JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 37, 779-792 (2021)
DOI: 10.6688/JISE.202107_37(4).0003

Successive Multitask GAN for Age Progression and Regression

RUI-CANG XIE, ZHI-TING CHEN
AND GEE-SERN (JISON) HSU
Department of Mechanical Engineering
National Taiwan University of Science and Technology
Taipei, 10607 Taiwan
E-mail: {birken1018;anson183785}@gmail.com; {jison}@mail.ntust.edu.tw

Due to recent progresses made by state-of-the-art deep learning approaches, the facial age progression and regression has become an attractive research topic in the fields of computer vision. Many existing approaches require paired data which refer to the face images of the same person at different ages. As the cost of collecting such paired datasets is expensive, some emerging approaches have been proposed to learn the facial age manifold from unpaired data. However, the images generated by these approaches suffer from the weakness in generating some age traits, for example wrinkles and creases. To generate better age traits, we propose the Successive Multitask GAN (SM-GAN) for age progression and regression. The SM-GAN consists of n triple networks, $[T_0, T_1, ..., T_{n-1}]$, and a face feature extractor C. Each triple network T_i consists of a generator G_i, a discriminator D_i and a multitask classifier M_i, *i.e.*, $T_i = [G_i, D_i, M_i]$. G_i is trained for transforming between neighboring age groups. D_i is trained to distinguish the generated faces from the real faces in each age group in the training set. M_i is trained for age and gender classification. The face feature extractor C warrants the identity consistency between the input and the generated output of G_i. The pixel-wise loss is also exploited to maintain the image attributes between the input and the generated output.To better define the age groups appropriate for successive age generation, we propose a facial age clustering approach to better determine the boundary ages needed for age segmentation. Experiments show that the proposed SM-GAN can generates better facial age images with better age traits compared with other contemporary approaches.

Keywords: generative adversarial network, face generation, facial age transformation, age progression and regression, face recognition

1. INTRODUCTION

Given a face as input, the facial age progression/regression refers to the generation of facial images at older/younger ages for the same input face in the sense that the identity of the input face can be well preserved in the output. This is a challenging task due to the intrinsic complexity of the facial appearance variation caused by the physical aging process, which can be related to physical condition, gender, race, makeup and other factors. It has received increasing attention in recent years because of the effectiveness of deep learning approaches, the availability of large facial aging datasets and commercial potentials. Some approaches are developed based on GAN. For example, the S^{2}GAN

Received September 20, 2020; revised November 28, 2020 & February 9, 2021; accepted March 3, 2021.
Communicated by Ching-Chun Huang.

[1] encodes the personalized aging basis and applies specific age transforms to create an age representation to decode aged faces. The IPCGAN [2] generates face images for different age groups using a conditional GAN, and preserves the input identity with an identity-preserving module.

The proposed architecture fuses an Adversarial Autoencoder (AAE) and a Generative Adversarial Network (GAN) for facial age progression and regression. The most relevant works include the Conditional Adversarial Autoencoder (CAAE) proposed by Zhang *et al*. [3] and the Disentangled Representation-learning GAN (DR-GAN) proposed by Luan *et al*. [4].

The CAAE is designed to transform an input image to a low-dimensional latent vector by an encoder, which can disentangle personality and age features, and to learn the facial age manifold by mapping the latent vector into a high-dimensional space for image generation. The low-dimensional latent vector is manipulated to control the facial age manifold to achieve smooth age progression and regression. The CAAE framework contains an adversarial network to make the generated image more realistic. However, the CAAE cannot handle non-frontal faces or faces with large illumination variation, and in many cases the generated images lose some age clues, for example, wrinkles.

The DR-GAN is built on the common two-player GAN setup, its generator explores a special encoder-decoder structure, leading to the desired disentanglement. The input to the encoder is a face image of any pose, the output of the decoder is a synthetic face at a target pose, and the learned facial representation connects the encoder and decoder. The discriminator follows the same discriminator design in the Categorical Generative Adversarial (CGA) network [5] which is trained to not only distinguish synthetic from real images, but also predict the identity and pose.

For facial age progression and regression, many approaches consider the age groups that are separated by age boundaries. The most common choice is the 10-year interval, *e.g.*, 21-30, 31-40, 41-50 and beyond 50 with age boundaries at 30, 40 and 50. We call these Regular Boundary (RB), which are postulated in an ad-hoc way without interpretation, and can be different one another in different works [1, 2, 6, 7] . To better determine the age boundaries, we propose the Clustering-based Boundary (CB) that clusters similar age features in each age group.

The novelties of the proposed network include the following:

1. The generator G_i is trained to disentangle the identity-preserving latent vector u_k from the age-dependent code u_a and gender-dependent code u_g. The disentangled representations allows to alter the age of the input image x to produce the desired age at the output $\hat{x}$. Different from the CAAE that computes the L_2 loss between the input and generated images for keeping the facial appearances similar, we extract the identity latent vector from the disentangled representation learning by computing the losses from the discriminators D_i.

2. Unlike the discriminator in the CAAE that only distinguishes the generated (fake) image from the real image, the multitask discriminator M_i in the proposed framework does not just learn to distinguish fake from real images, but also classifies the identity and age of the real and generated images. This makes the proposed network different from the DR-GAN, where the discriminator is trained to distinguish fake/real and classify the pose and identity.

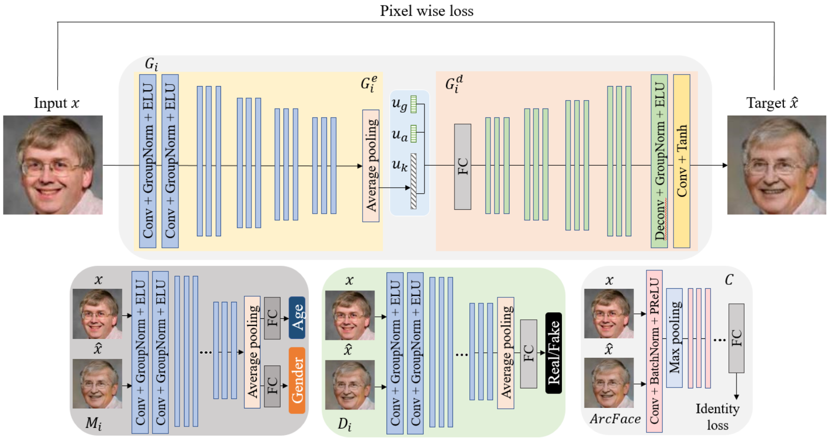


Fig. 1. Our proposed framework for facial age progression/regression consists of a generator G_i, a discriminator D_i and a multitask classifier M_i. G_i consists of the encoder G_i^e and decoder G_i^d. G_i^e encodes an input image to a latent vector u_k, which will be concatenated with an age code u_a and a gender code u_g, and $[u_k, u_a, u_g]$ enters G_i^d as input and generate $\hat{x}_k$. D_i aims to make $\hat{x}_k$ look realistic, and $M_i$constrain $\hat{x}_k$'s age and gender in the desired age and gender, and pixel wise loss make $\hat{x}_k$'s close to x_k to preserve image content.

3. To improve the training stability, we implement the Wasserstein loss in the discriminator D_i following the settings in the Wasserstein GAN with gradient penalty (WGAN-GP) instead of the cross-entropy loss considered in the common GANs.

The rest of the paper is organized as follows: we will present the proposed framework in details in Section 2. Section 3 presents the experimental protocols and results, followed by a conclusion given in Section 4.

2. PROPOSED APPROACH

The proposed framework consists of n triple networks denoted as $[T_0, T_1, ..., T_{n-1}]$ and a face feature extractor C. Each triple network T_i consists of a generator G_i, a discriminator D_i and a multitask classifier M_i, *i.e.*, $T_i = [G_i, D_i, M_i]$. The face feature extractor C warrants the identity consistency between the input and the generated output of G_i. The pixel-wise loss is made to maintain the similarity between the input and the generated output. The system configuration is shown in Fig. 1.

Given a face image x as input and a target age a_t, the piecewise generator $G = [G_0, G_1, ..., G_{n-1}]$ generates an output image $\hat{x} = G(x)$ that preserves the identity of x and shows the age traits of the target age a_t. The n triple networks $[T_i]_{i=0}^{n-1}$ are meant to characterize the facial appearance transformation across the $n+1$ age groups $[A_0, A_1, ..., A_n]$, where A_i is a set of faces within a specific age span, and the ages contained in A_i increase with i. Given the age groups $[A_0, A_1, ..., A_n]$, the learning of the framework involves the following states and settings:

1. The piecewise generator G is structured as a series of n successive generators, *i.e.*, $G = [G_0, G_1, ..., G_{n-1}]$. G_i is trained for transforming between the age groups A_i

and A_{i+1}. Each G_i consists of an encoder G_i^e and a decoder G_i^d. Given an image x_k, the encoder G_i^e is trained to encode x_k into a latent vector $u_k = G_i^e(x_k)$, which is made disentangled of the age and gender through training. u_k will be concatenated with an age code u_a and a gender code u_g to form $v_k = [u_k, u_a, u_g]$. The decoder G_i^d is trained to decode v_k to an image $\hat{x}_k = G_i^d(v_k)$ with the same identity as of x_k and at the target age defined by the age code u_a. $u_a \in \mathbb{R}^2$ is meant to transform the age of $\hat{x}_k = G_i(x_k)$ between A_i and A_{i+1}. The gender code u_g keeps the gender consistency between input and output. The identity is preserved by considering the identity loss evaluated by the face classifier C when training G_i. In addition to the identity loss, the adversarial loss of D_i, the classification loss of M_i and the pixel-wise loss are all considered when training G_i.

2. The discriminator D_i is trained to distinguish the real face images x_k from the generated $\hat{x}_k$.

3. The multitask classifier M_i is trained for age and gender classification. When updating the parameters in M_i during training, we only consider the real data. When updating G_i during training, we consider the classification loss computed from both the real and generated data.

4. The face feature extractor C is formed by the feature embedding layers of a pretrained model which delivers a state-of-the-art performance for face verification. The feature loss between x_k and $\hat{x}_k$ evaluated by C is considered when training G_i.

5. The pixel-wise loss computes the pixel-to-pixel loss of the input x_k and the output $\hat{x}_k$ images, the purpose is to preserve similarity between the input and output some variables, *e.g.*, the pose and illumination condition can be partially kept.

In summary, our framework can be split into two parts, the loss function network C and the pixel-wise loss, and the successive generation networks $[T_i]_i$. The loss function network is made by the feature embedding layers of the networks trained for face recognition and age estimation. The loss function network is *not* updated during training, and used purely as a loss function that computes the difference between the input image x and the generated image $\hat{x}$. The successive generation networks are designed to capture the successive transformation of the facial appearance across different age periods. The objective considered when training the successive generation networks includes the losses evaluated by the loss function networks. The details are presented in the next two sections.

2.1 Loss Function Networks

The facial feature extractor C is formed by the feature embedding layers of the ArcFace network [8]. The ArcFace considers the additive angular margin loss for making the highly discriminative feature for face recognition. The feature can be interpreted geometrically as the correspondence to geodesic distance on a hypersphere. The ResNet-50 and ResNet-100 are employed as the feature embedding network, followed by batch normalization (BN) [9], dropout, a fully connected layer and another BN, generating a 512D embedding feature. In our framework, we choose the ResNet-50 for a faster runtime speed (8.9 ms/face v.s. 15.4 ms/face on ResNet-100 [10]). The embedding feature vectors for the input image x and the generated $\hat{x} = G_i(x)$ can be written as $C(x)$ and the generated $C(\hat{x})$, respectively. The facial feature loss is defined as the cosine distance between $C(x)$ and $C(\hat{x})$,

$$L_c = \frac{C(x) \cdot C(\hat{x})}{||C(x)||||C(\hat{x})||}. \tag{1}$$

We compared the similarity between the input x and the generated $\hat{x}$ by computing the following pixel-wise loss,

$$L_p = \frac{1}{W * H * C}||x - \hat{x}||_2^2. \tag{2}$$

2.2 Successive Generation Network

The three component networks G_i, D_i and M_i are built on the same base net, which is modified from the CASIA network as the CASIA network is relatively simple in structure but delivers a satisfying performance for face verification [11]. To improve the learning properties, we have implemented the group normalization [12] and replaced the MaxPool and ReLU by the strided convolution and exponential linear unit (ELU) [13], respectively.

When an input image x_k is presented to G_i, G_i^e encodes it into a latent vector u_k, which is made disentangled of the age and gender through training. u_k will be concatenated with an age code u_a and a gender code u_g to form the disentangled feature $v_k = [u_k, u_a, u_g]$. The decoder G_i^d is trained to decode u_k to an image $\hat{x}_k = G_i^d(u_k)$ with the same identity as of x_k and at the target age defined by the age code u_a. The age code $u_a \in \mathbb{R}^2$ can control the generation of $\hat{x}_k = G_i(x_k)$ with age in either A_i or A_{i+1}. The gender code u_g aims to keep the gender consistency between input and output. The disentanglement can be achieved if the generated image $\hat{x}$ meets the following requirements: 1) a realistic quality, 2) the generated age trait strong enough to be claimed as for the target age, 3) the gender consistency with x. Additionally, our framework also requires the identity preservation and perceptual similarity with x. For requirement 1), we need to consider the adversarial loss for D_i. For requirements 2) and 3), we need to define the age and gender classification loss for M_i. For the additional requirements, we consider the face feature loss in Eq. (4) and pixel-wise loss when training G_i.

Because of the training instability of the common discriminator caused by the minimization of Jensen-Shannon divergence, we explore the Wasserstein Generative Adversarial Network (WGAN) [14]. The discriminator in the WGAN considers a cost function based on the Wasserstein-1 distance between the data distribution p_d and the model distribution p_g, denoted as $W(p_d, p_g)$, converting the problem to the cost of transporting the mass of p_g to that of p_d. We employ the WGAN with a gradient penalty (WGAN-GP) [14], where a constraint is imposed on the gradient norm of the discriminator's output and the following adversarial loss $L_{adv}^{D_i}$ is considered:

$$L_{adv}^{D_i} = \quad \mathbb{E}_{x \sim p_d}\left[D^i(x)\right] - \mathbb{E}_{\hat{x} \sim p_g}\left[D^i(\hat{x})\right] + \lambda \mathbb{E}_{\hat{x} \sim p_{\hat{x}}}\left[(||\nabla_{\hat{x}} D^i(\hat{x}) - 1||_2)^2\right] \tag{3}$$

where λ is the parameter to adjust the gradient penalty.

The multitask classifier M_i aims for enhancing the age classification and enforcing the gender consistency between the input and output. Instead of using the conventional softmax function to compute the classification loss, we employed the additive angular margin loss (known as ArcFace loss [8]), which is written as follows:

$$L_m = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\|\mathbf{x}_i\|\psi(\theta_{y_i,i}+m)}}{e^{\|\mathbf{x}_i\|\psi(\theta_{y_i,i}+m)}+\sum_{j\neq y_i}^{n}e^{\|\mathbf{x}_i\|\cos(\theta_{j,i})}} \tag{4}$$

where N is the batch size, n is the class number, $\theta_{j,i}$ is the angle between the weight w_j and the deep feature of $\mathbf{x}_i$ that belongs to the y_i-th class, m is the additive angular margin penalty between $\mathbf{x}_i$ and w_j. $\psi(\theta_{y_i,i})$ is a monotonic function, $\psi(\theta_{y_i,i}) = (-1)^k\cos(m_a\theta_{y_i,i}) - 2k$, where $\theta_{y_i,i} \in [\frac{k\pi}{m_a}, \frac{(k+1)\pi}{m_a}]$ and $k \in [0, m_a - 1]$. $m_a \geq 1$ is an integer that controls the angular margin. See [8] for details.

The ArcFace loss in (4) is employed to compute the age classification loss L_a for the age groups $[A_i]_i$, and the gender classification loss L_g for the gender groups, which are set as 2 groups by default. When training M_i, we optimize $L_a^{M_i}$ and $L_g^{M_i}$ by considering the real images $[x_i]_i$ only. When training G_i, we optimize $L_a^{G_i}$ and $L_g^{G_i}$ by considering both the real images $[x_i]_i$ and the generated $[\hat{x}_i]_i$, as G_i is trained to generate better $\hat{x}_i$ to be correctly classified in age and gender.

The above discussion can be summarized to describe the three losses: 1) The loss considered when training the generator G_i includes the adversarial loss $L_{adv}^{D_i}$, the age classification loss $L_a^{G_i}$, the gender classification loss $L_g^{G_i}$, the face classification loss L_c and the pixel-wise loss L_p; 2) The loss considered when training D_i is the adversarial loss $L_{adv}^{D_i}$ only; 3) The loss considered when training M_i includes the age classification loss $L_a^{M_i}$ and the gender classification loss $L_g^{M_i}$. Therefore, the loss functions for G_i, D_i and M_i can be written as follows:

$$L_{G_i} = L_{adv}^{D_i} + \lambda_a L_a^{G_i} + \lambda_g L_g^{G_i} + \lambda_c L_c + \lambda_p L_p \tag{5}$$

$$L_{D_i} = L_{adv}^{D_i} \tag{6}$$

$$L_{M_i} = L_a^{M_i} + L_g^{M_i} \tag{7}$$

2.3 Clustering for Age Segmentation

Instead of using the common ad-hoc ways for defining the age groups (as those reviewed in Section 1), a clustering approach is proposed for the desired age segmentation. The approach is developed based on the fact that the difference in facial appearance between two faces of different ages increases with the age gap, and the same also applies to two groups of faces of different ages. We first extract the facial age feature by using a fine-tuned Deep EXpectation (DEX) age estimator [15], and apply the Gaussian Mixture Model (GMM) [16] to cluster the extracted age features. Given a training set D_t and a validation set D_v, the proposed approach consists of the following steps:

1. We retrain the pretrained DEX age estimator on the training set D_t so that the retrained DEX model can better encode the faces in the MORPH and CACD datasets. We extract the 4096D features from the last fully-connected (fc) layer as the facial age feature.

2. As the 4096D feature vector is too high in dimension to be properly clustered,

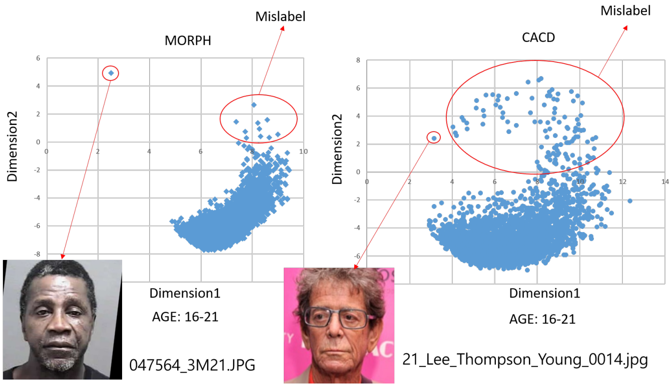


Fig. 2. MORPH and CACD mislabel distribution under age 16-22.

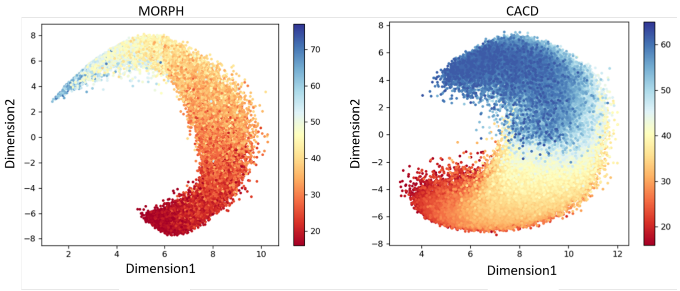
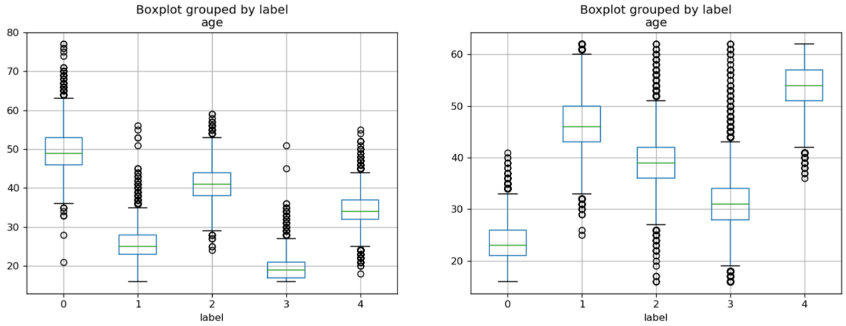


Fig. 3. MORPH and CACD reduce the dimension to a two-dimensional distribution.

we run an exhaustive search for the appropriate feature dimension by using the TruncatedSVD [17] together with the GMM (Gaussian Mixture Model) clustering. It is found that the 8D reduced feature vector is sufficient for characterizing the ages in both the MORPH and CACD.

3. It is known that the CACD [18] has a lot of labeling noises. Some samples of the noises can be seen using the 8D age feature vectors. To better observe these noise samples, we further reduce the dimension of the 8D age feature vector to 2D by using the aforementioned TruncatedSVD [17]. Fig. 2 shows some portion of such samples in the 2D feature space with two corresponding images displayed.
4. To remove the data with labeling noise, we only select the data with 8D age feature within a threshold for each age group in the feature space as the valid set. The threshold is formed by $r \cdot \sigma_i$, where r is a weight factor and σ_i is the standard deviation of the data within age group i. We ran experiments on r from 1 to 2, and selected 1.6 as the best. Fig. 3 shows the noise-removed data, reduced to 2D for better visualization.

Based on the above approach, the clustering of the first two dimensional components for the MORPH and CACD is shown in Fig. 4. We select the age groups: 16-22 23-30 31-37 38-45 46+ for the MORPH and 16-27 28-35 36-42 43-50 51+ for the CACD.



MORPH CACD

Fig. 4. MORPH and CACD result of clustering-based boundary.

3. EXPERIMENTS

3.1 Database and Setup

The MORPH [19] and the CACD [18] databases are used in our experiments. The MORPH is one of the largest publicly available longitudinal face database with mugshot images, and it includes the meta data for race, gender, date of birth, and date of acquisition. Our experiments were performed on MORPH Album-2 which contains 55,134 images of 13,000 individuals with age between 16 to 77 years. The CACD [18] contains 163,446 face images of 2,000 celebrities captured in much less controlled conditions. Besides large variations in pose, illumination, and expression, images in CACD are collected via Google Image Search, making it a very challenging dataset due to the mismatching between actual face presented in each image and associated labels provided (name and age). The FG-NET[20] dataset consists of 1,002 images of 82 individuals with age between 1 to 69. As only very limited images are available, we use the FG-NET for the validation experiment in which the generated images are compared against the real face images.

According to Clustering-based Boundary results, we divided each database into five age groups. The thresholds adopted in our face verification experiments were the same as those used in [21], *i.e.*, threshold = 76.5 for FAR = $1e-5$. We randomly select 80% of the images as the training set and the rest 20% as the testing set and ran five-fold cross validation for the experiments. For each run, four folds were used for training, and the remaining fold was used for performance evaluation. The average of the five outcomes was taken as the performance to report. The metric measurements for the performance are all conducted via the public APIs of Face++ [22]. To set up the experiments, we first aligned the faces by using the landmark detection algorithm and code offered in [23], and cropped each image to 100×100 pixels. When each image was entered to the framework, it was cropped randomly using a 96×96 window for data augmentation. We used the Adam optimizer to solve for the networks G_i and D_i with learning rate $2e^{-4}$ and batch size 64. The networks were trained from scratch. We updated the discriminator D_i for every 4 iterations on the generator G_i.

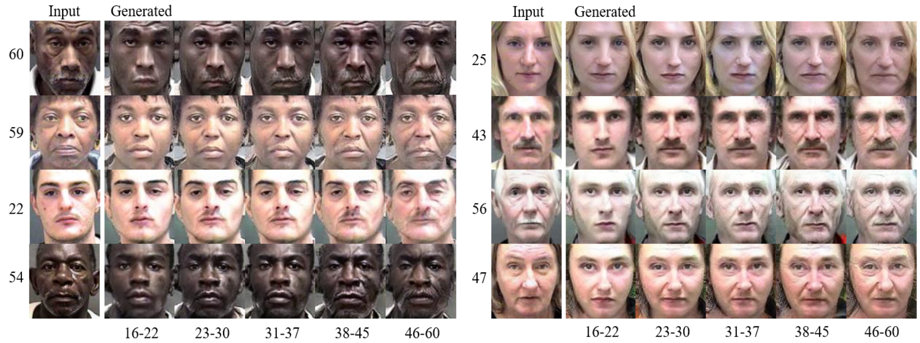


Fig. 5. The samples generated for age progression and regression from the Morph. The leftmost column shows the input images with corresponding age labels, and the rest columns are images generated for different age groups.

3.2 Performance Evaluation

3.2.1 Age transfer accuracy

Given an input face and target age label, the proposed network is trained to generate an output face with the same identity as input and in the age as specified by the age label. Some samples of the generated images and the given input faces are shown in Fig. 5. On the MORPH and CACD, the face images of the whole testing set are transferred to each age group. We estimated the ages of both the generated images and the originals in the dataset using the Face++ APIs for fair comparison. The performance in Table 1 shows the comparison of the proposed SM-GAN with three contemporary approaches, the CAAE [3], IPCGAN [2] and S^{2}GAN [1]. There are two settings for our SM-GAN in Table 1, one with the ad-hoc regular boundaries (RB) for the following age intervals: 16-20, 21-30, 31-40, 41-50 and 51+ and the Clustering-based Boundary (CB). Following the CB scheme presented in Section 2.3, we select the following age intervals for MORPH: 16-22, 23-30, 31-37, 38-45 and 46+; and 16-27, 28-35, 36-42, 43-50 and 51+ for the CACD. The upper part of Table 1 shows the estimated ages and the lower part shows the mean absolute errors (MAE). It demonstrates that the proposed SM-GAN with RB can be comparable in performance to the IPCGAN and S^{2}GAN, and it can outperform both with CB.

3.2.2 Identity preservation

In identity evaluation we only consider translations from the youngest group to the other age groups, same as [21]. To evaluate the performance of the proposed method objectively, all metric measurements are conducted via stable public APIs of Face++ [22]. Thresholds adopted in our face verification experiments (threshold = 76.5, FAR = $1e-5$) are the same as those used in [21]. Therefore, quantitative results of our experiments are comparable to those reported in [21]. Face verification experiments are conducted to check whether the identity information has been preserved during the face aging process. Similar to previous literature, comparisons between synthetic elderly face images from different age groups of the same subject are also conducted to inspect if the identity information is consistent among different separately trained age mappings. The results are shown in Table 2. It shows that our SM-GAN with RB can effectively retain the iden-

Table 1. Age evaluation on MORPH and CACD (Age).

	MORPH					CACD				
Age groups	-20	21-30	31-40	41-50	51+	-20	21-30	31-40	41-50	51+
CAAE [3]	22.03	25.1	30.14	34.69	40.6	25.42	29.64	32.61	37.92	41.9
IPCGAN [2]	21.42	28.74	35.91	44.8	51.11	19.73	27.43	36.22	44.7	52.53
S^{2}GAN [1]	18.26	25.83	35.44	45.22	53.64	17.61	24.05	36.07	45.73	55.31
Ours (RB)	20.97	28.25	35.42	45.2	52.97	16.64	28.08	36.9	45.31	52.43
Age groups	-22	23-30	31-37	38-45	46+	-27	28-35	36-42	43-50	51+
Ours (CB)	21.88	27.84	35.43	42.04	51.45	24.26	31.94	38.41	47.31	52.66
Absolute Difference between Mean Ages (MAE)										
CAAE [3]	2.72	1.62	7.7	12.74	14.65	6.8	3.33	1.91	9.48	11.81
IPCGAN [2]	2.11	2.02	1.93	2.63	4.14	1.11	1.12	1.7	2.7	1.18
S^{2}GAN [1]	**1.05**	0.89	2.4	2.21	**1.61**	**1.01**	2.26	1.55	1.67	1.6
Ours (RB)	1.66	1.53	2.42	2.23	2.28	1.98	1.77	2.38	2.09	1.28
Ours (CB)	1.39	**0.84**	**0.98**	**1.46**	1.89	**1.01**	**1.68**	**1.35**	**1.37**	**1.05**

Table 2. Face verification rates on MORPH and CACD (%).

	MORPH					CACD				
Age groups	-20	21-30	31-40	41-50	51+	-20	21-30	31-40	41-50	51+
CAAE [3]	57.5	53.9	58.7	6.0	5.6	57.6	61.8	43.8	37.9	11.0
IPCGAN [2]	62.3	47.8	76.3	79.7	54.2	73.8	80.3	72.4	73.1	77.2
S^{2}GAN [1]	95.1	93.3	92.3	95.0	89.3	96.4	97.2	94.9	**97.2**	95.2
Ours (RB)	**99.89**	98.49	97.78	**97.11**	**96.45**	99.84	97.43	**98.66**	97.18	96.51
Age groups	-22	23-30	31-37	38-45	46+	-27	28-35	36-42	43-50	51+
Ours (CB)	99.76	**99.52**	**98.83**	96.33	94.33	**99.88**	**97.9**	98.37	97.02	**96.72**

tity compared to other methods, and it is much higher than the IPCGAN [2] and S^{2}GAN [1]. It can also be seen that our SM-GAN with CB performs similarly, reflecting that the age boundaries can have a stronger influence on the facial age generation than on the identity preservation. However, to better transform facial age and preserve identity, the appropriately designed age boundaries are still needed.

3.2.3 Qualitative comparison with prior work

For qualitative comparison, we select several state-of-the-art approaches, including the face transformer (FT) [24], the coupled dictionary learning (CDL) [25], the hidden factor analysis (RFA) [26], the CAAE [3], the improved CAAE (CAAE++) [27], the contextual GANs (C-GAN) [28], the global and local consistent age GANs (GLCA-GAN) [7], and the Pyramid Architecture of GANs (PAGAN) [6]. Fig. 6 shows the comparison, with the input images on the top row, the images generated by other approaches in the middle row and the image generated by our approach at the bottom. The images generated by other approaches are directly cropped from their papers and pasted. It shows that our approach can generate clearer wrinkles, gray beards and more senior characteristics for progressing the age; it can also generate baby-like faces for regressing the age and generate good results in diverse backgrounds.

4. CONCLUSION

We propose the Successive Multitask GAN (SM-GAN) for age progression and regression. The SM-GAN is composed of n triple networks, $[T_0, T_1, ..., T_{n-1}]$, and a face

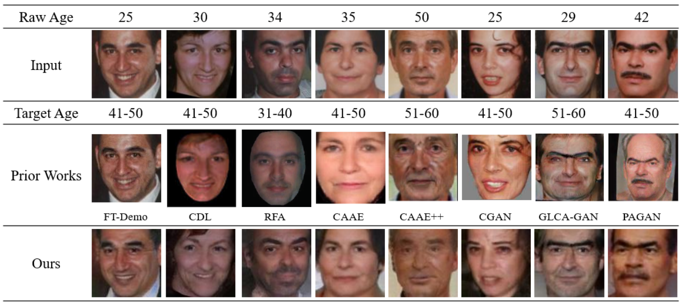


Fig. 6. Comparison with state-of-the-art approaches. Top row are input images, the images generated by SOTA in the middle row and those generated by the proposed approach are in the bottom.

feature extractor C. Each triple network T_i consists of a generator G_i, a discriminator D_i and a multitask classifier M_i, *i.e.*, $T_i = [G_i, D_i, M_i]$. G_i is trained together with D_i and M_i for transforming between neighboring age groups, distinguishing the generated faces from the real faces in each age group, and classifying the age and gender of the generated faces. To better determine the boundary ages required by the SM-GAN, we propose a Clustering-based Boundary which is verified effective in our experiments. The proposed network also explores the state-of-the-art facial feature extractor and the pixel-wise loss to preserve identity and image attributes of the input. Experiments have demonstrated the performance for facial age progression/regression with identity preservation and robustness against cluttered backgrounds.

REFERENCES

1. Z. He, M. Kan, S. Shan, and X. Chen, "S2gan: Share aging factors across ages and share aging trends among individuals," in *Proceedings of IEEE International Conference on Computer Vision*, 2019, pp. 9440-9449.
2. Z. Wang, X. Tang, W. Luo, and S. Gao, "Face aging with identity-preserved conditional generative adversarial networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7939-7947.
3. Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5810-5818.
4. L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1415-1424.
5. J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," *arXiv Preprint*, 2015, arXiv:1511.06390.
6. H. Yang, D. Huang, Y. Wang, and A. K. Jain, "Learning face age progression: A pyramid architecture of gans," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 31-39.

7. P. Li, Y. Hu, Q. Li, R. He, and Z. Sun, "Global and local consistent age generative adversarial networks," in *Proceedings of IEEE 24th International Conference on Pattern Recognition*, 2018, arXiv:1801.08390.
8. J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, arXiv:1801.07698.
9. S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv Preprint*, 2015, arXiv:1502.03167.
10. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770-778.
11. D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv Preprint*, 2014, arXiv:1411.7923.
12. Y. Wu and K. He, "Group normalization," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3-19.
13. D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv Preprint*, 2015, arXiv:1511.07289.
14. M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv Preprint*, 2017, arXiv:1701.07875.
15. R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proceedings of IEEE International Conference on Computer Vision WWW*, 2015, pp. 10-15.
16. D. A. Reynolds, "Gaussian mixture models," *Encyclopedia of Biometrics*, 2015, pp. 827-832.
17. P. C. Hansen, "The truncated SVD as a method for regularization," *BIT Numerical Mathematics*, Vol. 27, 1987, pp. 534-553.
18. B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Proceedings of the 13th European Conference on Computer Vision*, LNCS 8694, 2014, pp. 768-783.
19. K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 341-345.
20. G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes, "Overview of research on facial ageing using the fg-net ageing database," *IET Biometrics*, Vol. 5, 2016, pp. 37-46.
21. Y. Liu, Q. Li, and Z. Sun, "Attribute-aware face aging with wavelet-based generative adversarial networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11877-11886.
22. Megvii Inc., "Face++ research toolkit," http://www.faceplusplus.com.
23. 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 *Proceedings of International Conference on Computer Vision*, 2017, arXiv:1703.07332.
24. C. S. D. at Aberystwyth University, "Face of the future," http://cherry.dcs.aber.ac.uk/Transformer/index.html.

25. X. Shu, J. Tang, H. Lai, L. Liu, and S. Yan, "Personalized age progression with aging dictionary," in *Proceedings of IEEE International Conference on Computer Vision*, 2015, pp. 3970-3978.
26. H. Yang, D. Huang, Y. Wang, H. Wang, and Y. Tang, "Face aging effect simulation using hidden factor analysis joint sparse representation," *IEEE Transactions on Image Processing*, Vol. 25, 2016, pp. 2493-507.
27. J. Zeng, X. Ma, and K. Zhou, "Caae++: Improved caae for age progression/ regression," *IEEE Access*, 2018, pp. 66715-66722.
28. S. Liu, Y. Sun, D. Zhu, R. Bao, W. Wang, X. Shu, and S. Yan, "Face aging with contextual generative adversarial nets," in *Proceedings of the 25th ACM International Conference on Multimedia*, 2017, pp. 82-90.

Rui-Cang Xie received the B.S. degree in Department of Photonics from Feng Chia University, Taichung, Taiwan, in 2016, and the M.S. degree in Mechanical Engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2020. His research interest on deep learning and computer vision, in particular, generative adversarial network.

Zhi-Ting Chen received the B.S. degree in Mechanical Engineering from Chung Yuan Christian University, Taoyuan, Taiwan, in 2018. He is now pursuing his M.S. at National Taiwan University of Science and Technology. His research interest on deep learning and computer vision, in particular, age transformation and generative adversarial network.

Gee-Sern (Jison) Hsu received his dual M.S. degree in Electrical and Mechanical Engineering and his Ph.D. in Mechanical Engineering from the University of Michigan, Ann Arbor, in 1993 and 1995, respectively. From 1995 to 1996, he was a Post-Doctoral Fellow with the University of Michigan. From 1997 to 2000, he was a Senior Research Staff with the National University of Singapore. In 2001, he joined Penpower Technology, where he led research on face recognition and intelligent video surveillance. His team at Penpower Technology was a recipient of the Best Innovation and Best Product Award at the SecuTech Expo for 3 consecutive years (2005 2007). In 2007, he joined the Department of Mechanical Engineering, National Taiwan University of Science and Technology, where he is now an Associate Professor. His research interests include computer vision and pattern recognition. He received best paper awards in ICMT 2011, CVGIP 2013, CVPRW 2014, ARIS 2017 and CVGIP 2018. He is a senior member of IEEE and IAPR.