2021 IEEE/CVF International Conference on Computer Vision (ICCV) | 978-1-6654-2812-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCV48922.2021.00385

Disentangled Lifespan Face Synthesis

Sen He[1,2]*, Wentong Liao[3]*, Michael Ying Yang[4], Yi-Zhe Song[1,2], Bodo Rosenhahn[3], Tao Xiang[1,2]
[1]CVSSP, University of Surrey, [2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence,
[3]TNT, Leibniz University Hannover, [4]SUG, University of Twente

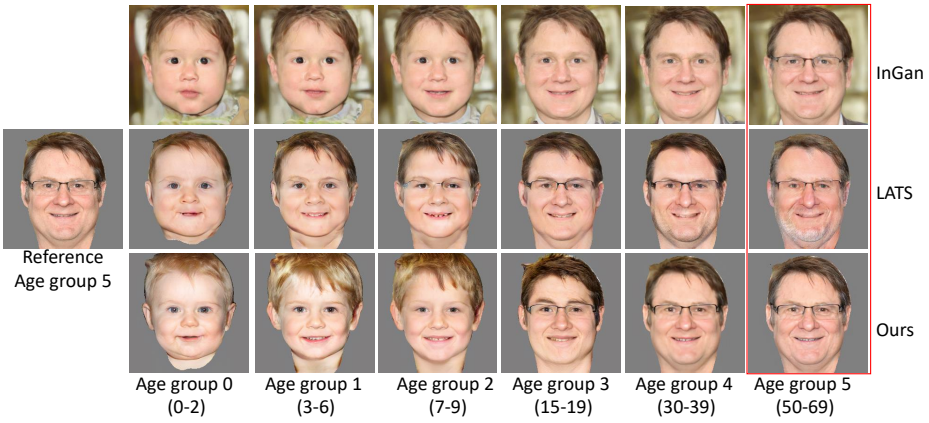

Figure 1: Examples of generated face images using our lifespan face synthesis model and two state-of-the-art alternatives. InGAN [43] generates valid texture transformation but fails in shape transformation and identity preservation. LATS [24] improves on shape transformation and identity preservation, but is poor in reconfiguration (the re-generated image in the same age group 5 as the reference looks very different). In contrast, our model overcomes all these limitations, yielding the most plausible effects of aging whilst being identity-preserving.

Abstract

A lifespan face synthesis (LFS) model aims to generate a set of photo-realistic face images of a person's whole life, given only one snapshot as reference. The generated face image given a target age code is expected to be age-sensitive reflected by bio-plausible transformations of shape and texture, while being identity preserving. This is extremely challenging because the shape and texture characteristics of a face undergo separate and highly nonlinear transformations w.r.t. age. Most recent LFS models are based on generative adversarial networks (GANs) whereby age code conditional transformations are applied to a latent face representation. They benefit greatly from the recent advancements of GANs. However, without explicitly disentangling their latent representations into the texture, shape and identity factors, they are fundamentally limited in modeling the nonlinear age-related transformation on texture and shape whilst preserving identity. In this work, a novel LFS model is proposed to disentangle the key face characteristics including shape, texture and identity so that the unique shape and texture age transformations can be modeled effectively. This is achieved by extracting shape, texture and identity features separately from an encoder. Critically, two transformation modules, one conditional convolution based and the other channel attention based, are designed for modeling the nonlinear shape and texture feature transformations respectively. This is to accommodate their rather distinct aging processes and ensure that our synthesized images are both age-sensitive and identity preserving. Extensive experiments show that our LFS model is clearly superior to the state-of-the-art alternatives. Codes and demo are available on our project website: https://senhe.github.io/projects/iccv_2021_lifespan_face.

1. Introduction

What would a young adult look like as an infant and what will she/he resemble in 20 or even 40 years time? Lifespan face synthesis or face aging and rejuvenation aims to answer these questions by synthesizing the face of a person's whole

*Equal contribution

978-1-6654-2812-5/21/$31.00 ©2021 IEEE
DOI 10.1109/ICCV48922.2021.00385

life given only a snapshot. It is an intriguing problem but also has many applications, *e.g*. cross-age face recognition [25] and finding lost children [36]. It has therefore attracted a great deal of attention recently [38, 37, 7, 31, 43, 24].

Lifespan face synthesis (LFS) is a challenging face attribute editing problem. Compared to other face attribute editing works [9, 21, 40] where many attributes such as glasses, hair style and smiling are manipulated with a single model, LFS focuses on one attribute only, namely age. However, age editing is arguably the hardest task out of all the attributes. It is thus typically studied on its own. This is because aging is an extremely complex face transformation process. In particular, over the lifespan of a person, the face experiences changes in both shape and texture [20]. Further, such changes are nonlinear over time and rather different for shape and texture: a face's appearance changes are first dominated by shape deformation from baby to young adult because of the growth of bones of skull; such changes then primarily take the form of texture transformation when an adult grows older, *e.g*. colors of beard and hair, wrinkles.

Therefore, an ideal LFS model must meet three requirements [7, 31, 43, 24]: (1) **Age-sensitive reflected by bio-plausible shape and texture transformations**: Given a reference face image and a random target age, the generated face image should have valid shape deformation as well as texture transformation compared to the reference face image. In particular, the highly nonlinear transformations mentioned above need to be respected. (2) **Identity-preserving**: no matter how large the age gap is between the target and reference, the generated image must depict the same person. (3) **Reconfigurable**: When the target age is the same as the age of the reference face image, the generated face image should be as similar to the reference image as possible.

However, despite the best efforts of researchers in the past two decades, none of the existing LFS models can meet all three requirements. Before the deep learning era, LFS models are either 'prototype' based [35, 34, 14] modeling mean age appearance for different age groups, or 'physical' based [33, 34] with explicitly modeling of the underlying biological aging mechanism. The former omits personalized information. While the later one requires images of same persons over the whole lifespan which is infeasible to scale. More recent methods [36, 22, 1, 42, 38, 37, 7, 31, 43, 24] benefit from the advancements in deep generative adversarial networks (GANs) [5, 2]. Using these methods, a latent face representation encompassing information of shape, texture and identity is transformed conditional on the target age before being fed into an image generator. Thanks to the recent breakthroughs in GANs such as Style-GAN [12, 13], these models can now generate incredibly high quality face images. But as shown in Fig. 1, they still fail in one or more of the three requirements.

This is because none of these models can effectively disentangle a face representation into shape, texture and identity related parts. Such a disentanglement is crucial for LFS because without the disentanglement, it is impossible to apply different age-conditional manipulations to these different representations to model the aforementioned nonlinear transformations in shape and texture appearance, whilst being identity-preserving. As a result, it is difficult to avoid unwanted editing. For example, identity can be changed as illustrated in the first row in Fig. 1. Furthermore, some incompatible transformation may occur, yielding unrealistic effects in the generated images (glasses start to appear in age group 2 of the middle row example).

In this work, for the first time we propose a LFS model that explicitly disentangles a learned latent face representation into shape, texture and identity. Our model is a conditional GAN with an encoder-decoder architecture. First, features of different layers of a shared CNN encoder are extracted and subject to different feature extraction modules. Second, to model the distinct nonlinear transformations on shape and texture with respect to age, two novel feature transformation modules are developed for shape and texture. They are based on conditional convolution and channel attention respectively to reflect the intrinsically different aging effects on shape and texture. Last but not least, to facilitate the disentanglement of shape and texture, a regularization loss is introduced on shape based on the intuition that shape changes are small when an adult is growing older [30]. As shown in Fig. 1, our disentanglement LFS model can effectively overcome the limitations of the state-of-the-art competitors and meet all three requirements simultaneously.

The contributions of this paper are as follows: (1) for the first time, we explicitly model the face's shape, texture and identity characteristics in an end-to-end trained lifespan face synthesis (LFS) model. (2) To model the separate nonlinear aging processes on shape and texture, we propose separate shape and texture transformation modules based on conditional convolution and channel attention respectively, as well as a shape regularization loss to facilitate the disentanglement. (3) Extensive experiments are carried out to demonstrate that our model is much superior to the state-of-the-art alternatives.

2. Related Works

Generative adversarial networks Generative adversarial networks (GANs) [5] are used by most recent image generation and manipulation methods. The developments of GANs can be mainly divided into two groups since the vanilla GANs [5]. One group tries to better measure the distribution divergence between the generated images and the original images [2]. The other group focuses on the architecture design, which has evolved from the original

fully connected networks to multi-scale convolutional architectures [11]. The most recent architectures is the style-GAN architecture [12, 13], where a random noise is first projected into a latent space and then used for convolution modulation. Style-GAN architecture has also been adopted in the recent state-of-the-art lifespan face synthesis models [31, 24, 43] as well as the proposed model in this work.

Face manipulation Face manipulation aims to edit a reference face image by changing some attributes, *e.g.*, age, smile and pose. The manipulated image is expected to contain intended attribute changes whilst preserving other attributes and identity. Recently, face manipulation has been studied intensively [9, 21, 40, 31, 8, 17, 43]. AttGAN [9] uses an attribute classification constraint to regularize the manipulated image. STGAN [21] selectively transfers the required attribute while keeping other factors unchanged. [31] learns the direction of each attribute in the latent space of style-GAN [12, 13], and then manipulates the exact latent code [43] of the reference image accordingly. Note that some generic face manipulation models do support age editing. However, they only manipulate a face image to be either younger or older (i.e., binary manipulation), an easier task than LFS.

Lifespan face synthesis Lifespan face synthesis (LFS) is the most challenging face manipulation task. Classical 'prototype' based methods [35, 34, 14] divide the continuous ages into several discrete age clusters, and then compute the mean face in each cluster for reference. In contrast, 'physical' based methods model the change of each aging factor in a parametric manner. [18] explored different parametric models (linear, quadratic and cubic) for the aging function. [32] uses a concatenational graph to model the aging process. Both groups of methods are based on manually designed rules, which is impossible to approximate the complex and nonlinear aging process. Further, they usually require images of the same person at different ages, which are very difficult to collect.

More Recent methods use conditional GANs for image generation. Yang *et al*. [38] propose a pyramid discriminator to penalize different factors in the aging progress. IPGAN [37] uses AlexNet [16] pre-trained on ImageNet [29] to enhance the identity preservation in the aging process. S2GAN [7] learns different transformation basis for different age groups. LATS [24] employs style-GAN architecture, in which the input is the encoded reference image and the style code is the embedded age representation. All these methods apply age-conditional transformation on an entangled latent representation of the reference image. They are thus intrinsically limited in modeling the distinct nonlinear transformations of shape and texture over ages. This limitation motivates the proposed disentanglement LFS approach.

Face disentanglement There are many existing efforts on disentangling the face into different latent factors, *e.g.*, identity, pose, shape and texture. Peng *et al*. [27] propose to disentangle face into identity and pose by reconstruction. Shen *et al*. [31] propose to disentangle the learned latent space into different attributes by supervised projection. Nitzan *et al*. [23] disentangle the identity information via latent space mapping. In this work, we propose to disentangle face into two aging-related factors, *i.e.*, shape and texture, as well as age-insensitive factor, namely identity. The key novelties over existing face disentanglement works are the two separate transformation modules designed to capture the distinct aging effects on shape and texture, and a shape regularization loss. These are crucial for effective age-sensitive disentanglement.

Dataset for lifespan face synthesis is hard to collect because ideally it should contain face images of the same person from baby to pensioner age. Most existing datasets miss face images of the 0 to 10 years age range. For example, the popular MORPH dataset [28] only has an age range of 16 to 77. The only dataset covering the whole range is the recently re-annotated FFHQ dataset [24], which contains ages ranging from 0 to 70 years old. It is thus used in this work. Overall existing LFS datasets are relatively small and insufficient for fine-grained age synthesis.

3. Method

3.1. Problem definition

Due to the dataset limitation, we follow [24] and divide the age range into 6 discrete groups (0 to 5). Each group has an age code $z_i \in \mathbb{R}^{|6N|}$, which is computed as:

$$z_i = \mathbf{1}_i + \mathbf{n}, \tag{1}$$

where $\mathbf{1}_i$ contains all ones on elements through iN to $(i+1)N$ and zeros elsewhere, $\mathbf{n} \in \mathbb{R}^{|6N|}$ is a Gaussian noise.

Given a reference face image $I_r \in \mathbb{R}^{H \times W \times 3}$ from the r_{th} age group and a target age code z_t for the t_{th} age group, a lifespan face synthesis model $\mathcal{F}$ aims to generate a target face image I_t, where $I_t = \mathcal{F}(I_r, z_t)$. The generated face image should have the same identity as the reference image but exhibit age-sensitive texture and shape changes according to the target age. Discrete lifespan face synthesis is done by traversing the age code of all age groups. To synthesize faces in a more fine-grained manner, the corresponding age code can be obtained by linear interpolation between the age codes of the neighboring two age groups.

As illustrated in Fig. 2, our model consists of five parts, an encoder ($\mathcal{E}$), a shape transformation module ($\mathcal{S}_t$), a texture transformation module ($\mathcal{T}_t$), an age embedding module ($\mathcal{A}_\mathcal{E}$), a generator ($\mathcal{G}$) and a discriminator ($\mathcal{D}$). Each of them will be detailed in the following sections.

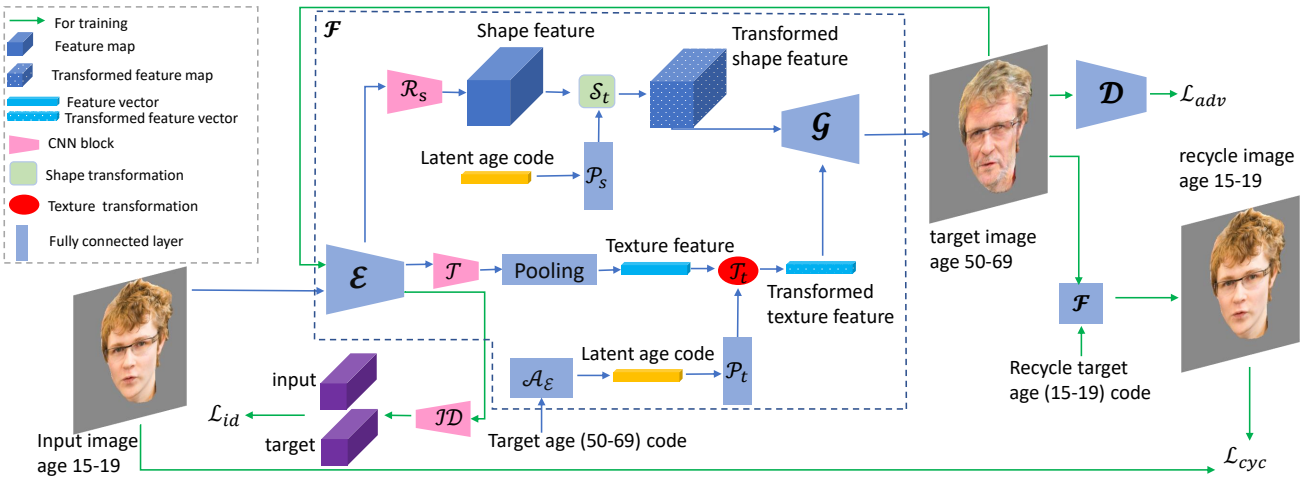

Figure 2: A schematic of our model. A latent representation of a reference image is disentangled into shape and texture relevant features, which are then transformed through separate transformation modules, conditioned on the target age. The transformed shape and texture feature are then fed into a style-GAN generator for the target image generation.

3.2. Feature extraction

Previous methods [37, 43, 24] extract entangled representation of face image and transform it according to the target age. Without factorizing the latent representation into shape, texture and identity relevant factors, it is impossible to model the transformations on shape and texture separately whilst preserving identity. Latent representation disentanglement is thus the key to effect LFS meeting all three requirements. To that end, we use our encoder ($\mathcal{E}$) to extract 3 distinct sets of features, *i.e.*, shape (f_s), texture (f_t) and identity (f_{id}). Inspired by the neural style transfer works [4, 10] which suggest that structure information can be extracted from the middle layers of a CNN and texture information from the deeper layers, we also propose to extract these three features from different layers of the encoder CNN ($\mathcal{E}$).

Concretely, the shape features ($f_s \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$) are extracted from the middle part of our encoder ($\mathcal{E}_m$):

$$f_s = \mathcal{R}_s(\mathcal{E}_m(I_r)), \tag{2}$$

where $\mathcal{R}_s$ is a residual block [6] to extract shape information for the raw features from the middle part of CNN. Both texture and identity features are extracted from the last layer of our encoder ($\mathcal{E}_d$). Specifically, the texture features ($f_t \in \mathbb{R}^{|C|}$) are computed as:

$$f_t = \mathcal{T}(\mathcal{E}_d(I_r)), \tag{3}$$

where T is a convolutional projection module that extracts the texture information and pools it into a vector. In the meanwhile, the identity features ($f_{id} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2C}$) are extracted by another convolutional projection module ($\mathcal{ID}$) with one downsampling layer:

$$f_{id} = \mathcal{ID}(\mathcal{E}_d(I_r)). \tag{4}$$

3.3. Shape and texture transformation

After feature extraction, we model the age-conditioned shape and texture transformation in a different way. For the shape transformation, we use conditional convolutions, where the convolution filters are modulated by the target age information:

$$\begin{aligned} f_s(z_t) &= \mathcal{S}_t(f_s, z_t) \\ &= \mathbf{conv}(f_s, \mathcal{M}(\mathbf{w}_s, \mathcal{P}_s(\mathcal{A}_\mathcal{E}(z_t))), \end{aligned} \tag{5}$$

where $\mathbf{w_s}$ is the filter weights, $\mathcal{A}_\mathcal{E}(z_t) \in \mathbb{R}^{|C|}$ encodes the target age code into a latent space, which is used for convolution modulation ($\mathcal{M}$), and $\mathcal{P}_s$ is a linear projection layer. With our formulation, the target age information will modulate the filter's weights and thus implicitly change shape information accordingly. Conditional convolution is adopted because shape transformation over age is often global and progressive; age code conditioned convolutional filters are well suited for capturing such transformation. Further, it is also flexible enough for learning at what age group shape changes become minimal.

For texture transformation, we use the age-conditioned channel attention, defined as:

$$f_t(z_t) = \mathcal{T}_t(f_t, z_t) = f_t \circ \mathcal{P}_t(\mathcal{A}_\mathcal{E}(z_t)), \tag{6}$$

where $\circ$ is element-wise multiplication and $\mathcal{P}_t$ is a linear projection layer. Again, this is determined by the nature of the texture changes caused by aging. In particular, as shown in [31, 43], different elements in face features f_t represent different attributes, *e.g.*, hair color and wrinkles. These attribute are present across ages but with different strengths. With age-conditioned channel attention, different aging attributes can be easily amplified or suppressed. For example,

the attention module will learn that wrinkles need to be suppressed by younger ages while amplified by older ages.

The transformed shape and texture information are then fed into a style-GAN [13] based generator $\mathcal{G}$ for target image generation:

$$I_t = \mathcal{G}(f_s(z_t), f_t(z_t)). \tag{7}$$

3.4. Shape regularization

Inspired by the previous finding [20] that the shape of an adult face usually remains unchanged, we propose a shape regularization to enforce this observation. This regularization would indirectly ensure that the extracted features f_s is indeed shape related and disentangled from the texture feature f_t which changes significantly when an adult is getting older. Concretely, for transforming a reference face image I_{r_e} in an adult group r_e (*e.g.* 40 years old) into an older age group t_e (*e.g.* 60 years old), the transformed face image I_{t_e} should have the same shape information:

$$\mathcal{L}_s = \left\| \mathcal{R}_s(\mathcal{E}_m(I_{r_e})) - \mathcal{R}_s(\mathcal{E}_m(I_{t_e})) \right\|^2, \tag{8}$$

where $\mathcal{L}_s$ measures shape difference and will be minimized.

3.5. Objectives

There are 5 learning objectives in our model's training. To ensure the identity preservation, an identity loss $\mathcal{L}_{id}$ is computed using the identity information between the reference image and the generated target image:

$$\mathcal{L}_{id} = \left\| \mathcal{ID}(\mathcal{E}_d(I_r)) - \mathcal{ID}(\mathcal{E}_d(I_t)) \right\|^2. \tag{9}$$

Meanwhile, a cycle consistency loss is applied to enhance the identity preservation:

$$\mathcal{L}_{cyc} = \left\| I_r - \mathcal{F}(I_t, z_r) \right\|^2. \tag{10}$$

To maintain the model's reconfiguration, a reconstruction loss is used when the target age is the same as the reference age:

$$\mathcal{L}_r = \left\| I_r - \mathcal{G}(f_s(z_r), f_t(z_r)) \right\|^2. \tag{11}$$

Furthermore, a conditional adversarial loss is used to improve the realism of the generated images:

$$\begin{aligned} \mathcal{L}_{adv} = {} & \mathbb{E}_{I^r_{im} \sim p^r_{data}(I^r_{im})}[log(\mathcal{D}(I^r_{im}|z)] \\ & + \mathbb{E}_{I^g_{im} \sim p^g_{data}(I^g_{im})}[1 - log(\mathcal{D}(I^g_{im}|z)]. \end{aligned} \tag{12}$$

The overall training objectives are summed together:

$$\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_r\mathcal{L}_r + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{id}\mathcal{L}_{id} + \lambda_s\mathcal{L}_s, \tag{13}$$

where λ_{adv}, λ_r, λ_{cyc}, λ_{id}, and λ_s denote the hyperparameters for balancing the 5 objectives.

4. Experiments

Datasets We train our model on the current largest face age dataset, *i.e.* FFHQ dataset [12], with age annotations covering all age groups. Due to label noise in the annotation, the original dataset with 70000 images is pruned to 28701 images [24]. The pruned dataset has 14232 training images and 198 testing images for male, and 14066 training images and 205 testing images for female. As per standard, we train male's model and female's model separately. Ages are divided into 6 discrete age groups (0-2, 3-6, 7-9 , 15-19, 30-39, 50-59) in the training dataset. The last two age groups are used to apply the shape regularization in Eq. (8). Following [24], the non-face regions in the input images are masked out using off-the-shelve face parsing model [3] trained on CelebAMask-HQ [19].

Implementation Details Our model is implemented in PyTorch. Each input image is resized to 256×256, the same size as in [24]. We set the batch size to 2 given the hardware available to us (a single Nvidia RTX 2080-Ti GPU). The length of age code is set to 300 ($N = 50$). The latent space dimension $C = 256$. The encoder $\mathcal{E}$ has two pooling layers in the first two blocks. And the generator $\mathcal{G}$ has two upsampling layers in the last two blocks. All parameters are trained using Adam optimizer [15]. The initial learning rate is set to 0.001, and decayed by 0.1 at epoch 50 and 100. The whole model is trained with 300 epochs. EMA [39] is used in the model training.

Evaluation Metrics We evaluate our model both automatically and manually (user study). In automatic evaluation, we use the off-the-shelve VGG-face [26] to evaluate the model's identity preservation. We also use LPIPS [41] to evaluate model's reconfiguration between the reference image and the re-generated image in the same age group. For manual evaluation, we run perceptual study on Amazon Mechanical Turk (AMT) to compare the quality of the generated lifespan face images from different models. Given a reference image, the generated images using different models are evaluated from 6 perspectives, namely, *identity preservation, shape transformation, texture transformation, reconfiguration, age error and age accuracy*. For identity preservation, the AMT workers were asked to judge how well the generated images preserved the identity in the reference image. For shape and texture transformation, they scored how plausible the transformations are. For reconfiguration, a worker was asked to score how similar the generated image in the same age group is to the reference image. For these 4 metrics, the scores are in 5 levels (1-5, higher being better). For age error and age accuracy, each worker was asked to evaluate whether the generated image belongs to the target age group and estimate the age difference to that group. Each AMT worker was randomly allocated 30 reference images. 10 workers participated in the evaluation for all 6 metrics.

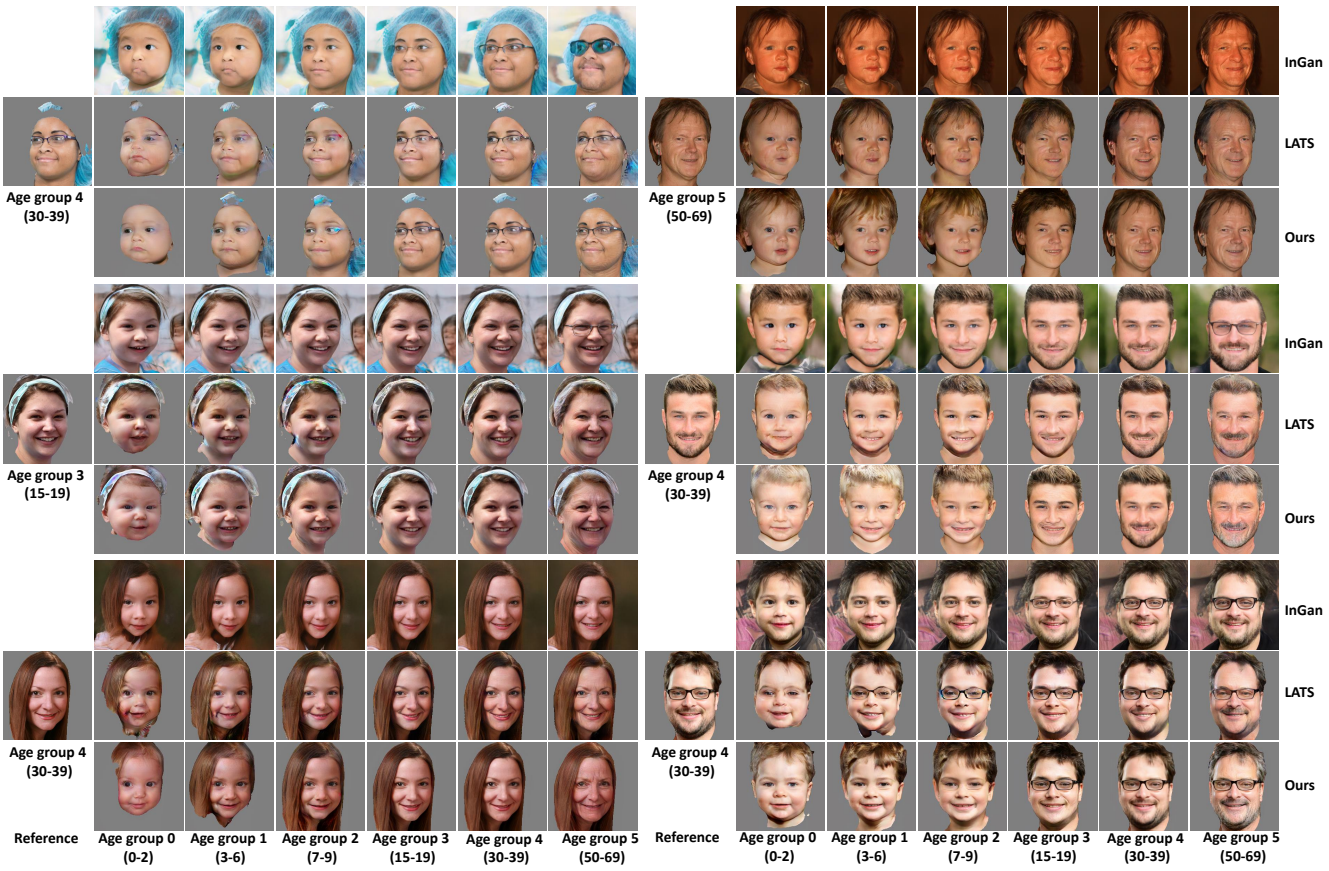

Figure 3: Qualitative results comparing our model against recent state-of-the-art models InGan [43] and LATS [24].

Methods	Identity preservation↑	Shape transformation↑	Texture transformation↑	Reconfiguration↑	Age error↓	Age accuracy↑
IPGAN [37]	**3.92±0.17**	2.38±0.42	2.50±0.12	3.93±0.01	11.33±0.89	27.0%
InGAN [43]	2.74±0.17	2.51±0.22	2.37±0.16	3.56±0.35	8.64±2.80	39.4%
LATS [24]	3.18±0.13	2.89±0.44	3.22±0.17	3.49±0.25	5.67±3.61	60.0%
Ours	3.07±0.19	**3.18±0.35**	**3.30±0.21**	**4.07±0.27**	**3.53±2.81**	**65.6%**

Table 1: User study results for different compared models.

Baselines We compare our model with two state-of-the-art lifespan face synthesis models, *i.e.*, LATS [24] and InGAN [43], both using styel-GAN based generators as ours. To synthesize lifespan faces in InGAN, the aging parameter is adjusted according to the reference images' age. We also compare with IPGAN [37], which uses a standard convolutional neural network as generator with a focus on identity preservation.

4.1. Main results

The results of user study and automatic evaluation are shown in in Tables 1 and 2, respectively. It can be seen that our model achieves significantly better overall performance using all evaluation metrics. It is noted that IPGAN [37] has very good identity preservation but fails completely at the main task, i.e., generating age-sensitive bio-plausible shape and texture transformations. As a result, its reconfiguration results are very strong in both tables. However, this is due to the fact that it hardly changes the face regardless what the target age is. LATS [24] is capable of good texture transformation, but is poor on reconfiguration and yields much lower age accuracy than our model. Note that for LFS, the identity preservation objective is often contradictory to those of shape and texture transformation. Table 1 shows that our model's identity preservation is marginally lower than LATS, but this is more then compensated by the much more superior performance on the other 5 metrics. We can see from the qualitative comparison in Fig. 3 that our model

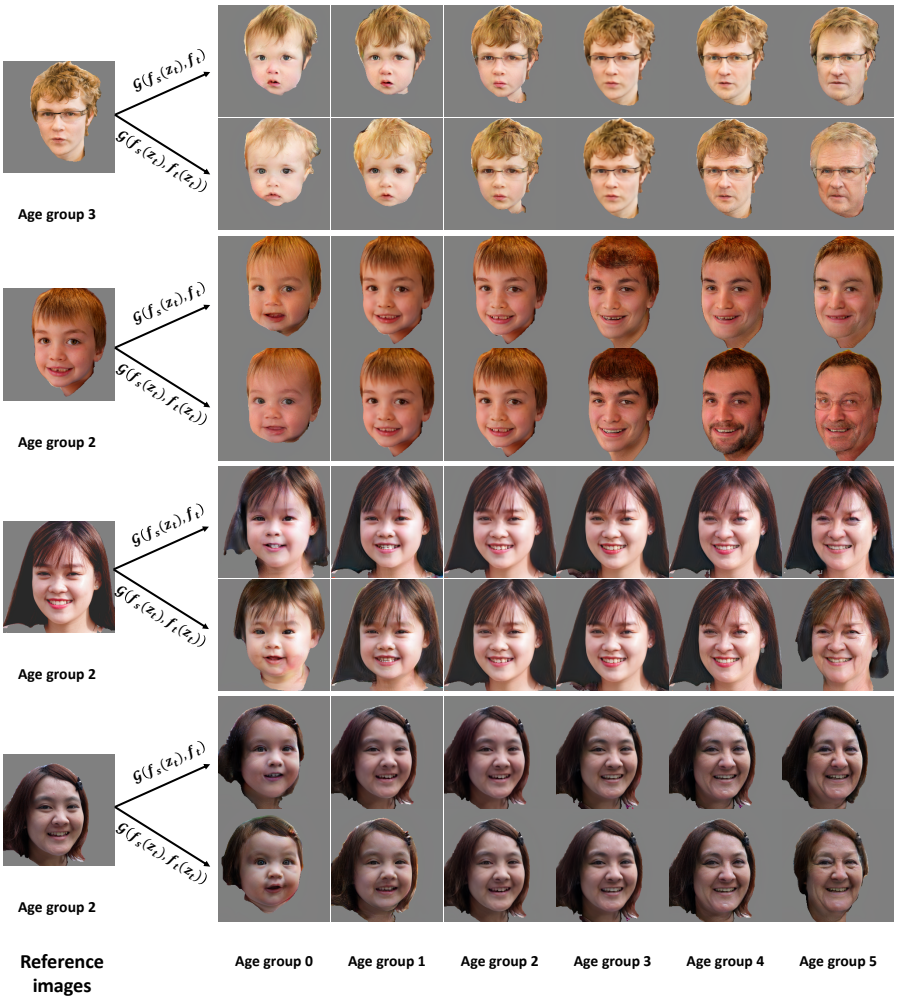

Figure 4: Validation of the disentanglement of shape and texture features in our model. In the first row of each example, the reference's texture information is fixed. Only shape information is transformed to synthesize the face images in different age groups. In the second row of each example, texture transformation is added to contrast its aging effects against shape.

Methods	ID (↑)	Reconfig (↓)
IPGAN [37]	**99.18%**	**0.03 ± 0.01**
InGAN [43]	92.35%	0.17± 0.09
LATS [24]	96.68%	0.11 ± 0.03
Ours	96.47%	0.07 ± 0.02

Table 2: Automatic evaluation on identity preservation (ID) and reconfiguration (Reconfig).

can synthesize lifespan face with (1) more significant shape deformation among young groups and better texture transformation among older groups, (2) better reconfiguration, and (3) more realistic images (*e.g.*, baby's face and glasses).

4.2. Ablation study

Disentangled Representations Our main idea is to learn disentangled representations. Did our model learn it? In this section, we qualitatively check, from two perspectives, whether the learned shape and texture representations have indeed been disentangled. First, we generate lifespan face images by only transforming f_s while keeping f_t fixed, i.e., $I_t = \mathcal{G}(f_s(z_t), f_t)$. We then transform both f_s and f_t to contrast their functionalities, *i.e.*, $I_t = \mathcal{G}(f_s(z_t), f_t(z_t))$. The result is shown in Fig. 4. It can be seen that the transformation of f_s yields significant shape deformation in the generated faces. However, it has little impact on the texture of the generated faces. Once the transformation of f_t is added in, we can see then significant texture changes from young to older adults (age group 4 to 5). Interestingly, the added f_t has little impact on the shape of the generated images. These results therefore validate the design that f_s and f_t are learning shape and texture information, respectively. More importantly, shape transformation is indeed more significant in younger groups (age group 0 to 3) while the added texture transformation mostly influences the texture

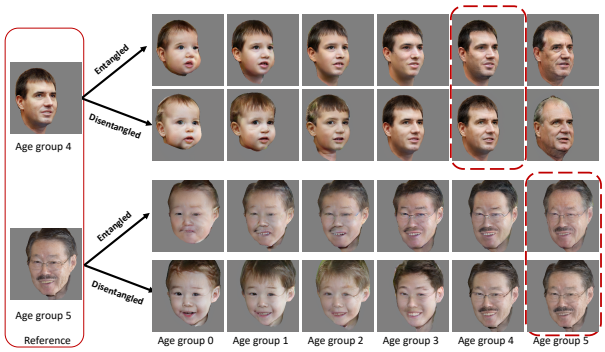

Figure 5: Qualitative comparison between entangled and disentangled lifespan face synthesis. In each example, the top row is the entangled model while the bottom row is disentangled. Red dotted boxes indicate that the generated images are within the same age group as the reference image.

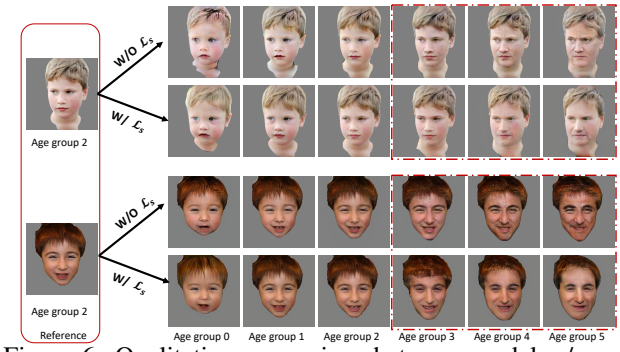

Figure 6: Qualitative comparison between model w/o and w/ shape regularization ($\mathcal{L}_s$).

in the older groups (age group 4 to 5). In other words, the nonlinear aging process has been learned in our proposed shape and texture transformation.

Disentangled vs. Entangled Representation How important is disentanglement for lifespan face synthesis? To answer this question, we use the same encoder in our model to extract an entangled feature representation f_{en} for the reference image. We then generate the target image using the same generator, conditioned on the target age, *i.e.*, $I_t = \mathcal{G}(f_{en}, \mathcal{A}_{\mathcal{E}}(z_t))$. From Fig. 5, it is clear that disentangled representation gives (1) better image quality, (2) more significant shape deformation in the younger groups, (3) better texture transformation for older age groups, and (4) better reconfiguration.

The Effectiveness of Shape Regularization To validate the effectiveness of shape regularization in Eq. (8), we trained a model without shape regularization. For comparison, we then generate the lifespan face images by only transforming f_s. As we can see from Fig. 6, without shape regularization, there are still significant texture transformation (wrinkles) among older groups, albeit only f_s is transformed and f_t is fixed. On the contrary, with shape regularization, the transformation of f_s has nearly no effect on the

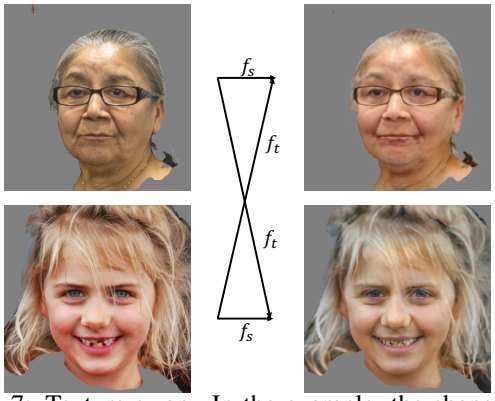

Figure 7: Texture swap. In the example, the shape information (f_s) of reference image is fixed while the texture information (f_t) is swapped with another image.

texture among older groups. This suggests that our shape regularization helps to clean the texture information in f_s, thus improves disentanglement of shape and texture.

Limitations Beyond the ablation study of the disentangled representations f_s and f_t in controlling the aging effects as shown in Fig. 4, we further examine their limitations. A texture swap experiment is conducted, where we use one image's f_s and the swapped f_t from another image to generate a new image. As we can see from the result in Fig. 7, f_t seems to be dominated by skin color, which is a key defining age-agnostic texture characteristic. As for age-related texture characteristic, e.g., wrinkles, it is not directly disentangled in f_t. However, as we can see from Fig. 4, the transformation of f_t can amplify or suppress wrinkles.

5. Conclusion

In this paper, we proposed a novel lifespan face synthesis model based on latent representation disentanglement. In contrast to previous methods that learn entangled face representation, our method disentangles a face representation into shape and texture. We proposed age-conditioned convolution and channel attention for shape and texture transformation, respectively to reflect the distinct aging effects on shape and texture. Extensive experiments and evaluations show the superiority of our method compared to previous state-of-the-art models.

Acknowledgment

This work has been supported by the Federal Ministry of Education and Research (BMBF), Germany, under the project LeibnizKILabor (grant no. 01DD20003), the Center for Digital Innovations (ZDIN) and the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC 2122).

References

[1] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *ICIP*, 2017. 2

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 2

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 5

[4] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. In *NeurIPS*, 2015. 4

[5] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NeurIPS*, 2014. 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[7] Zhenliang He, Meina Kan, Shiguang Shan, and Xilin Chen. S2gan: Share aging factors across ages and share aging trends among individuals. In *ICCV*, 2019. 2, 3

[8] Zhenliang He, Meina Kan, Jichao Zhang, and Shiguang Shan. Pa-gan: Progressive attention generative adversarial network for facial attribute editing. *arXiv preprint arXiv:2007.05892*, 2020. 3

[9] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *TIP*, 28(11), 2019. 2, 3

[10] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 4

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 3

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 3, 5

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2, 3, 5

[14] Ira Kemelmacher-Shlizerman, Supasorn Suwajanakorn, and Steven M Seitz. Illumination-aware age progression. In *CVPR*, 2014. 2, 3

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 3

[17] Jeong-gi Kwak, David K Han, and Hanseok Ko. Cafe-gan: Arbitrary face attribute editing with complementary attention feature. In *ECCV*, 2020. 3

[18] Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *TPAMI*, 24(4), 2002. 3

[19] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 5

[20] L.G.Farkas. *Anthropometry of the Head and Face*. 2007. 2, 5

[21] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *CVPR*, 2019. 2, 3

[22] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, Ngan Le, and Marios Savvides. Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition. In *ICCV*, 2017. 2

[23] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *TOG*, 39(6), 2020. 3

[24] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *ECCV*, 2020. 1, 2, 3, 4, 5, 6, 7

[25] Unsang Park, Yiying Tong, and Anil K Jain. Age-invariant face recognition. *TPAMI*, 32(5), 2010. 2

[26] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015. 5

[27] Xi Peng, Xiang Yu, Kihyuk Sohn, Dimitris N Metaxas, and Manmohan Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *CVPR*, 2017. 3

[28] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *FGR06*, 2006. 3

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3), 2015. 3

[30] Rosy Setiawati and Paulus Rahardjo. Bone development and growth. *Osteogenesis and bone regeneration*, 10, 2019. 2

[31] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 2, 3, 4

[32] Jinli Suo, Xilin Chen, Shiguang Shan, Wen Gao, and Qionghai Dai. A concatenational graph evolution aging model. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2083–2096, 2012. 3

[33] Jinli Suo, Song-Chun Zhu, Shiguang Shan, and Xilin Chen. A compositional and dynamic model for face aging. *TPAMI*, 32(3), 2009. 2

[34] Yusuke Tazoe, Hiroaki Gohara, Akinobu Maejima, and Shigeo Morishima. Facial aging simulator considering geometry and patch-tiled texture. In *SIGGRAPH*, pages 1–1. 2012. 2, 3

[35] Bernard Tiddeman, Michael Burt, and David Perrett. Prototyping and transforming facial textures for perception research. *CGA*, 21(5), 2001. 2, 3

[36] Wei Wang, Zhen Cui, Yan Yan, Jiashi Feng, Shuicheng Yan, Xiangbo Shu, and Nicu Sebe. Recurrent face aging. In *CVPR*, 2016. 2

[37] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. Face aging with identity-preserved conditional generative adversarial networks. In *CVPR*, 2018. 2, 3, 4, 6, 7

[38] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K Jain. Learning face age progression: A pyramid architecture of gans. In *CVPR*, 2018. 2, 3

[39] Yasin Yazici, Chuan-Sheng Foo, Stefan Winkler, Kim-Hui Yap, Georgios Piliouras, and Vijay Chandrasekhar. The unusual effectiveness of averaging in gan training. In *ICLR*, 2019. 5

[40] Weidong Yin, Ziwei Liu, and Chen Change Loy. Instance-level facial attributes transfer with geometry-aware flow. In *AAAI*, 2019. 2, 3

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[42] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017. 2

[43] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *ECCV*, 2020. 1, 2, 3, 4, 6, 7