AgeTransGAN for Facial Age Transformation with Rectified Performance Metrics

Gee-Sern Hsu, Rui-Cang Xie, Zhi-Ting Chen, and Yu-Hong Lin

National Taiwan University of Science and Technology, Taipei, Taiwan
{jison,m10703430,m10803432,m10903430}@mail.ntust.edu.tw

Abstract. We propose the AgeTransGAN for facial age transformation and the improvements to the metrics for performance evaluation. The AgeTransGAN is composed of an encoder-decoder generator and a conditional multitask discriminator with an age classifier embedded. The generator considers cycle-generation consistency, age classification and cross-age identity consistency to disentangle the identity and age characteristics during training. The discriminator fuses age features with the target age group label and collaborates with the embedded age classifier to warrant the desired target age generation. As many previous work use the Face++ APIs as the metrics for performance evaluation, we reveal via experiments the inappropriateness of using the Face++ as the metrics for the face verification and age estimation of juniors. To rectify the Face++ metrics, we made the Cross-Age Face (CAF) dataset which contains 4000 face images of 520 individuals taken from their childhood to seniorhood. The CAF is one of the very few datasets that offer far more images of the same individuals across large age gaps than the popular FG-Net. We use the CAF to rectify the face verification thresholds of the Face++ APIs across different age gaps. We also use the CAF and FFHQ-Aging datasets to compare the age estimation performance of the Face++ APIs and an age estimator that we made, and propose rectified metrics for performance evaluation. We compare the AgeTransGAN with state-of-the-art approaches by using the existing and rectified metrics.

1 Introduction

Facial age transformation refers to the generation of a new face image for an input face such that the generated face is the same identity as the input but at the age specified by the user. Facial age transformation is an active research topic in the fields of computer vision [1,5,23,13,17]. It is a challenging task due to the intrinsic complexity of the facial appearance variation caused by the physical aging process, which can be affected by physical condition, gender, race and other factors [4]. It has received increasing attention in recent years because of the effectiveness of the GANs [11,13,17,5,1], the availability of large facial age datasets and application potentials. It can be applied in the entertainment and cinema industry, where an actor's face often needs to appear in a younger or older age. It can also be applied to find missing juniors/seniors as the pictures for reference can be years apart.

Many approaches have been proposed in recent years [1,5,11,23,13,17]. However, many issues are yet to be addressed. The performance of the approaches is usually evaluated by target age generation and identity (ID) preservation. The former measures if the generated facial age reaches the target age, and the latter measures if the identity is kept well after the transformation. The best compromise between target age generation and ID preservation is hard to define because facial appearance does not change much across a small age gap, but it can change dramatically across a large age gap. For ID preservation, many approaches use a pretrained face model to make the generated face look similar to the input source [1,24,23], which constrains the age trait generation on the target face. Another big issue is the metrics for a fair performance evaluation. Many handle this issue by using a commercial software, e.g., the Face++ APIs [8], which is a popular choice [13,24,12]. Some turn to manual evaluation via a crowd-sourcing platform [17]. Some use proprietary age estimation models [6]. We address most of the above issues in this paper.

We propose the AgeTransGAN to handle bidirectional facial age transformation, i.e., progression and regression. The AgeTransGAN is composed of an encoder-decoder generator and a conditional multitask discriminator with an age classifier embedded. The generator explores cycle-generation consistency, age classification and cross-age identity consistency to disentangle the identity and age characteristics during training. The discriminator fuses age features with the target age group label and collaborates with the embedded age classifier to warrant the desired age traits made on the generated images.

To address the flaw of using Face++ APIs for performance evaluation and the constraint of using a pretrained face model for ID preservation, we made the Cross-Age Face (CAF) dataset which contains 4000 face images of 520 individuals taken from their childhood to seniorhood. Each face in the CAF has a ground-truth age label, and each individual has images across large age gaps. To the best of our knowledge, the CAF is one of the very few datasets that offer face images of the same individuals across large age gaps, and it contains more individuals and images than the popular FG-Net. We use the CAF to rectify the face verification thresholds of Face++ APIs which perform poorly when verifying junior and children faces. The cross-age face verification rate is an indicator for ID preservation. We also use the CAF and the FFHQ-Aging datasets to compare the age estimation performance of Face++ APIs and a tailor-made age estimator which will be released with this paper. We summarize the contributions of this paper as follows:

1. A novel framework, the AgeTransGAN, is proposed and verified effective for identity-preserving facial age transformation. The novelties include the network architecture and loss functions designed to disentangle the identity and age characteristics so that the AgeTransGAN can handle transformation across large age gaps.
2. The Conditional Multilayer Projection (CMP) discriminator is proposed to extract the multilayer age features and fuse these features with age class labels for better target age generation.

3. A novel database, the Cross-Age Face (CAF), is released with this paper. It is one of the very few databases that offer 4000 images of 520 individuals with large age gaps from early childhood to seniorhood. It can be used to rectify face verification across large age gaps and verify age estimators.
4. Experiments show that the AgeTransGAN demonstrates better performance than state-of-the-art approaches by using both the conventional evaluation metrics and the new metrics proposed in this paper. To facilitate related research, we release the trained models with this paper, https://github.com/AvLab-CV/AgeTransGAN.

In the following, we first review the related work in Sec. 2, followed by the details of the proposed approach in Sec. 3. Sec. 4 presents the experiments for performance evaluation, and a conclusion is given in Sec. 5.

2 Related Work

As our approach is related to high-resolution image generation, the conditional GAN and the facial age transformation, this review covers all these topics.

Motivated by the effectiveness of the adaptive instance normalization (AdaIN) [7], the StyleGAN [9] defines a new architecture for high-resolution image generation with attribute separation and stochastic variation. The generator is composed of a mapping network, a constant input, a noise addition, the AdaIN and the mixing regularization. The mapping network transfers the common latent space into an intermediate but less entangled latent space. Instead of using a common random vector as input, the StyleGAN uses the intermediate latent vector made by the mapping network. Given the intermediate latent vector, the learned affine transformations produce the styles that manipulate the layers of the generator via the AdaIN operation. To extract the styles from different layers, the generator processes the bias and noise broadcast within each style block, making the relative impact inversely proportional to the style magnitudes. The modified version, the StyleGAN2 [10], moves the bias and noise broadcast operations outside of the style block, applies a revised AdaIN to remove the artifacts made by StyleGAN, and uses the perceptual path length (PPL) as a quality metric to improve the image quality. The architectures for the generator and discriminator are modified by referring to other work for improvements.

The conditional GAN (cGAN) is considered a promising tool for handling class-conditional image generation [16]. Unlike typical GANs, the discriminators in the cGANs discriminate between the generation distribution and the target distribution given the pairs of generated data x and the conditional variable y. Most cGANs feed the conditional variable y into the discriminator by concatenating y to the input or to some feature vectors [14,18,25,21]. However, the cGAN with projection discriminator (PD) [15] considers a different perspective of incorporating the conditional information into the discriminator. It explores a projection scheme to merge the conditional requirement in the model. The discriminator takes an inner product of the embedded conditional vector y with

the feature vector, leading to a significant improvement to the image generation on the ILSVRC-2012 dataset.

A significant progress has been made in facial age transformation recently. The Identity-Preserved Conditional GAN (IPC-GAN) [22] explores a cGAN module for age transfer and an identity-preserved module to preserve identity with an age classifier to enhance target age generation. The Pyramid Architecture of GAN (PA-GAN) [23] separately models the constraints for subject-specific characteristics and age-specific appearance changes, making the generated faces present the desired aging effects while keeping the personalized properties. The Global and Local Consistent Age GAN (GLCA-GAN) [11] consists of a global network and a local network. The former learns the whole face structure and simulates the aging trend, and the latter imitates the subtle local changes. The wavelet-based GAN (WL-GAN) [13] addresses the matching ambiguity between young and aged faces inherent to the unpaired training data. The Continuous Pyramid Architecture of GAN (CPA-GAN) implements adversarial learning to train a single generator and multiple parallel discriminators, resulting in smooth and continuous face aging sequences.

Different from previous work that focuses on adult faces and considers the datasets such as MORPH [19] and CACD [2], the Lifespan Age Transformation Synthesis (LATS) [17] redefines the age transformation by considering a lifespan dataset, the FFHQ-aging, in which 10 age groups are manually labeled for ages between 0 and 70+. Built on the StyleGAN [9], the LATS considers the 10 age groups as 10 domains, and applies multi-domain translation to disentangle age and identity. The Disentangled Lifespan Face Synthesis (DLFS) [5] proposes two transformation modules to disentangle the age-related shape and texture and age-insensitive identity. The disentangled latent codes are fed into a StyleGAN2 generator [10] for target face generation. Considering aging as a continuous regression process, the Style-based Age Manipulation (SAM) [1] integrates four pretrained models for age transformation: a StyleGAN-based encoder for image encoding, the ArcFace [3] for identity classification, a VGG-based model [20] for age regression and the StyleGAN2 for image generation. Different from the above approaches that either directly use the pretrained StyleGAN or made minor modifications, the proposed AgeTransGAN makes substantial modifications to the overall StyleGAN2 architecture so that the generator can better disentangle age and identity characteristics, and the discriminator can better criticize the age traits made on the generated images.

3 Proposed Approach

The proposed AgeTransGAN consists of an encoder-decoder generator $G = [G_{en}, G_{de}]$, where G_{en} is the encoder and G_{de} is the decoder, and a conditional multitask discriminator D_p. The details are presented below.

3.1 Encoder and Decoder

The configuration of the generator $G = [G_{en}, G_{de}]$ is shown in Figure 1. The multitask encoder G_{en} takes the source image I_i and the target age group label

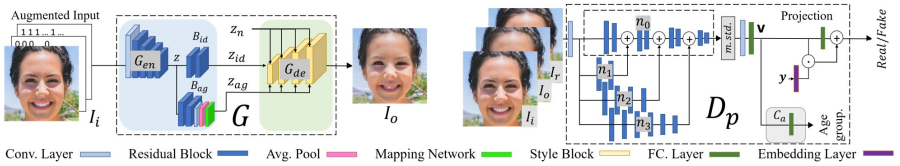

Fig. 1. [Left] The generator $G = [G_{en}, G_{de}]$ with networks B_{id} and B_{ag} for age and identity disentanglement. [Right] The Conditional Multitask Projection (CMP) discriminator D_p with four subnets $[\mathbf{n}_k]_{k=1}^4$ for multilayer feature extraction and an age classifier C_a. See supplementary document for details on network settings.

y_t as input, and generates the identity latent code z_{id} and the age latent code z_{ag}. z_n is a Gaussian noise that enters the decoder G_{de} after each convolution layer. The encoder G_{en} is developed based on a modification of the StyleGAN2 discriminator, which consists of an input layer, a convolution layer, 8 downsampling residual blocks and a scalar output layer.

We first fuse the 3-layer (RGB) I_i with the target age group label y_t by augmenting I_i with the one-hot array that represents y_t using N_a layers of 0's and 1's, where N_a is the number of all target age groups. The augmented input $\hat{I}_i$ is entered into the common layers of G_{en} to generate a facial representation z. The common layers include the input layer, the convolution layer and the downsampling residual blocks. z is further processed by two independent component networks, B_{id} and B_{ag}, which we propose to disentangle the identity and age characteristics by jointly minimizing a set of specifically designed loss functions. The component network B_{id}, composed of two residual blocks, transfers z to the identity latent code z_{id} that will enter the decoder G_{de}. The other component network B_{ag}, composed of two residual blocks, a convolution layer and a mapping network, transfers z to the age latent code z_{ag}. The training with the loss functions presented below makes z_{id} capture the identity characteristics and z_{ag} capture the age characteristics. We may write z_{id} and z_{ag} as $z_{id}(I, y)$ and $z_{ag}(I, y)$ to indicate their dependence on the input I and label y. The identity latent code z_{id} and the age latent code z_{ag} will be abbreviated as the ID code and the age code, respectively, in the rest of the paper for simplicity.

The decoder G_{de} is modified from the StyleGAN2 generator with three modifications: 1) Six additional loss functions considered at training, 2) The original constant input replaced by the ID code z_{id}, and 3) The multi-stream style signals that enter the upsampling style blocks via the AdaIN are replaces by the age code z_{ag}. The generator takes I_i and the target age group label y_o as input to generate the target age output I_o. As we consider the cycle consistency between input and output during training, we also enter the generated output I_o and the input age group label y_i to the generator to reconstruct the source input I_r during training.

The configurations of G_{en} and G_{de} are shown in Figure 1. The details of network settings are given in the supplementary document.

The loss functions considered for training the generator G include the adversarial loss, the identity loss, the cycle-consistency loss, the age class loss, the pixel-wise attribute loss and the perceptual path length regularization. The following adversarial loss $\mathcal{L}_G^{adv}$ warrants the desired properties of the generated faces.

$$\mathcal{L}_G^{adv} = \mathbb{E}_{I_i \sim p(I_i)} \log \left[D_p \left(G(I_i, y_t), y_t \right) \right] + \mathbb{E}_{I_o \sim p(I_o)} \log \left[D_p \left(G(I_o, y_i), y_i \right) \right] \quad (1)$$

The following identity (ID) loss $\mathcal{L}_{id}$ ensures the ID preservation at the output I_o by forcing the ID code z_{id} of the source I_i close to that of I_o.

$$\mathcal{L}_{id} = \| z_{id}(I_i, y_t) - z_{id}(I_o, y_i) \|_1 \quad (2)$$

The triplet loss $\mathcal{L}_t$, defined on the ID code z_{id} as shown in (3) below, moves the reconstructed ID code $z_{id}(I_r, y_i)$ closer to the source ID code $z_{id}(I_i, y_i)$ while moving the output ID code $z_{id}(I_o, y_t)$ further away from the source ID code $z_{id}(I_i, y_i)$.

$$\mathcal{L}_t = \| z_{id}(I_i, y_i) - z_{id}(I_r, y_i) \|_2^2 - \| z_{id}(I_i, y_i) - z_{id}(I_o, y_t) \|_2^2 + m_t \quad (3)$$

where m_t is the margin determined empirically.

Note that both the ID loss $\mathcal{L}_{id}$ in (2) and the triplet loss $\mathcal{L}_t$ in (3) are defined on the ID code z_{id}; but with the following differences: 1) $\mathcal{L}_{id}$ verifies the ID preservation for the transformation across all age groups/classes, i.e., $z_{id}(I_i, y_t), \forall y_t$ and $z_{id}(I_o, y_i), \forall y_i$; however, $\mathcal{L}_t$ only considers the within-class transformation, i.e., $z_{id}(I_i, y_i)$ and $z_{id}(I_o, y_t)$. 2) $\mathcal{L}_{id}$ aims to preserve the identity only between the source input and the generated output; while $\mathcal{L}_t$ aims to enhance the ID preservation between the source and the reconstructed source, and simultaneously penalize the ID preservation across age transformation.

The cycle-consistency loss $\mathcal{L}_{cyc}$ makes the age progression and regression mutually reversible, i.e., the input I_i can be reconstructed from the target I_t in the same way as the target I_t is generated from the input I_i. It is computed by the following L_1 distance between I_i and the reconstructed input $I_r = G(I_o, y_i)$.

$$\mathcal{L}_{cyc} = \| I_i - G(I_o, y_i) \|_1 \quad (4)$$

The following age class loss $\mathcal{L}_a$, which is the cross-entropy loss computed by using the age classifier C_a in the discriminator D_p, is considered when training G (and also when training D_p).

$$\mathcal{L}_a^{(g)} = \mathbb{E}_{I \sim p(I)} [- \log C_a(\mathbf{v}(I), y)] \quad (5)$$

where $\mathbf{v}(I)$ is a latent code generated within the discriminator D_p for image I, and more details are given in Sec.3.2. The following pixel-wise attribute loss $\mathcal{L}_{px}$ is need to maintain the perceptual attribute of I_i at the output I_o.

$$\mathcal{L}_{px} = \mathbb{E}_{I_i \sim p(I_i)} \frac{1}{w \times h \times c} \| I_o - I_i \|_2^2 \quad (6)$$

where w, h, and c are the image dimension. $\mathcal{L}_{px}$ is good at keeping the background, illumination and color conditions of I_i at the generated I_o. Similar losses are used in [11,13,24,1]. Without this loss, as the settings for LATS [17] and DLFS [5], we have to crop each input face during preprocessing.

To encourage that a constant variation in the style signal results in a constant scaled change in the image, the StyleGAN2 employs the following perceptual path length regularization $\mathcal{L}_{pl}$ to make the generator smoother. We apply the same regularization on the age code z_{ag}.

$$\mathcal{L}_{pl} = \mathbb{E}_{z_{ag}} \mathbb{E}_{I_o} \left(\left\| \mathbf{J}_{z_{ag}}^T I_o \right\|_2 - a_p \right)^2 \tag{7}$$

where $\mathbf{J}_{z_{ag}} = \partial G(I_i, y_t)/\partial z_{ag}$ is the Jacobian, and a_p is a constant. $\mathbf{J}_{z_{ag}}^T I_o$ can be written as $\nabla_{z_{ag}}(G(I_i, y_t) \cdot I_o)$ for a better implementation of the needed back propagation.

The 7 loss functions in (1)∼(7) are combined by the following weighted sum to train G.

$$\mathcal{L}_G = \mathcal{L}_G^{adv} + \lambda_{id}\mathcal{L}_{id} + \lambda_t \mathcal{L}_t + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_a^{(g)}\mathcal{L}_a^{(g)} + \lambda_{px}\mathcal{L}_{px} + \lambda_{pl}\mathcal{L}_{pl} \tag{8}$$

where $\lambda_{id}, \lambda_t, \lambda_{cyc}, \lambda_a^{(g)}, \lambda_{px}$ and λ_{pl} are the weights determined empirically.

3.2 CMP Discriminator

The Conditional Multitask Projection (CMP) discriminator D_p is proposed to not just distinguish the generated images from the real ones, but also force the facial traits on the generated faces close to the real facial traits shown in the training set. To attain these objectives, we make a substantial revision to the StyleGAN2 discriminator with three major modifications: 1) Embedding of a multilayer age feature extractor S_a, 2) Integration with a label projection module to make the age-dependent latent code conditional on the target age group label, and 3) Embedding of an age classifier C_a for supervising the target age generation. The configuration of D_p is shown in Figure 1.

Multilayer Age Feature Extractor: We keep the same input layer, the convolution layer and the mini-batch standard deviation as in the StyleGAN2 discriminator, but modify the 8 downsampling residual blocks (res-blocks) for multilayer feature extraction. The 8 res-blocks is used as the base subnet $\mathbf{n}_0$ to make other subnets for extracting multilayer features. We remove the smallest two res-blocks in $\mathbf{n}_0$ to make the 6-res-block subnet $\mathbf{n}_1$. Repeating the same on $\mathbf{n}_1$ makes the 4-res-block subnet $\mathbf{n}_2$, and repeating on $\mathbf{n}_2$ makes the 2-res-block subnet $\mathbf{n}_3$. The output features from $\mathbf{n}_1$, $\mathbf{n}_2$ and $\mathbf{n}_3$ are added back to the same dimension features in the corresponding layers in $\mathbf{n}_0$, as shown in Figure 1. Therefore, the feature output from $\mathbf{n}_0$ integrates the features from all subnets. The feature output is further processed by the mini-batch standard deviation, followed by a convolution layer and a fully-connected layer to generate an intermediate latent code $\mathbf{v}$. $\mathbf{v}$ can be written as $\mathbf{v}(I)$ as the input image I

can be the generator's input I_i and the generated I_o, which are both given to D_p during training.

Label Projection Module: We design this module to make the age-dependent latent code conditional on the target age group label. It has two processing paths. One path converts $\mathbf{v}(I)$ to a scalar $v(I)$ by a fully-connected layer. The other computes the label projection, which is the projection of the age group label $y \in \mathbf{I}^{N_a}$ onto the latent code $\mathbf{v}$, as shown in Figure 1. The computation takes the inner product of the embedded y and $\mathbf{v}(I)$, i.e., $(y^T E_m) \cdot \mathbf{v}(I)$, where E_m denotes the embedding matrix. The operation for the discriminator $D_p(I, y)$ can be written as follows:

$$D_p(I, y) = y^T E_v \cdot \mathbf{v}(I) + v(I) \tag{9}$$

where $y = y_t$ when $I = I_i$, and $y = y_i$ when $I = I_o$. The argument I in $\mathbf{v}(I)$ and $v(I)$ shows that both can be considered as the networks with I as input, i.e., $\mathbf{v}(\cdot)$ is the forward-pass of D_p without the last fully-connected layer, and $v(\cdot)$ is the forward-pass of D_p.

Age Classifier Embedding: The age classifier C_a in Figure 1 supervises the target age generation by imposing the requirement of age classification on the latent code $\mathbf{v}$. It is made by connecting $\mathbf{v}$ to an output layer made of a softmax function. The age class loss $\mathcal{L}_a^{(d)}$ is computed on C_a in the same way as given in (5), but revised for D_p.

We consider the adversarial loss, the age class loss and the R1 regularization when training D_p. The adversarial loss $\mathcal{L}_{D_p}^{adv}$ can be computed as follows.

$$\begin{aligned}\mathcal{L}_{D_p}^{adv} =& \mathbb{E}_{I_i \sim p(I_i)} \log\left[D_p\left(I_i, y_i\right)\right] + \mathbb{E}_{I_i \sim p(I_i)} \log\left[1 - D_p\left(G(I_i, y_t), y_t\right)\right] + \\ & \mathbb{E}_{I_o \sim p(I_o)} \log\left[1 - D_p\left(G(I_o, y_i), y_i\right)\right]\end{aligned} \tag{10}$$

The following R1 regularization $\mathcal{L}_{r1}$ is recommended by the StyleGAN [9] as it leads to a better FID score.

$$\mathcal{L}_{r1} = \mathbb{E}_{I_i \sim p(I_i)} \left[\left\|\nabla_{I_i} D_p\left(I_i, y_i\right)\right\|^2\right] \tag{11}$$

The overall loss for training D_p can be written as follows.

$$\mathcal{L}_{D_p} = \mathcal{L}_{D_p}^{adv} + \lambda_{r1}\mathcal{L}_{r1} + \lambda_a^{(d)}\mathcal{L}_a^{(d)} \tag{12}$$

where λ_{r1} and $\lambda_a^{(d)}$ are determined in the experiments.

4 Experiments

We first introduce the database and experimental settings in Sec.4.1. As the Face++ APIs [8] are used as the performance metrics in many previous work [13,24,12], we follow this convention for comparison purpose but reveal the inappropriateness by experiments. We address this issue in Sec.4.1 with proposed schemes to handle. Sec.4.2 reports an ablation study that covers a comprehensive comparison across different settings on the generator and discriminator. The comparison with state-of-the-art approaches is presented in Sec.4.3.

Table 1. Age estimation on FFHQ-aging and CAF by using Face++ API and our age estimator, better one in each category shown in boldface

Age group	0-2	3-6	7-9	10-14	15-19	20-29	30-39	40-49	50-69	70+
EAM (Estimated Age Mean) on whole FFHQ-aging Dataset										
Face++	10.19	20.31	24.64	25.92	26.10	29.64	39.93	54.34	67.81	**76.96**
Our Estimator	**1.50**	**5.08**	**8.96**	**13.17**	**18.94**	**24.27**	**32.08**	**42.57**	**57.57**	68.28
EAM / MAE (Mean Absolute Error) on whole CAF Dataset										
Real	1.17	4.53	7.93	12.00	17.13	24.04	33.76	43.82	56.61	72.35
Face++	19/17.29	27.68/22.15	29.67/20.93	28.70/16.23	27.39/10.09	29.31/6.52	36.20/7.27	45.17/7.35	56.92/7.33	70.65/7.56
Our Estimator	1.36/**1.28**	5.65/**2.16**	8.78/**3.23**	14.03/**5.68**	17.98/**5.43**	25.37/**3.46**	34.35/**3.50**	46.37/**5.57**	53.91/**5.74**	67.46/**6.71**

4.1 Databases and Experimental Settings

Due to page limit, we report our experiments on the FFHQ-Aging [17] in the main paper, and the experiments on the MORPH [19] and CACD [2] in the supplementary document. The FFHQ-Aging dataset [17] is made of ∼70k images from the FFHQ dataset [9]. Each image is labeled with an age group which is not based on ground-truth but on manual annotation via crowd-sourcing [17]. 10 age groups are formed: 0–2, 3–6, 7–9, 10–14, 15–19, 20–29, 30–39, 40–49, 50–69 and ≥70 years, labeled as $G_{10}0$, $G_{10}1$, ..., $G_{10}9$, respectively. We follow the same data split as in [9] that takes the first 60k images for training and the remaining 10k for testing. $G_{10}5$ ($20 \sim 29$) is taken as the source set and the other nine groups as the target sets.

The weights in (8) are experimentally determined as λ_{px}=10, λ_{pl}=2, λ_{cyc}=10, λ_t=0.1, λ_{id}=1 and λ_a^g=1; and those in (12) are λ_{r1}=10 and $\lambda_a^{(d)}$=1. We chose the Adam optimizer to train G and D_p at learning rate $2e^{-4}$ on an Nvidia RTX Titan GPU. See supplementary document for more details about data preprocessing, other training and testing settings.

Metrics for Performance Evaluation

Similar to the previous work [13,24,12], we also use the public Face++ APIs [8] as the metrics for evaluating the performance, but reveal via experiments the inappropriateness of using the Face++ APIs for the face verification and age estimation of subjects younger than 20. The Face++ APIs can estimate the age of a face and allow different thresholds for face verification. We first followed the same 1:1 face verification setup as in [13,24,12], where the generated face was verified against the input face with similarity threshold 76.5 for FAR 10^{-5}.

Using the same evaluation metrics allows a fair comparison with the previous work. However, using the same similarity threshold for face verification across the entire lifespan can be inappropriate, because the facial appearance does not change much across a small age gap, but it can change dramatically across a large age gap. The dramatic change would affect the ID preservation. This fact explains that we can sometimes be surprised to see someone's face has changed so much that we cannot recognize after tens of years of separation. Besides, we also found that the Face++ APIs reported large errors when estimating the ages of infants and young children, although it performed relatively well estimating the ages of adults older than 20 years.

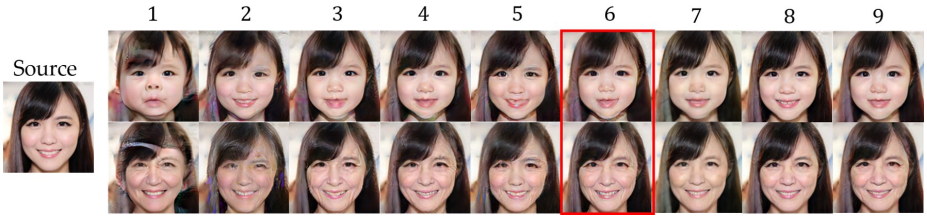

Fig. 2. Generated images with different settings: 1. B/L; 2. $B/L+\mathcal{L}_{id}^{pre}$; 3. $B/L+\mathcal{L}_{cyc}$; 4. $B/L+\mathcal{L}_{cyc}+\mathcal{L}_t$; 5. $B/L+\mathcal{L}_{cyc}+\mathcal{L}_t+\mathcal{L}_{id}^{pre}$; 6. (Best) $B/L+\mathcal{L}_{cyc}+\mathcal{L}_t+\mathcal{L}_{id}$ with D_p+C_a; 7.D_0+C_a; 8. $D_p w/o C_a$; and 9.$D_p+C_a(s)$

To better define the metrics needed for face verification across large age gaps and the age estimation for infants and children, we propose a rectification to the Face++ APIs usage and an age estimator that we made. For the rectification, we need a dataset with face images across large age gaps from early childhood for a sufficient number of individuals. The existing datasets cannot meet this requirement. The largest age gap for the same individual in the MORPH is less than 10 years, and it is 11 years in CACD. The FG-Net only has 1,002 images of 89 subjects although some are across large age gaps. To meet the requirement, we make the Cross-Age Face (CAF) dataset, which contains 4000 face images of 520 individuals. Each face has a ground-truth age, and each individual has images in at least 5 age groups across $G_{10}0 \sim G_{10}9$ ($0 \sim> 70$ years). The numbers of subjects in $G_{10}0 \sim G_{10}9$ are 341, 364, 312, 399, 469, 515, 435, 296, 195, and 67, respectively. More detail of the dataset is in supplementary document.

We use the CAF dataset for the following tasks: 1) Rectify the similarity thresholds given by the Face++ APIs for face verification across various age gaps. We define the thresholds by forming intra and inter pairs for each age gap, and selecting the threshold for an allowable FAR, e.g., 10^{-4}. See Supplementary Materials for more information about the CAF and MIVIA datasets.

The comparison of face verification with and without the proposed rectification is given in the following sections. Table 1 shows the age estimation on the FFHQ-Aging and CAF datasets by using the Face++ APIs and our age estimator. Note that each image in the FFHQ-Aging does not have a ground-truth age, and only has an age-group label by crowd-sourcing annotation, so we can only compute the Estimated Age Mean (EAM) for each age group. The EAM refers to the mean of the estimated ages of all images. But each image in the CAF has a ground-truth age so we compare in terms of both the EAM and MAE (Mean Absolute Error). Table 1 reveals that Face++ APIs consistently make large errors estimating the ages of the faces younger than 20 on both datasets. Our age estimator instead presents more reliable estimated ages.

4.2 Ablation Study

To compare the effectiveness of the losses considered in (8), we first define a baseline, denoted as B/L in Table 2, which only includes the adversarial loss $\mathcal{L}_G^{adv}$,

Table 2. Performance on FFHQ-Aging for transferring $G_{10}5$ (20−29) to other 9 groups with different settings on the loss function (Top) and discriminator D_p (Bottom). Both the rectified and common thresholds used for face verification, and used for our age estimator. Best one in each category shown in **boldface**, those in red show good verification rates but poor target age generation.

Age group	0-2	3-6	7-9	10-14	15-19	30-39	40-49	50-69	70+
Face Verification Rate (%), Rectified Threshold (Common Threshold)									
Threshold	61.8(76.5)	68.9(76.5)	72.7(76.5)	74.2(76.5)	76.6(76.5)	76.3(76.5)	71.7(76.5)	65.2(76.5)	65.2(76.5)
B/L	61.53.5)	82.9(25.9)	86.45(75.9)	88.5(77.2)	79.46(80.3)	72.7(71.6)	81.1(68.5)	75.5(31.7)	70.9(27.8)
$+ L_{id}^{pre}$	91.7(85.7)	99.2(91.3)	**100(100)**	**100(100)**	**100(100)**	**100(100)**	**100**(98.5)	98.7(97.4)	92.8(80.2)
$+ L_{id}$	63.1(5.7)	82.9(25.9)	89.4(77.7)	87.9(75.8)	80.88(81.4)	82.1(81.6)	80.9(68.2)	89.4(71.5)	78.3(29.9)
$+ L_{cyc}$	74.37(20.47)	93.1(81.7)	93.8(83.2)	93.3(83.3)	93.1(94.1)	95.5(94.3)	95.1(89.7)	93.9(80.3)	91.9(77.4)
$+ L_{cyc} + L_t$	77.4(23.7)	95.3(85.1)	95.5(85.1)	95.4(86.6)	93.7(94.6)	95.7(93.6)	96.9(93.2)	96.2(83.7)	95.9(82.7)
$+ L_{cyc} + L_t + L_{id}^{pre}$	**98.3(95.4)**	**100(98.4)**	**100**(98.6)	**100(100)**	**100(100)**	**100(100)**	**100(98.6)**	**100(97.4)**	**100(95.7)**
$+ L_{cyc} + L_t + L_{id}$	80.3(10.3)	96.5(86.3)	95.9(85.4)	95.8(86.7)	**100(100)**	**100(100)**	**100**(97.7)	97.6(85.7)	96.8(84.7)
$D_0 + C_a$	74.7(20.6)	92.5(80.8)	90.3(82.1)	92.4(83.5)	96.5(97.2)	**100(100)**	96.6(92.7)	94.3(81.5)	84.5(67.1)
D_p w/o C_a	75.2(22.3)	94.4(85.6)	91.2(82.4)	92.4(83.3)	95.8(96.5)	**100(100)**	95.6(90.3)	93.4(80.2)	83.8(66.6)
$D_p + C_a$	**80.3(10.3)**	**96.5(86.3)**	**95.2(85.4)**	**95.8(86.7)**	**100(100)**	**100(100)**	**100(97.7)**	**97.6(85.7)**	**96.8(84.7)**
$D_p + C_a(single)$	77.7(23.9)	94.8(85.7)	93.6(83.7)	94.3(85.3)	98.6(100)	**100(100)**	**100**(96.6)	95.2(82.9)	94.4(82.3)
EAM, Ours / Mean Error									
Raw data (Training set)	1.5/-	4.9/-	8.6/-	12.8/-	18.9/-	31.9/-	43.9/-	57.2/-	68.9/-
B/L	1.2/0.3	**4.5/0.4**	12.0/3.4	17.2/4.4	20.8/1.9	28.0/3.9	32.2/11.7	41.7/15.5	53.8/15.1
$+ L_{id}^{pre}$	7.9/6.4	6.6/1.7	12.4/3.8	17.4/4.6	21.5/2.6	26.1/5.8	30.8/13.1	40.6/16.6	53.2(61.1)/15.7
$+ L_{id}$	1.1/0.4	2.5/2.0	6.7/1.9	13.8/1.0	**18.6/0.3**	32.5/0.6	40.4/3.5	51.3/5.9	67.2/1.7
$+ L_{cyc}$	**1.4/0.1**	3.3/1.6	7.0/1.6	13.2/0.4	17.2/1.2	32.8/0.9	41.4/2.5	52.6/4.6	67.5/1.4
$+ L_{cyc} + L_t$	**1.4/0.1**	3.3/1.6	7.2/1.4	**12.6/0.2**	17.6/1.3	32.6/0.7	41.7/2.2	54.6/2.6	**69.0/0.1**
$+ L_{cyc} + L_t + L_{id}^{pre}$	8.9/7.4	7.9/3.0	11.9/1.5	15.5/3.4	20.2/1.3	29.9/2.0	37.0/6.9	41.0/16.2	48.7/24.2
$+ L_{cyc} + L_t + L_{id}$	1.1/0.4	**4.5/0.4**	**8.8/0.2**	13.5/0.7	18.7/0.2	**32.3/0.4**	**41.7/2.2**	**55.5/1.7**	68.4/0.5
$D_0 + C_a$	2.6/1.1	6.2/1.3	10.0/1.4	14.3/1.5	21.2/1.3	29.1/2.8	39.9/4.0	52.3/4.9	62.2/6.7
D_p w/o C_a	3.6/2.1	7.1/2.2	10.7/2.1	15.7/2.9	20.1/1.2	28.4/3.5	36.6/7.3	50.6/7.2	61.8/7.1
$D_p + C_a$	**1.1/0.4**	**4.5/0.4**	**8.8/0.2**	**13.5/0.7**	**18.7/0.2**	**32.3/0.4**	**41.7/2.2**	**55.5/1.7**	**68.4/0.5**
$D_p + C_a(single)$	2.3/0.8	5.8/0.9	9.3/0.7	14.0/1.2	19.3/0.4	30.6/1.3	38.2/3.7	53.5/3.7	64.4/4.5

the age class loss $\mathcal{L}_a^{(g)}$, the pixel-wise attribute loss $\mathcal{L}_{px}$, and the perceptual path length regularization $\mathcal{L}_{pl}$. $\mathcal{L}_G^{adv}$ is needed to warrant the quality of the generated images; $\mathcal{L}_a^{(g)}$ is needed for age classification; $\mathcal{L}_{px}$ is needed to preserve the source image attribute; and $\mathcal{L}_{pl}$ is needed for image quality improvement (A comparison of the baselines with and without these losses is given in the supplementary document). We compare the performance when combining the baseline with the identity loss $\mathcal{L}_{id}$, the triplet loss $\mathcal{L}_t$ and the cycle-consistency loss $\mathcal{L}_{cyc}$. We also compare with a general way to compute the identity loss by using an off-the-shelf pretrained face encoder [24,22], and we choose the pretrained ArcFace [3].

Table 2 shows the comparisons on the FFHQ-Aging by using the common and rectified Face++ thresholds for face verification and our age estimator,where $\mathcal{L}_{id}^{pre}$ denotes the identity loss computed using the pretrained ArcFace to replace L_{id}. The performance measures in parentheses are for the common threshold 76.5 and Face++ APIs, and those out of parentheses are for rectified thresholds and our age estimator. The results can be summarized as follows.

- When $\mathcal{L}_{id}^{pre}$ is included, the ID preservation is substantially upgraded, on the cost of much deteriorating target age generation, as shown by $B/L + \mathcal{L}_{id}^{pre}$ and $B/L + \mathcal{L}_{cyc} + \mathcal{L}_t + \mathcal{L}_{id}^{pre}$. The large errors in the estimated mean ages are shown in red. Clearly $\mathcal{L}_{id}^{pre}$ can well preserve identity, but badly damage the target age generation. Figure 2 shows the generated images.
- The triplet loss $\mathcal{L}_t$, which can only be computed with the cycle-consistency loss $\mathcal{L}_{cyc}$, demonstrates a balanced performance for ID preservation and age

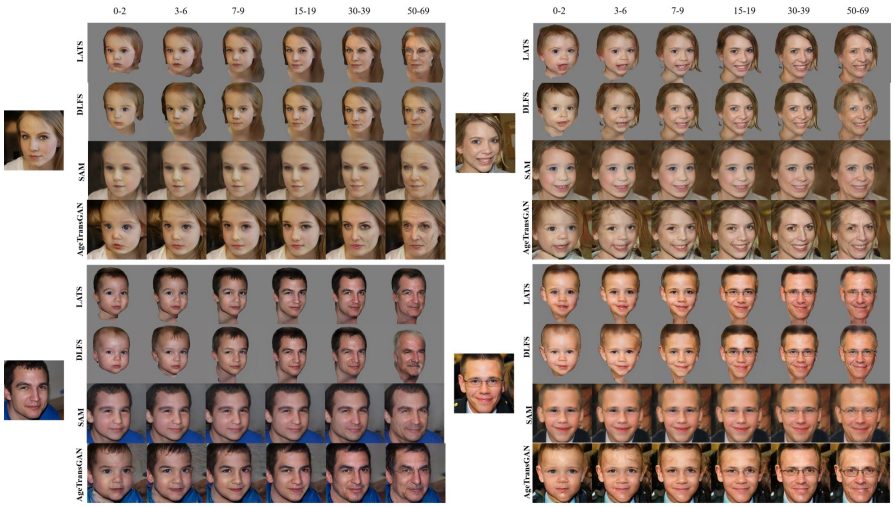

Fig. 3. Qualitative comparison with state-of-the-art methods for age transformation with the source faces on the left side.

transformation with $B/L + \mathcal{L}_{cyc} + \mathcal{L}_t$. The performance is further enhanced when $\mathcal{L}_{id}$ is added in, resulting in the final selected settings.

– With the selected $B/L + \mathcal{L}_{cyc} + \mathcal{L}_t + \mathcal{L}_{id}$, the face verification rates with rectified thresholds show more plausible results than the common constant threshold 76.5.

To better determine the settings for the CMP discriminator D_p, we compare the performance with and without the age classifier C_a, and the condition without the label projection. We also compare the performance of using multilayer and single-layer features in D_p. The bottom part of Table 2 shows the comparison of 1) C_a with D_0, where D_0 is the discriminator D_p without the label projection; 2) D_p without C_a; 3) D_p with C_a; 4) D_p with C_a but using single-layer features, i.e., only with the subnet $\mathbf{n}_0$ in Figure 1. Figure 2 shows the samples made by the four settings. The results can be summarized as follows:

– The performances of $D_0 + C_a$ and D_p w/o C_a are similar for both tasks of ID preservation and target age generation, although the former is slightly better for generating the children's ages.

– $D_p + C_a(single)$ with single-layer feature slightly outperforms $D_0 + C_a$ and D_p w/o C_a for both tasks.

– $D_p + C_a$ with multilayer feature outperforms $D_p + C_a(single)$ for both tasks with clear margins, especially on the youngest groups, i.e., $G_{10}0$ and $G_{10}1$.

The above comparisons have verified the settings with $D_p + C_a$, which is used for the comparison with other approaches. The margin m_t in (3) is experimentally determined as 0.5 out of a study reported in the supplementary document.

Table 3. Performance on FFHQ-Aging for transferring $G_{10}5$ (20−29) to other 9 groups (only 6 groups available by using LATS and DLFS), using both common and rectified thresholds for Face++ face verification (Top), and Face++ and our age estimator for age estimation (Bottom).

Age group	0-2	3-6	7-9	10-14	15-19	30-39	40-49	50-69	70+
Verification Rate(%) (Common Threshold)									
Threshold	61.8(76.5)	68.9(76.5)	72.7(76.5)	74.2(76.5)	76.6(76.5)	76.3(76.5)	71.7(76.5)	65.2(76.5)	65.2(76.5)
LATS [17]	51.5(5.2)	62.9(12.5)	82.7(78.9)	-	92.7(92.7)	92.7(91.5)	-	88.9(71.1)	-
DLFS [5]	52.8(12.4)	67.7(15.3)	81.9(75.2)	-	97.9(97.9)	97.5(96.8)	-	88.4(72.1)	
SAM [1]	**93.7(54.8)**	88.3(67.8)	85.9(74.8)	88.8(82.9)	89.8(90.0)	90.8(90.5)	87.7(76.7)	83.1(46.2)	68.9(23.6)
AgeTransGAN	80.3(10.3)	**96.5(86.3)**	**95.2(85.4)**	**95.8(86.7)**	**100(100)**	**100(100)**	**100(97.7)**	**97.6(85.7)**	**96.8(84.7)**
EAM, Ours/Mean Error									
Raw data	1.5	4.9	8.6	12.8	18.9	31.9	43.9	57.2	68.9
LATS [17]	**4.6/3.1**	**5.4/0.5**	**7.6/1.0**	-/-	**20.4/1.5**	31.6/0.3	-/-	52.1/5.1	-/-
DLFS [5]	**2.0/0.5**	**4.1/0.8**	**10.6/2.7**	-/-	**21.6/4.5**	**30.2/3.6**	-/-	49.5/7.1	-/-
SAM [1]	**5.4/3.9**	**7.3/2.4**	**10.4/1.8**	13.7/0.9	**20.3/1.4**	**32.3/0.4**	**43.2/0.7**	**58.7/1.6**	70.7/1.8
AgeTransGAN	**1.1/0.4**	**4.5/0.4**	**8.8/0.2**	**13.5/0.7**	**18.7/0.2**	**32.3/0.4**	41.7/2.2	55.5/1.7	**68.4/0.5**

Table 4. Performance on CAF for transferring $G_{10}5$ $(20-29)$ to other 9 groups (only 6 groups available by using LATS and DLFS), using rectified thresholds for Face++ face verification, and our age estimator for target age estimation.

Age group	0-2	3-6	7-9	10-14	15-19	30-39	40-49	50-69	70+
CAF									
Verification Rate (%)									
LATS [17]	66.5	72.9	73.7	-	98.1	82.7	-	83.2	-
DLFS [5]	54.7	69.4	83.2	-	100	100	-	85.3	
SAM [1]	**95.2**	88.3	85.8	88.7	89.6	90.9	87.9	83.5	69.4
AgeTransGAN	88.6	**97.9**	**99.7**	**100**	**100**	**100**	**100**	**100**	**100**
EAM									
LATS [17]	4.5	5.8	10.6	-	21.8	32.2	-	44.4	-
DLFS [5]	2.0	**4.2**	10.3	-	22.4	31.2	-	51.3	
SAM [1]	6.2	7.5	10.4	14.6	21.1	**33.4**	**44.8**	**55.2**	68.8
AgeTransGAN	**1.8**	5.4	**8.8**	**13.8**	**16.1**	32.1	43.5	54.0	**69.3**

4.3 Comparison with SOTA Methods

Table 3 shows the comparison with LATS [17], DLFS [5], and SAM [1], which all offer pretrained models in their GitHub sites. As revealed in Table 1, the Face++ APIs performs poorly for the estimation of younger ages and our age estimator performs well, we only use the latter for the comparison. The AgeTransGAN shows the best balanced performance for both ID preservation and target age generation for transforming to most age groups. Although the SAM performs best for ID preservation on $G_{10}0$, the corresponding target age generation is the worst with mean error 3.9 years. SAM also performs best for target age generation on $G_{10}7$ and $G_{10}8$, the corresponding verification rates for ID preservation are incomparable to those of the AgeTransGAN. Figure 3 shows a qualitative comparison. The AgeTransGAN demonstrates better age traits generated on faces of different age groups while maintaining plausible levels of similarities to the input (source) images. Note that the LATS and DLFS lack the attribute loss $\mathcal{L}_{px}$, all faces must be cropped during preprocessing, but the AgeTransGAN can process images with backgrounds.

Table 4 shows the comparison on the CAF dataset with faces of real ages. The AgeTransGAN outperforms other approaches for ID preservation on 8 age groups, and for target age generation on 5 age groups, showing the best overall balanced performance. The performance difference for age generation decreases considerably for the groups older than $G_{10}6$, showing that all approaches perform

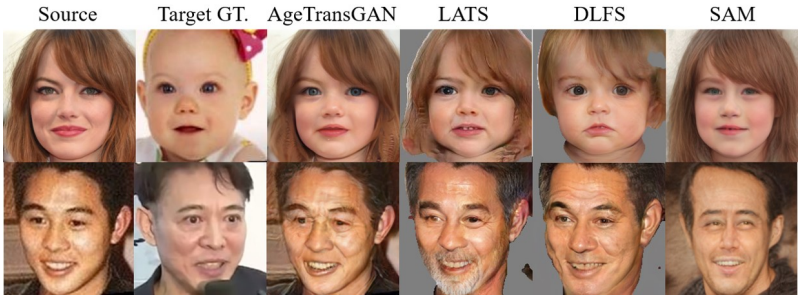

Fig. 4. Qualitative comparison with state-of-the-art methods for age regression (female) and progression (male) on CAF.

similarly well for generating adult faces. SAM performs best for ID preservation on $G_{10}0$, but is the worst for target age generation. Figure 4 shows the CAF samples with images generated for age progression (male) and regression (female) by all approaches. Although LATS makes beard, it does not generate sufficient wrinkles. The faces generated by SAM do not preserve some required levels of identity similarities to the source faces. The faces made by the AgeTransGAN show better qualities on identity similarity and target age traits.

5 Conclusion

We propose the AgeTransGAN for identity-preserving facial age transformation, and a rectification scheme for improving the usage of the popular metrics, Face++ APIs. The AgeTransGAN merges cycle-generation consistency, age classification and cross-age identity consistency to disentangle the identity and age characteristics, and is verified effective for balancing the performance for age transformation and identity preservation. The rectification scheme is offered with a new dataset, the CAF (Cross-Age Face), and an age estimator. We follow the conventional way to compare with other approaches, and highlight the issues with the existing metrics on the new FFHQ-Aging and CAF benchmarks. We address those issues through the rectification scheme and experiments, and verify the AgeTransGAN, the CAF dataset and our age estimator.

References

1. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Only a matter of style: Age transformation using a style-based regression model. ACM Transactions on Graphics (TOG) **40**(4), 1–12 (2021) 1, 2, 4, 7, 13
2. Chen, B.C., Chen, C.S., Hsu, W.H.: Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. TMM (2015) 4, 9
3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019) 4, 11
4. Fu, Y., Guo, G., Huang, T.S.: Age synthesis and estimation via faces: A survey. TPAMI (2010) 1

5. He, S., Liao, W., Yang, M.Y., Song, Y.Z., Rosenhahn, B., Xiang, T.: Disentangled lifespan face synthesis. In: ICCV (2021) 1, 2, 4, 7, 13
6. He, Z., Kan, M., Shan, S., Chen, X.: S2gan: Share aging factors across ages and share aging trends among individuals. In: ICCV (2019) 2
7. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017) 3
8. Inc., M.: Face++ research toolkit. http://www.faceplusplus.com 2, 8, 9
9. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019) 3, 4, 8, 9
10. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020) 3, 4
11. Li, P., Hu, Y., Li, Q., He, R., Sun, Z.: Global and local consistent age generative adversarial networks. In: ICPR (2018) 1, 2, 4, 7
12. Li, Z., Jiang, R., Aarabi, P.: Continuous face aging via self-estimated residual age embedding. In: CVPR (2021) 2, 8, 9
13. Liu, Y., Li, Q., Sun, Z.: Attribute-aware face aging with wavelet-based generative adversarial networks. In: CVPR (2019) 1, 2, 4, 7, 8, 9
14. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014) 3
15. Miyato, T., Koyama, M.: cgans with projection discriminator. arXiv preprint arXiv:1802.05637 (2018) 3
16. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: ICML (2017) 3
17. Or-El, R., Sengupta, S., Fried, O., Shechtman, E., Kemelmacher-Shlizerman, I.: Lifespan age transformation synthesis. In: ECCV (2020) 1, 2, 4, 7, 9, 13
18. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396 (2016) 3
19. Ricanek, K., Tesafaye, T.: Morph: A longitudinal image database of normal adult age-progression. In: FG (2006) 4, 9
20. Rothe, R., Timofte, R., Van Gool, L.: Dex: Deep expectation of apparent age from a single image. In: ICCV (2015) 4
21. Sricharan, K., Bala, R., Shreve, M., Ding, H., Saketh, K., Sun, J.: Semi-supervised conditional gans. arXiv preprint arXiv:1708.05789 (2017) 3
22. Wang, Z., Tang, X., Luo, W., Gao, S.: Face aging with identity-preserved conditional generative adversarial networks. In: CVPR (2018) 4, 11
23. Yang, H., Huang, D., Wang, Y., Jain, A.K.: Learning face age progression: A pyramid architecture of gans. In: CVPR (2018) 1, 2, 4
24. Yang, H., Huang, D., Wang, Y., Jain, A.K.: Learning continuous face age progression: A pyramid of gans. TPAMI (2019) 2, 7, 8, 9, 11
25. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017) 3