Facial Age and Expression Synthesis Using Ordinal Ranking Adversarial Networks

Yunlian Sun, Jinhui Tang, *Senior Member, IEEE*, Zhenan Sun, and Massimo Tistarelli

***Abstract*— Facial image synthesis has been extensively studied, for a long time, in both computer graphics and computer vision. Particularly, the synthesis of face images with varying ages, expressions and poses has received an increasing attention owing to several real-world applications. In this paper, facial age and expression synthesis are addressed. While previous and current research papers on facial age synthesis mostly adopt an age span of 10 years, this paper investigates face aging with a shorter time span. For expression synthesis, given a neutral face, we work on synthesizing faces with varying expression intensities (e.g., from zero to high). Note that both human ages and expression intensities are inherently ordinal. To fully exploit this ordinal nature, we devise ordinal ranking generative adversarial networks (ranking GAN). For each face, a one-hot label is assigned to define its age range/expression intensity. By exploiting the relative order information among age ranges/expression intensities, a binary ranking vector is further computed for each face. In ranking GAN, one-hot labels are used as the condition of the generator for synthesizing faces with target age groups/expression intensities. Moreover, we add a sequence of cost-sensitive ordinal rankers on top of several multi-scale discriminators, with the aim of minimizing age/intensity rank estimation loss when optimizing both the generator and discriminators. In order to evaluate the proposed ranking GAN, extensive experiments are carried out on several public face databases. As demonstrated by the experimental testing, this ranking scheme performs well even when the amount of available labeled training data is limited. The reported experimental results well demonstrate the effectiveness of ranking GAN on synthesizing face aging sequences and faces with varying expression intensities.**

***Index Terms*— Face image aging, facial expression synthesis, generative adversarial networks, ordinal ranking.**

Manuscript received July 8, 2019; revised December 14, 2019 and February 25, 2020; accepted March 2, 2020. Date of publication March 16, 2020; date of current version March 27, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001001, in part by the National Natural Science Foundation of China under Grant 61603391, Grant 61925204, and Grant 61427811, and in part by the Italian Ministry of Research under Grant PRIN 2015 and Grant SPADA. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Domingo Mery. *(Corresponding author: Jinhui Tang.)*

Yunlian Sun and Jinhui Tang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: yunlian.sun@njust.edu.cn; jinhuitang@njust.edu.cn).

Zhenan Sun is with the Center for Research on Intelligent Perception and Computing, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: znsun@nlpr.ia.ac.cn).

Massimo Tistarelli is with the Department of Science and Information Technology, University of Sassari, 07100 Sassari, Italy (e-mail: tista@uniss.it).

Digital Object Identifier 10.1109/TIFS.2020.2980792

I. Introduction

UNDERSTANDING and manipulating face images is an extensively studied topic in both vision and graphics communities. Particularly, face aging and expression synthesis have received tremendous attention. Applications like finding lost/wanted people, human-computer interaction, face recognition, animation, human cosmetic study and entertainment can all benefit from them. Facial age synthesis aims to predict future faces (i.e., age progression) or construct former faces (i.e., age regression) of an individual. The fact that aging causes pronounced changes in both the appearance and anatomy of human faces makes modelling this process a very difficult problem. Similarly, semantic manipulation of facial expressions is challenging due to the non-linear facial geometry variation caused by different expressions. Here, we list two common challenging issues faced by today's age and expression synthesis techniques. First, in order to well capture the aging/expression mechanism of facial features, a good algorithm generally needs sufficient labeled training data. However, existing databases either include very few people or provide very few personal face images. Second, different individuals usually have different aging/expression processes. For example, face aging can be affected by not only internal factors (e.g., gene, gender and ethnicity) but also external factors (e.g., working environment and living style). This diversity makes investigating the general aging/expression mechanism tougher. Although rather complicated, they have received significant attention. There has been great progress achieved on both topics [1]–[4]. Along with the rapid development of deep learning in various fields, using deep neural networks (DNN) to train face synthesis models has attracted increasing attention. In particular, generative adversarial networks (GANs) have been actively investigated and achieved impressive results for both face aging and expression synthesis [5].

It should be noted that most of the existing facial age synthesis literature focuses on long-term aging. For instance, the time span between neighbouring adult age groups in recent methods was always set to 10 years [6]–[13]. Simulating face aging with a shorter time span instead received far less attention. Suppose we use an age span of 10, and need to predict the facial appearance of a teenager at the age of 30. Should we apply the model of $20 \sim 30$ or $30 \sim 40$ to the teenager? Note that either model uses data with age differing by even 10 years from 30 for training. As a result, neither

of them could provide an accurate prediction of the subject's facial appearance. On the other hand, by adopting a shorter time span, a more accurate model of the facial aging process could be devised, leading to a more accurate prediction of the subject's appearance at the target age. However, using a shorter time span suffers from the decrease in training data and thus is prone to generate images with lower quality. For example, if we use an age span of 10, then we have sufficient training data for each age range/group. If we change the span to 5, there will be only about half left for each age range/group. In this study, we address this more challenging age synthesis task, including both age progression and regression. Specifically, we concentrate on synthesis over an age span of 5 years.

For expression synthesis, given a neutral face, existing work generally centers on synthesizing faces with 7 prototypical expressions, categorized as happiness, sadness, surprise, anger, fear, disgust and contempt. However, it is sometimes insufficient for real-world applications when we need further a happy face with a certain degree of strongness/strength. Similarly to short-term aging, synthesizing faces with varying intensities suffers from the lack of labeled data (i.e., data annotated with expression intensities). As a result, some research attempts to seek help from unsupervised techniques, e.g., by manually setting the expression code [14], [15] or linear interpolation of facial geometry parameters [16]. Unsupervised methods, however, cannot well capture facial variation caused by different expression intensities. In [17], [18], Action Units (AU) are investigated for continuous expression synthesis. Working in a supervised manner, they are competent for catching expression changes. However, they require face data labeled with AU. In this paper, we aim to find a supervised solution which can fully utilize limited labeled data and meanwhile effectively capture expression variation. Given a neutral face, our task is to synthesize an image sequence showing a certain expression with intensity from zero to high. Specifically, we consider 4 different intensities (i.e., zero, low, medium and high).

Human ages and expression intensities are inherently ordinal. They form a well-ordered set and thus have strong interrelationship. To fully utilize the ordering property and meanwhile take full advantage of limited labeled data, in this paper, we resort to learning-to-rank techniques for our synthesis tasks. Specifically, we present an ordinal ranking adversarial network and name it ranking GAN. Given a face, we first assign a one-hot label to it to indicate which age group/expression intensity it belongs to. By exploiting the relative order information among different age ranges/expression intensities, we further associate each face with a binary ranking vector. In ranking GAN, one-hot labels are used as the condition of the generator for synthesizing faces with target age groups/expression intensities. In addition, we add a sequence of cost-sensitive ordinal rankers on top of several multi-scale discriminators, with the aim of minimizing age/intensity rank estimation loss when optimizing both the generator and discriminators. This ranking scheme can work well even when training data is insufficient, since all training samples are exploited for building each age group/intensity's ranker. Moreover, learning-to-rank together with cost sensitivity enables our approach to well catch the correlation among different age ranges/expression intensities. The use of multi-scale discriminators further makes our ordinal rankers more robust, so that aging patterns/expression variation can be successfully captured. Experimental results on several public face databases demonstrate the effectiveness of ranking GAN on both capturing face aging patterns and synthesizing faces with varying intensities.

The main contributions of this work are:

1) We investigate face aging with a shorter time span, including both age progression and age regression.

2) We study expression synthesis with varying intensities in a supervised manner, in order to successfully capture facial variation caused by different intensities.

3) We present an ordinal ranking adversarial network for age and expression synthesis, attempting to fully utilize both the limited labeled training data and the well-ordering characteristic of human ages and expression intensities. To the best of our knowledge, this is the first attempt trying to exploit the ordinal nature of human ages and expression intensities for face synthesis.

4) We conduct extensive experiments to validate the effectiveness of ranking GAN on face aging and expression synthesis with varying intensities.

The rest of the paper is organized as follows: Section II gives some related work. The proposed ranking GAN is detailed in Section III. Section IV goes on to describe our implementation details and introduce face databases used to test ranking GAN. In Section V, we report experimental results. Finally, we conclude the whole work and further give some interesting future work in Section VI.

II. Related Work

A. Facial Age Synthesis

In early period, researchers generally exploit the biological structure and aging process of facial features such as cranium, muscles, and skin [19]–[23]. For example, in [19], Ramanathan and Chellappa developed a craniofacial growth model for young face aging. Through investigating the anatomy structure of facial skin, Wu *et al.* proposed a 3-layer dynamic skin model to synthesize wrinkles for face aging [22]. These physical methods usually require long personal aging sequences, so that complex modelling can be carried out. Prototype approaches instead do not utilize much of the biological prior knowledge [24]–[28]. They usually compute a prototype for each aging stage. The difference between prototypes of two aging stages is then considered as the aging pattern. In [25], an aging transform was derived by using shape and color differences between young and old male prototypes. Tiddeman *et al.* proposed a wavelet-based method for prototyping and transforming facial textures [26]. Owing to the availability of large amount of data and powerful computational hardware, there has been growing interest in employing DNN for age synthesis. In [6], Wang *et al.* proposed a recurrent neural network-based approach to model the aging pattern between neighbouring age groups. Duong *et al.* employed temporal restricted boltzmann machines for learning aging transformation [7]. Among various deep models, GAN has attracted the most attention. In [9], Antipov *et al.* introduced GAN to age synthesis tasks. Afterwards, various GAN-based approaches have been developed. For example, studies in [10]–[13] all made use of the image generation ability of GAN.

B. Facial Expression Synthesis

Early approaches on expression synthesis generally resort to techniques in computer graphics. In [29], Blanz *et al.* developed a system to create 3D animations from a single face image or a video. They transfer facial expression of a different person to the reconstructed face model by mapping geometric difference vectors. In [30], Pighin *et al.* proposed to create photorealistic textured 3D facial models from photographs of a subject, and to create smooth transitions between different expressions by morphing between different models. A second category of methods are example-based approaches which edit faces by either reusing sample patches of existing images or reordering images from an existing expression dataset. In [31], expression is mapped to a new face by comparing the user's face to all face images of the target person and returning the best match. Li *et al.* generated facial expression videos by retrieving frames that have similar expressions to the input ones [32]. In [33], a system combining face reordering with face warping was developed to edit expression in videos. With recent development of deep learning in various fields, researchers have turned their attention to using deep generative models for expression synthesis. In [34], deep belief net was adopted. Yeh *et al.* instead combined the generative ability of variational autoencoders with optical flow-based face manipulation [35]. GAN has also been successfully applied to expression synthesis [14]–[18], [36]. For example, in [16] and [36], facial geometry was used as a condition of the generator for generating faces with target expressions.

C. Ordinal Ranking for Face Analysis

Ordinal ranking, also called ordinal regression, is an interesting topic in machine learning community [37]. For face analysis, it has found applications in age estimation and expression intensity prediction. Note that the majority of existing age estimation approaches either employ a classifier to determine a coarse age range or use a regressor to calculate the exact age value. Human ages are inherently ordinal. They form a well-ordered set and thus have strong interrelationship. However, classification approaches simply treat ages as independent labels. In addition, human face matures in different ways at different ages, e.g., craniofacial growth in childhood and texture changes in adulthood. This property makes the process of face aging non-stationary in the feature space. Regression approaches consider ages as numerical values which utilize ordinal information. However, as manifested in [38], it is difficult for regressors to learn non-stationary functions which best fit the mapping from the feature space to the age space since they are prone to overfitting. To well adopt the ordering property of human ages, recent studies have turned to learning-to-rank approaches [38]–[41]. In these approaches, human ages are considered as a set of rank orders. For each age/rank order, they separate all the faces into two groups based on whether a face is elder than the given rank order. By doing this, they transform age estimation into a series of binary classification problems. Given that the cost of misclassification typically varies among different age pairs, researchers further investigated cost-sensitive learning [42]. Similarly to age estimation, ordinal ranking has been successfully applied to expression intensity prediction [43]–[45].

D. Generative Adversarial Networks

GAN offers a distinct and promising approach for training image synthesis models [5]. A classical GAN consists of a generator G and a discriminator D, which are trained alternatively via an adversarial process. The discriminator tries to distinguish real samples from fake ones. The generator instead attempts to synthesize fake samples that can fool the discriminator. GAN has achieved great success in various image generation tasks. However, it suffers from training instability. To solve this issue, various attempts have been made, including designing new network architectures [46], modifying learning objectives [47], employing regularization techniques [48] and so on. For face synthesis, the conditional GAN has particularly been widely used [49], where the generator and discriminator are conditioned on some extra information. Our approach is also a conditional GAN.

III. Ordinal Ranking Adversarial Networks

The proposed approach is a GAN-based network. It consists of a generator G and 3 multi-scale discriminators denoted as D_1, D_2 and D_3. We use one-hot labels as the condition to guide the generator for synthesizing faces with target age groups/expression intensities. Our discriminators share the same network architecture but operate at different image scales. Concretely, we create an image pyramid of 3 scales by downsampling the input with a factor of 2 and 4. D_1, D_2 and D_3 are then trained to distinguish real faces from generated ones at the 3 scales, resp. In addition, the 3 discriminators undertake the task of estimating inputs' age/intensity ranks, by using binary ranking vectors. We achieve this by adding a sequence of cost-sensitive ordinal rankers on top of them. Figure 1 illustrates our approach.

A. Problem Formulation

Suppose we have a total of M age ranges/expression intensities. Given a training set with N face images $\Omega = \{I_i | i = 1, \ldots, N\}$, we use $\boldsymbol{h}_i \in \mathbb{R}^M$ to denote the one-hot label of I_i.

By treating each age group/expression intensity as a rank, we can get a total of M rank orders, $l_m = m$, where $m = 1, \ldots, M$. For an image from the m-th range/intensity, we use l_m as its representative rank. Given a face image I_i, we use $y_i \in \{l_1, \ldots, l_M\}$ to denote its rank order. For each rank $l_m (1 \leq m < M)$, we separate the training set Ω into two subsets, Ω_m^+ and Ω_m^-, as follows:

$$\begin{aligned} \Omega_m^+ &= \{I_i | y_i > l_m\}, \\ \Omega_m^- &= \{I_i | y_i \leq l_m\}. \end{aligned} \tag{1}$$

Ω_m^+ and Ω_m^- are then used to train an ordinal ranker which aims to decide whether a sample's rank is larger than l_m. After this, we can get a total of $M - 1$ ordinal rankers which are placed on top of multi-scale discriminators.

For each sample I_i, by respectively comparing its rank with $\{l_1, \ldots, l_{M-1}\}$, we can get a binary ranking vector $\boldsymbol{r}_i \in \mathbb{R}^{M-1}$. The m-th dimension of $\boldsymbol{r}_i$ denotes whether I_i's rank is larger than l_m. Thus, $\boldsymbol{r}_i$ takes the form as:

$$\boldsymbol{r}_i^{(m)} = \begin{cases} 1, & y_i > l_m \\ 0, & y_i \leq l_m \end{cases}, \tag{2}$$

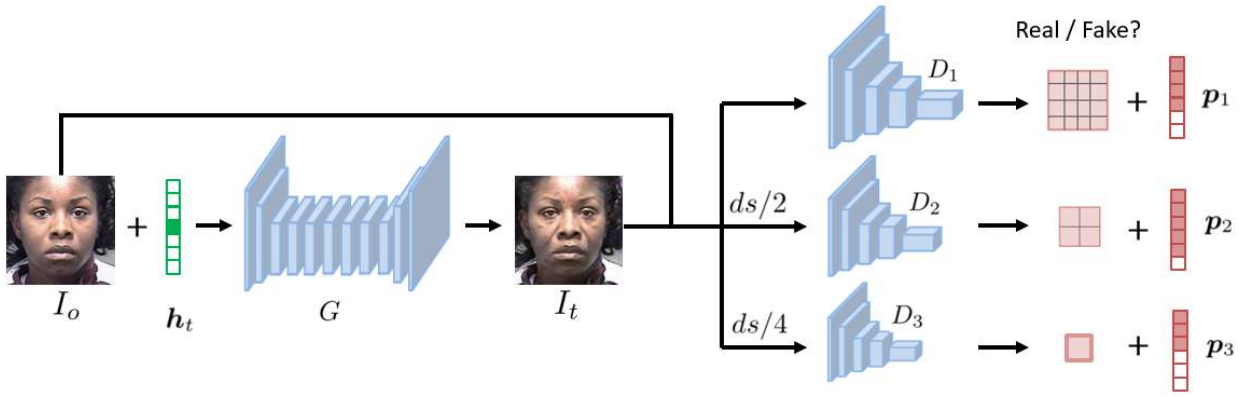

Fig. 1. Framework of ranking GAN. We use $\boldsymbol{h}_t \in \mathbb{R}^M$ to denote the target one-hot label, where M is the number of age groups/expression intensities. $ds/2$ and $ds/4$ represent operations of downsampling with a factor of 2 and 4, resp. We use $\boldsymbol{p}_1$, $\boldsymbol{p}_2$ and $\boldsymbol{p}_3 \in \mathbb{R}^{M-1}$ to denote ranking vectors predicted by the 3 discriminators. I_o and I_t are the input and generated faces with size $H \times W \times 3$. We broadcast the target condition $\boldsymbol{h}_t$ to a $H \times W \times M$ tensor and then concatenate it with I_o to form the input of G with size $H \times W \times (3+M)$. Since $\boldsymbol{h}_t$ is a one-hot label, in the M conditional maps $H \times W \times M$, only one map is filled with ones while the rest are all filled with zeros.

where $m = 1, \ldots, M-1$. Similarly to one-hot label $\boldsymbol{h}_i$, $\boldsymbol{r}_i$ encodes also I_i's age/intensity information.

As stated in Section II-C, age/intensity estimation is inherently a cost-sensitive problem. We should take cost sensitivity into consideration when designing our ordinal rankers. Therefore, we further associate I_i with a cost vector $\boldsymbol{c}_i \in \mathbb{R}^{M-1}$. The m-th dimension of $\boldsymbol{c}_i$ denotes the cost of misclassifying I_i with the m-th ordinal ranker. We define $\boldsymbol{c}_i$ as follows:

$$\boldsymbol{c}_i^{(m)} = \begin{cases} y_i - l_m, & y_i > l_m \\ l_m - y_i + 1, & y_i \leq l_m \end{cases}, \tag{3}$$

where $m = 1, \ldots, M-1$.

B. Ranking GAN Training

Now, for each sample, we have a corresponding one-hot label $\boldsymbol{h}$, a ranking vector $\boldsymbol{r}$ and a cost vector $\boldsymbol{c}$. Similarly, each age range/expression intensity has its own $\boldsymbol{h}$, $\boldsymbol{r}$ and $\boldsymbol{c}$. Samples belong to the same age range/expression intensity own the same $\boldsymbol{h}$, $\boldsymbol{r}$ and $\boldsymbol{c}$. Hence, one-hot labels, ranking vectors and cost vectors are in a one-to-one correspondence. In our work, one-hot labels are only used to tell which age range/expression intensity we want to synthesize, i.e., the target condition. Ranking and cost vectors instead are adopted to compute the age/intensity estimation loss, since they can capture the ordering characteristic of ages/expression intensities.

Given an input image I_o with its one-hot label $\boldsymbol{h}_o$, the goal is to synthesize a face image I_t with the same identity but from a different age group/expression intensity specified by $\boldsymbol{h}_t$. For I_o and I_t, we use $\boldsymbol{r}_o$ and $\boldsymbol{r}_t$ to denote their corresponding ranking vectors and use $\boldsymbol{c}_o$ and $\boldsymbol{c}_t$ to represent their cost vectors. During face generation, we randomly generate $\boldsymbol{h}_t$ as the condition of G so that it can flexibly synthesize new faces corresponding to different age groups/expression intensities. Since our discriminators are trained to not only distinguish between real and generated samples but also estimate their age/intensity ranks, we adopt a rank estimation loss in addition to the adversarial loss. In order to preserve the content of the input, we further employ a reconstruction loss. Finally, we perform a total variation regularization on synthesized faces with the aim of reducing unfavorable artifacts [50].

TABLE I

NETWORK ARCHITECTURE OF THE GENERATOR ("I", "D", "R", "U", AND "O" RESPECTIVELY DENOTE "INPUT", "DOWNSAMPLING", "RESIDUAL BLOCK", "UPSAMPLING", AND "OUTPUT")

	Layer	Output Size	Details
Input	ConvI	$128 \times 128 \times 64$	K7, S1, P3
Down-sampling	ConvD1	$64 \times 64 \times 128$	K4, S2, P1
	ConvD2	$32 \times 32 \times 256$	K4, S2, P1
Bottleneck	ConvR1	$32 \times 32 \times 256$	K3, S1, P1
	...	...	...
	ConvR6	$32 \times 32 \times 256$	K3, S1, P1
Up-sampling	DeconvU1	$64 \times 64 \times 128$	K4, S2, P1
	DeconvU2	$128 \times 128 \times 64$	K4, S2, P1
Output	ConvO	$128 \times 128 \times 3$	K7, S1, P3

1) Adversarial Loss: Ranking GAN is conditioned on the input image and a target one-hot label, adversarial losses for the generator and discriminators are thus defined as

$$L_{adv}^D = \frac{1}{3}\sum_{k=1}^{3}\Big\{ -\mathbb{E}_{I_o}\Big[log D_k(I_o)\Big] - \mathbb{E}_{I_o,\boldsymbol{h}_t}\Big[log\Big(1 - D_k\big(G(I_o,\boldsymbol{h}_t)\big)\Big)\Big]\Big\}, \tag{4}$$

$$L_{adv}^G = \frac{1}{3}\sum_{k=1}^{3}\mathbb{E}_{I_o,\boldsymbol{h}_t}\Big[log\Big(1 - D_k\big(G(I_o,\boldsymbol{h}_t)\big)\Big)\Big]. \tag{5}$$

2) Rank Estimation Loss: Apart from the adversarial loss, our discriminators attempt to minimize the estimation loss of age/intensity ranks to ensure both I_o and I_t are correctly classified into their corresponding age groups/expression intensities. To achieve this, we add a sequence of cost-sensitive ordinal rankers on top of multi-scale discriminators. This loss is adopted when optimizing both the generator and discriminators. The loss for optimizing discriminators is applied to input images and formulated as

$$L_{rank}^D = \frac{1}{3}\sum_{k=1}^{3}\Big\{\mathbb{E}_{I_o,\boldsymbol{r}_o}\Big[-\sum_{m=1}^{M-1}\boldsymbol{c}_o^{(m)}\Big(\boldsymbol{r}_o^{(m)} log\sigma(\boldsymbol{p}_k^{(m)}) + (1-\boldsymbol{r}_o^{(m)})log\big(1-\sigma(\boldsymbol{p}_k^{(m)})\big)\Big)\Big]\Big\}, \tag{6}$$

where $\boldsymbol{p}_k (k = 1, 2, 3)$ are predicted ranking vectors of I_o and $\sigma(x)$ is the sigmoid function, i.e., $\sigma(x) = 1/(1+e^{-x})$.

TABLE II

NETWORK ARCHITECTURE OF MULTI-SCALE DISCRIMINATORS ("M" IS THE NUMBER OF AGE GROUPS/EXPRESSION INTENSITIES. "I", "H", AND "O" RESPECTIVELY DENOTE "INPUT", "HIDDEN", AND "OUTPUT")

	Layer	Output Size			Details
		D_1	D_2	D_3	
Input	ConvI	$64 \times 64 \times 64$	$32 \times 32 \times 64$	$16 \times 16 \times 64$	K4, S2, P1
Hidden	ConvH1	$32 \times 32 \times 128$	$16 \times 16 \times 128$	$8 \times 8 \times 128$	K4, S2, P1
	ConvH2	$16 \times 16 \times 256$	$8 \times 8 \times 256$	$4 \times 4 \times 256$	K4, S2, P1
	ConvH3	$8 \times 8 \times 512$	$4 \times 4 \times 512$	$2 \times 2 \times 512$	K4, S2, P1
	ConvH4	$4 \times 4 \times 1024$	$2 \times 2 \times 1024$	$1 \times 1 \times 1024$	K4, S2, P1
Output	ConvO1	$4 \times 4 \times 1$	$2 \times 2 \times 1$	$1 \times 1 \times 1$	K3, S1, P1
	ConvO2	$1 \times 1 \times (M-1)$	$1 \times 1 \times (M-1)$	$1 \times 1 \times (M-1)$	K*, S1, P0

* Kernel sizes of D_1, D_2 and D_3 here are 4, 2 and 1, resp.

Note that $\boldsymbol{p}_k (k = 1, 2, 3)$ are estimated by the $M-1$ ordinal rankers which are trained using the two subsets of Eq 1. By minimizing this loss, our discriminators can learn to estimate I_o's age/intensity rank as $\boldsymbol{r}_o$. On the other hand, the loss used to optimize G is applied to synthesized images and defined as

$$L_{rank}^{G} = \frac{1}{3}\sum_{k=1}^{3}\Big\{\mathbb{E}_{I_t,\boldsymbol{r}_t}\Big[-\sum_{m=1}^{M-1}\boldsymbol{c}_t^{(m)}\Big(\boldsymbol{r}_t^{(m)} log\sigma(\boldsymbol{p}_k^{(m)}) + (1-\boldsymbol{r}_t^{(m)}) log\big(1-\sigma(\boldsymbol{p}_k^{(m)})\big)\Big)\Big]\Big\}, \quad (7)$$

where $\boldsymbol{p}_k (k = 1, 2, 3)$ are estimated age/intensity ranks of I_t. They are predicted by the $M-1$ ordinal rankers which are trained using the two subsets of Eq 1. As a result, our generator can learn to generate samples that have rank $\boldsymbol{r}_t$. With this loss, we can guarantee the aging/expression effect generation.

3) Reconstruction Loss: To ensure the synthesized image preserves the content of its input, we apply a reconstruction loss to G. It takes the form as

$$L_{rec} = \mathbb{E}_{I_o,\boldsymbol{h}_t,\boldsymbol{h}_o}\Big[\big\| I_o - G\big(G(I_o,\boldsymbol{h}_t),\boldsymbol{h}_o\big)\big\|_1\Big], \quad (8)$$

where G takes in $G(I_o, \boldsymbol{h}_t)$ and the original label $\boldsymbol{h}_o$ as inputs and attempts to reconstruct the original image I_o. We use L_1 norm to encourage less blurring output.

4) Overall Objective: Finally, our objective functions to optimize the generator and discriminators are weighted sums of all the above defined losses. They are written, respectively, as

$$L_D = L_{adv}^{D} + \lambda_{rank} L_{rank}^{D}, \quad (9)$$

$$L_G = L_{adv}^{G} + \lambda_{rank} L_{rank}^{G} + \lambda_{rec} L_{rec} + \lambda_{tv} L_{tv}, \quad (10)$$

where λ_{rank}, λ_{rec} and λ_{tv} are trade-off parameters. We use L_{tv} to denote the total variation regularization imposed on synthesized samples.

C. Network Architecture

For the generator, we draw lessons from CycleGAN [51] and StarGAN [52]. The idea of using multi-scale discriminators comes from pix2pixHD [53]. Our network receives input images with size $128 \times 128 \times 3$. In Tables I and II, we give detailed architectures of G and D, where "K", "S", and "P" denote the kernel size, stride size, and padding size, resp.

Following [12], [17] and [52], we broadcast the target condition to a $128 \times 128 \times M$ tensor and then concatenate it with the input image to form a tensor with size $128 \times 128 \times (3+M)$. The tensor is then sent to G. Since we adopt one-hot labels as the target condition, in the M conditional maps, only one map is filled with ones while the rest are all filled with zeros. We can also follow CAAE [10] and ExprGAN [14] which first encode the input image into a vector and then concatenate it with the one-hot label. However, as will be shown in Section V, both CAAE and ExprGAN perform poorly on our tasks. We thus choose broadcasting the target condition. Our generator contains two stride-2 convolution layers for downsampling, six residual blocks, and two stride-2 transposed convolution layers for upsampling. Instance normalization followed by ReLU activation is adopted in all layers except the last output layer, which employs Tanh.

The 3 discriminators respectively receive inputs with size $128 \times 128 \times 3$, $64 \times 64 \times 3$ and $32 \times 32 \times 3$. They share the same network architecture and employ PatchGANs. For each discriminator, we add two output layers ConvO1 and ConvO2 on top of them. ConvO1 differentiates real images from fake ones and outputs the probability of local patches to be real. ConvO2 instead implements rank estimation and outputs estimated age/intensity ranks. Note that no feature normalization but Leaky ReLU with a negative slope of 0.01 is applied to all layers of each discriminator.

IV. EXPERIMENTAL SETTINGS

A. Implementation Details

We choose to use Wasserstein GAN for stabilized training [47]. L_{adv}^{D} is therefore modified as

$$L_{adv}^{D} = \frac{1}{3}\sum_{k=1}^{3}\Big\{-\mathbb{E}_{I_o}\Big[D_k(I_o)\Big] + \mathbb{E}_{I_o,\boldsymbol{h}_t}\Big[D_k\big(G(I_o,\boldsymbol{h}_t)\big)\Big] + \lambda_{gp}\mathbb{E}_{\hat{I}}\Big[\Big(\big\|\nabla_{\hat{I}} D_k(\hat{I})\big\|_2 - 1\Big)^2\Big]\Big\}, \quad (11)$$

where $\hat{I}$ is sampled uniformly along a straight line between a real sample and its generated one. λ_{gp} is the coefficient of gradient penalty. We set it to be 10 in all our experiments. Coefficients of reconstruction loss and total variation regularization in Eq.10 are respectively set as $\lambda_{rec} = 10$ and $\lambda_{tv} = 0.0001$. We use Adam with a learning rate of 0.0001, $\beta_1 = 0.5$, and $\beta_2 = 0.999$ to train our network.

For face aging, we run a single optimization step for G every five optimization steps of discriminators and set $\lambda_{rank} = 4$. Using a batch size of 16, the training takes less than 14 hours using a single NVIDIA TITAN Xp GPU.

TABLE III
NUMBER OF TRAINING AND TEST IMAGES ON MORPH

Age Group	[16,20]	[21,25]	[26,30]	[31,35]	[36,40]	[41,45]	[46,50]
# train	6,592	6,324	4,983	6,007	6,182	5,595	3,400
# test	1,613	1,661	1,388	1,475	1,553	1,292	803

For expression synthesis, we optimize G every time we optimize discriminators. And λ_{rank} is set to 20. With a batch size of 8, the training takes about 15 hours using the same GPU. Note that these parameters are determined empirically.

B. Experimental Data

Since our proposed approach considers the ordering property of human ages, using inaccurately labeled data will interfere with the model training. To well evaluate the performance of ranking GAN on face aging, we thus perform experiments on the MORPH database [54] where all face images are labeled with accurate ages. For expression synthesis, we choose MUG [55], Oulu-CASIA [56] and CK+ [57] databases. For all face images, we first detect 68 facial landmarks and then use 3 of them (i.e., left eye center, right eye center, and mouth center) to perform alignment [58]. The final images are of size $128 \times 128 \times 3$. Although using a larger image size can get better results, it will lead to a very expensive training. For expression synthesis, we take synthesis of happiness and surprise as our case study. Thus, given a neutral face, we aim to synthesize a happiness/surprise sequence with intensity from zero to high.

1) MORPH: The MORPH database is a popular benchmark for facial age estimation. The dataset we choose consists of 52,099 color images. Subject ages range from 16 to 77 years old. Since there are very few people who are elder than 50, we do not consider all the ages. With an age span of 5, we define 7 age ranges/groups containing 48,868 images, i.e., $16 \sim 20$, $21 \sim 25$, $26 \sim 30$, $31 \sim 35$, $36 \sim 40$, $41 \sim 45$ and $46 \sim 50$. We randomly select around $1/5$ for test and use the remaining for training. Note that test subjects are disjoint from training subjects. We list the data configuration in Table III.

2) MUG: The MUG database consists of image sequences of happiness, sadness, surprise, anger, fear and disgust from 86 subjects, of which only 52 subjects are made publicly available. Each sequence contains 50 to 160 images showing a certain expression from neutral to apex then back to neutral. We randomly choose 47 subjects for training and use the rest for test. Since we consider only synthesis of happiness and surprise, we manually split each happiness/surprise sequence into 4 intensities: zero, low, medium and high.

3) Oulu-CASIA: The Oulu-CASIA database includes videos of 80 subjects with 6 expressions, i.e., happiness, sadness, surprise, anger, fear and disgust. Videos are captured under 3 different illumination conditions using both NIR and VIS imaging systems. We choose VIS images captured under strong illumination for our experiments. We randomly select $1/10$ for test and use the remaining for training. Each happiness/surprise video is then manually split into 4 intensities.

4) CK+: The CK+ database contains 593 sequences from 123 subjects with 7 universal expressions. Each sequence starts with a neutral emotion and ends with the peak of a certain expression. Note that there are some subjects without happiness/surprise sequence. For happiness, we have a total of 87 subjects. And the number of subjects with surprise is 82. To fully utilize the limited data, we use all these subjects for training, and use their corresponding remaining for test (i.e., 36 and 41 subjects for happiness and surprise, resp). Similarly, we manually split each sequence into 4 intensities.

TABLE IV
NUMBER OF TRAINING AND TEST IMAGES ON MUG, OULU-CASIA AND CK+ ("H" AND "S" RESPECTIVELY DENOTE "HAPPINESS" AND "SURPRISE")

		Intensity	Zero	Low	Medium	High
MUG	H	# train	4,191	1,514	2,420	2,485
		# test	420	-	-	-
	S	# train	4,499	1,815	2,269	2,413
		# test	312	-	-	-
Oulu-CASIA	H	# train	498	470	500	504
		# test	56	-	-	-
	S	# train	552	510	405	500
		# test	56	-	-	-
CK+	H	# train	684	462	512	442
		# test	77	-	-	-
	S	# train	925	393	404	432
		# test	110	-	-	-

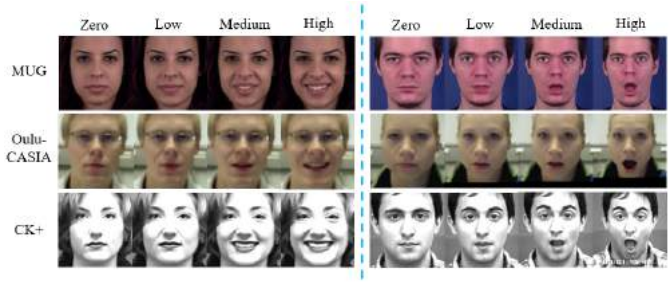

Fig. 2. Training examples of different intensities. Left and right parts respectively show happiness and surprise data.

In Table IV, we list the number of training and test images for each expression database. Note that we test only neutral faces in this study, i.e., images with zero intensity. For determining samples' intensities when preparing training data, we classify samples as "Low" if they show slight expressions, group them into "Medium" if they present obvious expressions and classify them as "High" if they have maximum expressions. In Figure 2, we give several examples of different intensities.

V. SYNTHESIS RESULTS

We first report synthesis results obtained by using ranking GAN on the 4 databases. Then we compare our approach with prior work to show its superiority. In order to completely evaluate ranking GAN, we further present several ablation studies.

TABLE V

OBJECTIVE AGE ESTIMATION RESULTS (IN YEARS) OBTAINED BY FACE++ ON MORPH WITH A SPAN OF 5 YEARS

Age Group	$16 \sim 20$	$21 \sim 25$	$26 \sim 30$	$31 \sim 35$	$36 \sim 40$	$41 \sim 45$	$46 \sim 50$
Real Face	24.21 ± 4.68	28.26 ± 5.41	31.75 ± 5.94	36.21 ± 6.56	40.83 ± 7.28	45.21 ± 7.64	51.06 ± 7.81
Ours	24.40 ± 4.99	27.39 ± 5.22	31.03 ± 5.62	37.40 ± 6.24	41.06 ± 6.22	46.88 ± 6.57	52.87 ± 6.39
CAAE [10]	22.31 ± 3.69	25.32 ± 4.47	28.61 ± 5.50	31.31 ± 6.02	33.71 ± 6.27	36.52 ± 6.89	40.17 ± 7.03
IPCGANs [12]	24.52 ± 5.51	27.32 ± 6.12	29.96 ± 6.66	34.11 ± 7.25	38.29 ± 7.62	44.10 ± 7.50	48.92 ± 7.06

TABLE VI

OBJECTIVE FACE VERIFICATION RESULTS OBTAINED BY FACE++ ON MORPH WITH A SPAN OF 5 YEARS

Age Group	$16 \sim 20$	$21 \sim 25$	$26 \sim 30$	$31 \sim 35$	$36 \sim 40$	$41 \sim 45$	$46 \sim 50$
Ours	95.58 ± 1.04	95.97 ± 0.68	96.00 ± 0.60	96.08 ± 0.51	95.95 ± 0.62	95.75 ± 0.81	94.93 ± 1.14
CAAE [10]	75.05 ± 8.96	76.11 ± 8.31	76.90 ± 7.63	77.42 ± 7.13	77.94 ± 6.98	77.33 ± 7.08	76.56 ± 7.38
IPCGANs [12]	94.29 ± 1.57	95.12 ± 1.08	95.41 ± 0.84	95.61 ± 0.59	95.46 ± 0.66	94.86 ± 1.22	93.88 ± 2.01

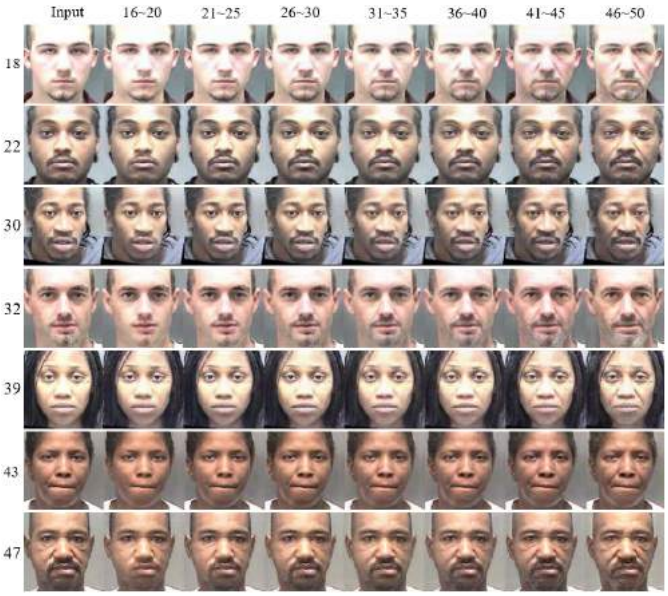

Fig. 3. Qualitative results obtained by using ranking GAN on MORPH. Input faces together with actual ages are shown in the leftmost column. For each row, synthesized faces in groups elder than the input age are progression results, while those in younger groups belong to regression results.

Finally, we examine the generalization capability of ranking GAN in synthesizing other basic expressions.

A. Age Synthesis

We show in Figure 3 some synthesis results on MORPH. As observed, using ranking GAN achieves smooth aging sequences and generates visually plausible results. For age progression, wrinkles appears gradually. Hair gets gray slowly. For some males, beard gets gray also in aged faces. For regression, wrinkles, mustache and beard get reduced or even removed. To well examine ranking GAN's ability in both capturing aging patterns and keeping identity cue, we further perform a quantitative analysis. For evaluating aging effect, we employ the online face analysis tool of Face++ [59] to estimate ages of both input faces (i.e., real faces) and their synthesized ones obtained by ranking GAN. For each age group, we report the mean and standard deviation of estimated ages. The results are shown in Table V. Although there exists deviation in Face++'s estimated ages from actual ages (e.g., 24.21 ± 4.68 vs $16 \sim 20$), the overall aging trend is relatively smooth. Note that, here, directly comparing ages of generated face with actual ages of real faces is not reasonable. Since ages of generated faces are estimated by Face++, for a fair comparison, real faces should be also estimated by Face++. This is reported in the "Real Face" row of Table V. By comparing estimated ages of generated faces with those of real faces, we can see that ranking GAN successfully captures face aging patterns. In all age groups, faces synthesized by ranking GAN have very close ages to real faces.

TABLE VII

OBJECTIVE AGE ESTIMATION AND FACE VERIFICATION RESULTS OBTAINED BY FACE++ ON MORPH WITH A SPAN OF 10 YEARS

Age Group	[16,25]	[26,35]	[36,45]	[46,55]
Real Face	26.26	34.05	42.82	52.86
Generated Age	25.86	34.66	43.85	55.40
Verification Result	95.73	96.00	95.82	94.02

Objective face verification is also conducted using Face++ in order to check whether identity is well preserved during face aging. We compare each test face with its corresponding synthesized ones. For each comparison, we can get a confidence value indicating the similarity of two faces. The confidence lies within [0,100]. Higher confidence indicates higher possibility that two faces belong to the same person. Finally, for each age group, we calculate the mean and standard deviation of confidence over all test faces. The results are reported in Table VI. As can be seen, we obtain very high confidence for all age groups. We can therefore conclude ranking GAN performs well in preserving identity during face aging.

To show the advantage of using a shorter time span, we further apply our method to face aging with a span of 10 years. We define 4 age groups on MORPH including $16 \sim 25$, $26 \sim 35$, $36 \sim 45$ and $46 \sim 55$. In Table VII, we report mean values of the results of age estimation and face verification. As observed, using ranking GAN achieves also promising results when the span of 10 is adopted. Now let us return to the question raised in Section I. If we want to know what a teenager will look like when he becomes 30, which of the 4 models should we use? Note that both mean ages of Model $26 \sim 35$ (34.66) and Model $16 \sim 25$ (25.86) deviate from 30 by more than 4 years. Therefore, neither of them can give a convincing result. When adopting a shorter time span, however, we can get more fine-grained models. For example, 7 models can be obtained when a span of 5 is used. From Table V, we can see Model $26 \sim 30$ can be adopted, since its mean age (31.03) are very close to 30.

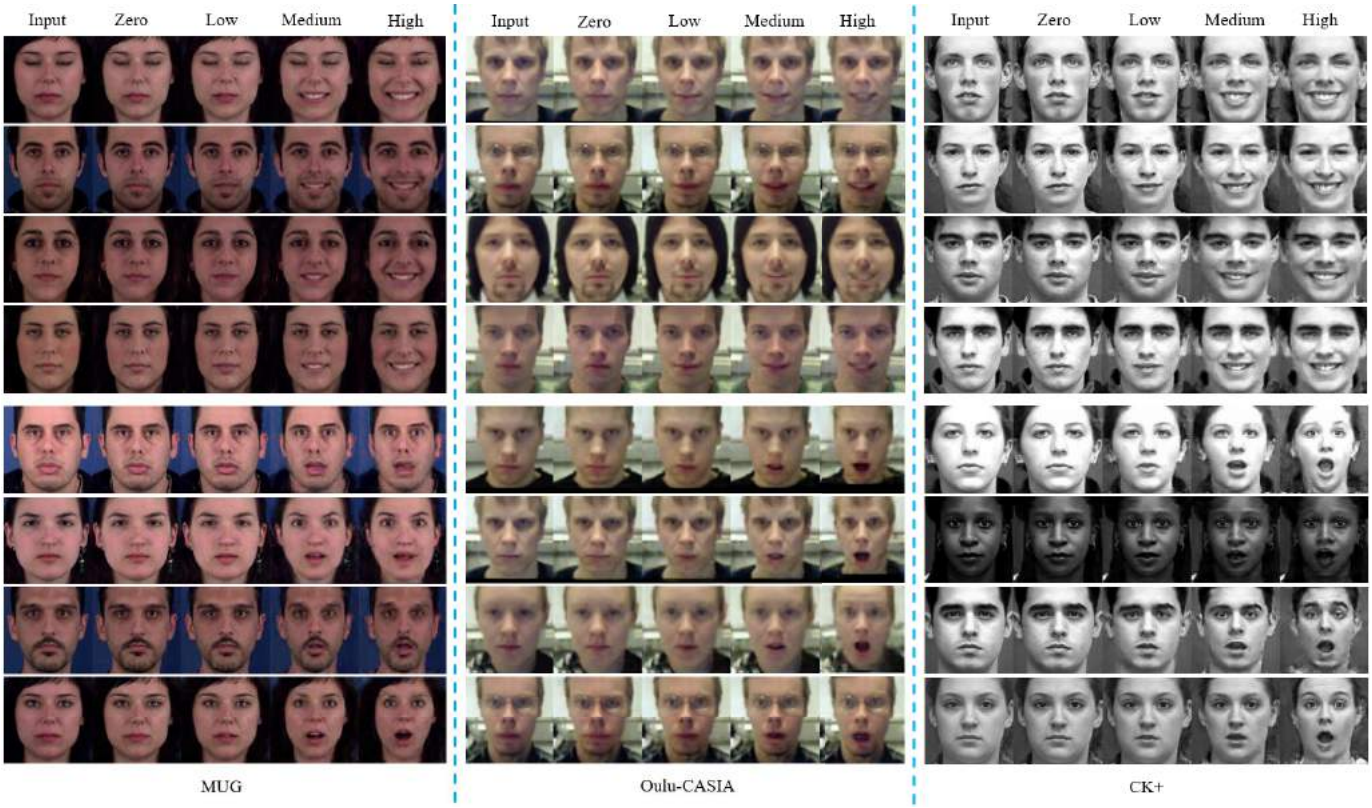

Fig. 4. Qualitative expression synthesis results obtained by using ranking GAN. Top and bottom parts respectively show happiness and surprise synthesis.

TABLE VIII
OBJECTIVE FACE VERIFICATION RESULTS OBTAINED BY FACE++ ON MUG, OULU-CASIA AND CK+. FACES ARE SYNTHESIZED BY RANKING GAN

	Happiness				Surprise			
	Zero	Low	Medium	High	Zero	Low	Medium	High
MUG	96.51 ± 0.19	96.20 ± 0.23	94.16 ± 0.68	92.41 ± 1.32	96.51 ± 0.28	96.13 ± 0.29	92.68 ± 1.74	90.26 ± 2.70
Oulu-CASIA	96.45 ± 0.27	95.92 ± 0.47	94.86 ± 0.71	93.35 ± 0.97	95.69 ± 0.48	95.13 ± 0.70	94.12 ± 0.87	85.86 ± 4.51
CK+	96.44 ± 0.89	95.45 ± 0.79	93.22 ± 1.45	89.70 ± 3.09	96.63 ± 0.60	95.31 ± 1.11	90.51 ± 3.78	81.16 ± 7.36

B. Expression Synthesis

We show expression synthesis results in Figure 4. As observed, for both happiness and surprise synthesis, ranking GAN achieves promising results. From zero to peak, the intensity gets higher and higher. Identity is also well preserved during expression process. Note that although training data in Oulu-CASIA and CK+ is very limited (as listed in Table IV), we obtain visually plausible results. For quantitative evaluation, we check only whether identity is well preserved. Similarly, we perform verification using Face++. The confidence values are reported in Table VIII. As observed, we achieve high confidence in most cases. Only when synthesizing surprise faces with high intensity on Oulu-CASIA and CK+, the identity is not well preserved. The main reason might be the large variation in facial geometry caused by big surprise. The very limited data might also contribute to this.

C. Comparison With Prior Work

As for the control methods, we use CAAE [10] and IPCGANs [12] for face aging and ExprGAN [14] for expression synthesis. We choose these methods because their codes are made publicly available on Github. We thus can easily implement them for our own tasks. For CAAE and IPCGANs, we use the same training and test data as ranking GAN's. For ExprGAN, following the original work, we consider 5 intensities and use images with the highest intensity from 6 expressions. Since many subjects in CK+ have less than 6 expressions, we implement ExprGAN only on MUG and Oulu-CASIA. On Oulu-CASIA, we use the last 3 frames of each video. And on MUG, we select around 10 frames with the highest intensity for each video.

Qualitative results are shown in Figures 5 and 6. From Figure 5, we can see that CAAE performs poorly in both identity preservation and aging effect generation. Moreover, its synthesized faces lack fine details. IPCGANs, by contrast, show superiority in keeping identity cue and generating fine details. However, for age progression of young faces, they achieve low quality faces for the several eldest groups. For example, the eye region is synthesized poorly. For expression synthesis, we observe that ExprGAN suffers from over-fitting severely. It performs poorly in keeping identity cue and modifies even input's gender. Moreover, when synthesizing faces with low intensity, ExprGAN generates very low quality images.

We further quantitatively compare our approach with the control methods. The results of CAAE and IPCGANs are

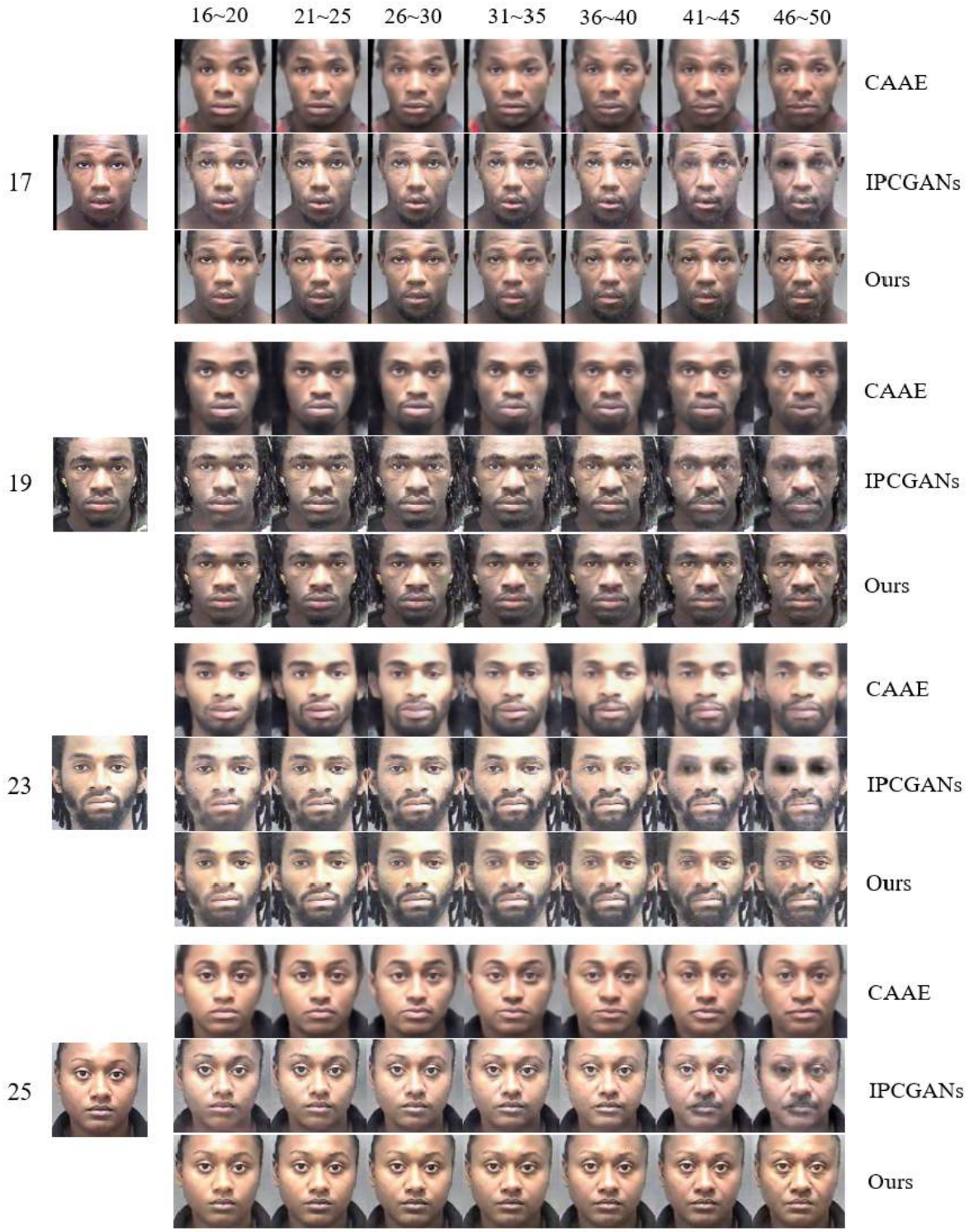

Fig. 5. Comparison with prior work: CAAE [10] and IPCGANs [12]. In the leftmost column we give input faces and their actual ages.

reported in Tables V and VI, while results of ExprGAN are shown in Table IX. As observed, for facial age synthesis, we achieve the best results in both learning aging patterns and keeping identity cue. By comparing Table IX with VIII, we can see ranking GAN performs much better in identity preservation than ExprGAN.

It should be noted that the training process of IPCGANs suffers from model collapse. The results reported here are thus obtained before model collapse. For expression synthesis, in order to generate different expressions (e.g., happiness, sadness, surprise, anger, fear and disgust), our approach needs to train different models. However, once we finish the model training, expression synthesis will be a very efficient process. On a single NVIDIA GTX1080Ti GPU, generating one expression sequence only costs about 12ms.

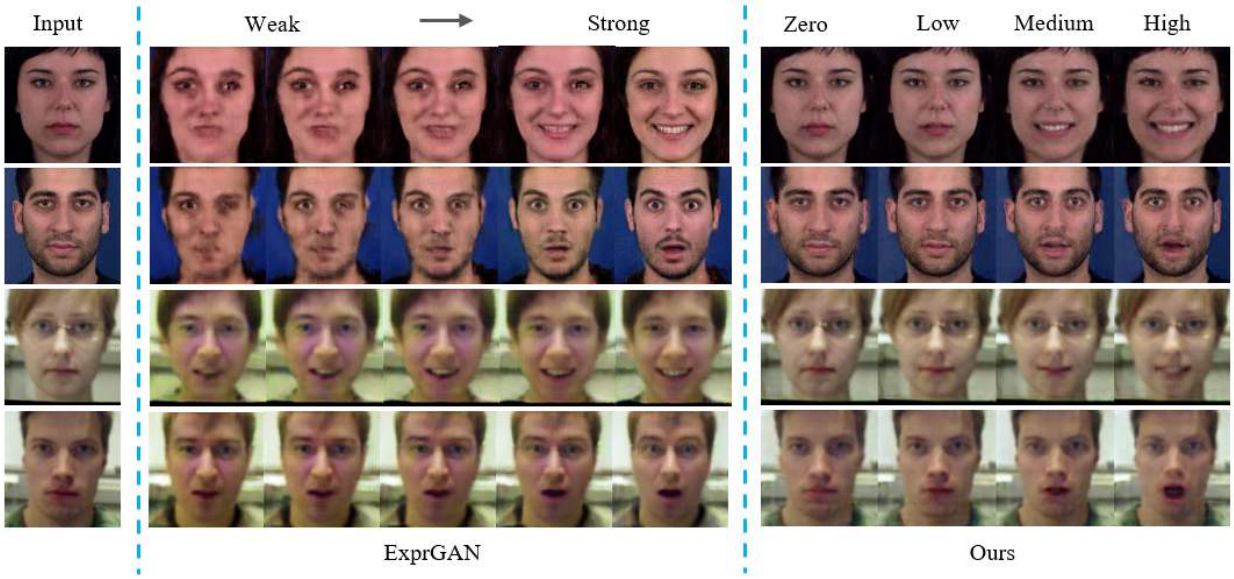

Fig. 6. Comparison with prior work: ExprGAN [14]. The first half rows show results of MUG, while the remaining are results of Oulu-CASIA.

TABLE IX

OBJECTIVE FACE VERIFICATION RESULTS OBTAINED BY FACE++ ON MUG AND OULU-CASIA. FACES ARE SYNTHESIZED BY EXPRGAN [14]

	Intensity	1	2	3	4	5
MUG	Happiness	52.94 ± 6.31	56.47 ± 6.31	61.28 ± 6.12	64.73 ± 6.36	63.06 ± 6.24
	Surprise	49.73 ± 2.98	46.16 ± 5.74	48.50 ± 7.36	56.07 ± 9.06	63.01 ± 9.00
Oulu-CASIA	Happiness	59.34 ± 9.79	62.58 ± 9.52	65.85 ± 10.64	67.79 ± 10.01	67.72 ± 9.78
	Surprise	56.96 ± 10.47	58.03 ± 10.50	58.87 ± 10.03	58.98 ± 10.11	60.21 ± 9.83

TABLE X

ABLATION STUDY RESULTS WITH OBJECTIVE AGES ESTIMATED BY FACE++ ON MORPH

Age Group	$16 \sim 20$	$21 \sim 25$	$26 \sim 30$	$31 \sim 35$	$36 \sim 40$	$41 \sim 45$	$46 \sim 50$
Real Face	24.21 ± 4.68	28.26 ± 5.41	31.75 ± 5.94	36.21 ± 6.56	40.83 ± 7.28	45.21 ± 7.64	51.06 ± 7.81
Ours	24.40 ± 4.99	27.39 ± 5.22	31.03 ± 5.62	37.40 ± 6.24	41.06 ± 6.22	46.88 ± 6.57	52.87 ± 6.39
No Ranking	24.85 ± 5.32	27.60 ± 5.12	32.87 ± 5.89	37.34 ± 6.38	44.51 ± 6.88	47.69 ± 7.11	$57.55 \sim 7.20$
No Multi-scale D	23.00 ± 4.50	26.68 ± 4.40	28.41 ± 4.69	34.27 ± 5.26	38.18 ± 5.64	50.89 ± 6.46	$57.52 \sim 6.97$
No Reconstruction	22.52 ± 4.14	26.31 ± 4.39	29.73 ± 5.19	35.99 ± 6.24	40.97 ± 6.38	47.15 ± 6.43	$56.87 \sim 6.82$

TABLE XI

ABLATION STUDY RESULTS WITH VERIFICATION CONFIDENCE CALCULATED BY FACE++ ON MORPH

Age Group	$16 \sim 20$	$21 \sim 25$	$26 \sim 30$	$31 \sim 35$	$36 \sim 40$	$41 \sim 45$	$46 \sim 50$
Ours	95.58 ± 1.04	95.97 ± 0.68	96.00 ± 0.60	96.08 ± 0.51	95.95 ± 0.62	95.75 ± 0.81	94.93 ± 1.14
No Ranking	94.77 ± 1.20	94.65 ± 1.16	94.50 ± 1.28	94.60 ± 1.31	94.50 ± 1.26	94.61 ± 1.29	$92.81 \sim 2.08$
No Multi-scale D	94.45 ± 1.32	93.72 ± 1.73	93.67 ± 1.82	94.35 ± 1.38	93.64 ± 1.79	93.06 ± 2.02	$92.17 \sim 2.48$
No Reconstruction	93.38 ± 2.10	94.32 ± 1.46	94.62 ± 1.07	94.89 ± 0.96	94.49 ± 1.20	93.84 ± 1.66	$91.39 \sim 2.64$

D. Ablation Study

In order to validate the contribution of different modules, next we conduct several ablation studies. The first study is used to examine the contribution of ordinal ranking, which is performed by replacing ordinal ranking with softmax loss-based age group classification. The second study is performed by using a single-scale discriminator in order to check the contribution of multi-scale discriminators. Finally, we check the contribution of the reconstruction loss. We perform age synthesis on MORPH for these ablation studies. Qualitative results are given in Figure 7. As observed, without ordinal ranking, generated faces present artifacts. Without multi-scale discriminators, synthesized faces look blurred. If removing the reconstruction loss, generated faces show inconsistent color. We further quantitatively check the contribution of the 3 modules. Quantitative measures include objective ages, verification confidence and SSIM [60]. We report these results in Tables X, XI and XII. From Table X, we can see that removing any of the 3 modules leads to poorer aging effect. For example, for Group $46 \sim 50$, ages of synthesized faces differ by even 6 years from real ages. For identity preservation and SSIM, as observed from Tables XI and XII, we achieve the best result.

To demonstrate the effectiveness of the 3 multi-scale discriminators, we further investigate different numbers of discriminators and downsample rates. Specifically, we consider

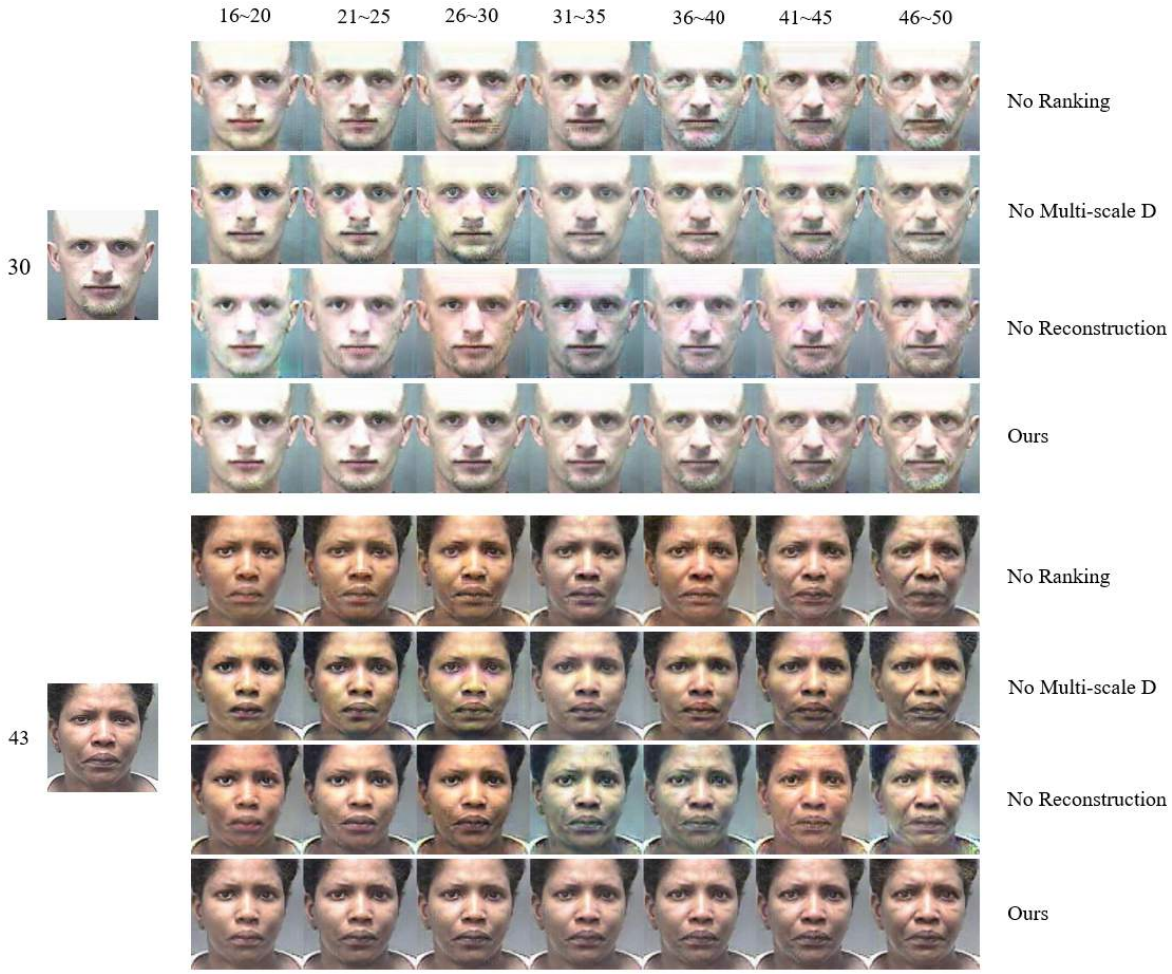

Fig. 7. Ablation study results on MORPH. Input faces and their actual ages are put in the leftmost column.

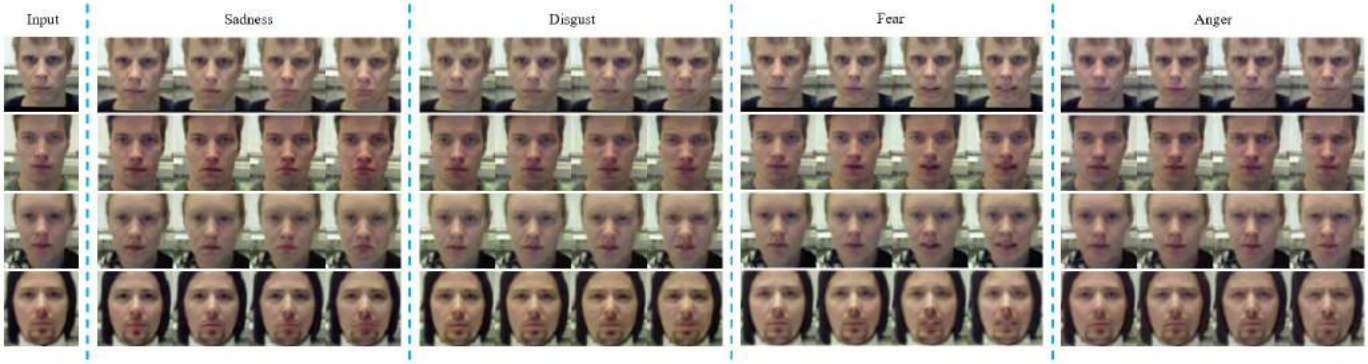

Fig. 8. Sadness, disgust, fear and anger synthesis obtained by using ranking GAN on Oulu-CASIA.

TABLE XII

ABLATION STUDY RESULTS WITH SSIM ON MORPH

Age Group	$16 \sim 20$	$21 \sim 25$	$26 \sim 30$	$31 \sim 35$	$36 \sim 40$	$41 \sim 45$	$46 \sim 50$
Ours	0.881 ± 0.02	0.887 ± 0.02	0.879 ± 0.02	0.884 ± 0.02	0.880 ± 0.02	0.876 ± 0.02	0.862 ± 0.02
No Ranking	0.831 ± 0.03	0.818 ± 0.03	0.801 ± 0.03	0.774 ± 0.08	0.803 ± 0.03	0.810 ± 0.03	$0.791 \sim 0.04$
No Multi-scale D	0.877 ± 0.02	0.867 ± 0.02	0.843 ± 0.03	0.869 ± 0.03	0.861 ± 0.02	0.851 ± 0.02	$0.849 \sim 0.03$
No Reconstruction	0.782 ± 0.04	0.791 ± 0.03	0.789 ± 0.03	0.802 ± 0.03	0.786 ± 0.03	0.781 ± 0.03	$0.739 \sim 0.03$

the use of 1, 2 and 3 discriminators. For downsample rates, we consider 2 and 4. Therefore, we can get a total of 4 schemes. They are respectively D_1, $D_1 + D_2$, $D_1 + D_3$ and $D_1 + D_2 + D_3$. We perform happiness synthesis on CK+ for

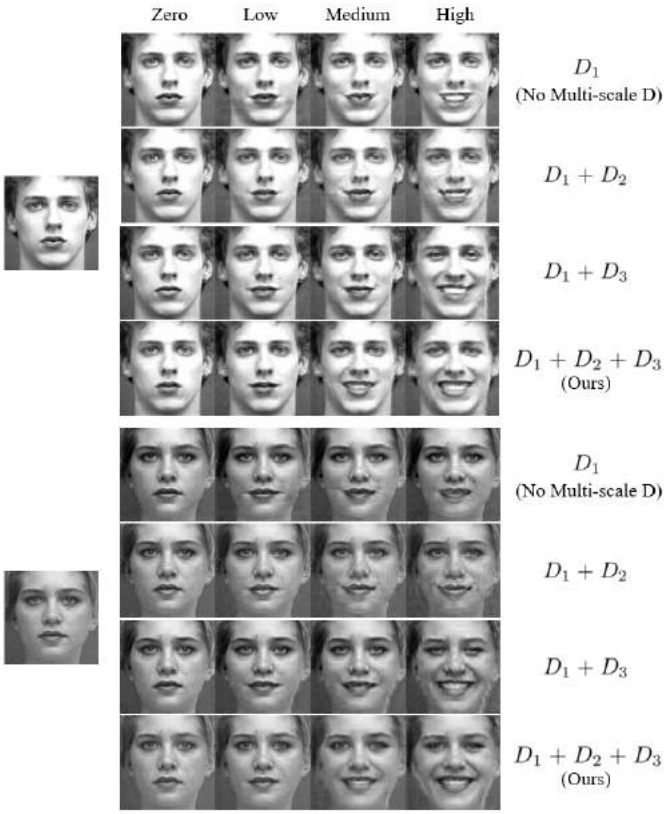

Fig. 9. Ablation study results on CK+. Input faces are put in the leftmost column.

this study. The results are given in Figure 9. Note that we report only qualitative results, since it is easy to observe from these results the superiority of our scheme ($D_1 + D_2 + D_3$). For example, when adopting the other 3 schemes, faces with low and medium intensities look very similar to each other. Faces with medium intensities do not even open the mouth and show teeth.

E. Generalization Capability Study

Finally, we examine the generalization capability of ranking GAN in synthesizing other basic expressions including sadness, disgust, fear and anger. We conduct the experiments on Oulu-CASIA. In Figure 8, we give several examples. As observed, for all the 4 expressions , we obtain promising results. Identity is well preserved during expression process. The intensity also increases gradually from zero to peak.

VI. Conclusion and Future Work

In this paper, to fully utilize the well-ordering property of human ages and expression intensities, we developed an ordinal ranking adversarial network for face aging and expression synthesis with varying intensities. This ranking scheme can take full advantage of limited training data. Moreover, learning-to-rank together with cost sensitivity enables our approach to well catch the correlation among different age ranges/expression intensities. The use of multi-scale discriminators further makes our ordinal rankers more robust, so that aging patterns/expression variation can be successfully captured. We conducted extensive experiments to evaluate the performance of ranking GAN. Promising results demonstrated the effectiveness of our approach on synthesizing face aging sequences and faces with varying expression intensities.

Similar to human ages and expression intensities, facial poses are inherently ordinal and form a well-ordered set. As a future work, we would investigate ordinal ranking and deep generative models for face rotation. In addition, with craniofacial growth, facial growth shows large differences from birth to adulthood. It thus will be more significant to study face aging with a shorter time span for this early stage.

Acknowledgment

The authors would like to thank the associate editor and anonymous reviewers for their useful suggestions and significant efforts spent to help them for further improving their article.

References

[1] N. Ramanathan, R. Chellappa, and S. Biswas, "Computational methods for modeling facial aging: A survey," *J. Vis. Lang. Comput.*, vol. 20, no. 3, pp. 131–144, Jun. 2009.

[2] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955–1976, Nov. 2010.

[3] F. I. Parke and K. Waters, *Computer Facial Animation*. Boca Raton, FL, USA: CRC Press, 2010.

[4] Z. Deng and J. Noh, "Computer facial animation: A survey," in *Data-Driven 3D Facial Animation*. London, U.K.: Springer, 2008, pp. 1–28.

[5] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[6] W. Wang *et al.*, "Recurrent face aging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2378–2386.

[7] C. N. Duong, K. Luu, K. G. Quach, and T. D. Bui, "Longitudinal face modeling via temporal deep restricted Boltzmann machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5772–5780.

[8] C. N. Duong, K. G. Quach, K. Luu, T. H. N. Le, and M. Savvides, "Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3755–3763.

[9] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2089–2093.

[10] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5810–5818.

[11] H. Yang, D. Huang, Y. Wang, and A. K. Jain, "Learning face age progression: A pyramid architecture of GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 31–39.

[12] Z. Wang, X. Tang, W. Luo, and S. Gao, "Face aging with identity-preserved conditional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7939–7947.

[13] P. Li, Y. Hu, R. He, and Z. Sun, "Global and local consistent wavelet-domain age synthesis," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 11, pp. 2943–2957, Nov. 2019.

[14] H. Ding, K. Sricharan, and R. Chellappa, "Exprgan: Facial expression editing with controllable expression intensity," in *Proc.32nd AAAI Conf. Artif. Intell.*, Apr. 2018, pp. 6781–6788.

[15] G. Gu, S. Tae Kim, K. Kim, W. J. Baddar, and Y. Man Ro, "Differential generative adversarial networks: Synthesizing non-linear facial variations with limited number of training data," 2017, *arXiv:1711.10267*. [Online]. Available: http://arxiv.org/abs/1711.10267

[16] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *Proc. ACM Multimedia Conf. Multimedia Conf. (MM)*, 2018, pp. 627–635.

[17] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: Anatomically-aware facial animation from a single image," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 818–833.

[18] Y. Zhou and B. E. Shi, "Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Oct. 2017, pp. 370–376.

[19] N. Ramanathan and R. Chellappa, "Modeling age progression in young faces," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jan. 2006, pp. 387–394.

[20] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 442–455, Apr. 2002.

[21] Y. Tazoe, H. Gohara, A. Maejima, and S. Morishima, "Facial aging simulator considering geometry and patch-tiled texture," in *Proc. ACM SIGGRAPH Posters (SIGGRAPH)*, 2012, p. 90.

[22] Y. Wu, N. M. Thalmann, and D. Thalmann, "A dynamic wrinkle model in facial animation and skin ageing," *J. Vis. Comput. Animation*, vol. 6, no. 4, pp. 195–205, Oct. 1995.

[23] Z. Liu, Z. Zhang, and Y. Shan, "Image-based surface detail transfer," *IEEE Comput. Graph. Appl.*, vol. 24, no. 3, pp. 30–34, May 2004.

[24] D. M. Burt and D. I. Perrett, "Perception of age in adult Caucasian male faces: Computer graphic manipulation of shape and colour information," in *Proc. Roy. Soc. London. B, Biol. Sci.*, vol. 259, pp. 137–143, Feb. 1995.

[25] D. A. Rowland and D. I. Perrett, "Manipulating facial appearance through shape and color," *IEEE Comput. Graph. Appl.*, vol. 15, no. 5, pp. 70–76, 1995.

[26] B. Tiddeman, M. Burt, and D. Perrett, "Prototyping and transforming facial textures for perception research," *IEEE Comput. Graph. Appl.*, vol. 21, no. 4, pp. 42–50, 2001.

[27] Y. Fu and N. Zheng, "M-face: An appearance-based photorealistic model for multiple facial attributes rendering," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 830–842, Jul. 2006.

[28] I. Kemelmacher-Shlizerman, S. Suwajanakorn, and S. M. Seitz, "Illumination-aware age progression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3334–3341.

[29] V. Blanz, C. Basso, T. Poggio, and T. Vetter, "Reanimating faces in images and video," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 641–650, Sep. 2003.

[30] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, "Synthesizing realistic facial expressions from photographs," in *Proc. ACM SIGGRAPH Courses (SIGGRAPH)*, 2005, p. 19-es.

[31] I. Kemelmacher-Shlizerman, A. Sankar, E. Shechtman, and S. M. Seitz, "Being John Malkovich," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 341–353.

[32] K. Li, F. Xu, J. Wang, Q. Dai, and Y. Liu, "A data-driven approach for facial expression synthesis in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 57–64.

[33] F. Yang, L. Bourdev, E. Shechtman, J. Wang, and D. Metaxas, "Facial expression editing in video using a temporally-smooth factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 861–868.

[34] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson, "Generating facial expressions with deep belief nets," in *Affective Computing*. London, U.K.: IntechOpen, 2008.

[35] R. Yeh, Z. Liu, D. B Goldman, and A. Agarwala, "Semantic facial expression editing using autoencoded flow," 2016, *arXiv:1611.09961*. [Online]. Available: http://arxiv.org/abs/1611.09961

[36] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang, "Geometry-contrastive GAN for facial expression transfer," 2018, *arXiv:1802.01822*. [Online]. Available: http://arxiv.org/abs/1802.01822

[37] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," in *Advances in Large Margin Classifiers*. Cambridge, MA, USA: MIT Press, 2000, pp. 115–132.

[38] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Ordinal hyperplanes ranker with cost sensitivities for age estimation," in *Proc. CVPR*, Jun. 2011, pp. 585–592.

[39] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4920–4928.

[40] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5183–5192.

[41] K. Li, J. Xing, C. Su, W. Hu, Y. Zhang, and S. Maybank, "Deep cost-sensitive and order-preserving feature learning for cross-population age estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 399–408.

[42] L. Li and H. T. Lin, "Ordinal regression by extended binary classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 865–872.

[43] M. Kim and V. Pavlovic, "Structured output ordinal regression for dynamic facial emotion intensity prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 649–662.

[44] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung, "Intensity rank estimation of facial expressions based on a single image," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2013, pp. 3157–3162.

[45] O. Rudovic, V. Pavlovic, and M. Pantic, "Context-sensitive dynamic ordinal regression for intensity estimation of facial action units," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 944–958, May 2015.

[46] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–16.

[47] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[48] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[49] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[50] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[51] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[52] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain Image-to-Image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[53] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[54] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit. (FGR)*, 2006, pp. 341–345.

[55] N. Aifanti, C. Papachristou, and A. Delopoulos, "The MUG facial expression database," in *Proc. 11th Int. Workshop Image Anal. Multimedia Interact. Services*, 2010, pp. 1–4.

[56] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.

[57] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, Jun. 2010, pp. 94–101.

[58] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, p. 4.

[59] Megvii Inc. (2019). *Face++*. [Online]. Available: https://www.faceplusplus.com/

[60] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.