Neurocomputing 451 (2021) 167–180

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Attention-aware conditional generative adversarial networks for facial age synthesis

Xiahui Chen [a], Yunlian Sun [a,*], Xiangbo Shu [a], Qi Li [b,c]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[b] National Laboratory of Pattern Recognition, Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[c] Artificial Intelligence Research, Chinese Academy of Sciences, Qingdao 266300, China

ARTICLE INFO

Article history:
Received 10 February 2020
Revised 13 April 2021
Accepted 16 April 2021
Available online 24 April 2021
Communicated by Zidong Wang

Keywords:
Facial age synthesis
Channel attention
Attention mask

ABSTRACT

Generative adversarial networks (GANs) have recently achieved impressive results in facial age synthesis. However, these methods usually select an autoencoder-style generator. And the bottleneck layer in the encoder-decoder generally gives rise to blurry and low-quality generation. To address this limitation, we propose a novel attention-aware conditional generative adversarial network (ACGAN). First, we utilize two different attention mechanisms to improve generation quality. On one hand, we integrate channel attention modules into the generator to enhance the discriminative representation power. On the other hand, we introduce a position attention mask to well-process images captured with various backgrounds and illuminations. Second, we deploy a local discriminator to enhance the central face region with informative details. Third, we adopt three types of losses to achieve accurate age generation and preserve personalized features: 1) The adversarial loss aims to synthesize photo-realistic faces with expected aging effects; 2) The identity loss intends to keep identity information unchanged; 3) The attention loss tries to improve the accuracy of attention mask regression. To assess the effectiveness of the proposed method, we conduct extensive experiments on several public aging databases. Experimental results on MORPH, CACD, and FG-NET show the effectiveness of the proposed framework.

1. Introduction

Facial age synthesis, including age progression and regression, aims to reconstruct faces with natural aging or rejuvenating effects [11,44]. It has wide applications, including finding missing children, improving face recognition accuracy, and social entertainment. Besides, it can also be used for data augmentation.

Synthesizing faces with desired aging effects, and simultaneously keeping identity stable is a troublesome task. On one hand, internal (e.g., genes) and external factors (e.g., living environment and health status) usually make different people age differently. It thus becomes difficult to model such personalized and uncontrollable aging processes accurately. On the other hand, the lack of labeled aging data and large facial variations also affect the process of face aging.

Many approaches have been proposed to tackle this issue. Traditional aging approaches can be roughly divided into two categories: physical model-based approaches and prototype model-based approaches. Physical model-based approaches usually utilize parametric models to model the physical and biological mechanisms of the face [62,23,55,20,49]. These approaches usually require masses of samples of the same person covering a wide range of ages, which are time-consuming to collect. And the complex aging mechanism found does not generalize well either. Prototype model-based approaches generally extract differences between age groups as universal aging patterns [3,43,40,18]. Different people have different aging processes. Thus, these methods are prone to ignore this diversity.

Apart from these above approaches, recently, generative adversarial networks (GANs) [13] have exhibited remarkable ability in synthesizing face aging sequences through adversarial learning [47,63,53,57]. These methods usually adopt the encoder-decoder architecture, where spatial or downsampling is essential to obtain high-level abstract representation. One drawback of these methods is that they generally treat each channel-wise feature equally, lacking discriminative learning ability across feature channels. Consequently, it may weaken the network's representation of power [61] and mislead the model to discard some vital information that can not be entirely recovered by transposed convolutions [27]. To

* Corresponding author.
E-mail address: yunlian.sun@njust.edu.cn (Y. Sun).

https://doi.org/10.1016/j.neucom.2021.04.068
0925-2312/

illustrate this visually, we combine the generator of [57] with the discriminator of [53] to train a model for face aging. Typical failure cases are shown in Fig. 1. (a). Synthesized faces look quite blurry, and some critical details are lost. We also list some failure cases of [24] in Fig. 1. (b). Although faces generated by these methods present the effect of aging, the visual quality is severely degraded by lost details (e.g., hair, background, and the face center). Thus, it is necessary to enhance the latent feature representation's discrimination ability to generate visually plausible images for facial age synthesis.

As a special convolutional mechanism, attention can not only tell the network where to focus but also improve the representation of interests [54]. It improves the representation by focusing on valuable features and suppressing worthless ones. Attention has recently shown improved performance in image classification [54,15], image synthesis [61], semantic segmentation[10,7,16], and image captioning [6]. Considering this merit, we propose an attention-aware conditional generative adversarial networks (ACGAN) for facial age synthesis to improve the visual fidelity of generated images. Specifically, we introduce channel attention to the commonly used autoencoder-style generator to improve the representation of regions of interest. Apart from this, we also utilize a spatial attention mask to constrain image translations. Considering the face center usually contains more pure and age-related features, we further define a local discriminator to encourage more photo-realistic results [9]. To sum up, our proposed method can selectively learn age-related features and thus modify corresponding age-related regions. The main contributions of this paper could be summarized as follows:

1) We propose an attention-aware conditional generative adversarial network to enhance the discriminative representation power. To the best of our knowledge, this is the first time that channel attention is adopted for face aging.

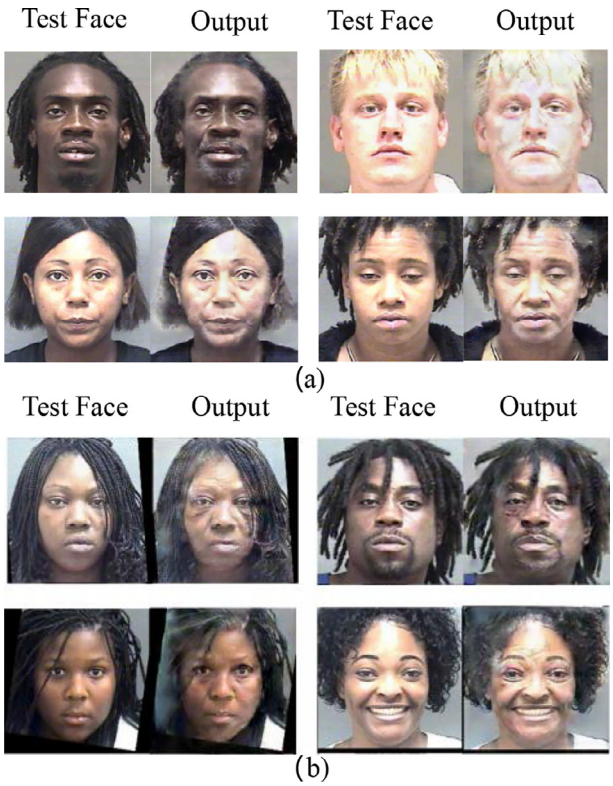


Fig. 1. Typical failure cases. For each pair of images, the left is the input face under 30 years old, and the right is the corresponding result over 50 years old.

2) We highlight the significance of the inner face and introduce a local discriminator to improve the visual fidelity of synthesized faces.
3) Extensive experiments on MORPH, CACD, and FG-NET are conducted to evaluate the proposed method. Both qualitative and quantitative experimental results demonstrate the effectiveness of our approach in synthesizing faces at desired ages with identity information being well-preserved.

2. Related work

2.1. Facial age synthesis

Existing facial age synthesis approaches can be roughly divided into physical model-based methods, prototype model-based methods, and deep learning-based methods.

Physical model-based methods [62,23,3,55,20,49] usually simulate face aging by modeling facial biological and physical mechanisms. As the earliest method, Todd et al. [49] introduced a revised cardioid-strain transformation to model facial appearance transition. Later, Wu et al. [55] proposed a dynamic model to simulate expressive wrinkles in 3D facial animation and skin aging. And Ramanathan et al. [40] adopted a shape transformation and an image gradient-based texture transformation towards modeling facial aging in adults. However, one drawback is that these methods are usually complex and require many images of the same person covering a long age span.

Prototype-based methods [43,39,56,40,18] usually divide faces into different age groups and calculate the average face of each age group as its prototype. After that, differences among prototypes are referred to as aging patterns. In [48], Tiddeman et al. adopted wavelet transform to capture features of texture details at multiple scales. In [18], Kemelmacher-Shlizerman et al. presented a prototype model-based method to capture differences in shape and texture under any desired illumination. However, one shortcoming of these methods is that using averaged faces as prototypes may eliminate personalized facial features, inclined to lose identity information. To well preserve personalized features, Shu et al. [45] proposed using paired residual dictionary learning to learn a set of age-specific dictionaries.

Recently, deep generative models [36,47,51,24,1] are used to solve these above problems. For example, Wang et al. [51] proposed a recurrent facial aging (RFA) framework based on a recursive neural network. This method can effectively eliminate ghost-artifacts presented in previous models. However, it requires the test data labeled with real ages to determine changes that need to be learned in the aging process. Antipov et al. [1], for the first time, introduced conditional generative adversarial networks (cGANs) [35] to synthesize faces with target ages. Zhang et al. [63] proposed a conditional adversarial autoencoder (CAAE) to achieve age progression and regression by traversing in a low-dimensional feature manifold. Yang et al. [57] designed multi-path discriminators to refine the results of facial age synthesis. And Li et al. [26] investigated spatial attention to constraining modifications to those regions highly relevant to face aging. Unlike these deep learning-based methods, we exploit channel attention to improve the latent feature representation's discrimination ability in this paper.

2.2. Generative adversarial networks

A typical generative adversarial network usually contains two modules: a generator G and a discriminator D. The generator G seeks to capture the real data distribution, while the discriminator D attempts to distinguish between real and fake data as much as possible. Since Ian Goodfellow et al. [13] proposed generative

adversarial networks (GANs) in 2014, they have been widely applied in various fields of computer vision, such as image super-resolution [52,60,22], style transfer [8,66], and facial synthesis [57,25,53,12,4]. Despite these successes, the training of GANs still suffers from some challenges, such as mode collapse and unstable training. Many efforts have been made to overcome these problems. For example, Conditional generative adversarial networks (cGANs) adds additional information to the generator and discriminator to guide the data generation process [35]. Wasserstein GAN (WGAN) [2], and subsequent WGAN with gradient penalty (WGAN-GP) [14] respectively adopted weight clipping and gradient penalty to constrain discriminator parameters. And the Least Squares GAN (LSGAN) [34] replaced the regular GAN loss with the least-squares loss, which can promote the generation of high-quality samples.

2.3. Attention mechanism

Attention plays a significant role in human perception [17]. Generally, the human visual system does not attempt to process the entire scene at one time. Instead, it selectively focuses on salient parts to capture visual structures [21]. Driven by the human perception process, attention attempts to learn from massive information selectively. It tries to select critical information and ignores useless information. Recently, many efforts have been made to apply attention mechanisms to various tasks [54,15,10,7,16,26,58,31,30]. In image recognition, Wang et al. [50] proposed a residual attention network built by stacking multiple attention modules to generate attention-aware features. Hu et al. [15] designed squeeze-and-excitation blocks by investigating the relationship between channels to improve the representation of CNN features. And [31,29,30] applied attention mechanisms to image set-based recognition. In semantic segmentation, the dual attention network (DANet) [10] summed the output of both the position attention module and the channel attention module to improve feature representation, which could contribute to more precise segmentation results. In facial expression synthesis, Pumarola et al. [38] exploited a spatial attention mechanism to make their facial expression synthesis model more robust to the change of background and illumination.

3. Methods

Considering that age-independent modification in the image context (e.g., background) may result in ghost artifacts, inspired by [38,26], we utilize a position attention mask regressed by the generator to restrict alternation within specific regions highly related to aging. However, different from [26], the proposed ACGAN further exploits channel attention to focus on informative features in channel dimensions. Apart from this, we employ a local discriminator to make better use of information in the face center. Note that only one model needs to be trained to address age progression and regression. Fig. 2 presents the overall framework of our proposed method.

3.1. Problem formulation

We define $\boldsymbol{x}_s$ whose age label is $\boldsymbol{c}_s$ as the input. Our goal is to train a generator G to learn the following translation: Given $\boldsymbol{x}_s$, we expect to generate a new face $\boldsymbol{x}_t$ with target age label $\boldsymbol{c}_t$ but from the same identity. In this paper, age information is encoded into a one-hot vector with the target age cluster being 1. And we divide all faces into $k = 4$ age clusters. Therefore, the age vector $\boldsymbol{c}_s, c_t \in \Re^{1\times k}$.

3.2. Attention-based generator

Most existing methods choose an hourglass-shaped full convolutional network as the generator [57,53,47]. The generator usually consists of two parts: an encoder G_{enc} and a decoder G_{dec}. The enco-

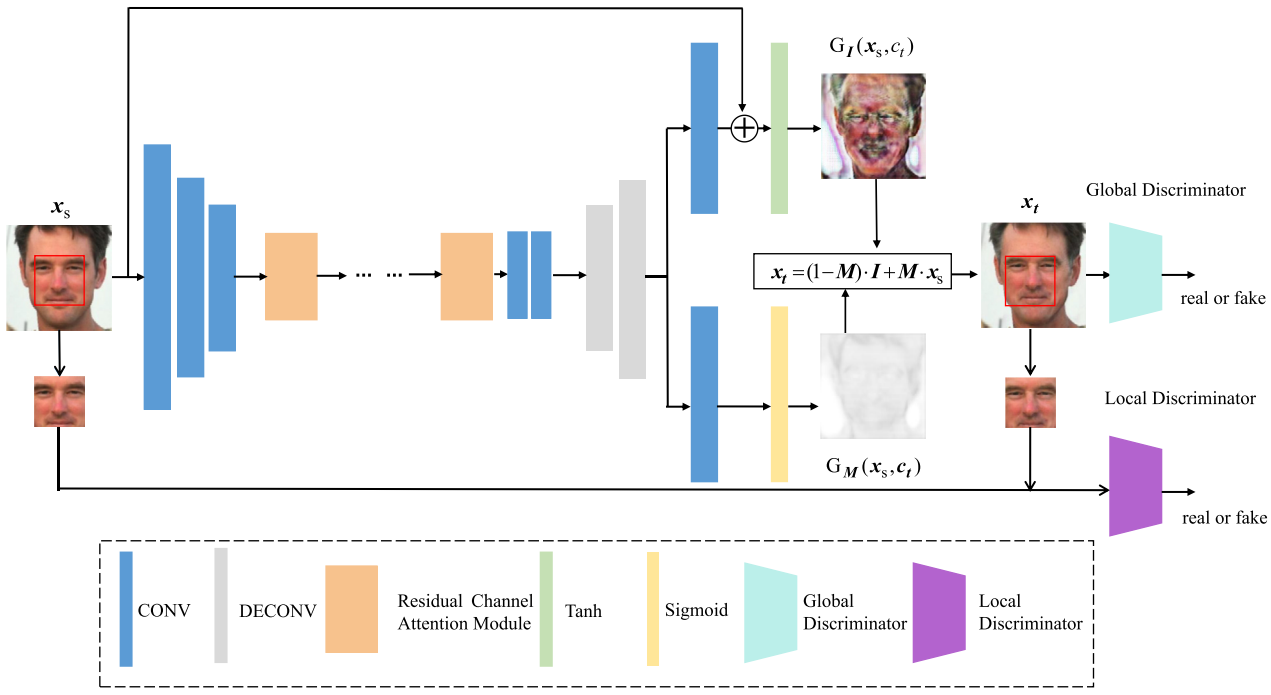


Fig. 2. Framework of ACGAN. Our network consists of a generator and two discriminators. The generator takes the face $\boldsymbol{x}_s$ and the target age label $\boldsymbol{c}_t$ as inputs and generates two outputs: an aged image I and an attention mask M.

der G_{enc} first encodes the input face $\boldsymbol{x}_s$ into a latent representation. The decoder G_{dec} then takes the latent representation and the target age label $\boldsymbol{c}_t$ as inputs to achieve age transformation [63]. Note that the target image $\boldsymbol{x}_t$ is generated step by step, and each step utilizes the information generated by the previous step. Consequently, the intermediate feature representation can largely affect the final output. However, previous work [57,53,47] usually treats each channel-wise feature equally, which may hinder the representation ability. Taking these into consideration, we investigate channel attention to tackle this issue.

This study proposes a novel attention-based generator by modeling the relationship between the latent space and age attribute. First, two convolutional layers with stride 2 are adopted for downsampling. Then, four residual channel attention modules followed by two convolutional layers with stride 1 constitute the bottleneck layers. Downsampling layers and bottleneck layers together establish the G_{enc}. Following IPC-GAN [53], we inject the age condition into the output of G_{enc} to guide the synthesis. Finally, two deconvolutional layers with stride 2 for upsampling build the decoder G_{dec}. To make the generator focus on modeling age differences, as [24] did, a shortcut connection between the input of G_{enc} and the output of G_{dec} is formed. The encoder G_{enc} and the decoder G_{dec} together form the auto-encoder architecture. And the network architecture of the generator is given in Table 1. More details can be found in the following.

3.2.1. Residual channel attention module

Previous work usually treats each channel-wise feature equally, which is not flexible for real cases. In order to make the network focus on more informative features, we exploit the interdependencies among feature channels. Designing an effective module to make the network focus on more informative features is a crucial step. On one hand, the information in the input space contains abundant low-frequency and high-frequency components that should be treated differently. On the other hand, convolutional layers with local receptive fields make subsequent layers unable to exploit the contextual information. Based on these analyses, we build a residual channel attention module to learn which components should be emphasized and suppressed. Specifically, we introduce channel attention to the residual block to capture feature dependencies in channel dimensions. Each channel of a feature map can be regarded as a feature detector. With channel attention, we can focus on more valuable channels and suppress useless features. As a result, our encoder can selectively extract features most relevant to age translations.

Both Fig. 3 and Table 2 describe the architecture of our residual channel attention module. We use $\boldsymbol{F}_{in} \in \Re^{h\times w\times c}$ to represent an intermediate feature representation, where h, w respectively denote the height and width of each feature map, and c represents the number of channels/feature maps. Thus, the final output of the residual channel attention module $\boldsymbol{F}_{out}$ can be formulated as:

$$\boldsymbol{F}_{out} = \boldsymbol{F}_{in} + \boldsymbol{U}. \tag{1}$$

To get $\boldsymbol{U}$, we first feed $\boldsymbol{F}_{in}$ into two convolutional layers whose kernel size is 3×3 and get $\boldsymbol{V} \in \Re^{h\times w\times c}$. And then, we calculate a channel attention map $\boldsymbol{Z} \in \Re^{1\times 1\times c}$ to adaptively rescale $\boldsymbol{V}$. This operation can be formulated as:

$$\boldsymbol{U} = \boldsymbol{Z} \otimes \boldsymbol{V}, \tag{2}$$

where $\boldsymbol{U} \in \Re^{h\times w\times c}$ and $\otimes$ denotes the element-wise multiplication. During multiplication, each value of $\boldsymbol{Z}$ is broadcasted along the spatial dimension to a $h \times w$ map.

To calculate $\boldsymbol{Z}$, we first squeeze the spatial dimension of $\boldsymbol{V}$. Two ways can be selected to aggregate the spatial information, i.e., average-pooling P_{avg} and max-pooling P_{max}. [65] suggests that average-pooling can be employed to learn the extent of the target

Table 1
Network architecture of the generator. There are some notations: N: the number of output channels, K: kernel size, S: stride size, P: padding size, IN: instance normalization, H, W: the height and width of the input image.

Part	Input	Layers Information	Output Shape
Downsampling	$(H, W, 3)$	CONV0-(N32, K9x9, S1, P3), IN, ReLU	$(H, W, 32)$
	$(H, W, 32)$	CONV1-(N64, K4x4, S2, P1), IN, ReLU	$(H/2, W/2, 64)$
	$(H/2, W/2, 64)$	CONV2-(N128, K4x4, S2, P1), IN, ReLU	$(H/4, W/4, 128)$
Bottleneck	$(H/4, W/4, 128)$	Residual Channel Attention Module1	$(H/4, W/4, 128)$
	$(H/4, W/4, 128)$	Residual Channel Attention Module4	$(H/4, W/4, 128)$
	$(H/4, W/4, 128)$	CONV3-(N128, K3x3, S1, P1), IN, ReLU	$(H/4, W/4, 128)$
	$(H/4, W/4, 128)$	CONV4-(N128, K3x3, S1, P1), IN, ReLU	$(H/4, W/4, 128)$
Upsampling	$(H/4, W/4, 128)$	DECONV1-(N64, K4x4, S2, P1), IN, ReLU	$(H/2, W/2, 64)$
	$(H/2, W/2, 64)$	DECONV2-(N32, K4x4, S2, P1), IN, ReLU	$(H, W, 32)$
Output Image Layer $\boldsymbol{I}$	$(H, W, 32)$	CONV5-(N3, K9x9, S1, P3), Tanh	$(H, W, 3)$
Output Attention Mask Layer $\boldsymbol{M}$	$(H, W, 32)$	CONV6-(N1, K4x4, S2, P1), Sigmoid	$(H, W, 1)$

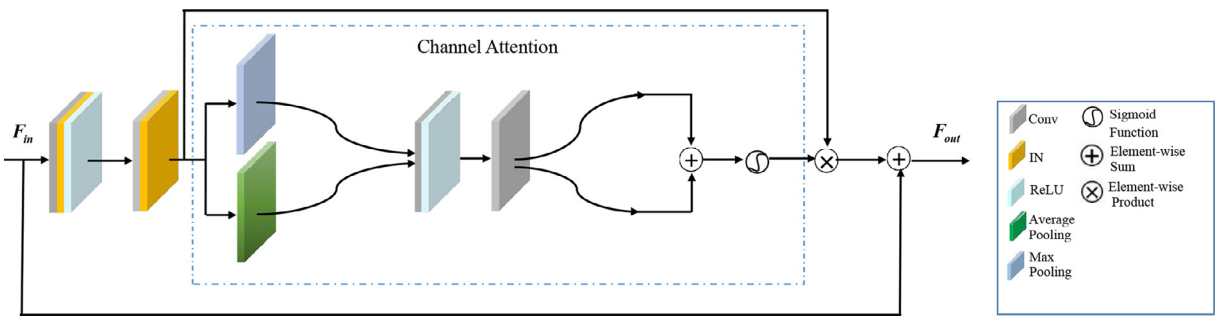


Fig. 3. Architecture of the channel attention module. $\oplus$ denotes matrix summation. $\otimes$ represents element-wise product. $\boldsymbol{F}_{in}$ and $\boldsymbol{F}_{out}$ are the input and output of the channel attention module, respectively.

Table 2
Channel attention module. There are some notations: h, w: the height and width of the input feature map, c: the number of channels, Avg-pool: an average-pooling operation, Max-pool: a max-pooling operation.

Part	Input	Layers Information	Output Shape
Body	(h, w, c)	CONV0-(N128, K3x3, S1, P1), IN, ReLU	(h, w, c)
	(h, w, c)	CONV1-(N128, K3x3, S1, P1), IN	(h, w, c)
Channel Attention Layer	(h, w, c)	Avg-pool	$(1, 1, c)$
	(h, w, c)	Max-pool	$(1, 1, c)$
	$(1, 1, c)$	CONV2-(N8, K1x1, S1, P0), ReLU	$(1, 1, c/16)$
	$(1, 1, c/16)$	CONV3-(N128, K1x1, S1, P0)	$(1, 1, c)$
	$(1, 1, c)$	Sigmoid	$(1, 1, c)$

object efficiently. And [54] argues that max-pooling can gather another information clue about distinctive object features. In our work, we utilize both operations. The aggregated information is then forwarded into a shared network, consisting of two convolutional layers, i.e., a channel-downscaling layer and a channel-upscaling layer. We use $\boldsymbol{W}_{down} \in \Re^{\frac{c}{r}\times c}$ and $\boldsymbol{W}_{up} \in \Re^{c\times\frac{c}{r}}$, where r is the reduction ratio, to respectively represent the weights of the channel-downscaling and channel-upscaling layers. We adopt $r = 16$ in this study. Next, we apply an element-wise summation. And the channel statistics $\boldsymbol{Z}$ are finally calculated as:

$$\begin{aligned}\boldsymbol{Z} = \quad & \sigma(\boldsymbol{W}_{up}(\boldsymbol{W}_{down}(P_{avg}(\boldsymbol{V})))) \\ & +\boldsymbol{W}_{up}(\boldsymbol{W}_{down}(P_{max}(\boldsymbol{V})))),\end{aligned} \tag{3}$$

where σ denotes the sigmoid function. Note that $\boldsymbol{W}_{down}$ is followed by a ReLU function.

3.2.2. Spatial attention module

To make the model robust to distracting environmental factors, such as background with complex textures, we add a new branch to G_{dec} to regress a position attention mask to localize specific regions for face editing. In this way, G_{dec} produces two outputs: one is a color mask $\boldsymbol{I}$, and the other is a position attention mask $\boldsymbol{M}$ to restrict the alternation of $\boldsymbol{I}$ within essential regions. The final output can be calculated by:

$$\boldsymbol{x}_t = (1 - \boldsymbol{M}) \cdot \boldsymbol{I} + \boldsymbol{M} \cdot \boldsymbol{x}_s, \tag{4}$$

where $\boldsymbol{M} = G_{\boldsymbol{M}}(\boldsymbol{x}_s, \boldsymbol{c}_t) \in \{0, \ldots, 1\}^{H\times W}$ and $\boldsymbol{I} = G_{\boldsymbol{I}}(\boldsymbol{x}_s, \boldsymbol{c}_t) \in \Re^{H\times W\times 3}$.

3.3. Global and local discriminators

We adopt a pair of discriminators to encourage generated images to be indistinguishable from real ones. Specifically, we first train a global discriminator D_{global} to determine the whole face's visual fidelity. There are two main reasons why we further introduce a local discriminator D_{local} in the age synthesis task: (1) The outer face often contains noise, which may interfere with the discriminator in extracting specific age features; (2) The inner face is more informative, gathering massive pure features, e.g., identity. The benefit of combining D_{global} and D_{local} is: D_{global} determines the authenticity globally, while D_{local} provides additional feedback to the generator to make generated textures within the face center more photo-realistic.

D_{global} and D_{local} have similar network structures. The only difference is that the input size of D_{local} is smaller than that of D_{global}. Thus, we employ one less convolutional layer. We take D_{global} as an example to describe network architecture details. A series of convolutional layers whose kernel size is 4 are employed to extract age-specific features. And each convolutional layer deploys a batch normalization for accelerated convergence except the last one. To ensure the age label of the generated face is consistent with $\boldsymbol{c}_t$, we add additional guidance to the discriminator. Following [53], we replicate the age vector and concatenate it to the output of the first convolutional layer. This strategy's effectiveness in promoting the visual quality of generated samples has been proved in [37]. Tables 3 and 4 present the network architectures of D_{global} and D_{local}, respectively.

Table 3
Network architecture of the global discriminator. There are some notations: BN: batch normalization, LReLU: Leaky ReLU.

Part	Input	Layers Information	Output Shape
Input Layer	$(H, W, 3)$	CONV0-(N64, K4x4, S2, P1), BN, LReLU	$(H/2, W/2, 64)$
Hidden Layer	$(H/2, W/2, 64 + k)$	CONV1-(N128, K4x4, S2, P1), BN, LReLU	$(H/4, W4, 128)$
	$(H/4, W/4, 128)$	CONV2-(N256, K4x4, S2, P1), BN, LReLU	$(H/8, W/8, 256)$
	$(H/8, W/8, 256)$	CONV3-(N512, K4x4, S2, P1), BN, LReLU	$(H/16, W/16, 512)$
	$(H/16, W/16, 512)$	CONV4-(N1024, K4x4, S2, P1), BN, LReLU	$(H/32, W/32, 1024)$
	$(H/32, W/32, 1024)$	CONV5-(N2048, K4x4, S2, P1), BN, LReLU	$(H/64, W/64, 2048)$
Output Layer	$(H/64, W/64, 2048)$	CONV6-(N1, K4x4, S2, P1)	$(H/128, W/128, 1)$

Table 4
Network architecture of the local discriminator.

Part	Input	Layers Information	Output Shape
Input Layer	$(H/2, W/2, 3)$	CONV0-(N64, K4x4, S2, P1), BN, LReLU	$(H/4, W/4, 64)$
Hidden Layer	$(H/4, W/4, 64 + k)$	CONV1-(N128, K4x4, S2, P1), BN, LReLU	$(H/8, W/8, 128)$
	$(H/8, W/8, 128)$	CONV2-(N256, K4x4, S2, P1), BN, LReLU	$(H/16, W/16, 256)$
	$(H/16, W/16, 256)$	CONV3-(N512, K4x4, S2, P1), BN, LReLU	$(H/32, W/32, 512)$
	$(H/32, W/32, 512)$	CONV4-(N1024, K4x4, S2, P1), BN, LReLU	$(H/64, W/64, 1024)$
Output Layer	$(H/64, W/64, 1024)$	CONV5-(N1, K4x4, S2, P1)	$(H/128, W/128, 1)$

3.4. Loss function

3.4.1. Adversarial loss

We add age conditions to the generator and the discriminator to generate photo-realistic faces with the expected aging effects. In other words, apart from distinguishing generated faces from real faces, our discriminators need to determine further whether the given sample is consistent with the given age label [41,53]. For each input $\boldsymbol{x}_s$, we randomly select $\boldsymbol{c}_t$ ($\boldsymbol{c}_t \neq \boldsymbol{c}_s$) as the target age label. The generated image is denoted as $\boldsymbol{x}_t$, i.e., $\boldsymbol{x}_t = G(\boldsymbol{x}_s, \boldsymbol{c}_t)$. Besides, we randomly select a face $\boldsymbol{f}_t$ from the target age group. $\boldsymbol{f}_t$ together with $\boldsymbol{c}_t$ forms the only positive sample of our discriminators. To improve the visual quality of synthesized faces, we employ the LSGAN loss [34] rather than the regular GAN loss. Our adversarial loss functions can be defined as follows:

$$L_{G_{global}} = E_{\boldsymbol{x}_s,c_t}[D_{global}(G(\boldsymbol{x}_s, \boldsymbol{c}_t), \boldsymbol{c}_t) - 1]^2, \tag{5}$$

$$L_{G_{local}} = E_{\boldsymbol{o}_{xt},c_t}[D_{local}(\boldsymbol{o}_{xt}, c_t) - 1]^2, \tag{6}$$

$$\begin{aligned} L_{D_{global}} = \; & E_{\boldsymbol{f}_t,c_t}\left[D_{global}(\boldsymbol{f}_t, \boldsymbol{c}_t) - 1\right]^2 \\ & + \tfrac{1}{2}(E_{\boldsymbol{x}_s,c_t}\left[D_{global}(G(\boldsymbol{x}_s, \boldsymbol{c}_t), c_t)\right]^2 \\ & + \tfrac{1}{2}(E_{\boldsymbol{f}_t,c_s}\left[D_{global}(\boldsymbol{f}_t, \boldsymbol{c}_s)\right]^2 \\ & + E_{\boldsymbol{x}_s,c_t}[D_{global}(\boldsymbol{x}_s, \boldsymbol{c}_t)]^2)), \end{aligned} \tag{7}$$

$$\begin{aligned} L_{D_{local}} = \; & E_{\boldsymbol{o}_{ft},c_t}[D_{local}(\boldsymbol{o}_{ft}, \boldsymbol{c}_t) - 1]^2 \\ & + \tfrac{1}{2}(E_{\boldsymbol{o}_{xt},c_t}[D_{local}(\boldsymbol{o}_{xt}, c_t)]^2 \\ & + \tfrac{1}{2}(E_{\boldsymbol{o}_{ft},c_s}[D_{local}(\boldsymbol{o}_{ft}, \boldsymbol{c}_s)]^2 \\ & + E_{\boldsymbol{o}_{xs},c_t}[D_{local}(\boldsymbol{o}_{xs}, \boldsymbol{c}_t)]^2)), \end{aligned} \tag{8}$$

where $\boldsymbol{o}_{ft}, \boldsymbol{o}_{xt}$ and $\boldsymbol{o}_{xs}$ respectively represent the central part of $\boldsymbol{f}_t, G(\boldsymbol{x}_s, \boldsymbol{c}_t)$, and $\boldsymbol{x}_s$. And λ_{local} is the penalty coefficient for the authenticity of the face center.

3.4.2. Identity loss

Without ground-truth supervision, we cannot ensure generated images well preserve the identity of inputs. Thus, it is necessary to introduce an identity loss. Considering that l_1-norm can encourage less ambiguous results [64], we choose it for defining our identity loss. The objective of identity loss L_{id} can be formulated as follows:

$$L_{id} = E_{\boldsymbol{x}_s,c_t,c_s}[||G(G(\boldsymbol{x}_s, \boldsymbol{c}_t), \boldsymbol{c}_s) - \boldsymbol{x}_s||_1]. \tag{9}$$

3.4.3. Attention loss

Minimizing the adversarial loss and identity loss can make the attention mask $\boldsymbol{M}$ easily saturate to 1, leading to $\boldsymbol{x}_s$ = G($\boldsymbol{x}_s, \boldsymbol{c}_t$). In such a case, the generator will have no effect. To prevent this, we regular the position attention mask $\boldsymbol{M}$ with a l_2-weight penalty. To further suppress artifacts, we perform a total variation regularization over $\boldsymbol{M}$. The attention loss thus takes the form as:

$$\begin{aligned} L_{att} = \; & \lambda_{tv} E_{\boldsymbol{x}_s,c_t}[\sum_{i,j}^{H,W}[(M_{i+1,j} - M_{i,j})^2 + (M_{i,j+1} - M_{i,j})^2]] \\ & + E_{\boldsymbol{x}_s,c_t}[||\boldsymbol{M}||_2], \end{aligned} \tag{10}$$

where λ_{tv} is a penalty coefficient, $\boldsymbol{M} = G_{\boldsymbol{M}}(\boldsymbol{x}_s, \boldsymbol{c}_t)$, and $\boldsymbol{M}_{i,j}$ is the i,j element of $\boldsymbol{M}$.

3.4.4. Overall loss function

The total loss function L is built by linearly combining all the losses mentioned above. It can be written as follows:

$$\begin{aligned} L_G &= \lambda_{gan}(L_{G_{global}} + \lambda_{local} L_{G_{local}}) + \lambda_{id} L_{id} + \lambda_{att} L_{att}, \\ L_D &= L_{D_{global}} + \lambda_{local} L_{D_{local}}, \end{aligned} \tag{11}$$

where λ_{gan}, λ_{local}, λ_{id} and λ_{att} are hyperparameters for balancing different losses.

4. Experiments

4.1. Dataset

Our experiment selects three datasets to test our method's performance.

MORPH [42], which includes approximately 13k individuals, contains 55k color images with age, gender, and ethnicity information. Images from MORPH are captured under uniform and moderate illumination with a simple background. And faces have near-frontal poses and neutral expressions.

CACD [5] contains 163k images of 2k celebrities, with ages ranging from 16 to 62. Compared to MORPH, images from CACD are obtained in more complex environments, showing variations in pose, illumination, make-up, and expression. Note that these images are downloaded from the internet through the Google search engine. Therefore, mismatches between faces and corresponding labels pose a challenge for facial age synthesis.

FG-NET [20] contains only 1,002 images from 82 people with ages ranging from 0 to 69. It is generally used for a fair comparison with previous work.

4.2. Implementation details

Faces are cropped and aligned to 256×256 according to five facial landmarks detected by multi-task cascaded convolutional networks (MTCNN) [59]. To further improve image generation and make our model more robust, we delete some unmatched images, such as age label mismatching and identity label mismatching. Finally, we get a total of 43,916 images from MORPH and 151,201 images from CACD. Following previous studies [25,57,32], we divide all faces into four age groups, i.e., 30-, 31–40, 41–50, and 51+. 4/5 of the data is randomly selected for training, and the rest is utilized for test. The number of faces in the training set and test set on MORPH and CACD is shown in Table 5.

All models are trained by choosing Adam [19] with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and learning rate $lr = 1 \times 10^{-4}$ to be the optimizer of the generator G and discriminator D. The batch size is set to 8. For data enhancement, we flip images horizontally with a probability of 0.5. We update the discriminator every two times we update the generator. As for trade-off parameters, we empirically set $\lambda_{gan} = 500, \lambda_{local} = 4, \lambda_{id} = 10, \lambda_{att} = 0.1$, and $\lambda_{tv} = 1e-4$. On MORPH and CACD, we train the models with 50,000 and 90,000 iterations, respectively. And it takes about 56 min per 10k

Table 5
Numbers of training and test images on MORPH and CACD.

	MORPH				CACD			
Age Group	30-	31–40	41–50	51+	30-	31–40	41–50	51+
Number of Training Images	13,106	11,553	8,408	2,419	37,461	31,537	30,088	21,961
Number of Test Images	3,007	3,689	2,123	611	9,436	8,671	6,743	5,304

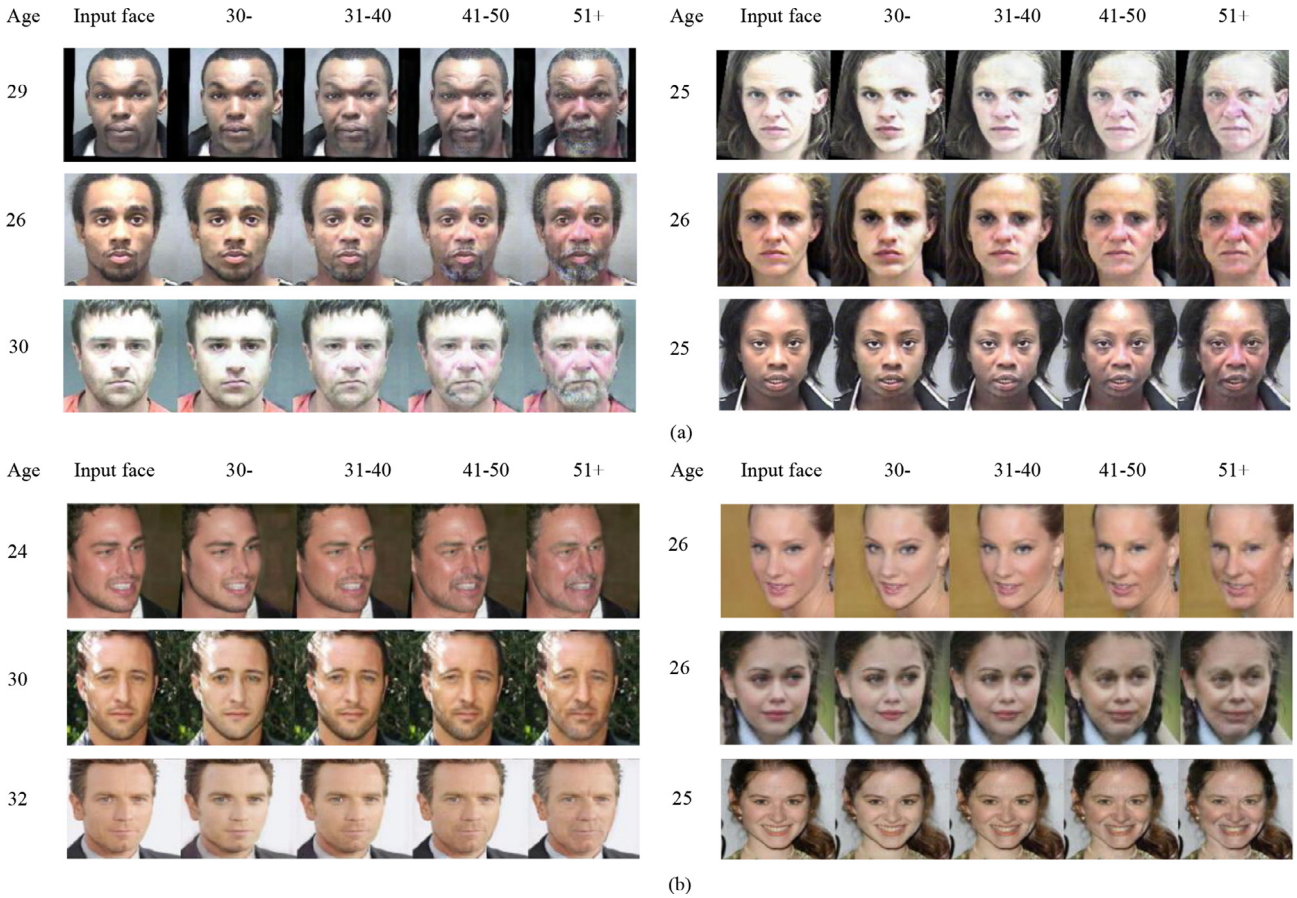


Fig. 4. Age progression results on MORPH and CACD. For each subject, the first column is the input face, and the remaining four columns denote the corresponding synthesized face under 30, 31–40, 41–50, and 51+ years old, respectively.

iterations. All experiments are conducted on an NVIDIA TITAN RTX GPU.

4.3. Qualitative results

Sample results are provided to make a qualitative evaluation. Besides, to show the superiority of our method, we further make a comparison with the existing methods.

Visual Fidelity: Age progression results are presented in Fig. 4, while age regression results are shown in Fig. 5. For each figure, (a) and (b) are the results on MORPH and CACD, respectively. For each subject, the first column is the input face, and the following four are synthesized faces corresponding to the age group 30-, 31–40, 41–50, and 51+ years old, respectively. As observed, from 30 to 50 years old, the main perceptible change is skin aging. Wrinkles become deeper, skin thinner and duller. And the hairline gradually rises. Besides, the hair may turn gray over 50. More results can be found in appendix.

Comparison with Prior Work: To make a fair comparison with prior work, we conduct testing on FG-NET with CACD as the training set. Six methods [45,51,63,57,25,24] are considered as the control methods. As shown in Fig. 6, ghost artifacts can be observed in the results of [45,51]. And faces synthesized by [63] are over-smooth. Compared with recent GAN-based approaches [57,25,24], our method can generate more subtle aging details (e.g., the hair of the woman in the first row gets gray). Overall, our method can improve the quality of the generated aging details while bringing less color distortion.

4.4. Quantitative results

We apply age progression on faces under or equal to 30 years old on MORPH and CACD, synthesizing their corresponding faces in the other three age groups, to quantitatively evaluate the performance of face aging. For the evaluation of visual fidelity, we resort to the inception score (i.e., IS) to quantitatively evaluate how real the synthesized faces are. The results are given in Table 6. We report both mean and standard deviation in Table 6. Note that the higher the inception score is, the more photo-realistic the synthesized faces are. From the results, we can conclude that our method can synthesize photo-realistic faces.

There are two underlying requirements of facial age synthesis, i.e., aging accuracy and identity permanence. Thus, we then calculate the metrics of aging accuracy and identity preservation. To make a fair comparison with the previous work [63,25,57,53], all metrics for quantitative analyses are computed based on the online face analysis tools Face++. Note that the results of [63,25] are directly reported in [32,57] in their paper, and [53] in [33].

Aging Accuracy: Following PAG-GAN [57], we estimate ages of both real faces and synthesized faces by Face++ to assess the performance on age translation accuracy. For each age group, the smaller the difference between the ages of synthesized faces and the ages of real faces is, the better the performance is. In other words, we expect a face aging model whose synthesized faces have ages close to real faces. To make a clearer comparison with other state-of-the-art methods, we further show the difference (i.e., age deviation from ages of actual faces) in Table 7. Following PAG-GAN, we use faces under or equal to 30 as test samples. And gen-

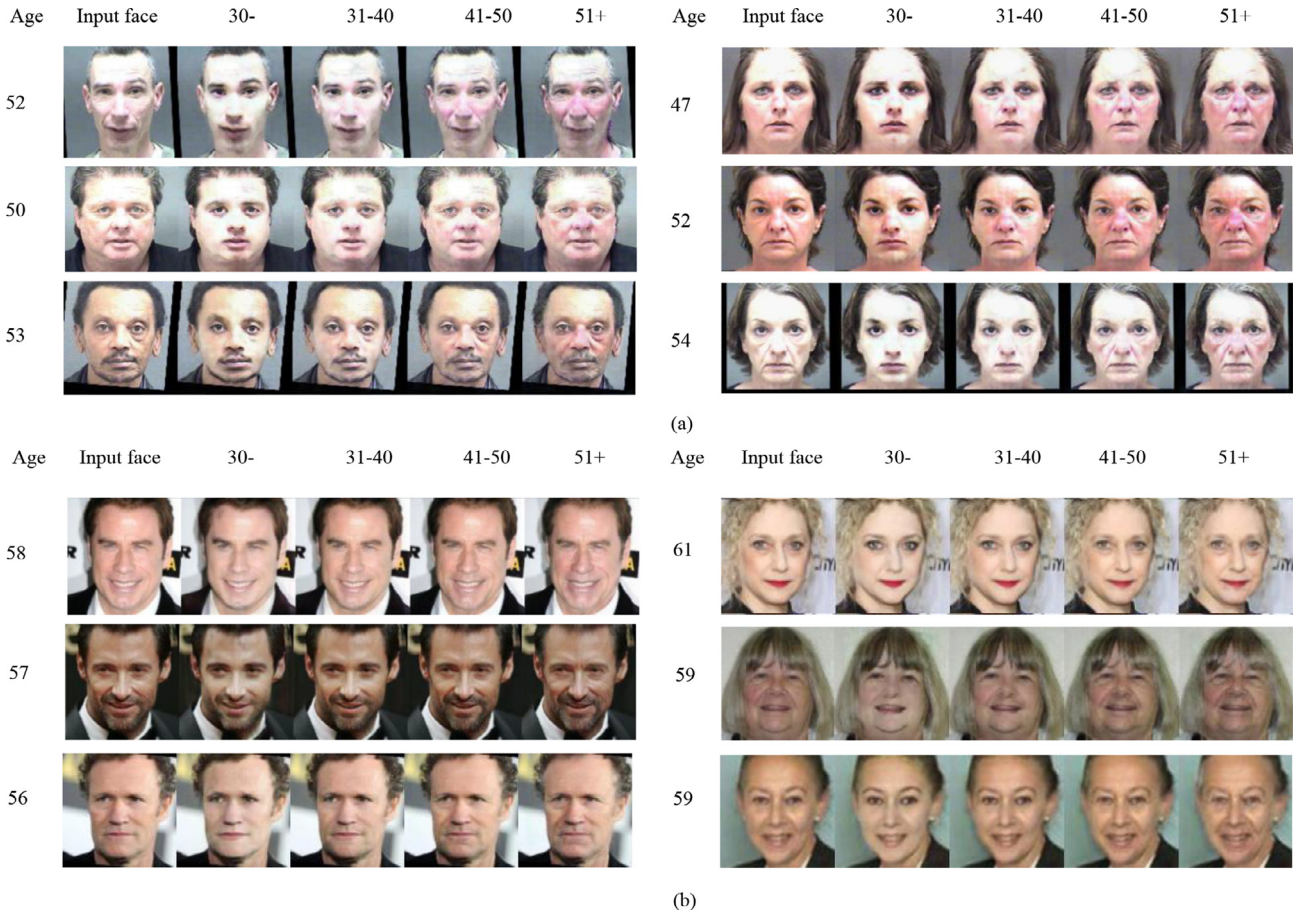


Fig. 5. Age regression results on MORPH and CACD. For each subject, the first column is the input face, and the remaining four columns denote the corresponding synthesized faces under 30, 31–40, 41–50, and 51+ years old, respectively.

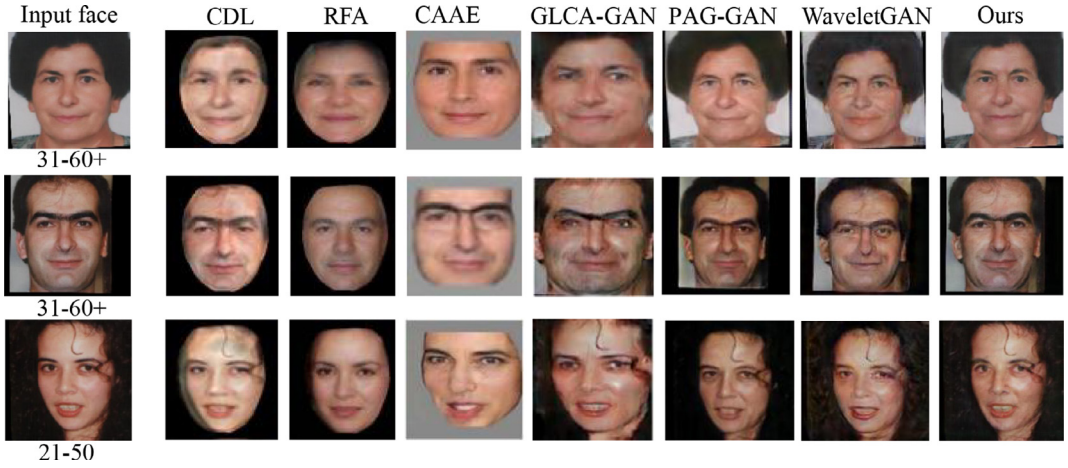


Fig. 6. Comparison with prior work CDL [45], RFA [51], CAAE [63], GLCA-GAN [25], PAG-GAN [57], WaveletGAN [24].

Table 6
Quantitative results computed with IS on MORPH and CACD.

	Input Faces (Real)	Aged Faces (Synthesized)		
Age Group	30-	31–40	41–50	51+
MORPH	2.31 ± 0.12	2.27 ± 0.09	2.19 ± 0.08	2.09 ± 0.07
CACD	2.90 ± 0.11	2.72 ± 0.10	2.64 ± 0.09	2.60 ± 0.08

Table 7
Age estimation results obtained by Face++ on MORPH and CACD.

	MORPH				CACD			
Age Group	30-	31–40	41–50	51+	30-	31–40	41–50	51+
Real Faces	28.69	38.80	47.75	58.14	30.32	38.04	46.51	54.07
Generated Faces	–	38.47	46.37	58.04	–	35.43	47.65	54.04

erated faces are their corresponding synthesized faces in the other three age groups (i.e., 31–40, 41–50, 51+). It can be seen in Table 7 that the estimated age distributions are well-matched with the real distributions in all age groups. We also compare the proposed method with the previous work [63,25,53,57] in terms of deviations between mean ages to validate the superiority. Table 8 shows the comparison results, in which the smallest and second smallest age deviations in each age group are marked in **red** and blue, respectively. We can see that our proposed method can achieve impressive results on MORPH. On the CACD database, PAG-GAN shows a little better performance than ours in aging effect generation. Their specially designed pyramid discriminator might contribute to this. However, this structure needs more training time. Taking the training on Morph as an example, our method takes approximately 5 h when running 50000 iterations, while PAG-GAN needs 8 h. In addition, we achieve better results than PAG-GAN in identity preservation on both MORPH and CACD, which is shown in Table 9. Finally, our approach does not need any pre-trained models, while PAG-GAN needs a pre-trained VGG model [46].

Identity Preservation: We conduct face verification experiments to measure the similarity between real faces and their corresponding synthesized faces in the other three age groups. There are two main evaluation metrics: the face verification confidence and face verification rate. The higher the face verification confidence is, the more similar the two faces are. Table 9 shows the face verification confidence results. As the time interval between two age groups increased, aging brings more changes in facial appearance naturally. Consequently, the verification performance decreases gradually. The higher the face verification rate is, the better the identity preservation is. We report TAR@FAR $= 1e-5$ with a threshold of 76.5 and compare our results with the previous work [63,25,53,57] in face verification rates. According to the results shown in Table 10, we can conclude that our method achieves the state-of-the-art performance in identity preservation. Note that [53] also performs well in preserving identity, but the signs of aging are not obvious enough.

4.5. Ablation study

We select faces below or equal to 30 years old on MORPH as testing samples, synthesizing faces in all age groups. Qualitative and quantitative studies are performed to explore the contributions of the proposed components.

Table 8
Age deviation from ages of real faces on MORPH and CACD (in absolute value). The smallest and the second smallest age deviation are respectively marked in **red** and blue.

	MORPH			CACD		
Age group	31-40	41-50	51+	31-40	41-50	51+
CAAE [63]	10.08	15.49	21.42	5.76	11.53	17.93
GLCA-GAN [25]	**0.23**	3.61	9.61	1.72	2.07	2.85
IPC-GAN [53]	2.15	1.87	1.62	**0.46**	1.22	2.08
PAG-GAN [57]	0.38	**0.52**	1.48	0.70	**0.22**	0.57
Ours	0.33	1.38	**0.10**	2.61	1.14	**0.03**

Table 9
Face verification confidence obtained by Face++ on MORPH and CACD.

	MORPH			CACD		
Age Group	31–40	41–50	51+	31–40	41–50	51+
30-	94.36	93.00	87.04	94.09	91.60	91.26
31–40	–	95.09	93.54	–	95.43	93.16
41–50	–	–	92.42	–	–	94.09

Table 10
Face verification rate (%) obtained by Face++ on MORPH and CACD (threshold = 76.5,FAR = 1e−5). Better results are in bold.

	MORPH			CACD		
Age Group	31–40	41–50	51+	31–40	41–50	51+
CAAE [63]	15.07	12.02	8.22	4.66	3.41	2.40
GLCA-GAN [25]	97.66	96.67	91.85	97.72	94.18	92.29
IPC-GAN [53]	**100.00**	**100.00**	**99.48**	**100.00**	97.95	97.36
PAG-GAN [57]	**100.00**	98.91	93.09	99.99	99.91	98.28
Ours	**100.00**	**100.00**	97.80	99.98	**99.96**	**99.56**

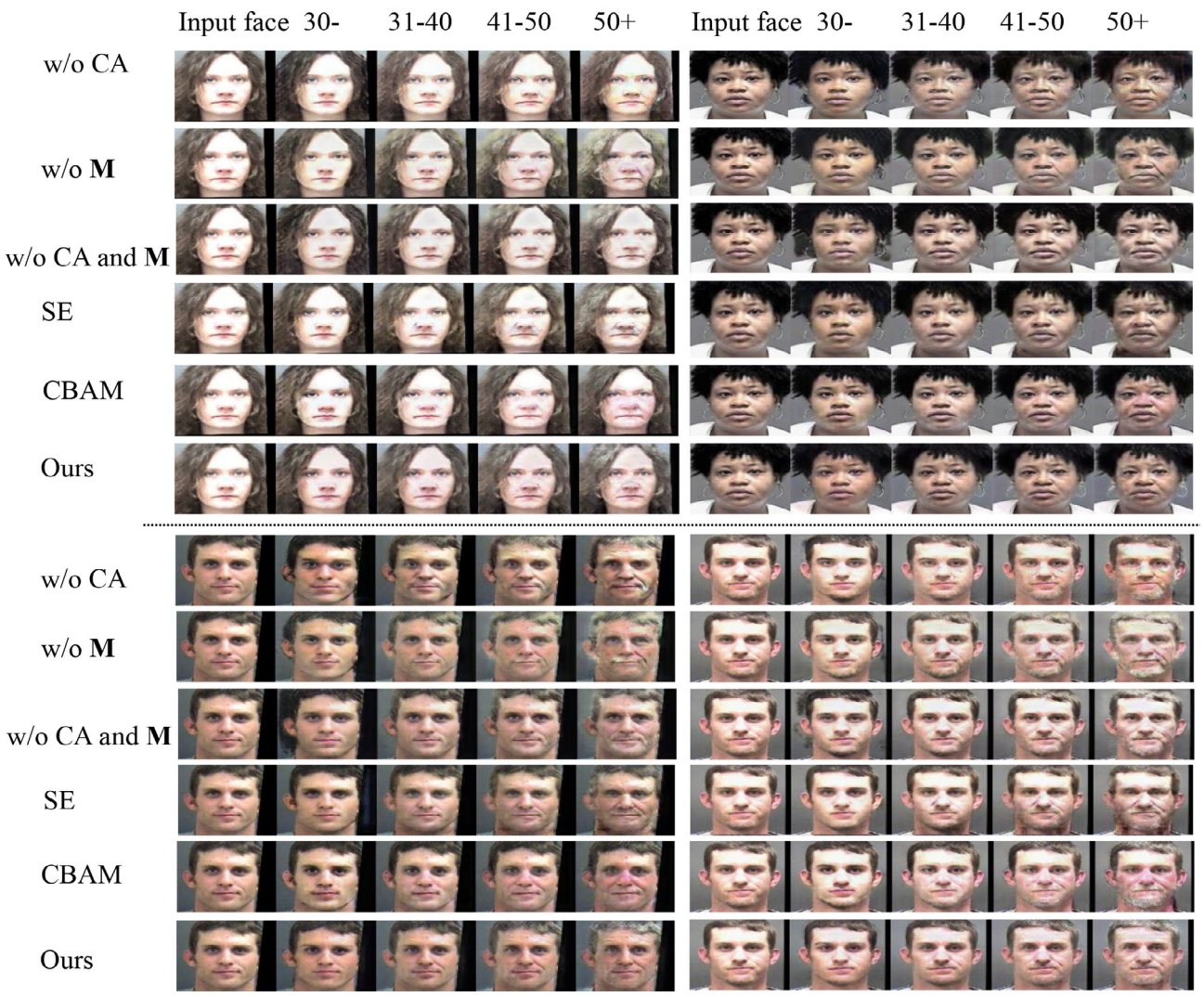


Fig. 7. Ablation study results of attention modules on MORPH. There are some notations: CA: channel attention, **M**: the attention mask $\boldsymbol{G}_M$, SE: the Squeeze and Exitation block [15], CBAM: the convolution block attention module [54].

Table 11
Age deviation from ages of real faces (in absolute value) obtained by Face++ on MORPH. There are some notations: CA: channel attention modules, M: the position attention mask ***M***, Local ***D***: the local discriminator.

Age Group	30-	31–40	41–50	51+
w/o CA	3.05	1.59	2.33	1.22
w/o M	3.22	2.38	0.90	4.88
w/o CA and M	3.38	7.27	8.52	9.34
SE [15]	5.01	6.17	2.27	0.72
CBAM [54]	5.28	3.59	**0.20**	0.42
w/o Local *D*	4.50	7.59	7.20	5.60
w/o shortcut connection	5.88	6.15	4.01	2.70
w/o identity loss	**0.33**	1.46	3.86	8.58
w/o age condition to *D*	14.58	5.40	2.95	13.02
Ours	0.35	**0.33**	1.38	**0.10**

4.5.1. Contribution of attention modules

We mainly analyze the contribution of the proposed attention modules from two aspects. On one hand, we exploit the gain of attention modules. On the other hand, we replace our attention modules with previous channel attention modules and compare the results to verify the proposed method's superiority. Visual illustrations are given in Fig. 7. For each subject, the first column is the input face whose age is 30-, and the following four are the synthesized counterpart in the age group 30-, 31–40, 41–50, and 51+, respectively. Corresponding quantitative results are reported in Table 11.

We experimentally explore the gain of attention modules. Without the channel attention module, the texture of the synthesized is over-smooth. And undesired freckles appear near cheeks. Without the attention mask, even though changes in hair, beard, and pouch are obvious, ghost artifacts might appear either. Without these two attention modules, synthesized faces may be blurred. In contrast, our model can generate the most visually

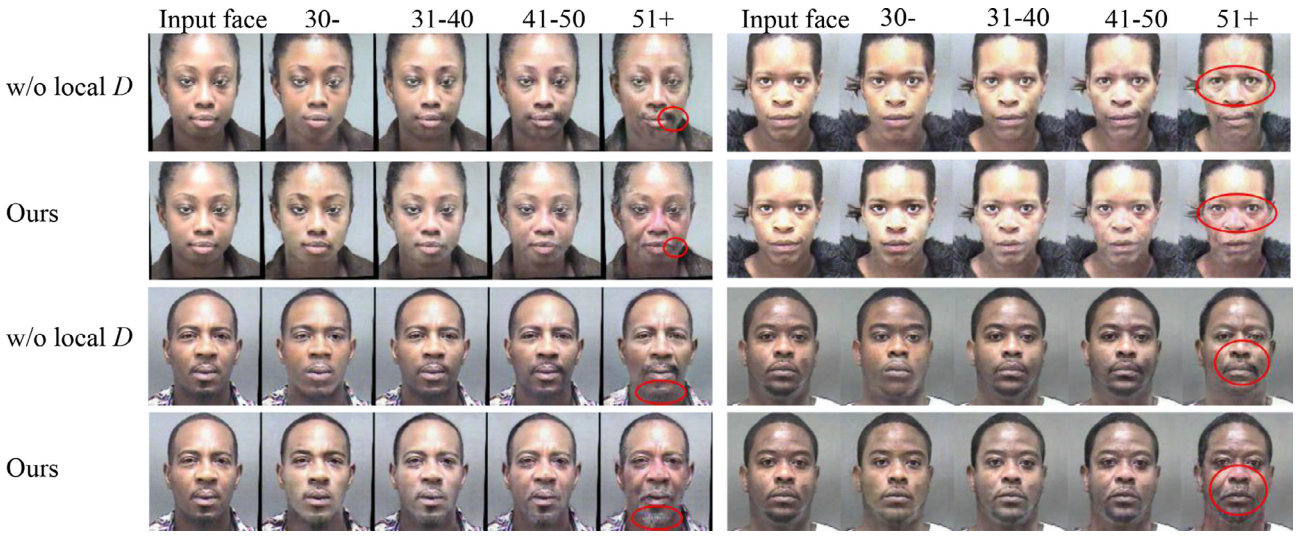


Fig. 8. Ablation study results of the local discriminator on MORPH. There is a notation: local *D*: the local discriminator.

Table 12
Age deviation from the ages of real faces (in absolute value) obtained by Face++ on MORPH under different values of λ_{att}.

Age Group	30-	31–40	41–50	51+
0.01	4.36	4.66	4.71	4.78
0.1 (selected)	**0.35**	**0.33**	1.38	**0.10**
1	3.17	2.53	**0.36**	3.37

Table 13
Age deviation from the ages of real faces (in absolute value) obtained by Face++ on MORPH under different values of λ_{local}.

Age Group	30-	31–40	41–50	51+
0.4	2.62	6.83	5.61	7.34
4 (selected)	**0.35**	**0.33**	**1.38**	**0.10**
40	1.75	8.49	9.23	3.99

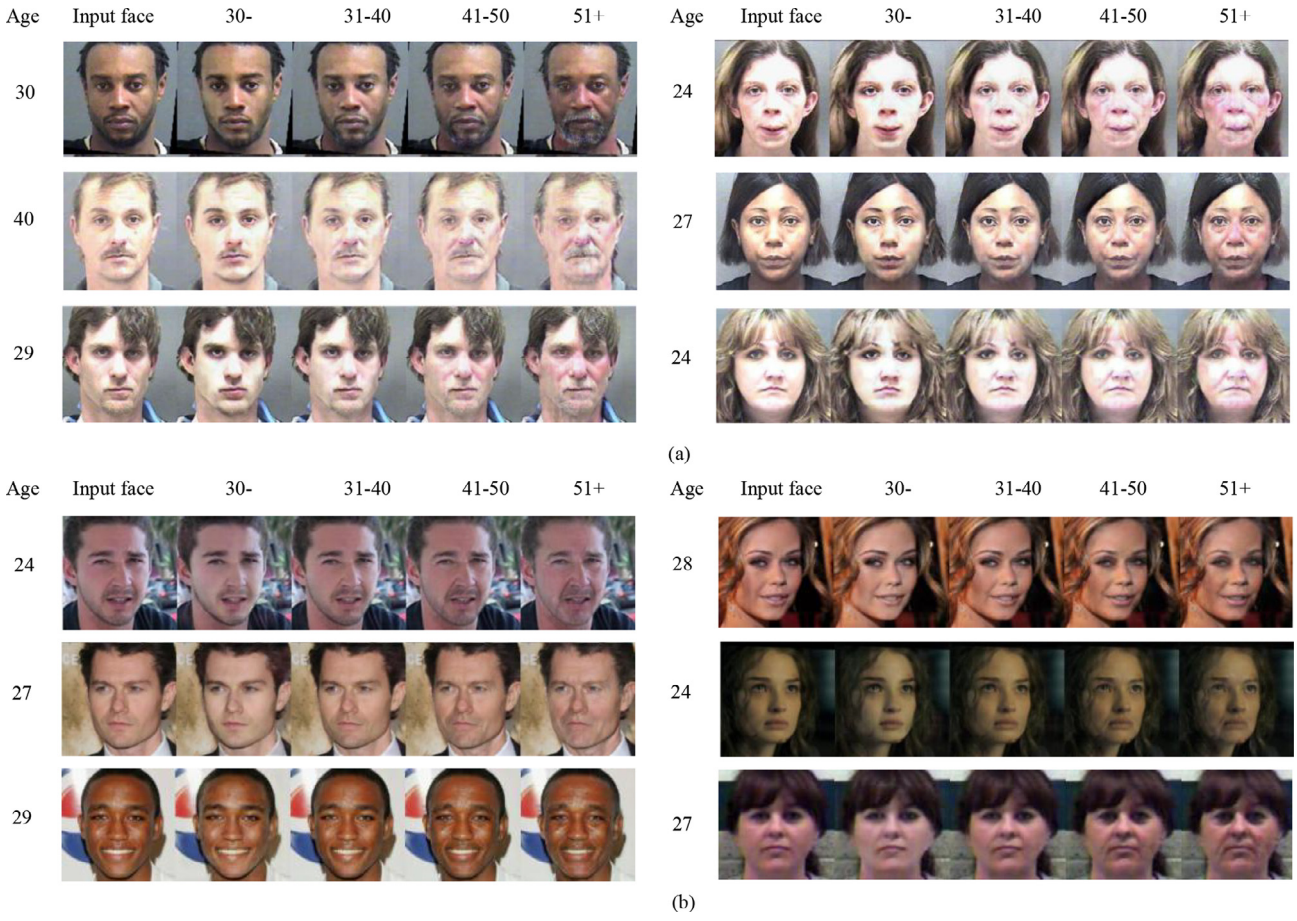


Fig. 9. Age progression results on MORPH and CACD. For each subject, the first column is the input face, and the remaining four columns denote the corresponding synthesized faces under 30, 31–40, 41–50, and 51+ years old, respectively.

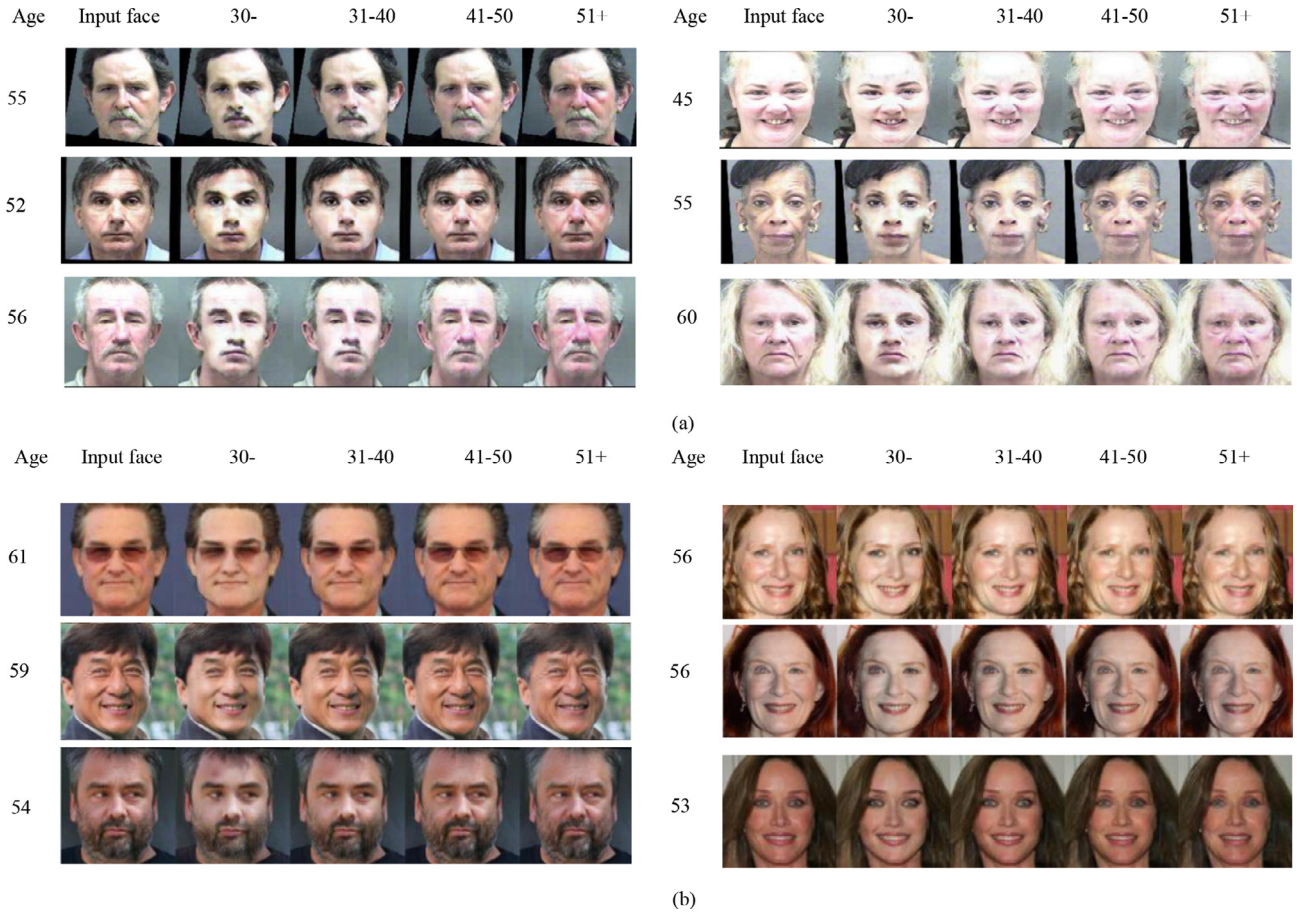


Fig. 10. Age regression results on MORPH and CACD. For each subject, the first column is the input face, and the remaining four columns denote the corresponding synthesized faces under 30, 31–40, 41–50, and 51+ years old, respectively.

plausible results. It can illustrate that both attention modules can encourage the aging effect generation, making the synthesized faces more visual plausible. And the same conclusion can be drawn from Table 11.

To further verify the superiority of our proposed attention module, the proposed attention modules are replaced by previous channel attention modules, such as the Squeeze and Excitation block (SE) [15] and the convolution block attention module (CBAM) [54]. On one hand, we replace the channel attention module with SE. As can be seen, synthesized results are inaccurate (much younger than expected). Notably, the textures of the synthesized faces in the age group 31–40 are over-smooth. Furthermore, the aged face of the fourth group seems to have lost the personality feature. On the other hand, we replace the proposed attention module with CBAM. Although signs of aging are obvious, from Table 11, quantitative results show that aging details in the 1^{st} two age groups are wiped out. Our results are better than these above attention modules in visual quality and aging accuracy. The reason may be that, based on SE, our channel attention modules further consider the max-pooling to enhancement discriminative features representation capability. Besides, the attention mask regressed by the generator can also make the network focus on regions highly related to aging while introducing a few additional parameters.

4.5.2. Contribution of the local discriminator

In theory, the inner face contains more informative features than the outer face. Thus, D_{local} can force the synthesized face to be more realistic and have expected aging characteristics within the central face region. We investigate the impact of including/excluding the local discriminator on the proposed method. We remove the local discriminator D_{local} by setting $\lambda_{local} = 0$ and keeping the rest of the settings unchanged. Visual comparison results are shown in Fig. 8. To highlight these differences, we mark the region where the contrast is clear with a red circle. It can be seen that, when D_{local} is excluded, some important details (e.g., identity) may be lost. Table 11 shows the quantitative results. Without D_{local}, the method achieves much lower age translation accuracy than our proposed method. According to these qualitative and quantitative results, we can conclude that the local discriminator can encourage a more accurate age translation.

4.5.3. Contribution of other components

Note that the ablation study about attention mechanisms and the local discriminator has already been conducted. In the revised manuscript, we investigate the contribution of other components (i.e., shortcut connection between $\boldsymbol{x}_s$ and $\boldsymbol{G}_l$, identity loss, and age condition to the discriminator D) on ACGAN. Quantitative results are presented in Table 11. According to the table, the aging effect of synthesized faces gets inaccurate if removing any one of them. On the contrary, our proposed method achieves better performance.

4.6. Analysis of hyperparameters

Note that λ_{gan}, λ_{tv} and λ_{id} are the hyperparameters that commonly used in previous GAN-based facial age synthesis work. We thus examine only λ_{att} and λ_{local} (i.e., our proposed modules) to

see whether the performance is sensitive to them. Specifically, we set different values for both parameters and assess the performance using aging accuracy. We report the results in Tables 12 and 13. From the experimental results, we can see that the performance is sensitive to λ_{att} and λ_{local}.

5. Conclusion and future directions

In this paper, we propose an attention-aware conditional generative adversarial network (ACGAN) for facial age synthesis. Specifically, channel attention is first employed to enhance the discrimination ability of the latent feature representation. A position attention mask is then regressed, which allows the network to focus on regions highly relevant to face aging. Apart from the regular global discriminator, a local discriminator is employed. It can encourage the generator to generate rich details within the face center region. And experimental results on MORPH, CACD, and FG-NET confirm the effectiveness of the proposed method in facial age synthesis.

Although the proposed method can synthesize faces with photo-realistic aging effects, there is room for improvement. These methods can not well handle young face aging where craniofacial growth and development are more dominant than skin aging. Thus, children aging will be our main research direction in the future. Besides, since we do not consider the interrelationship among ages/age groups, our proposed approach is vulnerable to noisy age labels. Inspired by [28], we would take this issue as a consideration to further reduce age gaps.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the anonymous reviewers for spending time on our work. This work was supported by the National Natural Science Foundation of China under Grant 62076131, Grant 62072245, and Grant 62076240.

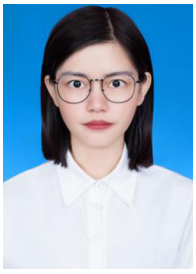
Appendix A

More experimental results on MORPH and CACD are listed in Figs. 9 and 10. For each figure, (a) and (b) are the results on MORPH and CACD, respectively. For each subject, the first column is the input face, and the following four are synthesized faces in the age group 30-, 31–40, 41–50, and 51+ years old, respectively.

References

[1] G. Antipov, M. Baccouche, J.L. Dugelay, Face aging with conditional generative adversarial networks, in: 2017 IEEE International Conference on Image Processing, IEEE, 2017, pp. 2089–2093.
[2] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, 2017, pp. 214–223.
[3] D.M. Burt, D.I. Perrett, Perception of age in adult caucasian male faces: Computer graphic manipulation of shape and colour information, The Royal Society London (1995) 137–143.
[4] J. Cao, Y. Hu, B. Yu, R. He, Z. Sun, 3d aided duet gans for multi-view face image synthesis, IEEE Transactions on Information Forensics and Security 14 (2019) 2028–2042.
[5] B.C. Chen, C.S. Chen, W.H. Hsu, Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset, IEEE Transactions on Multimedia 17 (2015) 804–815.
[6] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.S. Chua, Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5659–5667.
[7] L.C. Chen, Y. Yang, J. Wang, W. Xu, A.L. Yuille, Attention to scale: Scale-aware semantic image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3640–3649.
[8] Y. Choi, M. Choi, M. Kim, J.W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.
[9] J. Deng, S. Cheng, N. Xue, Y. Zhou, S. Zafeiriou, Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7093–7102.
[10] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.
[11] Y. Fu, G. Guo, T.S. Huang, Age synthesis and estimation via faces: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2010) 1955–1976.
[12] Z. Geng, C. Cao, S. Tulyakov, 3d guided fine-grained face manipulation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9821–9830.
[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, 2017, pp. 5767–5777.
[15] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[16] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 603–612.
[17] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis & Machine Intelligence (1998) 1254–1259.
[18] I. Kemelmacher-Shlizerman, S. Suwajanakorn, S.M. Seitz, Illumination-aware age progression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3334–3341.
[19] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980..
[20] A. Lanitis, C.J. Taylor, T.F. Cootes, Toward Automatic Simulation of Aging Effects on Face Images, IEEE, 2002, pp. 442–455.
[21] H. Larochelle, G.E. Hinton, Learning to combine foveal glimpses with a third-order boltzmann machine, in: Advances in Neural Information Processing Systems, 2010, pp. 1243–1251.
[22] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.
[23] F.R. Leta, A. Conci, D. Pamplona, I. Itanguy, Manipulating facial appearance through age parameters, in: Proc. Ninth Brazilian Symp. Computer Graphics and Image Processing, 1996, pp. 167–172..
[24] P. Li, Y. Hu, R. He, Z. Sun, Global and local consistent wavelet-domain age synthesis, IEEE Transactions on Information Forensics and Security 14 (2019) 2943–2957.
[25] P. Li, Y. Hu, Q. Li, R. He, Z. Sun, Global and local consistent age generative adversarial networks, in: 2018 24th International Conference on Pattern Recognition, IEEE, 2018, pp. 1073–1078.
[26] Q. Li, Y. Liu, Z. Sun, Age progression and regression with spatial attention modules, 2019, arXiv preprint arXiv:1903.02133..
[27] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, S. Wen, Stgan: A unified selective transfer network for arbitrary image attribute editing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3673–3682.
[28] X. Liu, F. Fan, L. Kong, Z. Diao, W. Xie, J. Lu, J. You, Unimodal regularized neuron stick-breaking for ordinal classification, Neurocomputing (2020).
[29] X. Liu, Z. Guo, S. Li, P. Jia, L. Kong, J. You, B.V.K.V. Kumar, Permutation-invariant feature restructuring for correlation-aware image set-based recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4986–4996.
[30] X. Liu, Z. Guo, J. You, B.V.K.V. Kumar, Dependency-aware attention control for image set-based face recognition, IEEE Transactions on Information Forensics and Security 15 (2020) 1501–1512.
[31] X. Liu, B. Vijaya Kumar, C. Yang, Q. Tang, J. You, Dependency-aware attention control for unconstrained face recognition with image sets, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 548–565.
[32] Y. Liu, Q. Li, Z. Sun, Attribute-aware face aging with wavelet-based generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11877–11886.
[33] Y. Liu, Q. Li, Z. Sun, T. Tan, A3gan: An attribute-aware attentive generative adversarial network for face aging, 2019, arXiv preprint arXiv:1911.06531..
[34] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2794–2802.

[35] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014, arXiv preprint arXiv:1411.1784..
[36] S. Palsson, E. Agustsson, R. Timofte, L. Van Gool, Generative adversarial style transfer networks for face aging, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 2084–2092.
[37] G. Perarnau, J. Van De Weijer, B. Raducanu, J.M. Álvarez, Invertible conditional gans for image editing, 2016, arXiv preprint arXiv:1611.06355..
[38] A. Pumarola, A. Agudo, A.M. Martinez, A. Sanfeliu, F. Moreno-Noguer, Ganimation: Anatomically-aware facial animation from a single image, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 818–833.
[39] N. Ramanathan, R. Chellappa, Modeling age progression in young faces, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE, 2006, pp. 387–394..
[40] N. Ramanathan, R. Chellappa, Modeling shape and textural variations in aging faces, in: 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, IEEE, 2008, pp. 1–8..
[41] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, 2016, arXiv preprint arXiv:1605.05396..
[42] K. Ricanek, T. Tesafaye, Morph: A longitudinal image database of normal adult age-progression, in: 7th International Conference on Automatic Face and Gesture Recognition (FGR06), IEEE, 2006, pp. 341–345..
[43] D.A. Rowland, D.I. Perrett, Manipulating facial appearance through shape and color, IEEE Computer Graphics and Applications 15 (1995) 70–76.
[44] M.M. Sawant, K.M. Bhurchandi, Age invariant face recognition: a survey on facial aging databases, techniques and effect of aging, Artificial Intelligence Review 52 (2019) 981–1008.
[45] X. Shu, J. Tang, H. Lai, L. Liu, S. Yan, Personalized age progression with aging dictionary, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3970–3978.
[46] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556..
[47] J. Song, J. Zhang, L. Gao, X. Liu, H.T. Shen, Dual conditional gans for face aging and rejuvenation, in: IJCAI, 2018, pp. 899–905..
[48] B. Tiddeman, M. Burt, D. Perrett, Prototyping and transforming facial textures for perception research, IEEE Computer Graphics and Applications 21 (2001) 42–50.
[49] J.T. Todd, L.S. Mark, R.E. Shaw, J.B. Pittenger, The perception of human growth, Scientific American 242 (1980) 132–145.
[50] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.
[51] W. Wang, Z. Cui, Y. Yan, J. Feng, S. Yan, X. Shu, N. Sebe, Recurrent face aging, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2378–2386.
[52] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy, Esrgan: Enhanced super-resolution generative adversarial networks, in: Proceedings of the European Conference on Computer Vision, 2018.
[53] Z. Wang, X. Tang, W. Luo, S. Gao, Face aging with identity-preserved conditional generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7939–7947.
[54] S. Woo, J. Park, J.Y. Lee, I. So Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 3–19.
[55] Y. Wu, N.M. Thalmann, D. Thalmann, A plastic-visco-elastic model for wrinkles in facial animation and skin aging, in: Fundamentals of Computer Graphics, World Scientific, 1994, pp. 201–213.
[56] Y. Wu, N.M. Thalmann, D. Thalmann, A dynamic wrinkle model in facial animation and skin ageing, The Journal of Visualization and Computer Animation 6 (1995) 195–205.
[57] H. Yang, D. Huang, Y. Wang, A.K. Jain, Learning face age progression: A pyramid architecture of gans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 31–39.
[58] G. Zhang, M. Kan, S. Shan, X. Chen, Generative adversarial network with spatial attention for face attribute editing, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 417–432.
[59] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Processing Letters 23 (2016) 1499–1503.
[60] W. Zhang, Y. Liu, C. Dong, Y. Qiao, Ranksrgan: Generative adversarial networks with ranker for image super-resolution, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3096–3105.
[61] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 286–301.
[62] Y. Zhang, T. Sim, Realistic and efficient wrinkle simulation using an anatomy-based face model with adaptive refinement, in: International 2005 Computer Graphics, IEEE, 2005, pp. 3–10.
[63] Z. Zhang, Y. Song, H. Qi, Age progression/regression by conditional adversarial autoencoder, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5810–5818.
[64] H. Zhao, O. Gallo, I. Frosio, J. Kautz, Loss functions for image restoration with neural networks, IEEE Transactions on Computational Imaging 3 (2016) 47–57.
[65] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[66] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.



Xiahui Chen is a Master Candidate at the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. Her research interests include biometrics, pattern recognition, and computer vision.

Yunlian Sun received the M.E. degree in computer science and technology from the Harbin Institute of Technology, China, in 2010, and the Ph.D. degree in ingegneria elettronica, informatica e delle telecomunicazioni from the University of Bologna, Italy, in 2014. She is currently an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. Her research interests include biometrics, pattern recognition, and computer vision.

Xiangbo Shu received the Ph.D. degree from the Nanjing University of Science and Technology, China, in 2016. He is currently an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. From 2014 to 2015, he worked as a Visiting Scholar at the Department of Electrical and Computer Engineering, National University of Singapore. His research interests include computer vision, multimedia computing, and deep learning. He is a member of the ACM and CCF. He has received the Excellent Doctoral Dissertation of CAAI, the Excellent Doctoral Dissertation of Jiangsu Province, the Best Student Paper Award in MMM2016, and the Best Paper Runner-up in ACM MM 2015.

Qi Li received the B.E. degree in automation from China University of Petroleum, Qingdao, China, in 2011, the Ph.D. degree in pattern recognition and intelligent systems from National Laboratory of Pattern Recognition, CASIA, Beijing, China, in 2016. He is currently an Associate Professor in the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, CASIA. His research interests include face recognition, computer vision and machine learning.