the CelebA-HQ [Karras et al. 2017] test set. For translating our images to the age groups in LIFE, we set the target age equal to the median age of each group. Best viewed zoomed-in. Additional results can be found in Appendix D.

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 [Deng et al. 2019] 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 5, as we move away from the age of the query, 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 a large variations in age. Following the above experiments and the conclusions from [Burton et al. 2016; Jenkins et al. 2011; Mileva et al. 2020], 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 [Clutterbuck and Johnston 2002; Megreya and Burton 2006, 2008]. 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. [2020] and Yao et al. [2020], which we refer to as *LIFE* and *HRFAE*, respectively. For both LIFE and HRFAE we use their official implementations and pre-trained models for evaluations. Note that each method

supports a different age range (0 – 70 for LIFE, 20 – 70 for HRFAE, and 5 – 100 for SAM). Therefore, when comparing two methods, we consider only ages supported by both methods.

Qualitative Evaluation. We provide a visual comparison with LIFE in Figure 6a. One can see that although LIFE is able to successfully capture the change in head shape across the different ages (see row 2), 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 6b 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 due to the wide bottleneck present in their architecture. In contrast, our approach 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.

Notice that all methods generally focus on modeling changes in the face regions. For example, all methods struggle in modeling changes in hair color and capturing receding hair lines with the progression of age. Specifically, 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 these possible changes over time. These techniques are explored in Section 4.4.

Quantitative Evaluation. We note that since LIFE operates using pre-defined age groups, we are unable to compare with it by directly translating images to specific target ages. However, LIFE is able to

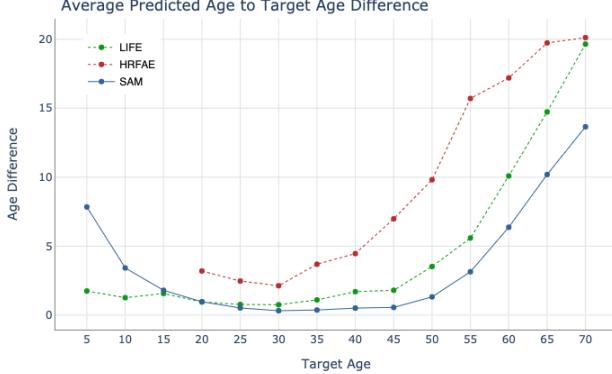


Fig. 7. Here, we examine each method’s ability to generate a full lifespan of images. Note that the lower the better.

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 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.

Having performed the above selection process, we now evaluate each method’s ability to accurately generate images for a wide-range of 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 losses, we use the Microsoft Azure Face API for age estimation. All metrics are computed using 1,000 test samples taken from the CelebA-HQ [Karras et al. 2017] test dataset.

Note that approximately 80% of the 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 5–10. 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 difference is most notable when translating to the older age range (60+).

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 worker 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 (i.e.

Table 1. Each cell indicates the percent of respondents who preferred the corresponding method for the evaluated metric. When compared to LIFE [Or-El et al. 2020] and HRFAE [Yao et al. 2020], SAM achieves comparable or superior aging accuracy and image quality across the different target ages.

Target	5		30		65		
	Method	Age	Quality	Age	Quality	Age	Quality
HRFAE	-	-	-	12.56	17.45	9.24	15.97
LIFE	55.1	18.88	7.25	12.11	11.14	11.14	11.64
SAM	44.9	81.12	80.19	70.44	79.62	72.39	

Human Evaluation I

Table 2. We asked respondents to identify 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. As shown, all three methods perform comparably.

Target	5		30		65	
	Method	ID Recall				
HRFAE	-	90.1	88.9			
LIFE	76.3	86.2	80.4			
SAM	89.6	91.1	86.7			

Human Evaluation II

300 responses for each target age). Note, all images were randomly selected for this evaluation.

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 worker to identify the individual shown in the transformed image. To increase the difficulty, the question is asked as an open question (i.e., the worker is 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 three methods are similar in their ability to faithfully 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 [Collins et al. 2020; Karras et al. 2019; Shen et al. 2020; Yang et al. 2020] 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.



Fig. 8. A visual comparison with InterFaceGAN [Shen et al. 2020] and StyleFlow [Abdal et al. 2020b] on real face images. Results for InterFaceGAN and StyleFlow are obtained by inverting the input into StyleGAN’s latent space using pSp [Richardson et al. 2020] and traversing along the learned age path.

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 (see Figure 9).

We compare our results with two state-of-the-art latent space approaches: InterFaceGAN [Shen et al. 2020] and StyleFlow [Abdal et al. 2020b]. 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.

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 direction. For StyleFlow, we use their official implementation for generating the transformed images. To edit a given *real* image, we

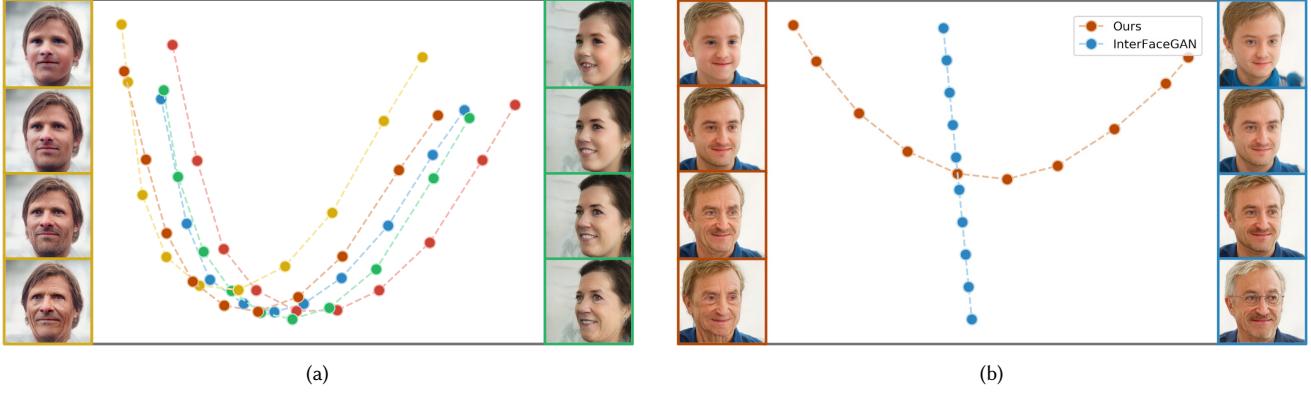


Fig. 9. In (a) we project the learned latent paths of five different images using PCA. Illustrated on the sides are intermediate outputs of two of the five images obtained by traversing along the corresponding colored path. As can be seen, the non-linearity of the learned paths are better suited to the 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).

first encode the image into the \mathcal{W}^+ latent space using pSp [Richardson et al. 2020], 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 8.

For completeness, we provide additional results for InterFaceGAN in Appendix D using their official implementation and age boundary. As their original paper and official implementation uses a StyleGAN1 [Karras et al. 2019] generator, we use IDInvert [Zhu et al. 2020] 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 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 9a, 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



Fig. 10. Images generated from randomly sampled w vectors that are traversed along the age manifold direction learned by SAM.

InterFaceGAN and SAM for the same real input face image. As can be seen in Figure 9b, 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 learned path using our approach is not trivial as the path’s behavior changes at different parts of the age progression.

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 10. 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 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.

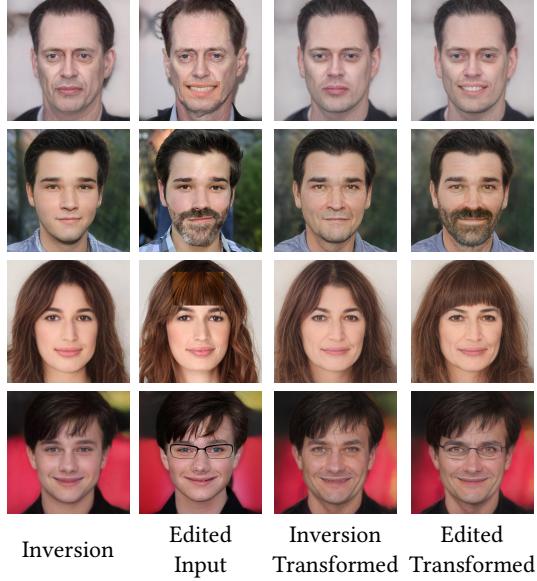


Fig. 11. SAM with patch editing allows for more fine-grained control over specific features such as expression, facial hair, hair style, and glasses.

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 (e.g., the hair of most individuals naturally becomes gray or recedes with age). It is therefore desirable to be able to control such attributes when modeling age progression. Here, we show that by using simple editing techniques, SAM provides additional control over the resulting image.

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 21, 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 [Karras et al. 2019], the fine styles

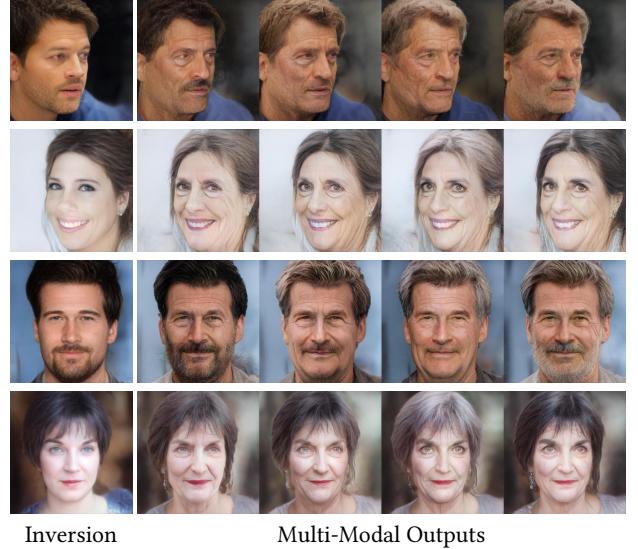


Fig. 12. Performing style-mixing on the age-transformed outputs of SAM allows additional editing control over features such as hair color and facial hair. Here, for each input image we perform style-mixing with four references images on layers 8 – 9 to obtain multi-modal results.

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 12 doing so enables fine-grained control in generating multiple plausible age transformation results.

5 DISCUSSION AND 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 representation, 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. Likewise, embedding an image into a set of vectors may make it more difficult to faithfully preserve input features such as the image background.

In our evaluations, we showed how our proposed method successfully disentangles age and other attributes such as hair color and hair style. Yet, such attributes naturally change as we age. To model these changes, we therefore proposed two editing techniques for controlling global changes (e.g., hair color) and local changes (e.g., the presence of eye glasses and facial hair). With that, capturing more complex changes such as receding hair lines and changes in skin color still remain a challenge with methods available today in part because they may be more difficult to capture.



Fig. 13. Limitations of SAM. Our approach may struggle when faced with extreme expressions or out-of-domain inputs not seen during the training of StyleGAN.

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 presented in Section 3.4. We present several challenging cases in Figure 13.

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 demonstrate 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.

ACKNOWLEDGMENTS

We would like to thank Elad Richardson, Kfir Goldberg, Ohad Fried, Yotam Nitzan, and Zongze Wu for their fruitful discussions and early feedback. We would also like to thank the anonymous reviewers for their insightful comments and constructive remarks. This work was supported in part by the Israel Science Foundation under Grant No. 2366/16 and Grant No. 2492/20.

REFERENCES

- Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space?. In *Proceedings of the IEEE international conference on computer vision*. 4432–4441.
- Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020a. Image2StyleGAN++: How to Edit the Embedded Images?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8296–8305.
- Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. 2020b. StyleFlow: Attribute-conditioned Exploration of StyleGAN-Generated Images using Conditional Continuous Normalizing Flows. arXiv:2008.02401 [cs.CV]
- Grigory Antipov, Mozej Baccouche, and Jean-Luc Dugelay. 2017. Face Aging With Conditional Generative Adversarial Networks. arXiv:1702.01983 [cs.CV]
- David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. 2019. Semantic Photo Manipulation with a Generative Image Prior. *ACM Trans. Graph.* 38, 4, Article 59 (July 2019), 11 pages. <https://doi.org/10.1145/3306346.3323023>
- John Bauld. 2019. Image taken by John Bauld and can be found here. License: Attribution 2.0 Generic (CC BY 2.0).
- Baylies. 2019. stylegan-encoder. <https://github.com/pbaylies/stylegan-encoder>. Accessed: January 2021.
- Georges Biard. 2016. (2016). Image taken by Georges Biard and can be found here. License: Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0).
- A. M. Burton, R. S. Kramer, K. L. Ritchie, and R. Jenkins. 2016. Identity From Variation: Representations of Faces Derived From Multiple Instances. *Cogn Sci* 40, 1 (Jan 2016), 202–223.
- Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8789–8797.
- Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8188–8197.
- Ruth Clutterbuck and Robert A Johnston. 2002. Exploring Levels of Face Familiarity by Using an Indirect Face-Matching Measure. *Perception* 31, 8 (2002), 985–994. <https://doi.org/10.1088/p3335> arXiv:<https://doi.org/10.1088/p3335> PMID: 12269591.
- Edo Collins, Raja Bala, Bob Price, and Sabine Süsstrunk. 2020. Editing in Style: Uncovering the Local Semantics of GANs. arXiv:2004.14367 [cs.CV]
- Antonia Creswell and Anil Anthony Bharath. 2018. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems* 30, 7 (2018), 1967–1974.
- Gorup de Besanez. 1990. Image taken by Gorup de Besanez and can be found here. License: Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).
- Jaqueleine de Souza. 2019. Image taken by Jaqueleine de Souza and can be found here. License: Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).
- Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4690–4699.
- Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. 2019. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439* (2019).
- Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D. Bui. 2018. Longitudinal Face Aging in the Wild - Recent Deep Learning Approaches. arXiv:1802.08726 [cs.CV]
- H. Fang, W. Deng, Y. Zhong, and J. Hu. 2020. Triple-GAN: Progressive Face Aging with Triple Translation Loss. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 3500–3509. <https://doi.org/10.1109/CVPRW50498.2020.000410>
- Yun Fu, Guodong Guo, and Thomas Huang. 2010. Age Synthesis and Estimation via Faces: A Survey. *IEEE transactions on pattern analysis and machine intelligence* 32 (11 2010), 1955–76. <https://doi.org/10.1109/TPAMI.2010.36>
- Angela George. 2012. Image taken by Angela George and can be found here. License: Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0).
- Markos Georgopoulos, James Oldfield, Mihalis A. Nicolaou, Yannis Panagakis, and Maja Pantic. 2020. Enhancing Facial Data Diversity With Style-Based Face Aging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. 2019. GANalyze: Toward Visual Definitions of Cognitive Image Properties. arXiv:1906.10112 [cs.CV]
- Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. GANSpace: Discovering Interpretable GAN Controls. *arXiv preprint arXiv:2004.02546* (2020).
- Z. He, M. Kan, S. Shan, and X. Chen. 2019. S2GAN: Share Aging Factors Across Ages and Share Aging Trends Among Individuals. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9439–9448. <https://doi.org/10.1109/ICCV.2019.00953>
- Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2018. AttGAN: Facial Attribute Editing by Only Changing What You Want. arXiv:1711.10678 [cs.CV]
- Xun Huang and Serge Belongie. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. arXiv:1703.06868 [cs.CV]
- Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal Unsupervised Image-to-image Translation. In *ECCV*.

- Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2018. Image-to-Image Translation with Conditional Adversarial Networks. arXiv:1611.07004 [cs.CV]
- Rob Jenkins, David White, Xandra Van Montfort, and A. Mike Burton. 2011. Variability in photos of the same face. *Cognition* 121, 3 (2011), 313 – 323. https://doi.org/10.1016/j.cognition.2011.08.001
- Robert A. Johnston, Masami Kanazawa, Takashi Kato, and Masaomi Oda. 1997. Exploring the Structure of Multidimensional Face-space: The Effects of Age and Gender. *Visual Cognition* 4, 1 (1997), 39–57. https://doi.org/10.1080/713756750 arXiv:https://doi.org/10.1080/713756750
- Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).
- Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4401–4410.
- Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8110–8119.
- Korush and Millie. 2020. Image taken by Korush and Millie and can be found here. License: Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).
- Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Fader Networks: Manipulating Images by Sliding Attributes. arXiv:1706.00409 [cs.CV]
- Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Kumar Singh, and Ming-Hsuan Yang. 2020. DRIT++: Diverse Image-to-Image Translation via Disentangled Representations. *International Journal of Computer Vision* (2020), 1–16.
- Peipei Li, Huaibo Huang, Yibo Hu, Xiang Wu, Ran He, and Zhenan Sun. 2019. UVA: A Universal Variational Framework for Continuous Age Analysis. arXiv:1904.00158 [cs.CV]
- Peipei Li, Huaibo Huang, Yibo Hu, Xiang Wu, Ran He, and Zhenan Sun. 2020. Hierarchical Face Aging Through Disentangled Latent Characteristics. In *European Conference on Computer Vision*. Springer, 86–101.
- Alan Light. 1989. Image taken by Alan Light and can be found here. License: Attribution 2.0 Generic (CC BY 2.0).
- Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.
- Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019b. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265* (2019).
- Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised Image-to-Image Translation Networks. In *Advances in Neural Information Processing Systems*. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., 700–708. https://proceedings.neurips.cc/paper/2017/file/dc6a6489640ca02b0d42dabeb8e46bb7-Paper.pdf
- Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. 2019a. Few-Shot Unsupervised Image-to-Image Translation. In *IEEE International Conference on Computer Vision (ICCV)*.
- Yunfan Liu, Qi Li, Zhenan Sun, and Tieniu Tan. 2020. Style Intervention: How to Achieve Spatial Disentanglement with Style-based Generators? arXiv:2011.09699 [cs.CV]
- Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
- Ahmed M. Megreya and A. Mike Burton. 2006. Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition* 34, 4 (01 Jun 2006), 865–876. https://doi.org/10.3758/BF03193433
- A. M. Megreya and A. M. Burton. 2008. Matching faces to photographs: Poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, 14(4) (2008), 364–372. https://doi.org/10.1037/a0013464
- Mila Mileva, Andrew W Young, Rob Jenkins, and A Mike Burton. 2020. Facial identity across the lifespan. *Cognitive psychology* 116 (2020), 101260.
- Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. arXiv:1411.1784 [cs.LG]
- Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. 2020. Face Identity Disentanglement via Latent Space Mapping. *ACM Trans. Graph.* 39, 6, Article 225 (Nov. 2020), 14 pages. https://doi.org/10.1145/3414685.3417826
- Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. 2020. Lifespan Age Transformation Synthesis. arXiv:2003.09764 [cs.CV]
- Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. 2020. Adversarial Latent Autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14104–14113.
- Narayanan Ramanathan, Rama Chellappa, and Soma Biswas. 2009. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages & Computing* 20, 3 (2009), 131–144.
- Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2020. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. arXiv:2008.00951 [cs.CV]
- Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. DEX: Deep EXpectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*.
- Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2018. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision* 126, 2-4 (2018), 144–157.
- David Shankbone. 2008. Image taken by David Shankbone and can be found here. License: Attribution-Share Alike 3.0 Unported.
- David Shankbone. 2010. Image taken by David Shankbone and can be found here. License: Attribution 3.0 Unported (CC BY 3.0).
- Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9243–9252.
- Yujun Shen and Bolei Zhou. 2020. Closed-Form Factorization of Latent Semantics in GANs. *arXiv preprint arXiv:2007.06600* (2020).
- Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]
- Hao Tang, Hong Liu, Dan Xu, Philip H. S. Torr, and Nicu Sebe. 2020. AttentionGAN: Unpaired Image-to-Image Translation using Attention-Guided Generative Adversarial Networks. arXiv:1911.11897 [cs.CV]
- X. Tang, Z. Wang, W. Luo, and S. Gao. 2018. Face Aging with Identity-Preserved Conditional Generative Adversarial Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 7939–7947. https://doi.org/10.1109/CVPR.2018.00828
- Ayush Tewari, Mohamed Elgharib, Mallikarjun B R, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2020. PIE: Portrait Image Embedding for Semantic Control. arXiv:2009.09485 [cs.CV]
- Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an Encoder for StyleGAN Image Manipulation. arXiv:2102.02766 [cs.CV]
- Government U.S. 2014. Image taken from here. Licensed under the Public Domain as a work of the U.S. federal government.
- Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. 2020. StyleGAN2 Distillation for Feed-forward Image Manipulation. *arXiv preprint arXiv:2003.03581* (2020).
- Andrey Voynov and Artem Babenko. 2020. Unsupervised Discovery of Interpretable Directions in the GAN Latent Space. *arXiv preprint arXiv:2002.03754* (2020).
- W. Wang, Z. Cui, Y. Yan, J. Feng, S. Yan, X. Shu, and N. Sebe. 2016. Recurrent Face Aging. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2378–2386. https://doi.org/10.1109/CVPR.2016.261
- Zongze Wu, Dani Lischinski, and Eli Shechtman. 2020. StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation. arXiv:2011.12799 [cs.CV]
- Weihao Xia, Yulin Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. 2021. GAN Inversion: A Survey. arXiv:2101.05278 [cs.CV]
- Ceyuan Yang, Yujun Shen, and Bolei Zhou. 2020. Semantic Hierarchy Emerges in Deep Generative Representations for Scene Synthesis. arXiv:1911.09267 [cs.CV]
- Hongyu Yang, Di Huang, Yunhong Wang, and Anil K. Jain. 2019. Learning Face Age Progression: A Pyramid Architecture of GANs. arXiv:1711.10352 [cs.CV]
- Xu Yao, Gilles Puy, Alasdair Newson, Yann Gousseau, and Pierre Hellier. 2020. High Resolution Face Age Editing. *CoRR* abs/2005.04410 (2020).
- Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. 2868–2876. https://doi.org/10.1109/ICCV.2017.310
- Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. 2019. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*. 9597–9608.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.
- Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. arXiv:1702.08423 [cs.CV]
- Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049* (2020).
- Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. 2016. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*. Springer, 597–613.
- Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017a. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.
- Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017b. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*.

Appendix

A IMPLEMENTATION DETAILS

Architectures. As in [Richardson et al. 2020], we use a pre-trained ResNet-IR architecture from [Deng et al. 2019] for our encoder backbone. We use a *fixed* StyleGAN2 generator pre-trained on the FFHQ [Karras et al. 2019] dataset. Identity embeddings are extracted using a pre-trained ArcFace network from [Deng et al. 2019]. As in [Yao et al. 2020], we use the age classifier from [Rothe et al. 2015, 2018] 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 [Or-El et al. 2020].

Training. 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 [Liu et al. 2019b] with Lookahead [Zhang et al. 2019], with a constant learning rate of 0.001. The input image resolution is 256×256 . Finally, 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}_{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}_{age} , the images are resized to 224×224 before being fed into the age classifier. Further, we set the loss lambdas as follows: λ_{I2} and λ_{lpips} are set to $\lambda_{I2} = 1$ and $\lambda_{lpips} = 0.6$ in the center region of the image and $\lambda_{I2} = 0.25$ 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 COMPARISON WITH MULTI-DOMAIN METHODS

Here, we compare our method with the current state-of-the-art in multi-domain image-to-image translation: FUNIT [Liu et al. 2019a] and StarGANv2 [Choi et al. 2020]. 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 [Or-El et al. 2020] and use the default training settings.

Qualitative Evaluation. We qualitatively compare the three methods in Figure 14. For performing inference on FUNIT and StarGANv2, we randomly sample a reference image from the FFHQ dataset 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 methods the reference image has a large effect on the translated image and alters various attributes besides the age.

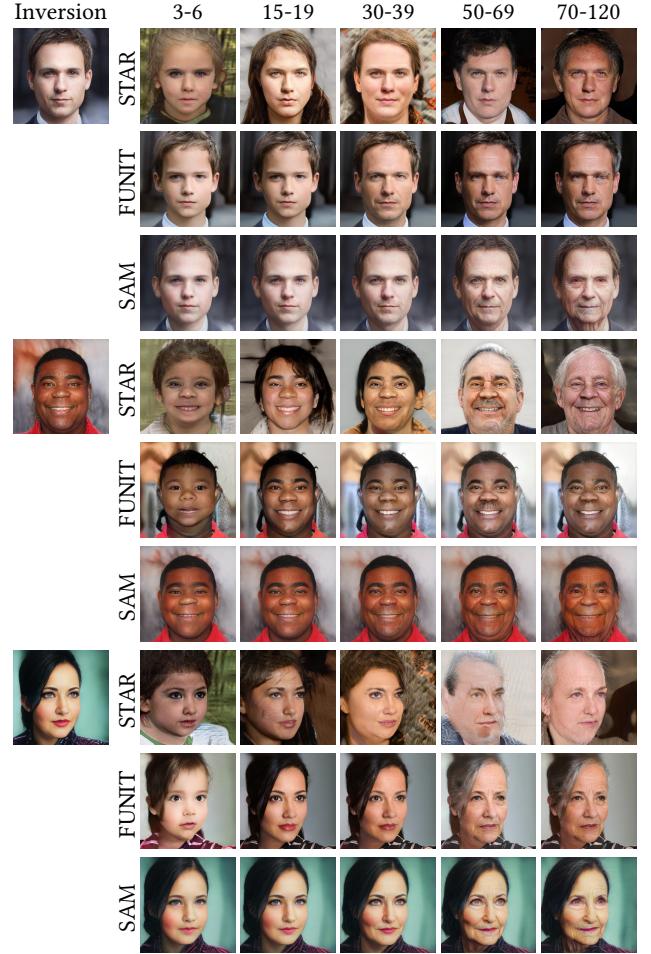


Fig. 14. Qualitative Comparison with FUNIT [Liu et al. 2019a] and StarGANv2 [Choi et al. 2020] (STAR). Note that FUNIT and StarGANv2 translate each source image using the same reference image. For translating images using SAM we set the target age equal to the median age of each group.

For example, one can see that the texture of the translated images is mostly taken from the reference image, 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 [Huang and Belongie 2017] 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 here where the texture, pose, lighting, etc. should be taken from the source. 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.

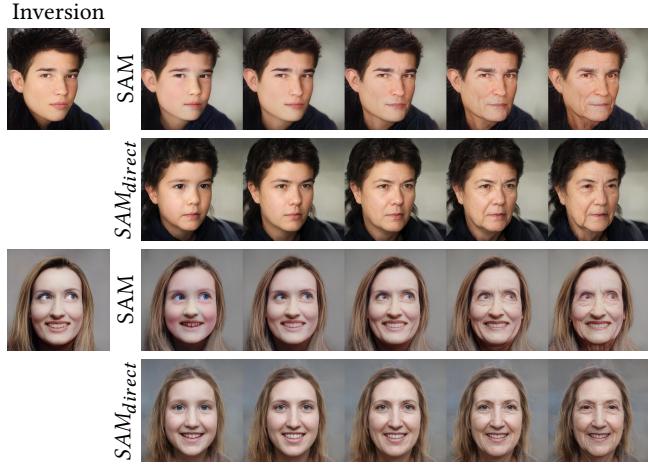


Fig. 15. Qualitative comparison between our SAM method and the SAM_{direct} variant. As can be seen, our SAM method is able to more accurately preserve identity and key facial features while faithfully modeling lifelong age transformation.

Table 3. Quantitative evaluation of the two SAM variants. Our residual-based SAM compares favorably with SAM_{direct} . Note, lower is better.

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

Ablation Study: Average Age Difference

C ABLATION STUDY

In this section we perform an ablation study on the training formulation of SAM and the loss objectives employed during training.

Training Formulation. We begin by validating the training formulation of SAM. Recall that the output of SAM is defined as

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

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 [Richardson et al. 2020] encoder. That is,

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

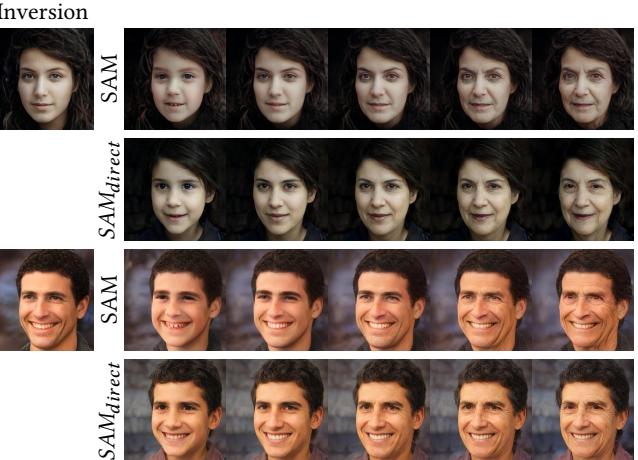
In this formulation, we train our SAM encoder to learn the *residual* vector between the latent code of the original input image and the latent code of the age-transformed image. In a sense, SAM is tasked with learning a *shift* in the latent space with respect to \mathbf{w}^* .

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

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

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.



In Figure 15 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, 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 better model the hair style of the top right image.

To verify that our residual-based approach does not harm the aging accuracy of the transformations we repeat the quantitative evaluation performed in Section 4.2 and compare 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. Full results are provided in Table 3. As shown, our residual-based approach does not restrict the SAM’s ability to generate images across a wide range of ages.

Loss Objectives. We now move to explore the importance of each of the loss objectives used in the training of SAM. Specifically, we show the importance of introducing a cycle pass during training. We additionally show that the \mathcal{L}_2 and LPIPS losses should be employed in the forward pass as well as in the cycle pass. Finally, we explore the influence of the w -regularization on the generated outputs.

In Figure 16, we show a qualitative comparison comparing SAM when trained with and without each of the above losses. For each input image, we translate the image to three different target ages: 5, 45, and 85. As can be seen, no one loss objective has a significant impact on the generated results. We first observe that when the \mathcal{L}_2 and LPIPS losses are not employed in the forward passes, SAM is better able to model changes in hair color (e.g., see example 2). However, we do find that our final configuration is able to better capture the younger age groups (see examples 2 and 3) and model changes to facial features such as wrinkles (see examples 1 and 2). Overall, we find that SAM is not overly sensitive to any one loss objective and the selected hyper-parameters as the results of all variants are plausible.

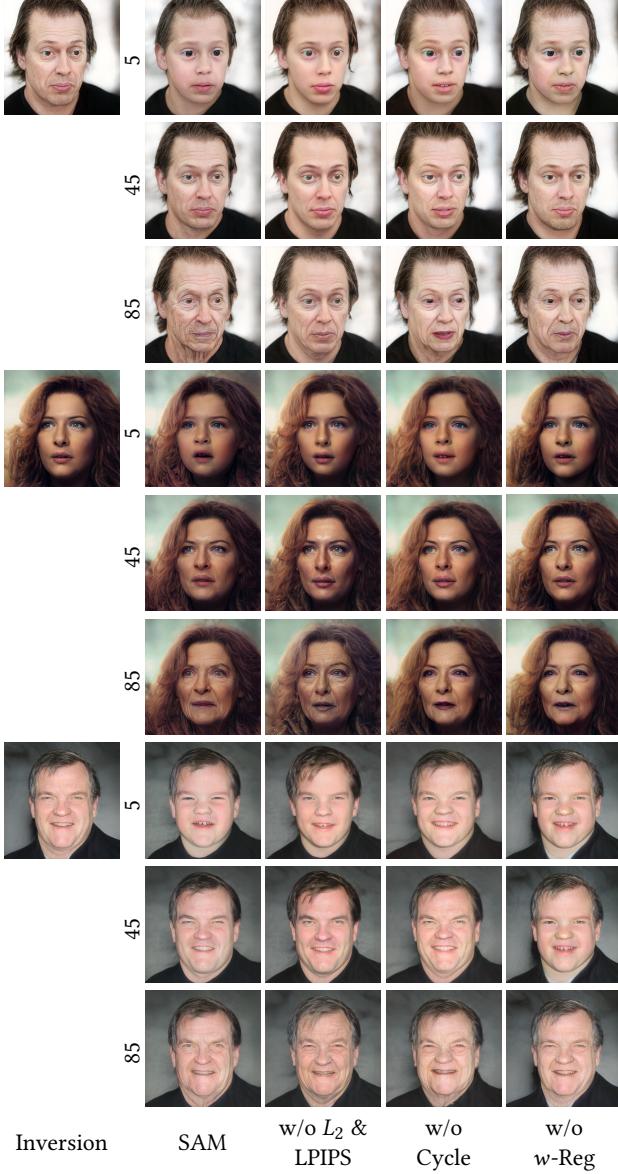


Fig. 16. Ablation study on the loss objectives employed during training of SAM. For each column, we train SAM using the same configuration without the specified loss objective. As shown, although no one variant is significantly better than the others, we find that our final configuration is able to better capture the target age of 5 and changes in face wrinkles when translating to the age of 85.

D ADDITIONAL RESULTS

The remainder of this document contains additional visual comparisons and results generated by SAM in the following Figures. Note that, as in the main manuscript, we are unable to display the original input images from the CelebA-HQ [Karras et al. 2017; Liu et al. 2015] dataset due to their licensing agreement. For each input, we therefore display the reconstruction image obtained by inverted each input using pSp [Richardson et al. 2020].

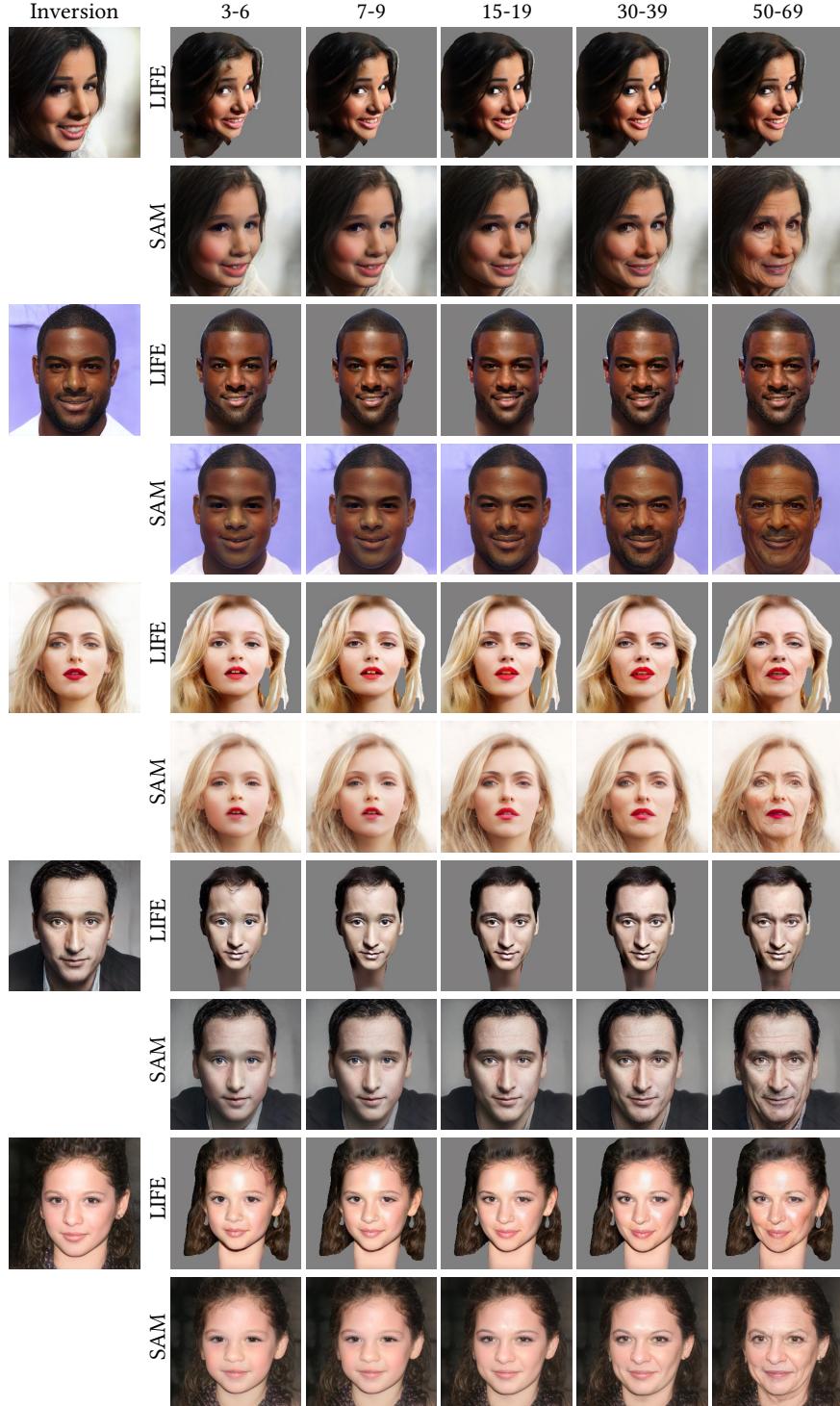


Fig. 17. Additional qualitative comparisons of age transformation results with LIFE [Or-El et al. 2020] on the CelebA-HQ [Karras et al. 2017] test set. For translating our images to the age groups in [Or-El et al. 2020], we set the target age equal to the median age of each group.

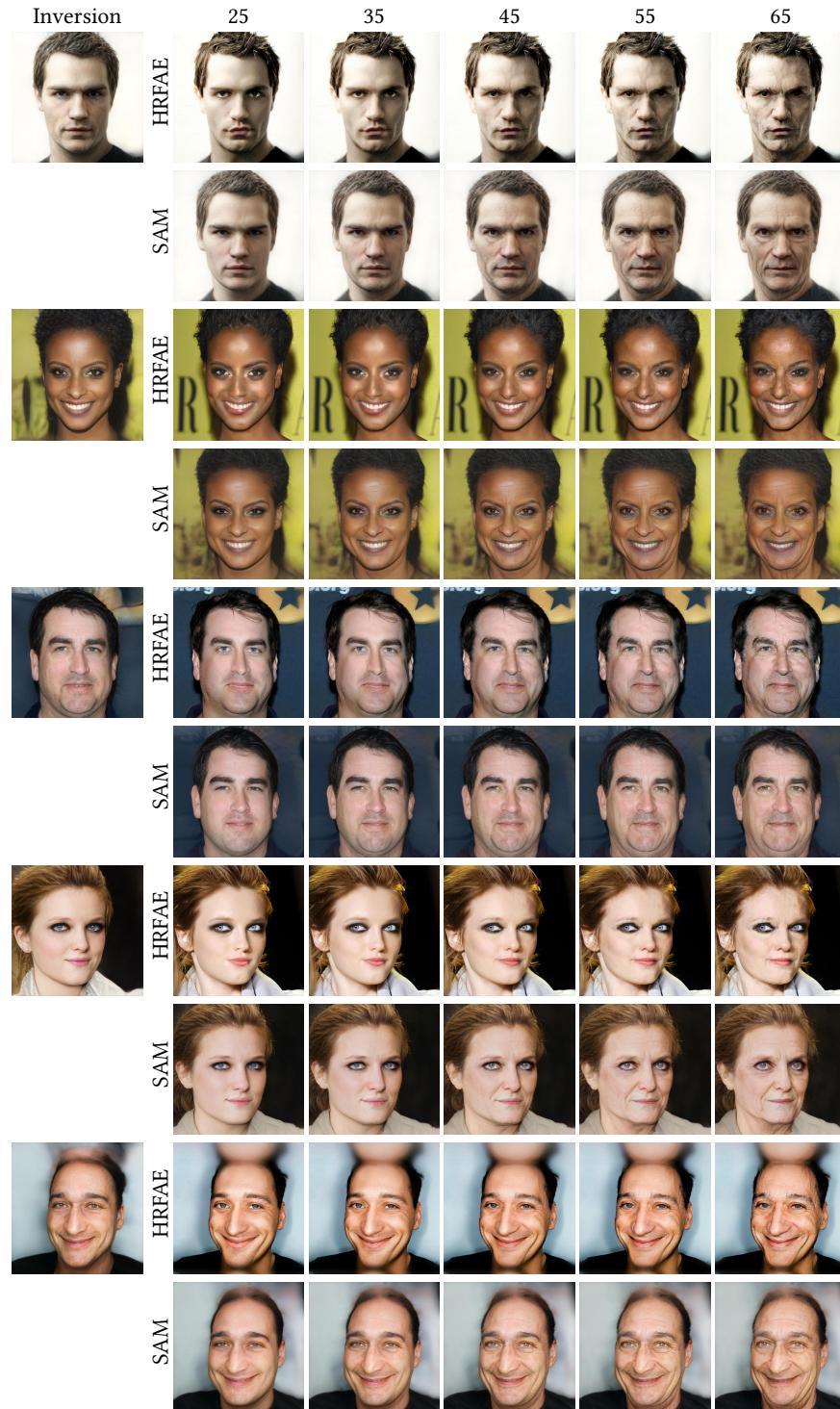


Fig. 18. Additional qualitative comparisons of age transformation results with HRFAE [Yao et al. 2020] on the CelebA-HQ [Karras et al. 2017] test set.



Fig. 19. Additional visual comparisons with InterFaceGAN [Shen et al. 2020] and StyleFlow [Abdal et al. 2020b] on real face images. Results for InterFaceGAN are obtained using IDInvert [Zhu et al. 2020] for inversion into the StyleGAN1 latent space. StyleFlow is performed on latents obtained by inverting the input using a pSp [Richardson et al. 2020] encoder into the StyleGAN2 latent space.



Fig. 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.

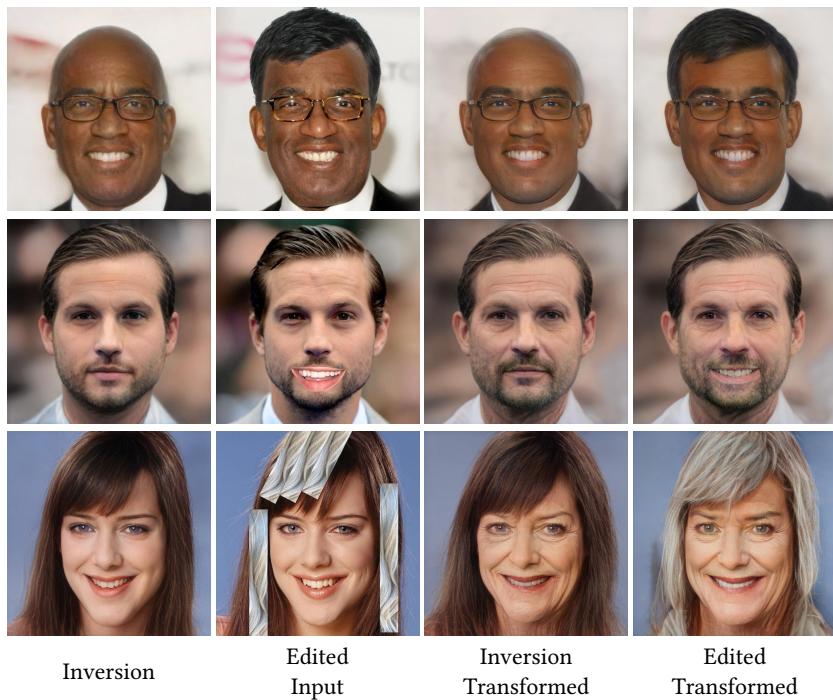


Fig. 21. Additional patch editing results generated using SAM.



Fig. 22. Additional age transformation results generated using SAM.

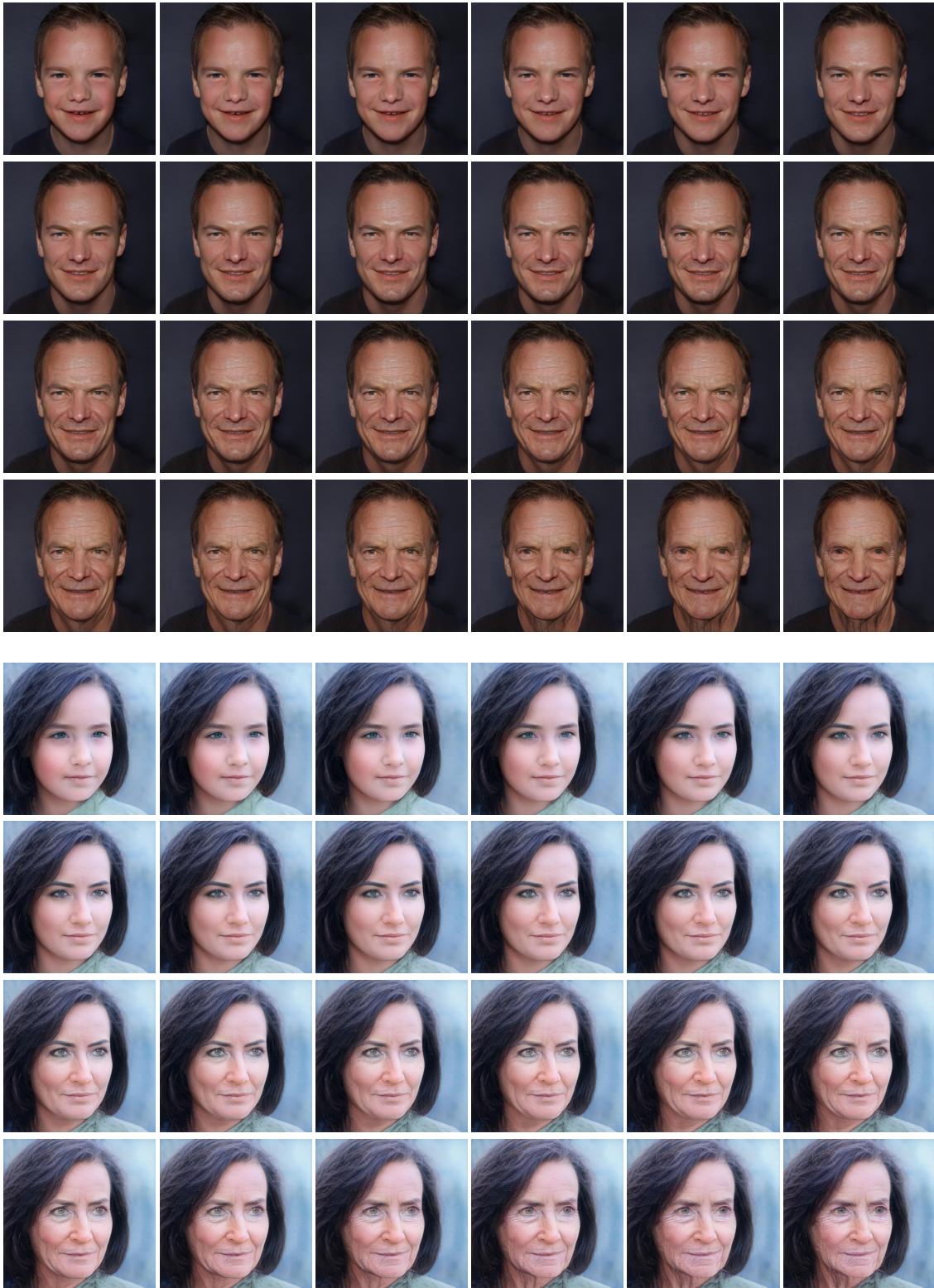


Fig. 23. Full lifespan aging results generated using SAM.