Age-Oriented Face Synthesis with Conditional Discriminator Pool and Adversarial Triplet Loss

Haoyi Wang, Victor Sanchez, *Member, IEEE,* Chang-Tsun Li, *Senior Member, IEEE*

***Abstract*—The vanilla Generative Adversarial Networks (GANs) are commonly used to generate realistic images depicting aged and rejuvenated faces. However, the performance of such vanilla GANs in the age-oriented face synthesis task is often compromised by the mode collapse issue, which may produce poorly synthesized faces with indistinguishable visual variations. In addition, recent age-oriented face synthesis methods use the L1 or L2 constraint to preserve the identity information on synthesized faces, which implicitly limits the identity permanence capabilities when these constraints are associated with a trivial weighting factor. In this paper, we propose a method for the age-oriented face synthesis task that achieves high synthesis accuracy with strong identity permanence capabilities. Specifically, to achieve high synthesis accuracy, our method tackles the mode collapse issue with a novel Conditional Discriminator Pool, which consists of multiple discriminators, each targeting one particular age category. To achieve strong identity permanence capabilities, our method uses a novel Adversarial Triplet loss. This loss, which is based on the Triplet loss, adds a ranking operation to further pull the positive embedding towards the anchor embedding to significantly reduce intra-class variances in the feature space. Through extensive experiments, we show that our proposed method outperforms state-of-the-art methods in terms of synthesis accuracy and identity permanence capabilities, both qualitatively and quantitatively.**

***Index Terms*—age-oriented face synthesis, generative adversarial networks, mode collapse, triplet loss**

I. Introduction

AGE-ORIENTED face synthesis (AOFS) is a generative task aiming to generate older and younger faces by rendering facial images with natural aging and rejuvenating effects. An efficient AOFS method can be integrated into a wide range of forensic and commercial applications (e.g., tracking suspects or missing children over a long time span, predicting the outcomes of cosmetic surgeries, and generating special visual effects on characters of video games, films and dramas [1], [2]). The synthesis in recent works [3]–[6] is usually conducted among age categories (e.g., the 30s, 40s, 50s) rather than specific ages (e.g., 32, 35, 39) since there is no noticeable visual change of a face over a few years. Recently, this practical yet intriguing problem has gained more and more attention from the research community.

The vanilla Generative Adversarial Network (GAN) [7] is commonly used as the backbone of several state-of-the-art AOFS methods [3], [8]–[11]. One of the biggest advantages of the vanilla GAN over other generative methods (e.g., the Variational Autoencoder [12]) is that it can generate sharp and realistic images by playing a minimax game between the generator and the discriminator. However, the vanilla GAN suffers from the mode collapse issue caused by a vanishing gradient due to the negative log-likelihood loss [13]. Specifically, once the discriminator converges, the loss does not penalize the generator further [14]. This allows the generator to find a specific mode (i.e., distribution) that can easily fool the discriminator [15]. The mode collapse issue may also occur in the AOFS task, where a mode is represented by an age category. Within this context, the vanilla GAN may generate faces with indistinguishable visual variations as exemplified in Fig. 1. This results in poor synthesis accuracy.

On the other hand, recent AOFS methods use the L1 or L2 constraint to preserve the identity information on synthesized faces. One disadvantage of these constraints is that they only penalize mean values and yield sparse results (i.e., features clusters with high intra-variances) [16]. Thus, the identity permanence capabilities of recent AOFS methods are compromised, especially when these constraints are associated with trivial weighting factor.

To boost the state-of-the-art performance in the AOFS task, this work proposes an AOFS method that includes two novel components, a Conditional Discriminator Pool (CDP) and an Adversarial Triplet Loss. The proposed CDP helps to achieve high synthesis accuracy by alleviating the mode collapse issue. Specifically, it allows learning multiple modes (i.e., age categories) explicitly and independently to generate realistic faces with a wide range of visual variations. Our CDP comprises multiple feature-level discriminators that learn the transformations from the source age category to the target age category. For each transformation, only the feature-level discriminator associated with the target age category is used. As a result, each feature-level discriminator only needs to learn one age category throughout the entire training process. The proposed Adversarial Triplet loss helps to preserve the identity information in the synthesized faces. This loss, which extends the Triplet loss [53], uses an additional ranking operation that can further optimize the distances within a triplet of feature embeddings comprising an *anchor*, a *positive*, and a *negative*. Specifically, it helps to bring the *positive* much closer to the *anchor*, while forcing the distance between the *anchor* and the *negative* to be larger than that between the *anchor* and the *positive*. The additional ranking operation forces the triplets to a play zero-sum game [5] during training. As a result, our Adversarial Triplet loss yields high-density clusters with dramatically reduced intra-class variances in the feature space.

H. Wang and V. Sanchez are with the Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK (e-mail: h.wang.16@warwick.ac.uk, v.f.sanchez-silva@warwick.ac.uk.)

C-T. Li is with the School of Information Technology, Deakin University, Geelong VIC 3216, Australia (e-mail: changtsun.li@deakin.edu.au.)

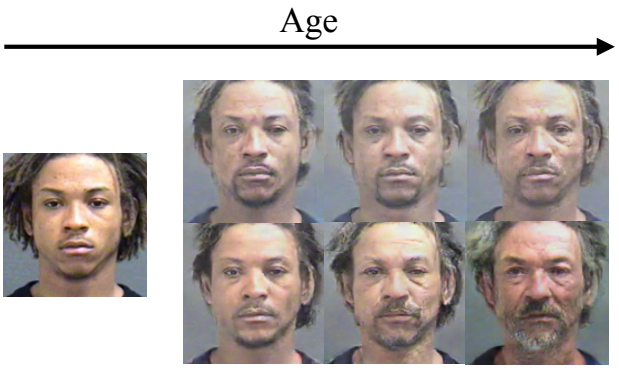


Fig. 1. A demonstration of face aging. The top row depicts images generated by a vanilla GAN suffering from the mode collapse issue. The bottom row depicts images generated by the proposed AOFS method.

Our contributions can be summarized as follows.

- We study the mode collapse issue in the AOFS task. To the best of our knowledge, our work is the first to tackle the AOFS task from the aspect of mode learning.
- We address the mode collapse issue in the vanilla GAN and attain high synthesis accuracy by proposing the CDP, which allows our AOFS method to learn multiple modes explicitly and independently.
- We propose the Adversarial Triplet loss to preserve the identity information in the synthesized images. Smaller intra-class variance can be achieved by forcing triplets to play zero-sum games during training.
- We evaluate the proposed AOFS method on several benchmark datasets to demonstrate its effectiveness in synthesizing realistic face images and preserving the identity information.

The rest of this paper is organized as follows. In Section 2, we review the related works on GANs, especially those tackling the mode collapse issue. In this Section, we also review the Triplet loss and the state-of-the-art AOFS methods. In Section 3, we present details of the proposed AOFS method including the CDP and the Adversarial Triplet loss. In Section 4, we explain the experimental settings and discuss the performance on several AOFS benchmark datasets. Finally, we conclude in Section 5.

II. Related Work

To motivate of our work, we next discuss the mode collapse issue of the vanilla GAN along with a number of previously proposed solutions. Then, we review the Triplet loss and show the main differences between the proposed Adversarial Triplet loss and other variations. Finally, we discuss some state-of-the-art AOFS methods.

A. Mode collapse in GANs

The vanilla GAN, introduced by Goodfellow *et al.* [7], can learn to generate sharp and realistic images by playing a minimax game between a generator and a discriminator. When training a vanilla GAN, the generator and the discriminator try to reach a Nash equilibrium [17] by minimizing the negative log-likelihood loss and the JS-divergence [18]. However, the involvement of the negative log-likelihood loss may cause the discriminator to converge faster than the generator [19]. Once the discriminator converges, the loss function stops penalizing the generator [14]. This is also known as the vanishing gradient problem [13], [20], [21] and is the main cause for the mode collapse issue. Since the parameters in the discriminator are not further updated, the generator may then find a specific mode that can easily fool the discriminator. When such an issue occurs, such a vanilla GAN can only generate samples of limited variation. Solving this mode collapse issue is a top trending research topic on GANs.

Since the mode collapse issue is caused by the vanishing gradient problem due to the negative log-likelihood loss, one strategy is to alleviate it by using an alternative loss function that minimizes a different divergence. Nowozin *et al.* [22] show that the optimization of GANs can be done by minimizing any f-divergence [23], which is a family of divergences aiming to minimize the distance between two distributions. Some commonly used members of the f-divergence family are the JS-divergence, the Kullback-Leibler divergence (KL-divergence) [24], the squared Hellinger divergence, and the Pearson χ^2 divergence [25]. The authors show that GANs trained with other divergences, like the KL-divergence or the squared Hellinger divergence, can generate images with more noticeable visual variations compared to those generated by a vanilla GAN. Although the work in [22] does not tackle the mode collapse issue directly, it shows the possibility of using other loss functions to optimize GANs.

Arjovsky *et al.* [26] propose the Wasserstein GAN (WGAN), which uses the Earth-Mover (EM) distance to calculate the distance between distributions of the real and synthesized data. Intuitively, the EM distance computes the cost of transforming one distribution to another, which is more sensitive to the differences between two distributions [26]. Therefore, even if the discriminator is well-trained, it can still keep rejecting the data synthesized by the generator. The Least Square GAN (LSGAN) [27], on the other hand, replaces the negative log-likelihood loss by the L1 loss. Minimizing the L1 loss is equivalent to minimizing the Pearson χ^2 divergence, which can produce overdispersed approximations and thus makes the LSGAN less mode-seeking [28], [29].

Although the methods discussed before may alleviate the mode collapse issue, the discriminator still learns from all the modes. Therefore, recently proposed methods focus on modifying the GAN structure. For example, Nguyen *et al.* [30] propose the Dual Discriminator Generative Adversarial Nets (D2GAN), where each discriminator favors data from a different distribution. By using this strategy, the method computes the KL and reverse KL divergences simultaneously, which increases the variety of samples. Based on this idea, Zhang *et al.* [31] propose a D2GAN variation with two customized discriminators. Specifically, one discriminator consists of residual blocks that increase the variety of generated samples. The other discriminator uses the scaled exponential linear unit (SELU) function [32] as the non-linear activation function. Adopting the SELU function guarantees that the

discriminator produces a non-zero value even if the distributions of the synthesized and real data are similar. The authors further propose the D2PGGAN [33] to stabilize the training by leveraging the idea of progressively increasing the complexity of the generator [34]. Durugkar *et al.* [35] propose a GAN with multiple discriminators. Their method alleviates the mode collapse issue to an extent since the generator has to fool a set of discriminators, which makes the generated samples more diverse. It is important to note that by introducing additional discriminators, in parallel, the aforementioned methods are more computationally complex than their counterparts (i.e., a vanilla GAN). On the contrary, by selecting a particular discriminator from a discriminator pool, our CDP only uses one discriminator for each transformation, which does not increase the computational complexity.

B. Triplet Loss

The Triplet loss [36] aims to learn feature embeddings by optimizing the geometric relationship in the feature space with a triplet of an *anchor*, a *positive* and a *negative*. In this context, the *anchor* and *positive* represent feature embeddings of the same class, while the *negative* of a different class. The goal is to minimize the distance between the *anchor* and the *positive* while simultaneously pushing the *negative* further away from the *anchor*. Several variations to this loss have been proposed. For instance, Chen *et al.* use an additional *negative* to form a quadruplet [37]. Huang *et al.* implement three ranking operations in total by using an *anchor*, a *negative* and three *positives* [38]. Ye *et al.*, on the other hand, adopt additional samples from other modalities [39]. It is worth noting that all of these variants leverage additional samples from the same or different modalities. Therefore, these losses can no longer help to optimize the geometric relationship within a triplet. This is explained in detail in Section III.D.

We find that the original Triplet loss produces clusters with large intra-class variances that can be further optimized. To produce high-density clusters, we add another ranking operation and propose an Adversarial Triplet loss to pull the *positive* closer to the *anchor*. It is worth noting that, compared to the aforementioned Triplet loss variants, our Adversarial Triplet loss still focuses on optimizing distances within triplets without added samples.

C. Age-Oriented Face Synthesis

The first AOFS methods can be traced back to [40]–[42] which study craniofacial growth in young faces. In the early stage, geometric-based methods were popular in research, and one of the most representative methods is the Active Shape Model (ASM) [43]. The authors model the shape of faces by adjusting the position of a number of points. Each point marks one part of the face, such as the position of the eyes or the boundary of the face. Synthetic facial images of different shapes and ages can then be obtained by adjusting the position of these points. Another approach to render aging or rejuvenating effects is to directly synthesize or remove wrinkles on a given facial image [44]–[48]. Later, Ramanathan and Chellappa [49] propose an aging-focused method called the craniofacial growth model for synthesizing elderly faces by leveraging facial landmark movements. Another early AOFS method is [50], where the authors use dictionary-based learning to encapsulate a personalized aging process, and associate a dictionary per subject to represent their aging characteristics.

With an increased popularity in deep learning, several attempts have been made to tackle the AOFS problem with a variety of network architectures. Wang *et al.* [4] and Zhang *et al.* [6] use conditional adversarial learning [51] to synthesize aged faces. Wang *et al.* further employ an age category classifier to boost the synthesis accuracy and an L2 constraint on the identity-specific features to preserve the identity information. Yang *et al.* [5] propose a GAN framework by implementing a customized discriminator with a pyramid architecture, which leads to more realistic results than a conventional discriminator as images can be discriminated based on multi-scale features. They adopt a pre-trained identity classifier to further preserve the identity in the synthesized images. Recently proposed AOFS methods use the Wavelet transform to enhance the texture information in the frequency domain so that richer aging and rejuvenating effects can be synthesized [3], [52]. He *et al.* [53] implement a GAN model with a customized generator, where a number of decoders are leveraged, one per age category. All of the decoders are associated with a weight factor to control the relative importance. Since all of the decoders are trained simultaneously, the computational complexity of these methods are proportional to the number of age categories being learned.

Our work is different from the aforementioned deep learning-based methods as it tackles the AOFS problem from a different angle (i.e., mode learning). Our method can achieve high synthesis accuracy by learning multiple modes explicitly and independently. Additionally, compared to the L1 loss, the L2 loss, and the simple classifiers used in those methods, our AOFS method uses the proposed Adversarial Triplet loss to keep the identity information unaltered in the synthesized facial images.

III. Proposed AOFS Method

In this section, we explain the proposed method by formulating the problem, and then explaining the pre-trained Multi-Task Feature Extractor (MTFE) used to extract age- and identity-specific features. We then present the proposed CDP and the Adversarial Triplet loss. Finally, we explain the overall loss used to train our method.

A. Problem Formulation

Since the transformation is conducted among age categories rather than specific ages, like [3], [5], [52], we divide the data into four age categories (i.e., 30^{-}, $31-40$, $41-50$, and 51^{+}). Each category is denoted by $\{C_i | i \in [1,4]\}$.

To render aging and rejuvenating effects, the proposed AOFS method accepts two faces, $x \in C_X$ and $y \in C_Y$, and the age label of y, l^y_{age}, where $X \neq Y$. Specifically, x is the face that is to be aged or rejuvenated, and y carries the target age information. Our method then aims to generate an aged or rejuvenated x, denoted by $\tilde{x}$, where $\tilde{x}$ is expected to belong

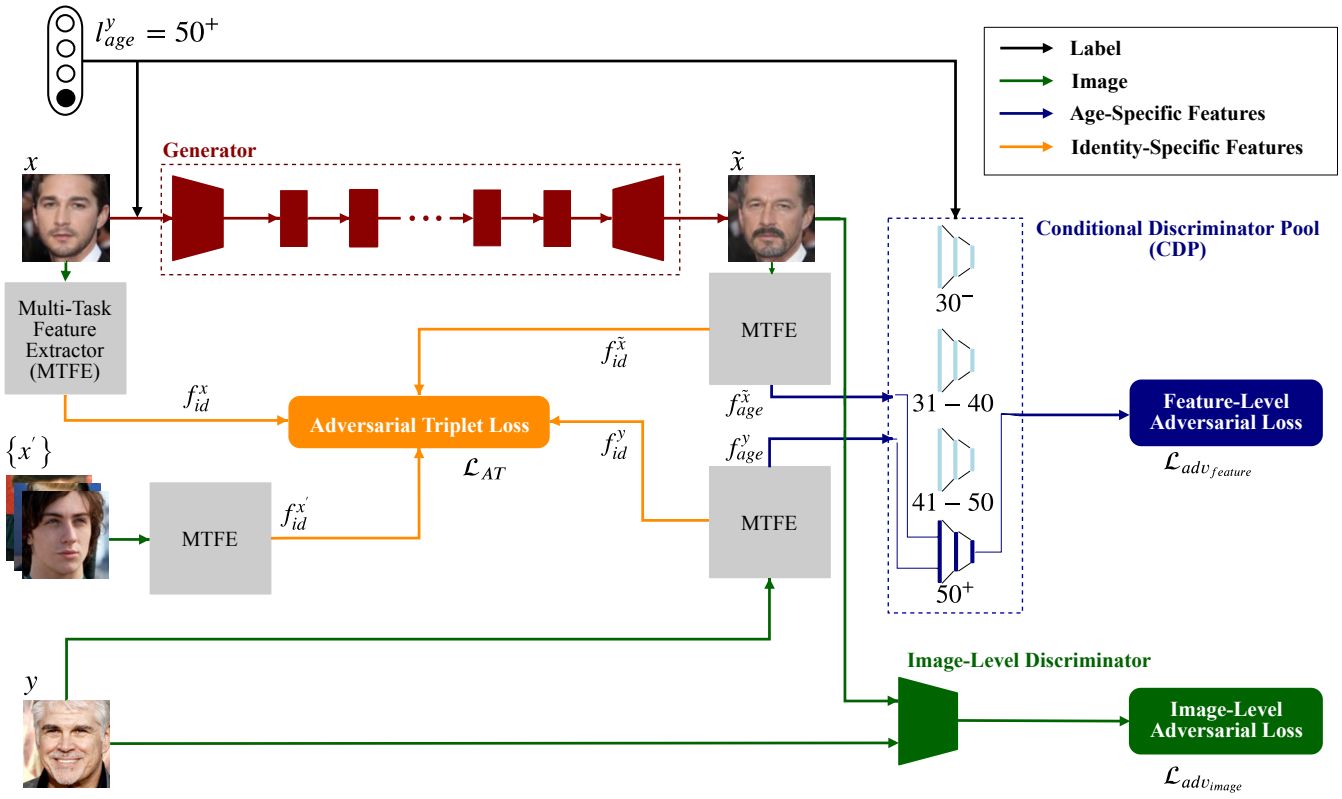


Fig. 2. Architecture of the proposed AOFS method. It consists of a generator with residual blocks (red rectangles), an image-level discriminator, and a CDP that contains several feature-level discriminators. The number of feature-level discriminators equals the number of age categories that the method should learn. Two adversarial losses are used to synthesize realistic aged and rejuvenated faces. To further optimize the identity features in the synthesized image, $\tilde{x}$, we leverage additional input images, $\{x'\}$, that are within the same age category as the source image, x. Image y carries the target age information for $\tilde{x}$.

to the same age category as y. Moreover, to ensure that the identity information is effectively preserved in $\tilde{x}$, our method also uses other images in the same batch, $\{x'\}$, to define the Adversarial Triplet loss. It is worth noting that x' and y do not share the same identity information of x.

In summary, the proposed method achieves three goals, simultaneously: 1) to generate realistic aged and rejuvenated faces; 2) to force the synthesized faces to be within the target age category; 3) to preserve the identity information in the synthesized image. Our architecture is shown in Fig. 2.

B. Multi-Task Feature Extractor (MTFE)

The CDP and the Adversarial Triplet loss of the proposed AOFS method use age- and identity-specific features from input images and synthesized images. To extract and disentangle these features, we use the decomposition method proposed in [54]. Specifically, we use a ResNet-50 [55] as the backbone for feature extractor (see Fig. 3). This model decomposes all of the features extracted from a facial image into two components based on a spherical coordinate system as

$$f_{sphere} := \{r; \boldsymbol{theta}\}, \tag{1}$$

where f_{sphere} is the set of features after the decomposition in which the angular component $\boldsymbol{theta} = \{\theta_1, \theta_2, ..., \theta_k\}$ indicates the identity-specific features for k identities, and the radial component r encodes the age-specific features.

We replace the regression loss used to learn age-specific features in [54] with an age regression model [56] to supervise the age-specific learning process. This has been shown to achieve better performance for the age estimation task [57]. We observe that feature extractors trained in this multi-tasking manner can achieve higher accuracy on both the age and identity classification tasks, opposed to single-task networks. Additionally, we use our proposed Adversarial Triplet loss to learn identity-specific features.

C. Conditional Discriminator Pool (CDP)

A vanilla GAN with a single image-level discriminator has a loss function for face synthesis that is usually formulated as

$$\begin{aligned}\mathcal{L}_{adv} =& \mathbb{E}_y[logD(y)] \\ &+ \mathbb{E}_x[log(1 - D(G(x))],\end{aligned} \tag{2}$$

where the generator G tries to minimize the loss, and the discriminator D tries to maximize the loss. As mentioned, GANs based on this loss function suffer from the mode collapse issue. To force the network to learn each mode independently and, thus, alleviate this issue, one can directly add more discriminators. However, such a strategy leads to a higher computational complexity with redundancy during

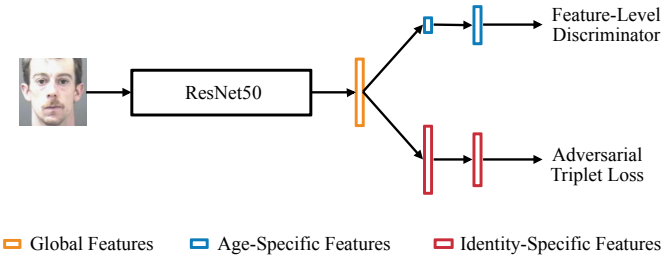


Fig. 3. Architecture of our MTFE. After the decomposition, we resize each set of task-specific features to be used by the corresponding feature-level discriminator of the CDP or the Adversarial Triplet loss.

training. Such complexity and redundancy arise because all the discriminators are expected to back-propagate the loss during each transformation. Therefore, we propose a mechanism to select the corresponding discriminator for each transformation based on the input label that represents the target age. Let us recall that the proposed AOFS method treats each age category as a mode, which results in four modes in total. We use the input label, l^y_{age}, to select the corresponding discriminator that learns the target age category. Our proposed method implements this mechanism on discriminators at the feature level, which are used to synthesize aging and rejuvenating effects. Therefore, we assemble four feature-level discriminators with an identical architecture to form our CDP. Each feature-level discriminator targets one mode. Our method additionally uses an image-level discriminator to remove artificial effects from the synthesized faces. Our method leverages the selected feature-level discriminator alongside the image-level discriminator per transformation (Fig. 2).

An alternative way to select the feature-level discriminator is to employ an additional classifier. However, within the context of AOFS, the accuracy of classifying the age categories may be very low, (i.e., from 25% to 60%). This depends on the specific age category for different AOFS benchmark datasets [4], [52]. Employing such a low-accuracy classifier results in selecting a discriminator that learns an incorrect mode. Instead, we directly use l^y_{age} to select discriminators to guarantee that each transformation is associated with the target mode. We then formulate the feature-level adversarial loss as follows:

$$\begin{aligned}\mathcal{L}_{adv_{feature}} &= \mathbb{E}_{f^y_{age}}[log(FD_{C_i}(f^y_{age})|l^y_{age})] \\ &+\mathbb{E}_{f^{\tilde{x}}_{age}}[log(1-(FD_{C_i}(f^{G(x|l^y_{age})}_{age})|l^y_{age}))],\end{aligned} \tag{3}$$

where FD_{C_i} is the selected feature-level discriminator trying to maximize the loss, f^y_{age} denotes the age-specific features extracted from the target image, y, and $f^{G(x|l^y_{age})}_{age}$ denotes the age-specific features extracted from the synthesized image $\tilde{x}$. Also, $G(x|l^y_{age})$ is the generator that produces $\tilde{x}$ conditioned on l^y_{age}. Finally, l^y_{age} is a one-hot encoded vector indicating the label for the target age category C_i.

D. Adversarial Triplet Loss

The Triplet loss [36] with three feature embeddings is formulated as

$$\mathcal{L}_{Triplet}(a,p,n) = \sum_{a,p,n}[m + Dist_{a,p} - Dist_{a,n}]_+, \tag{4}$$

where $Dist_{j,k}$ indicates the Euclidean distance between embeddings j and k in the feature space and a, p, n are the indices of the *anchor*, the *positive*, and the *negative*, respectively. This loss forces $Dist_{a,n}$ to be larger than $Dist_{a,p}$ by a minimal margin m. However, once this criterion is satisfied, $Dist_{a,p}$ cannot be further minimized, which may lead to large intra-class variances. To overcome this problem, we add another ranking operation to Eq. (4) to force $Dist_{a,n}$ to be larger than the distance between n and p, (i.e., $Dist_{n,p}$). This helps to further bring p closer to a by forcing different triplets with the same a and p, but different n, to play a zero-sum game. The Adversarial Triplet loss is then formulated as

$$\begin{aligned}\mathcal{L}_{AT}(a,p,n) = &\sum_{a,p,n}[m + Dist_{a,p} - Dist_{a,n}]_+ \\ &+ [Dist_{n,p} - Dist_{a,n}].\end{aligned} \tag{5}$$

Let us assume there are several triplets with the same a and p, but different n, where each distinct n is denoted by n_i. Under this assumption, Eq. (4) (i.e., Triplet loss) can be minimized provided that $Dist_{a,n_i} > Dist_{a,p} + m$, which may result in clusters with larger intra-class variances. To reduce these variances, $Dist_{a,n_i}$ should be larger than $Dist_{n_i,p}$. Let us take the triplets $a-p-n_1$ and $a-p-n_3$ in Fig. 4 as an example, with n_1, n_2, n_3, and n_4 from different classes. n_1 and n_3 should then maintain their relative position with respect to the $a-p$ cluster to be farther from its neighboring clusters. In other words, n_1 and n_3 should not get near to either n_2 or n_4. In this case, $\mathcal{L}_{AT}(a,p,n_1)$ tries to pull p towards n_1 and minimize $Dist_{n_1,p}$, while $\mathcal{L}_{AT}(a,p,n_3)$ tries to pull p towards n_3 and minimize $Dist_{n_3,p}$. Therefore, $\mathcal{L}_{AT}(a,p,n_1)$ and $\mathcal{L}_{AT}(a,p,n_3)$ play a zero-sum game as minimizing one loss increases the other. This is also true for $\mathcal{L}_{AT}(a,p,n_2)$ and $\mathcal{L}_{AT}(a,p,n_4)$. To minimize all of these losses (i.e., to have a total loss equal to zero), p should be in the same position as a so that $Dist_{a,n_i} = Dist_{n_i,p}$. In practice, however, our Adversarial Triplet loss pulls p to a position very close to a so that $Dist_{a,n_i} \approx Dist_{n_i,p}$.

Fig. 5 demonstrates the performance of the Adversarial Triplet loss on a real dataset. In this example, the feature distribution of the MNIST dataset for classification is presented. To this end, we employ an Alexnet [58] as the deep network, but replace all the fully-connected layers, except the output layer, by a single linear layer with two neurons for visualization purposes. From the figure, we can observe that the features learned by the Adversarial Triplet loss dramatically reduce the intra-class variances compared to the features learned by the Triplet loss. The classification accuracy attained by each loss is tabulated in Table I.

One of the most critical issues in the Triplet loss is that as the number of triplets grows, many triplets can easily satisfy the constraint in Eq. (4), which in turn may lead to poor convergence [36]. To overcome this issue in the Adversarial Triplet loss, we adopt a hard negative mining strategy [59]. Specifically, we use an online hard sample mining method in which each batch consists of samples from T classes, and each class has S samples within one batch, for a batch size of $B = TS$. In this method, each sample in a batch acts as the *anchor* for one triplet, thus, there are a total of B triplets

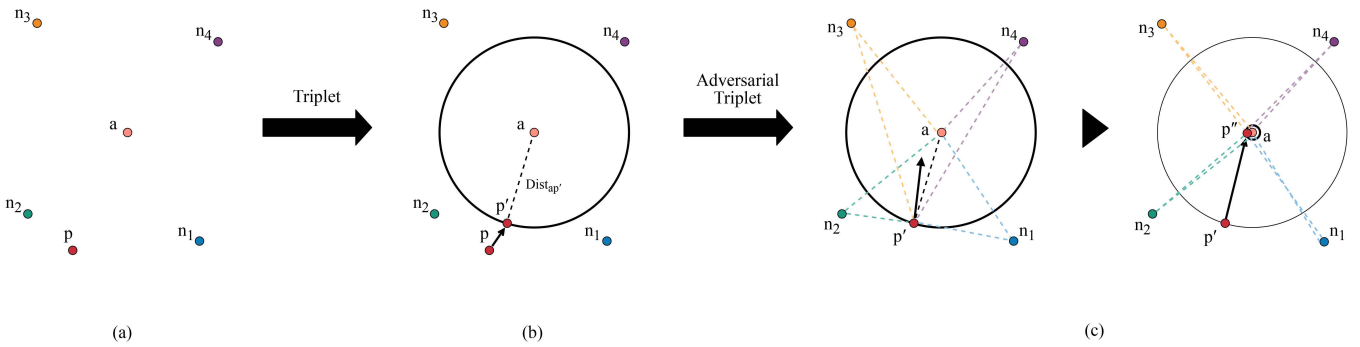


Fig. 4. An example showing how the Adversarial Triplet loss works. The a (*anchor*) and p (*positive*) are feature embeddings representing the same class. The *negatives* n_1, n_2, n_3, and n_4 indicate feature embeddings from other classes, each one from a distinct class. (a) Original positions of these feature embeddings. (b) By using the Triplet loss, p can move towards p' when minimizing Eq. (4). (c) Our Adversarial Triplet loss guarantees that for each n_i, where $i \in [1, 2, 3, 4]$, $Dist_{an_i} \approx Dist_{n_i p}$ by adding an additional operation as formulated in Eq. (5). In this case, p' may continue moving towards a and end up at a location that is extremely close to a, i.e., p''.

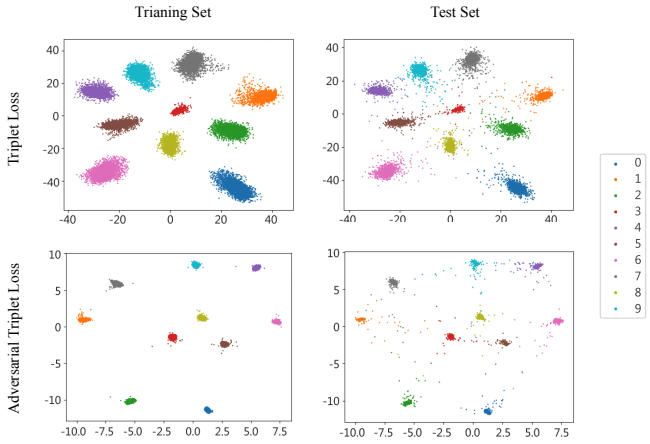


Fig. 5. Comparison of feature distribution of the MNIST dataset for classification with the Triplet loss and the Adversarial Triplet loss.

within one batch. For each *anchor*, a hardest *positive* sample with the largest distance and a hardest *negative* sample with the smallest distance are selected to form a triplet. This method does not require pre-defining the triplets and can generate hard triplets in an online manner. After incorporating this hard sample mining strategy, our Adversarial Triplet loss in Eq. (5) is as follows:

$$\begin{aligned}\mathcal{L}_{AT}(a,p,n) = &\sum_{t=1}^{T}\sum_{s=1}^{S}[m + \max_{p} Dist_{a,p} - \min_{n} Dist_{a,n}]_+ \\ &+ [Dist_{n,p} - \min_{n} Dist_{a,n}],\end{aligned} \tag{6}$$

where t is the class index and s is the image index for each class in one batch.

Since we are trying to optimize the identity-specific features on the synthesized faces when training our AOFS method, we use the identity-specific features, f_{id}^{x}, from the source image as the *anchor* and the identity-specific features, $f_{id}^{\tilde{x}}$, from the synthesized image as the *positive*. In addition, we use all other images in the same batch that do not share the same identity

TABLE I
CLASSIFICATION ACCURACY (%) ON THE MNIST DATASET.

Loss	Triplet	Adversarial Triplet
Accuracy	99.43	**99.67**

with the source image as the *negatives*. The Adversarial Triplet loss of our AOFS method with the hard sample mining strategy is then formulated as

$$\begin{aligned}\mathcal{L}_{AT}(f_{id}^{x}, f_{id}^{\tilde{x}}, \left\{f_{id}^{x'}, f_{id}^{y}\right\}) = &\sum_{t=1}^{T}\sum_{s=1}^{S} \\ &[m + Dist_{f_{id}^{x}, f_{id}^{\tilde{x}}} - \min_{\left\{f_{id}^{x'}, f_{id}^{y}\right\}} Dist_{f_{id}^{x}, \left\{f_{id}^{x'}, f_{id}^{y}\right\}}]_+ \\ &+ [Dist_{\left\{f_{id}^{x'}, f_{id}^{y}\right\}, f_{id}^{\tilde{x}}} - \min_{\left\{f_{id}^{x'}, f_{id}^{y}\right\}} Dist_{f_{id}^{x}, \left\{f_{id}^{x'}, f_{id}^{y}\right\}}],\end{aligned} \tag{7}$$

where $\left\{f_{id}^{x'}\right\}$ are the identity-specific features of images within the same age category as the source image but carrying different identity information, and f_{id}^{y} are the identity-specific features of images within the target age category. It is worth noting that the above equation does not have the max operation as in Eq. (6) since the *positive* in this case, $f_{id}^{\tilde{x}}$, is synthesized thus cannot be selected.

E. Overall Loss

The image-level adversarial loss in our AOFS method is formulated as

$$\begin{aligned}\mathcal{L}_{adv_{image}} = &\mathbb{E}_y[logD(y)] \\ &+ \mathbb{E}_x[log(1 - D(G(x|l_{age}^{y}))].\end{aligned} \tag{8}$$

The overall loss function, $\mathcal{L}_{overall}$, to train our method is a weighted summation of several losses, with $\mathcal{L}_{adv_{image}}$ removing ghost artifacts, $\mathcal{L}_{adv_{feature}}$ synthesizing ageing and rejuvenating effects and attaining high synthesis accuracy, and $\mathcal{L}_{AT}$ preserving the identity information as follows:

$$\begin{aligned}\mathcal{L}_{overall} = &\mathcal{L}_{adv_{image}} + \lambda_{adv_{feature}}\mathcal{L}_{adv_{feature}} \\ &+ \lambda_{AT}\mathcal{L}_{AT},\end{aligned} \tag{9}$$

TABLE II
ARCHITECTURE OF THE GENERATOR.

Encoder			
#Layer	Convolution	Normalization	Non-linear
1	*k=7, s=1, p=1*	Instance	ReLU
2	*k=3, s=2, p=1*	Instance	ReLU
Residual Block ($\times$ 6)			
#Layer	Convolution	Normalization	Non-linear
1	*k=3, s=2, p=1*	Instance	ReLU
2	*k=3, s=2, p=1*	Instance	ReLU
Decoder			
#Layer	Deconvolution	Normalization	Non-linear
1	*k=3, s=2, p=1*	Instance	ReLU
2	*k=3, s=2, p=1*	Instance	Tanh

TABLE III
ARCHITECTURE OF THE DISCRIMINATORS.

Feature-Level ($\times$ 4)			
#Layer	Fully-Connected	Normalization	Non-linear
1	*128*	Instance	LeakyReLU
2	*64*	Instance	LeakyReLU
3	*32*	Instance	LeakyReLU
4	*16*	Instance	LeakyReLU
5	*1*	-	-
Image-Level			
#Layer	Convolution	Normalization	Non-linear
1	*k=3, s=2, p=1*	Instance	LeakyReLU
2	*k=3, s=2, p=1*	Instance	LeakyReLU
3	*k=3, s=2, p=1*	Instance	LeakyReLU
4	*k=3, s=2, p=1*	Instance	LeakyReLU
5	*k=3, s=1, p=1*	-	-

where $\lambda_{adv_{feature}}$ and λ_{AT} control the relative importance among learning objectives.

IV. EXPERIMENTS

In this section, we first briefly describe the two AOFS benchmark datasets used in our experiments followed by the implementation details of our method. Then, we compare our method with state-of-the-art methods and conduct ablation studies, both qualitatively and quantitatively, to show that our method can achieve high synthesis accuracy while preserving the identity information on the synthesized facial images.

A. AOFS benchmark datasets

We use the MORPH II dataset [60] and the Cross-Age Celebrity Dataset (CACD) [61] to train the MTFE and evaluate our method. The MORPH II dataset contains about 55,000 facial images of individuals with ages ranging from 16 to 77. The CACD contains more than 160,000 facial images of individuals with ages ranging from 16 to 62. Most of the images in the MORPH II dataset are mugshots, while images in the CACD contain Pose, Illumination, and Expression (PIE) variations. Each image in both datasets is associated with an age label and an identity label.

All images are cropped to 128$\times$128 pixels and aligned based on the location of the eyes. Since not all images can be aligned by using this technique, in the end, 55,062 images from the MORPH II dataset and 159,226 images from the CACD are used in our experiments. For each dataset, we use 80% of the images for training and the remaining 20% for testing. The number of training images for each age category in the MORPH dataset is 19,949, 12,496, 8,982, and 2,622, for the categories $\{30^-, 31-40, 41-50, 51^+\}$, respectively. For the CACD, the number of training images of each age category is 39,416, 33,742, 30959, and 23,262, respectively. There is no identity overlap between the training and test sets.

We conduct a five-fold cross validation for all our experiments. For the MORPH II dataset, each fold has about 2,550 subjects with 3,989, 2,499, 1,796, and 524 images within each age category, respectively. For the CACD, each fold contains about 400 subjects with 7,883, 6,748, 6,191 and 4,652 images within each age category, respectively.

B. AOFS quality criteria

There are two criteria commonly used to measure the quality of synthesized images [5], [52] in the AOFS task. Under the first criterion, synthesized images are fed into an age category classifier to evaluate whether the depicted face has been transformed to the target age category. The second criterion measures identity permanence and relies on face verification models to validate whether the synthesized image and the source image depict the same person.

To evaluate our method and demonstrate its robustness, we use another two benchmark datasets to train two separate validation networks, one for each criterion. In particular, we use the AgeDB dataset [62] to train a network that evaluates the synthesis accuracy and a face recognition benchmark dataset, the VGGFace2 dataset [63], to train a network that evaluates the identity permanence capabilities. We use the commonly used ResNet-50 as the backbone for both evaluation networks.

C. Network architecture

The details of the architectures of the generator and discriminators in our AOFS method are tabulated in Tables II and III, respectively. We employ the architecture from [64] for our generator and the patch discriminator from [65] for our image-level discriminator. In both tables, for each convolutional and deconvolutional layer, k indicates the kernel size, s indicates the stride, and p indicates the padding size. In Table III, the second column for the feature-level discriminators tabulates the dimensions of the corresponding layer.

D. Data augmentation

When training the MTFE and validation networks, we use a combination of rotation, flip, and crop operations to augment the data. Specifically, we first randomly rotate each image by a angle between +10 deg. and -10 deg., and then randomly flip the rotated image with a probability of 0.5. Finally, we pad the image on all sides with 10 pixels and crop the padded image at a random location to the original image size (i.e. 128 $\times$ 128 pixels). When training the proposed AOFS

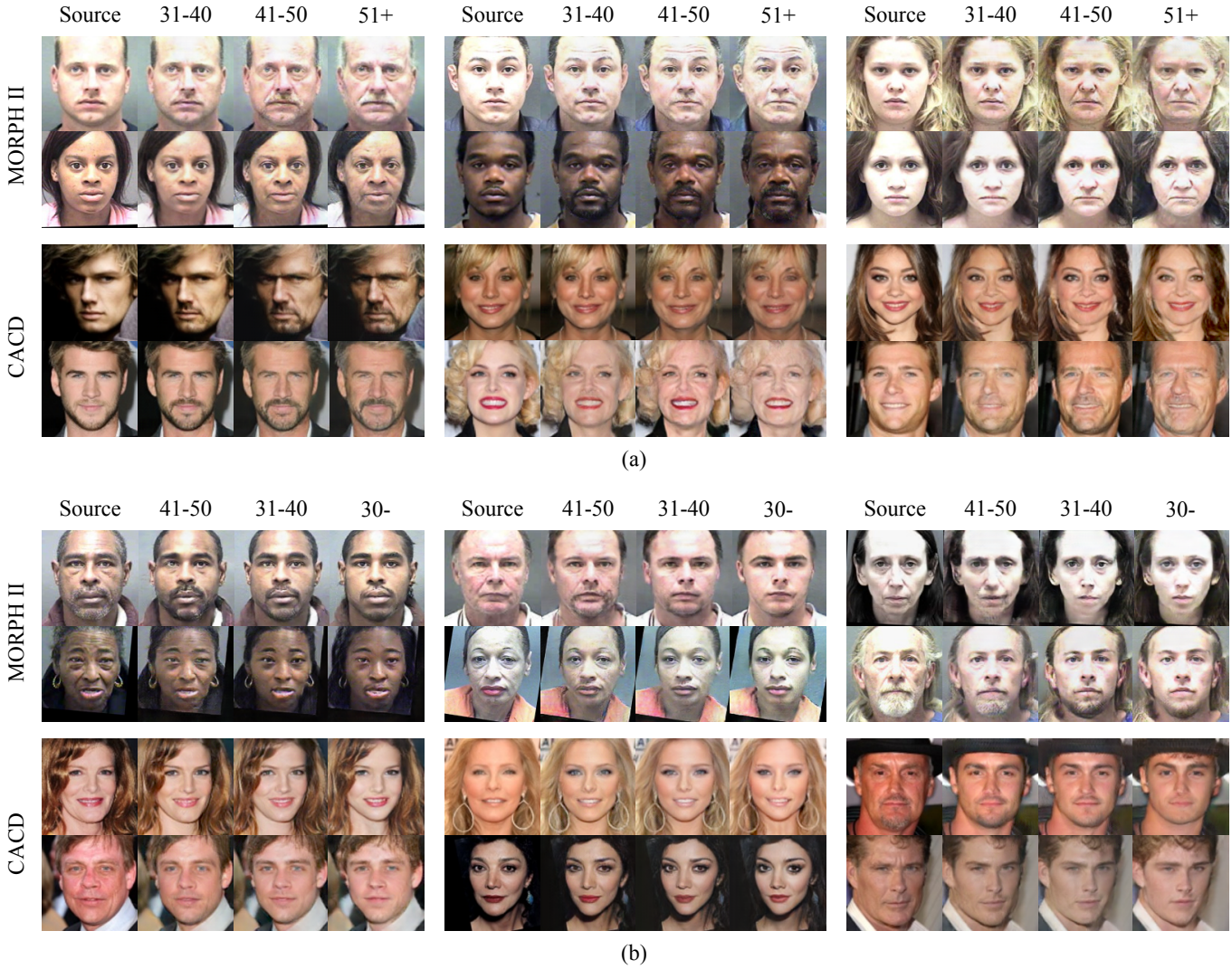


Fig. 6. Visual results for (a) the aging process and (b) the rejuvenating process. In each sub-figure, the top two rows show the synthesized results on the MORPH II dataset, and the bottom two rows show the synthesized results on the CACD.

method, in order to increase the size of the training set without introducing additional variance to the dataset, we only use the flip operation.

E. Hyper-parameter setting

When training the MTFE, we set the batch size to 128 and the initial learning rate to 0.002 for both datasets. We train it for 500 epochs while decreasing the learning rate by 0.1 every 150 epochs. When training the AOFS method, we set the batch size to 8 and the initial learning rate to 0.0002. The learning rate decreases linearly after the first 25 epochs. We empirically set $\lambda_{adv_{feature}}$ to 1 and λ_{AT} to 0.001. The margin hyper-parameter, m in Eq. (7), is set to 0.3. We use the PyTorch framework [66] for the implementation and run each experiment for 50 epochs. All experiments are run on a single NVIDIA GTX2080Ti GPU.

F. Synthesis accuracy

We first qualitatively evaluate the synthesized facial images based on their visual quality. We then present quantitative results based on age classification accuracy, image quality and the degree of mode collapse. We perform these evaluations for our AOFS and several state-of-the-art methods.

1) Visual Quality: Fig. 6 shows some sample images synthesized by our AOFS method. Fig. 6 (a) shows aging results for 6 subjects from the MORPH II dataset and 6 from the CACD using a source image from the youngest category (30^{-}). We can see that our method turns hair gray or white, introduces forehead wrinkles and nasolabial folds, and makes the skin to appear rough. Fig. 6 (b) shows rejuvenating results for 6 subjects from each dataset using a source image from the oldest category (51^{+}). We can see that for these cases, our method removes wrinkles and gray/white hair.

We also evaluate six state-of-the-art methods, namely the method by Antipov *et al.* [8], the Identity-Preserving Con-

TABLE IV
AGE CLASSIFICATION ACCURACY (%) ON THE IMAGES SYNTHESIZED FOR MORPH II AND CACD FOR THE AGING PROCESS.

	MORPH II			CACD		
Age Category	31-40	41-50	51^+	31-40	41-50	51^+
Natural Faces	59.04 ± 2.42	58.68 ± 2.18	58.83 ± 2.23	37.91 ± 5.09	37.34 ± 4.79	34.46 ± 4.92
Antipov *et al.* [8]	39.56 ± 2.28	39.79 ± 2.10	35.22 ± 2.50	20.29 ± 4.58	20.49 ± 5.04	18.43 ± 5.40
IPCGAN [4]	44.67 ± 2.25	44.70 ± 2.43	41.84 ± 1.77	24.90 ± 4.29	27.70 ± 4.25	28.49 ± 5.00
S^2GAN [53]	52.97 ± 2.65	52.46 ± 1.84	51.30 ± 1.98	29.25 ± 4.88	29.05 ± 4.62	26.33 ± 4.81
Liu *et al.* [52]	52.12 ± 1.97	53.85 ± 1.92	54.82 ± 1.45	29.31 ± 5.16	31.87 ± 4.95	**32.79 ± 4.88**
Li *et al.* [3]	51.22 ± 2.15	53.60 ± 1.74	54.61 ± 1.97	28.61 ± 4.41	31.02 ± 4.19	32.46 ± 4.75
Yang *et al.* [5]	53.24 ± 1.67	53.23 ± 2.86	53.20 ± 1.73	30.68 ± 4.12	30.85 ± 4.43	31.64 ± 4.38
w/o CDP	43.52 ± 1.73	41.53 ± 1.82	41.93 ± 1.45	25.01 ± 5.52	25.06 ± 4.89	25.55 ± 5.18
w/o ATL	56.49 ± 1.93	**55.48 ± 1.84**	54.58 ± 1.97	33.63 ± 3.96	**33.79 ± 4.34**	32.48 ± 4.61
Proposed	**56.60 ± 1.91**	55.42 ± 1.80	**54.63 ± 1.98**	**33.73 ± 3.91**	33.77 ± 4.32	32.54 ± 4.61

TABLE V
AGE CLASSIFICATION ACCURACY (%) ON THE IMAGES SYNTHESIZED FOR MORPH II AND CACD FOR THE REJUVENATING PROCESS.

	MORPH II			CACD		
Age Category	30^-	31-40	41-50	30^-	31-40	41-50
Natural Faces	63.08 ± 1.81	59.04 ± 2.42	58.68 ± 2.18	43.82 ± 4.06	37.91 ± 5.09	37.34 ± 4.79
Antipov *et al.* [8]	50.55 ± 2.32	44.71 ± 2.45	44.77 ± 1.84	28.41 ± 3.92	26.36 ± 5.87	26.17 ± 4.71
IPCGAN [4]	57.33 ± 1.82	52.03 ± 1.79	52.32 ± 2.21	32.67 ± 4.43	31.89 ± 4.50	31.41 ± 5.08
S^2GAN [53]	58.18 ± 1.83	54.11 ± 2.04	54.24 ± 1.43	33.36 ± 4.01	32.30 ± 4.38	32.63 ± 3.89
Liu *et al.* [52]	59.06 ± 2.41	55.33 ± 1.61	55.54 ± 2.01	36.65 ± 4.31	34.25 ± 4.34	34.26 ± 4.69
Li *et al.* [3]	58.87 ± 2.30	55.21 ± 2.18	55.06 ± 1.94	37.84 ± 4.66	34.95 ± 4.86	34.30 ± 4.26
Yang *et al.* [5]	60.79 ± 2.21	56.99 ± 2.17	**56.65 ± 2.39**	39.09 ± 4.72	35.62 ± 4.83	35.89 ± 4.61
w/o CDP	53.67 ± 2.35	51.41 ± 2.33	51.96 ± 2.45	29.17 ± 5.05	28.42 ± 5.39	28.67 ± 5.31
w/o ATL	61.15 ± 1.43	57.04 ± 1.33	56.57 ± 2.22	41.18 ± 4.09	**36.92 ± 4.13**	36.55 ± 4.82
Proposed	**61.20 ± 1.41**	**57.12 ± 1.36**	56.55 ± 2.23	**41.24 ± 4.12**	36.86 ± 4.10	**36.59 ± 4.81**

ditional Generative Adversarial Networks (IPCGAN) [4], the S^2GAN [53], and the methods by Liu *et al.* [52], Li *et al.* [3], and Yang *et al.* [5]. To have a fair comparison, we replace the feature extractors in these methods with our pre-trained MTFE and use the same number of residual blocks in their generator except for the method in [8], as there is no residual block originally involved in this particular method.

Since the synthesis accuracy of our AOFS method depends on the CDP, we also evaluate a baseline model without the CDP (hereinafter called *w/o CDP*) as part of an ablation study. The *w/o CDP* model replaces the CDP with a simple feature-level discriminator, which makes this model similar to a vanilla GAN but with two discriminators, one at the feature level and the other at the image level. In addition, we include another baseline model, namely *w/o ATL* in which the Adversarial Triplet loss is deleted to see whether this loss affects the age classification accuracy.

Fig. 7 depicts the visual results of these evaluations. Note that it is visually evident that the results generated by the *w/o CDP* model do not contain much aging and rejuvenating effects as this model suffers from the mode collapse issue. On the contrary, our proposed method can synthesize the aging and rejuvenating effects realistically. Among all of these evaluated state-of-the-art methods, Yang *et al.* [5] is able to synthesize the most realistic effects due to the use of a multi-level feature discriminator.

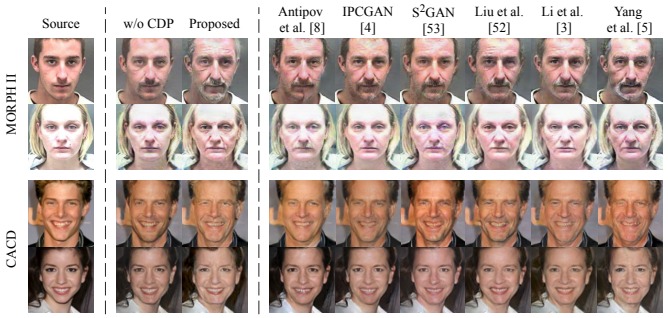


Fig. 7. Visual comparison of the proposed and six state-of-the-art methods on two benchmarks. The input image is in the youngest category and the results are expected to be in the eldest category.

2) Age category classification accuracy: Table IV and V tabulate the age category classification accuracies of various methods on the synthesized images when images from the 30^-

TABLE VI
RESNET SCORE AND FRÉCHET RESNET DISTANCE ON MORPH II.

Model	RS	FRD
Antipov *et al.* [8]	27.83 ± 1.34	31.72 ± 0.60
IPCGAN [4]	36.70 ± 1.18	28.08 ± 0.44
S^2GAN [53]	38.92 ± 1.14	25.64 ± 0.32
Liu *et al.* [52]	39.14 ± 1.23	25.57 ± 0.42
Li *et al.* [3]	39.26 ± 1.22	25.51 ± 0.41
Yang *et al.* [5]	43.35 ± 1.36	22.30 ± 0.59
w/o CDP	30.19 ± 1.26	28.62 ± 0.49
Proposed	**44.04 ± 1.25**	**21.93 ± 0.46**

TABLE VII
RESNET SCORE AND FRÉCHET RESNET DISTANCE ON CACD.

Model	RS	FRD
Antipov *et al.* [8]	24.71 ± 2.04	33.83 ± 0.95
IPCGAN [4]	33.21 ± 1.82	30.18 ± 0.79
S^2GAN [53]	34.24 ± 1.75	27.01 ± 0.61
Liu *et al.* [52]	34.54 ± 1.86	26.99 ± 0.63
Li *et al.* [3]	35.00 ± 1.91	26.91 ± 0.67
Yang *et al.* [5]	37.39 ± 2.09	24.62 ± 0.87
w/o CDP	30.87 ± 1.87	30.71 ± 0.82
Proposed	**38.55 ± 1.90**	**23.98 ± 0.73**

and 51^+ categories are used as source images, respectively. In these tables, the *Natural Faces* row tabulates the accuracy attained when using the original facial images. Since [8] uses a relatively shallow generator compared to other works, its performance is hence below others by a significant margin. IPCGAN uses the age labels as conditions in the GAN learning process and incorporates an age category classification loss. However, due to the fact that the classification error is high (the classifier is noisy), the gradient for the age information is not accurate. As a result, although its performance is higher than that of [8], it is still lower than the one attained on the original facial images by a large margin. The recently proposed S^2GAN attains a higher accuracy by implementing a customized generator where each age category is associated with a decoder. The methods of Liu *et al.* [52] and Li *et al.* [3] achieve similar accuracy since both use the Wavelet transform. Among all the other evaluated methods, the one proposed by Yang *et al.* [5] achieves the best performance by using a multi-level feature discriminator. By adding a feature-level discriminator to the vanilla GAN, the baseline *w/o CDP* model achieves a comparable performance to that achieved by IPCGAN. Additionally, we can see that the involvement of the Adversarial Triplet loss affects the age category classification accuracy subtly since our method uses disentangled features for both CDP and this loss. Overall, our proposed AOFS method outperforms all evaluated methods for the majority of age categories.

3) Image Quality: The synthesis accuracy is also related to the quality of the generated images [4]. The quality and diversity of the synthesized images are usually measured in terms of the Inception Score (IS) and the Fréchet Inception Distance (FID). IS measures the image quality and diversity by computing the KL divergence between the real and the generated class distributions. On the other hand, FID uses a multivariate Gaussian distribution to model the data distribution and the mean and the covariance from two distributions to compute their distance. Since we use a ResNet-50 to evaluate the identity permanence capabilities (see Section IV.G), we rename these two metrics as the ResNet Score (RS) and the Fréchet ResNet Distance (FRD). The RS and FRD are tabulated in Table VI and Table VII, respectively, for our AOFS method and several state-of-the-art methods. Since our AOFS method can render more realistic aging and rejuvenating effects than other evaluated methods and has stronger identity permanence capabilities, it achieves the best performance for both metrics, especially for the FRD, which is sensitive to the mode collapse issue.

TABLE VIII
DEGREE OF MODE COLLAPSE AS MEASURED BY THE KL DIVERGENCE.

Model	MORPH II	CACD
Antipov *et al.* [8]	1.86 ± 0.10	1.93 ± 0.13
IPCGAN [4]	0.64 ± 0.15	0.68 ± 0.21
S^2GAN [53]	0.59 ± 0.08	0.62 ± 0.11
Liu *et al.* [52]	0.55 ± 0.09	0.57 ± 0.13
Li *et al.* [3]	0.55 ± 0.11	0.58 ± 0.14
Yang *et al.* [5]	0.49 ± 0.04	0.52 ± 0.05
w/o CDP	1.19 ± 0.09	1.30 ± 0.14
Proposed	**0.37 ± 0.04**	**0.42 ± 0.07**

4) Degree of Mode Collapse: Since our method tackles the AOFS task from the aspect of mode learning, we also measure the degree of mode collapse by computing the KL divergence between the distributions of the synthesized images and the real images. We compute this divergence for all synthesized images within each fold.

The proposed AOFS method significantly outperforms the baseline model and the method in [8], which use the negative log-likelihood loss from the vanilla GAN (Table VIII). By using different discriminators to learn different modes, our method yields a lower divergence value compared to other methods that leverage the least square loss from the LSGAN.

G. Identity permanence

To evaluate the identity permanence on the synthesized images, we design a new baseline, the *Triplet* model. Specifically, we replace the Adversarial Triplet loss with the original Triplet loss to directly compare these two loss functions. Again, we include the baseline model *w/o CDP* here to evaluate whether the involvement of the CDP affects the identity permanence capabilities. These capabilities are measured in terms of the face verification accuracy, i.e., whether the synthesized image and the original image depict the same person. To this end, we define three input settings based on three different target age categories for each synthesis process. Specifically, the query images are the original facial images from the datasets, while the gallery images are the synthesized images that are expected to be within the target age category, as tabulated in Table

TABLE IX
FACE VERIFICATION RESULTS IN TERMS OF ACCURACY (%) FOR MORPH II AND CACD. THE QUERY IMAGES ARE THE ORIGINAL FACIAL IMAGES, AND THE GALLERY IMAGES ARE THE SYNTHESISED IMAGES GENERATED BY EACH CORRESPONDING MODEL.

Gallery Image		Aging			Rejuvenating		
		S31-40	S41-50	S51$^+$	S41-50	S31-40	S30$^-$
MORPH II	Antipov *et al.* [8]	94.46 ± 0.16	93.57 ± 0.12	91.24 ± 0.20	95.33 ± 0.16	93.54 ± 0.13	92.48 ± 0.27
	IPCGAN [4]	94.56 ± 0.23	93.87 ± 0.19	91.63 ± 0.22	94.91 ± 0.28	93.83 ± 0.20	92.21 ± 0.27
	S^{2}GAN [53]	94.88 ± 0.09	93.65 ± 0.17	91.44 ± 0.12	95.50 ± 0.11	94.72 ± 0.19	92.54 ± 0.18
	Liu *et al.* [52]	94.22 ± 0.28	93.49 ± 0.26	91.28 ± 0.21	95.63 ± 0.22	94.84 ± 0.23	93.23 ± 0.27
	Li *et al.* [3]	95.08 ± 0.11	93.99 ± 0.14	91.87 ± 0.15	95.40 ± 0.14	94.05 ± 0.16	92.52 ± 0.17
	Yang *et al.* [5]	94.29 ± 0.22	93.34 ± 0.27	91.18 ± 0.28	95.76 ± 0.21	94.40 ± 0.22	93.76 ± 0.29
	Triplet	97.87 ± 0.07	97.01 ± 0.09	94.86 ± 0.17	98.14 ± 0.06	98.23 ± 0.11	97.71 ± 0.14
	w/o CDP	99.05 ± 0.02	98.70 ± 0.06	**95.61 ± 0.14**	**99.61 ± 0.02**	**99.39 ± 0.09**	97.83 ± 0.10
	Proposed	**99.06 ± 0.03**	**98.73 ± 0.06**	95.58 ± 0.11	**99.61 ± 0.03**	99.36 ± 0.08	**97.85 ± 0.09**
CACD	Antipov *et al.* [8]	92.06 ± 0.27	88.46 ± 0.35	85.40 ± 0.56	92.67 ± 0.23	89.30 ± 0.28	86.24 ± 0.42
	IPCGAN [4]	92.29 ± 0.30	88.77 ± 0.33	85.22 ± 0.57	93.93 ± 0.25	89.32 ± 0.32	85.35 ± 0.50
	S^{2}GAN [53]	92.39 ± 0.35	88.94 ± 0.55	85.87 ± 0.59	93.32 ± 0.33	89.60 ± 0.42	86.29 ± 0.54
	Liu *et al.* [52]	92.25 ± 0.26	88.51 ± 0.32	85.46 ± 0.48	93.21 ± 0.23	89.50 ± 0.32	85.02 ± 0.47
	Li *et al.* [3]	93.33 ± 0.24	89.04 ± 0.38	85.91 ± 0.45	94.52 ± 0.21	89.47 ± 0.36	85.31 ± 0.39
	Yang *et al.* [5]	92.24 ± 0.29	88.58 ± 0.48	85.54 ± 0.57	92.80 ± 0.20	89.07 ± 0.39	86.91 ± 0.42
	Triplet	93.89 ± 0.17	92.73 ± 0.21	89.15 ± 0.24	94.79 ± 0.15	93.46 ± 0.17	90.31 ± 0.23
	w/o CDP	94.97 ± 0.12	94.14 ± 0.11	90.74 ± 0.15	95.05 ± 0.13	**94.58 ± 0.12**	91.66 ± 0.19
	Proposed	**94.98 ± 0.10**	**94.16 ± 0.14**	**90.77 ± 0.18**	**95.08 ± 0.11**	94.56 ± 0.14	**91.68 ± 0.15**

IX with the column headings *S31-40*, *S41-50*, and *S51*$^+$ for the aging process and headings *S41-50*, *S31-40*, and *S30*$^-$ for the rejuvenating process. For example, *S31-40* refers to the synthesized images expected to be within the $31-40$ category. We use the *cosine similarity* to measure the distance of each pair of query and gallery images.

All the state-of-the-art methods achieve a similar accuracy since they all use a similar strategy, namely, minimizing the distance between two identity-specific features using the L1 or L2 loss. Li *et al.* [3] slightly outperforms other methods as it uses a combination of these two losses (Table IX). The similar performance achieved among these methods may also be due to the quality of the images, since the identity information may be distorted in images of poor quality. By replacing the L1 or L2 loss with the Triplet loss, the identity permanence capability can be remarkably boosted by about 3% on both datasets. Thanks to the disentangled features and discriminative ability of the Adversarial Triplet loss, the effects of the CDP on the identity permanence capabilities is subtle (i.e., ± 0.03). As a result, this loss, which reduces intra-class variances within each age category in the feature space and is employed in the proposed AOFS method, achieves the highest accuracy among all identity preserving methods.

V. CONCLUSION

In this paper, we tackle the Age-Oriented Face Synthesis (AOFS) task via a mode-specific learning. Specifically, we present an AOFS method that incorporates a novel Conditional Discriminator Pool (CDP) to alleviate the mode collapse issue in the vanilla GAN. We also incorporate a novel Adversarial Triplet loss to attain strong identity permanence capabilities. By using the proposed CDP, only the target feature-level discriminator that learns the corresponding mode is deployed. Note that our mode-specific learning method (e.g., CDP) does not increase the computational complexity during training. Hence, our CDP allows learning multiple modes explicitly and independently. As a result, our AOFS method outperforms several state-of-the-art methods on AOFS benchmark datasets. In the future, we will aim to improve the aging and rejuvenating effects by including the synthesis and removal of wrinkles and face shape manipulation. We hypothesize that by improving these aspects of the synthesis process, synthesizing more realistic younger and older face images can be achieved.

VI. ACKNOWLEDGMENT

This work is supported by the EU Horizon 2020 - Marie Sklodowska-Curie Actions through the project Computer Vision Enabled Multimedia Forensics and People Identification (Project No. 690907, Acronym: IDENTITY).

REFERENCES

[1] Yun Fu, Guodong Guo, and Thomas S Huang, "Age synthesis and estimation via faces: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 11, pp. 1955–1976, 2010.
[2] Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 442–455, 2002.
[3] Peipei Li, Yibo Hu, Ran He, and Zhenan Sun, "Global and local consistent wavelet-domain age synthesis," *IEEE Transactions on Information Forensics and Security*, 2019.
[4] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao, "Face aging with identity-preserved conditional generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7939–7947.

[5] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K Jain, "Learning face age progression: A pyramid architecture of gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 31–39.

[6] Zhifei Zhang, Yang Song, and Hairong Qi, "Age progression/regression by conditional adversarial autoencoder," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 2.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[8] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay, "Face aging with conditional generative adversarial networks," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2089–2093.

[9] Angelo Genovese, Vincenzo Piuri, and Fabio Scotti, "Towards explainable face aging with generative adversarial networks," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3806–3810.

[10] Evangelia Pantraki and Constantine Kotropoulos, "Face aging as image-to-image translation using shared-latent space generative adversarial networks," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2018, pp. 306–310.

[11] Jian Zhao, Yu Cheng, Yi Cheng, Yang Yang, Fang Zhao, Jianshu Li, Hengzhu Liu, Shuicheng Yan, and Jiashi Feng, "Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 9251–9258.

[12] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *International Conference on Learning Representations*, 2014.

[13] Martin Arjovsky and Leon Bottou, "Towards principled methods for training generative adversarial networks," *International Conference on Learning Representations*, 2017.

[14] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li, "Mode regularized generative adversarial networks," *International Conference on Learning Representations*, 2017.

[15] Christopher M Bishop, *Pattern recognition and machine learning*, springer, 2006.

[16] Joseph P Robinson, Yuncheng Li, Ning Zhang, Yun Fu, and Sergey Tulyakov, "Laplace landmark localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10103–10112.

[17] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein, "Unrolled generative adversarial networks," *International Conference on Learning Representations*, 2017.

[18] Jianhua Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991.

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.

[20] William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M Dai, Shakir Mohamed, and Ian Goodfellow, "Many paths to equilibrium: Gans do not need to decrease a divergence at every step," *arXiv preprint arXiv:1710.08446*, 2017.

[21] Alexia Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard gan," *International Conference on Learning Representations*, 2019.

[22] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka, "f-gan: Training generative neural samplers using variational divergence minimization," in *Advances in Neural Information Processing Systems*, 2016, pp. 271–279.

[23] Imre Csiszár, Paul C Shields, et al., "Information theory and statistics: A tutorial," *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 4, pp. 417–528, 2004.

[24] Solomon Kullback and Richard A Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[25] Karl Pearson, "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[26] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[27] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2813–2821.

[28] Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei, "Variational inference via χ upper bound minimization," in *Advances in Neural Information Processing Systems*, 2017, pp. 2732–2741.

[29] Xudong Mao, Qing Li, Haoran Xie, Raymond Yiu Keung Lau, Zhen Wang, and Stephen Paul Smolley, "On the effectiveness of least squares generative adversarial networks," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[30] Tu Nguyen, Trung Le, Hung Vu, and Dinh Phung, "Dual discriminator generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2017, pp. 2670–2680.

[31] Zhaoyu Zhang, Mengyan Li, and Jun Yu, "On the convergence and mode collapse of gan," in *SIGGRAPH Asia 2018 Technical Briefs*. ACM, 2018, p. 21.

[32] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter, "Self-normalizing neural networks," in *Advances in neural information processing systems*, 2017, pp. 971–980.

[33] Zhaoyu Zhang, Mengyan Li, and Jun Yu, "D2pggan: Two discriminators used in progressive growing of gans," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3177–3181.

[34] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *International Conference on Learning Representations*, 2018.

[35] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan, "Generative multi-adversarial networks," *arXiv preprint arXiv:1611.01673*, 2016.

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[37] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 403–412.

[38] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 5375–5384.

[39] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen, "Visible thermal person re-identification via dual-constrained top-ranking.," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 1092–1099.

[40] Leonard S Mark and James T Todd, "The perception of growth in three dimensions," *Attention, Perception, & Psychophysics*, vol. 33, no. 2, pp. 193–196, 1983.

[41] Leonard S Mark, James T Todd, and Robert E Shaw, "Perception of growth: A geometric analysis of how different styles of change are distinguished.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 7, no. 4, pp. 855, 1981.

[42] James T Todd, Leonard S Mark, Robert E Shaw, and John B Pittenger, "The perception of human growth," *Scientific american*, vol. 242, no. 2, pp. 132–145, 1980.

[43] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[44] Yosuke Bando, Takaaki Kuratate, and Tomoyuki Nishita, "A simple method for modeling wrinkles on human skin," in *10th Pacific Conference on Computer Graphics and Applications, 2002. Proceedings.* IEEE, 2002, pp. 166–175.

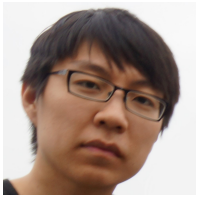
[45] Zicheng Liu, Zhengyou Zhang, and Ying Shan, "Image-based surface detail transfer," *IEEE Computer Graphics and Applications*, vol. 24, no. 3, pp. 30–35, 2004.

[46] Shigeru Mukaida and Hiroshi Ando, "Extraction and manipulation of wrinkles and spots for facial image synthesis," in *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*. IEEE, 2004, pp. 749–754.

[47] Yin Wu, Prem Kalra, Laurent Moccozet, and Nadia Magnenat-Thalmann, "Simulating wrinkles and skin aging," *The visual computer*, vol. 15, no. 4, pp. 183–198, 1999.

[48] Yin Wu, Nadia Magnenat Thalmann, and Daniel Thalmann, "A dynamic wrinkle model in facial animation and skin ageing," *The journal of visualization and computer animation*, vol. 6, no. 4, pp. 195–205, 1995.

[49] Narayanan Ramanathan and Rama Chellappa, "Modeling age progression in young faces," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 1, pp. 387–394.
[50] Xiangbo Shu, Jinhui Tang, Zechao Li, Hanjiang Lai, Liyan Zhang, Shuicheng Yan, et al., "Personalized age progression with bi-level aging dictionary learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 905–917, 2018.
[51] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
[52] Yunfan Liu, Qi Li, and Zhenan Sun, "Attribute-aware face aging with wavelet-based generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11877–11886.
[53] Zhenliang He, Meina Kan, Shiguang Shan, and Xilin Chen, "S2gan: Share aging factors across ages and share aging trends among individuals," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9440–9449.
[54] Yitong Wang, Dihong Gong, Zheng Zhou, Xing Ji, Hao Wang, Zhifeng Li, Wei Liu, and Tong Zhang, "Orthogonal deep features decomposition for age-invariant face recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 738–753.
[55] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[56] Haoyi Wang, Xingjie Wei, Victor Sanchez, and Chang-Tsun Li, "Fusion network for face-based age estimation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2675–2679.
[57] Rasmus Rothe, Radu Timofte, and Luc Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, vol. 126, no. 2-4, pp. 144–157, 2018.
[58] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[59] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.
[60] Karl Ricanek and Tamirat Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*. IEEE, 2006, pp. 341–345.
[61] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *European conference on computer vision*. Springer, 2014, pp. 768–783.
[62] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou, "Agedb: the first manually collected, in-the-wild age database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 51–59.
[63] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 67–74.
[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint*, 2017.
[65] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.
[66] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.



Haoyi Wang received the BSc degree in engineering from North China University of Technology, China, and University of Central Lancashire, UK, in 2013, the MSc degree from the University of Manchester, UK, in 2014. He is currently pursuing a PhD in computer science at the University of Warwick, UK. His research interests include deep supervised and unsupervised learning, generative models, and computer vision.

Victor Sanchez received the M.Sc. degree from the University of Alberta, Canada, in 2003, and the Ph.D. degree from the University of British Columbia, Canada, in 2010. From 2011 to 2012, he was with the Video and Image Processing Laboratory, University of California at Berkeley, as a Post-Doctoral Researcher. In 2012, he was a Visiting Lecturer with the Group on Interactive Coding of Images, Universitat Autonoma de Barcelona. From 2018 to 2019, he was a Visiting Scholar at the School of Electrical and Information Engineering, University of Sydney, Australia. He is currently an Associate Professor with the Department of Computer Science, University of Warwick, U.K. His main research interests are in the area of signal and information processing with applications to multimedia analysis, image and video coding, security, and communications. He has authored several technical papers in these areas and co-authored a book (Springer, 2012). His research has been funded by Consejo Nacional de Ciencia y Tecnologia, Mexico, the Natural Sciences and Engineering Research Council of Canada, the Canadian Institutes of Health Research, the FP7 and H2020 programs of the European Union, the Engineering and Physical Sciences Research Council, U.K., and the Defence and Security Accelerator, U.K.

Chang-Tsun Li received the BSc degree in electrical engineering from National Defence University (NDU), Taiwan, in 1987, the MSc degree in computer science from U.S. Naval Postgraduate School, USA, in 1992, and the PhD degree in computer science from the University of Warwick, UK, in 1998. He was an associate professor of the Department of Electrical Engineering at NDU during 1998-2002 and a visiting professor of the Department of Computer Science at U.S. Naval Postgraduate School in the second half of 2001. He was a professor of the Department of Computer Science at the University of Warwick (UK) until January 2017 and a professor of Charles Sturt University, Australia,) from January 2017 to February 2019. He is currently a professor of the School of Information Technology of Deakin University, Australia. His research interests include multimedia forensics and security, biometrics, data mining, machine learning, data analytics, computer vision, image processing, pattern recognition, bioinformatics, and content-based image retrieval. The outcomes of his multimedia forensics research have been translated into award-winning commercial products protected by a series of international patents and have been used by a number of police forces and courts of law around the world. He is currently the EURASIP Journal of Image and Video Processing (JIVP) and Associate Editor of IET Biometrics. He involved in the organisation of many international conferences and workshops and also served as member of the international program committees for several international conferences. He is also actively contributing keynote speeches and talks at various international events.