When Age-Invariant Face Recognition Meets Face Age Synthesis: A Multi-Task Learning Framework

Zhizhong Huang[1] Junping Zhang[1] Hongming Shan[2,3]*
[1] Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200433, China
[2] Institute of Science and Technology for Brain-inspired Intelligence and MOE Frontiers Center for Brain Science, Fudan University, Shanghai 200433, China
[3] Shanghai Center for Brain Science and Brain-inspired Technology, Shanghai 200031, China
{zzhuang19, jpzhang, hmshan}@fudan.edu.cn

Abstract

To minimize the effects of age variation in face recognition, previous work either extracts identity-related discriminative features by minimizing the correlation between identity- and age-related features, called age-invariant face recognition (AIFR), or removes age variation by transforming the faces of different age groups into the same age group, called face age synthesis (FAS); however, the former lacks visual results for model interpretation while the latter suffers from artifacts compromising downstream recognition. Therefore, this paper proposes a unified, multi-task framework to jointly handle these two tasks, termed MTLFace, which can learn age-invariant identity-related representation while achieving pleasing face synthesis. Specifically, we first decompose the mixed face features into two uncorrelated components—identity- and age-related features—through an attention mechanism, and then decorrelate these two components using multi-task training and continuous domain adaption. In contrast to the conventional one-hot encoding that achieves group-level FAS, we propose a novel identity conditional module to achieve identity-level FAS, with a weight-sharing strategy to improve the age smoothness of synthesized faces. In addition, we collect and release a large cross-age face dataset with age and gender annotations to advance AIFR and FAS. Extensive experiments on five benchmark cross-age datasets demonstrate the superior performance of our proposed MTLFace over state-of-the-art methods for AIFR and FAS. We further validate MTLFace on two popular general face recognition datasets, showing competitive performance for face recognition in the wild. The source code and dataset are available at https://github.com/Hzzone/MTLFace.

*Corresponding author

Figure 1: Sample results by our MTLFace. First row: the real faces of the same person at different ages with estimated age labels underneath. Remaining rows: the synthesized faces when given input faces in the red boxes.

1. Introduction

Face recognition has been a hot research topic in computer vision for many years. Recently, deep-learning-based methods achieve excellent performance, even surpassing humans in several scenarios, by empowering the face recognition models with deep neural networks [12, 20, 41]. The traditional wisdom is to utilize the margin-based metrics to increase the intra-class compactness and train the models with a massive amount of data to improve face recognition performance [51].

Despite the remarkable success of general face recognition (GFR), *how to minimize the effects of age variation* is a lingering challenge for current face recognition systems to correctly identify faces in many practical applications such as finding lost children. Therefore, it is of great significance

to achieve face recognition without age variation, *i.e*., age-invariant face recognition or AIFR. However, AIFR remains extremely challenging in the following three aspects. First, when the age gap becomes large in cross-age face recognition, age variation can largely affect the facial appearance, compromising the face recognition performance. Second, face age synthesis (FAS) is a complex process involving face aging/rejuvenation (*a.k.a* age progression/regression) since the facial appearance drastically changes over a long time and differs from person to person. Last, it is infeasible to obtain a large paired face dataset to train a model in rendering faces with natural effects while preserving identities.

To overcome these issues, current methods for AIFR can be roughly divided into two categories: generative and discriminative models. Given a face image, the generative models [8, 21, 33] aim to transform the faces of different ages into the same age group in order to assist the face recognition. Recently, generative adversarial networks (GANs) [11] have been successfully used to enhance the image quality of synthesized faces [23, 26, 49, 54, 55]; they typically use the one-hot encoding to specify the target age group. However, the one-hot encoding represents the age group-level face transformation, ignoring the identity-level personalized patterns and leading to unexpected artifacts. As a result, the performance of AIFR cannot be significantly improved due to the unpleasing synthesized faces and unexpected changes in identity. On the other hand, the discriminative models [4, 47] focus on extracting age-invariant features by disentangling the identity-related information from the mixed information so that only the identity-related information is expected for the face recognition systems. Although achieving promising performance in AIFR, they cannot provide users, for example policemen, with visual results as the generative methods to further verify the identities, which can compromise the model interpretability in the decision-making processes of many practical applications.

To further improve the image quality for generative models and provide the model interpretability for discriminative models, we propose a unified, multi-task learning framework to simultaneously achieve AIFR and FAS, termed MTLFace, which can enjoy the best of both worlds; *i.e*., learning age-invariant identity-related representation while achieving pleasing face synthesis. More specifically, we first decompose the mixed high-level features into two uncorrelated components—identity- and age-related features—through an attention mechanism. We then decorrelate these two components in a multi-task learning framework, in which an age estimation task is to extract age-related features while a face recognition task is to extract identity-related features; in addition, a continuous cross-age discriminator with a gradient reversal layer [7] further encourages the identity-related age-invariant features. Moreover, we propose an identity conditional module to achieve identity-level transformation patterns for FAS, with a weight-sharing strategy to improve the age smoothness of synthesized faces; *i.e*., the faces are aged smoothly. Extensive experiments demonstrate superior performance over existing state-of-the-art methods for AIFR and FAS, and competitive performance for general face recognition in the wild. Fig. 1 presents an example of age progression/regression of the same person from our MTLFace, showing that our framework can synthesize photorealistic faces while preserving identity.

Our contributions are summarized as follows. *First*, we propose a unified, multi-task learning framework to jointly handle AIFR and FAS, which can learn age-invariant identity-related representation while achieving pleasing face synthesis. *Second*, we propose an attention-based feature decomposition to separate the age- and identity-related features on high-level feature maps, which can constrain the decomposition process in contrast to the previous unconstrained decomposition on feature vectors. Age estimation and face recognition tasks are incorporated to supervise the decomposition process in conjunction with a continuous domain adaption. *Third*, compared to previous one-hot encoding achieving age group-level face transformation, we propose a novel identity conditional module to achieve identity-level face transformation, with a weight-sharing strategy to improve the age smoothness of synthesized faces. *Fourth*, extensive experiments demonstrate the effectiveness of the proposed framework for AIFR and FAS on five benchmark datasets, and competitive performance on two popular GFR datasets. *Last*, we collect and release a large cross-age dataset of millions of faces with age and gender annotations, which can advance the development of the AIFR and FAS. In addition, it is expected to be useful for other face-related research tasks; *e.g*., pretraining for face age estimation.

2. Related Work

Age-invariant face recognition (AIFR). Prior studies usually tackle age variation by disentangling age-invariant features from mixed features. For example, [9] adopted the hidden factor analysis (HFA) to factorize the mixed features and then reduce the age variation in identity-related features. [50] extended HFA [9] into a deep learning framework with the latent factor guided convolutional neural network (LF-CNN). At the same time, [57] introduced an age estimation task to guide the AIFR. Most recently, CNNs-based discriminative methods have achieved promising results for AIFR. OE-CNN [47] adapted a modified softmax loss [25] for AIFR by decomposing the facial embeddings into two orthogonal components such that the identity- and age-related features are represented as the angular and radial directions, respectively. Similarly, DAL [43] achieved

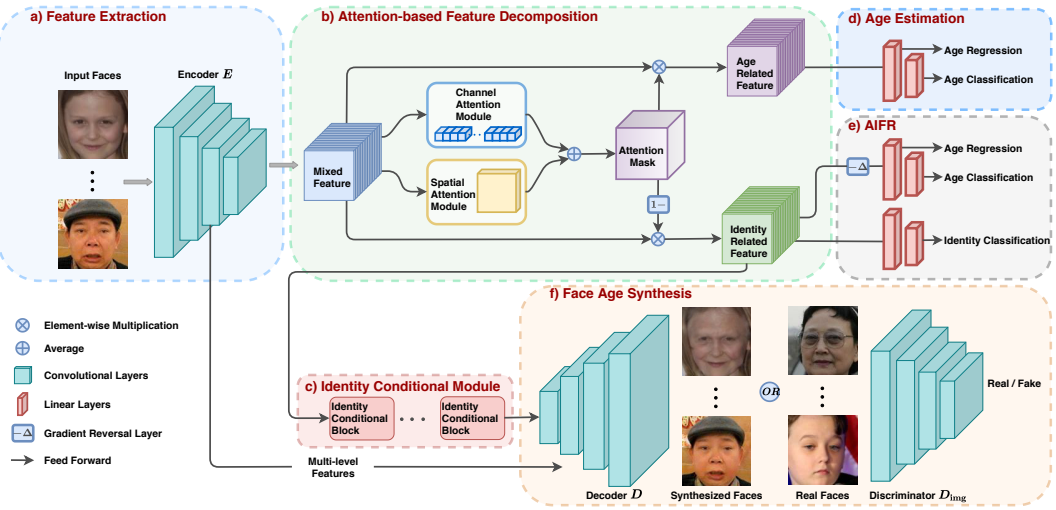

Figure 2: An overview of the proposed MTLFace including two tasks. AIFR: The encoder E first extracts the mixed feature maps from input faces, which are then decomposed into two disjoint identity- and age-related feature maps by the multi-task training and continuous domain adaption. FAS: The decoder D produces synthesized faces through identity conditional module based on multi-level features; the PatchDiscriminator D_{img} penalizes the framework for better visual quality.

the feature decomposition in an adversarial manner under the assumption that the two components are uncorrelated.

The work related to ours is [56], in which a cGANs-based model, with cross-age domain adversarial training extracting age-invariant representations, is adopted to achieve the two tasks simultaneously. However, it generates over-smoothed faces with subtle changes. Different from [56], our framework has following advantages: 1) our feature decomposition is done on feature maps through an attention mechanism; 2) a continuous domain adaption with gradient reversal layer is used to learn age-invariant identity-related representation; and 3) identity conditional module can achieve identity-level face synthesis and improve the age smoothness of synthesized faces.

Face age synthesis (FAS). Existing methods for FAS can be roughly divided into physical model-, prototype-, and deep generative model-based methods. Physical model-based methods [35, 36, 42] mechanically model the changes of appearance over time, but they are computationally expensive and require massive paired images of the same person with a long time. Prototype-based methods [19, 39] achieve face aging/rejuvenation using the average of faces in each age group, hence the identity cannot be well preserved. The deep generative model-based methods [32, 46] exploit the deep neural network for this task. For example, recurrent face aging (RFA) [46] used a recurrent neural network to model the intermediate transition states of age progression/regression, traversing on which a smooth face aging process can be achieved. Inspired by the powerful capability of generative adversarial networks (GANs) [11], especially conditional GANs (cGANs) [28], in generating high-quality images, many recent studies [16, 55, 49, 54] resort to them to improve the visual quality of synthesized faces and train models with unpaired age data. For example, [55] used a conditional adversarial autoencoder (CAAE) to achieve both age progression/regression by traversing on a low-dimensional face manifold. [49] introduced the perceptual loss to preserve the identities during face aging/rejuvenation. [54] designed a discriminator with the pyramid architecture to enhance the aging details.

However, these methods mainly aim at improving the visual quality of generated faces, and hardly improve the performance of AIFR due to the artifacts resulting from group-level face transformation, and the unexpected change in identity. Our method differs in the following aspects: 1) the proposed MTLFace achieves AIFR and FAS simultaneously to enhance the visual quality with identity-related information from AIFR; 2) the proposed identity conditional module (ICM) achieves an identity-level face age synthesis in contrast to the previous group-level face age synthesis; and 3) a weight-sharing strategy in ICM can improve the age smoothness of synthesized faces.

3. Methodology

Fig. 2 presents the architecture of the proposed MTLFace, which will be detailed in the following subsections.

3.1. Attention-based Feature Decomposition

As the faces change a lot over time, the critical problem of AIFR is that the age variation usually introduces the increasing intra-class distances. As a result, it is chal-

lenging to correctly recognize two faces of the same person with a large gap, since the mixed facial representations are severely entangled with unrelated information such as facial shape and texture changes. Recently, Wang *et al*. design a linear factorization module to decompose the feature vectors into two unrelated components [43]. Formally, the feature vector $\boldsymbol{x} \in \mathbb{R}^d$ extracted from an input image $\boldsymbol{I} \in \mathbb{R}^{3\times H\times W}$ can be decomposed as [43]:

$$\boldsymbol{x} = \boldsymbol{x}_{\text{age}} + \boldsymbol{x}_{\text{id}}, \tag{1}$$

where $\boldsymbol{x}_{\text{age}}$ and $\boldsymbol{x}_{\text{id}}$ denote the age- and identity-related components, respectively. This decomposition is implemented through a residual mapping. However, it has the following drawbacks: 1) this decomposition performs on one-dimensional feature vector, the resultant identity-related component lacks spatial information of face, not suitable for FAS; and 2) this decomposition is unconstrained, which may lead to unstable training.

To address these drawbacks, we instead propose to decompose the mixed feature-maps in a high-level semantic space through an attention mechanism, termed attention-based feature decomposition or AFD. The main reason is that manipulating on the feature vectors is more complicated than on the feature maps since the aging/rejuvenation effects, such as beards and wrinkles, appear in the semantic feature space but lose in the one-dimensional features. Formally, we use a ResNet-like backbone as encoder E to extract mixed feature maps $\boldsymbol{X} \in \mathbb{R}^{C\times H'\times W'}$ from an input image $\boldsymbol{I}$, *i.e*. $\boldsymbol{X} = E(\boldsymbol{I})$, the AFD can be defined as:

$$\boldsymbol{X} = \underbrace{\boldsymbol{X} \circ \sigma(\boldsymbol{X})}_{\boldsymbol{X}_{\text{age}}} + \underbrace{\boldsymbol{X} \circ \big(1 - \sigma(\boldsymbol{X})\big)}_{\boldsymbol{X}_{\text{id}}}, \tag{2}$$

where $\circ$ denotes element-wise multiplication and σ represents an attention module. In doing so, the age-related information in the feature maps can be separated through the attention module supervised by an age estimation task, and the residual part, regarded as the identity-related information, can be supervised by a face recognition task. As a result, the attention mechanism constrains the decomposition module, better at detecting the age-related features in semantic feature maps. We note that $\boldsymbol{X}$ is assumed to only contain the age and identity information as driven by the two corresponding tasks, the remaining information such as background is important for FAS, which is preserved by skip connections from encoder to decoder. Fig. 2(b) details the proposed AFD.

In this paper, we adopt the average of channel attention (CA) [14] and spatial attention (SA) [52] to highlight age-related information at both channel and spatial levels. Note that the outputs of these two attentions have different sizes, we first stretch each of them to the original input size and then average them. Different attention modules such as CA, SA, and CBAM [52] are also investigated in Sec. 4.

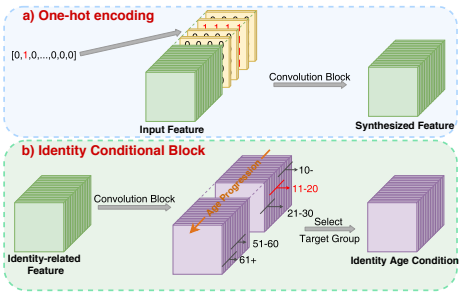

Figure 3: Comparison between one-hot encoding and ICB.

3.2. Identity Conditional Module

The mainstream face aging studies [23, 26, 49, 54, 55] usually split the ages into several non-overlapping age groups, since the changes over time are minor with a small age gap. These methods typically use one-hot encoding to specify the target age group to control the aging/rejuvenation process [23, 49, 55] as illustrated in Fig. 3(a). Consequently, a group-level aging/rejuvenation pattern, such as people having a beard when they are 30 years old, is learned for each age group due to the use of one-hot age condition. Its drawbacks are twofold: 1) one-hot encoding represents the age group-level aging/rejuvenation pattern, ignoring identity-level personalized pattern, particularly for different genders and races; and 2) one-hot encoding may not ensure the age smoothness of synthesized faces.

To address these issues raised by one-hot encoding, we propose an identity conditional block (ICB) to achieve identity-level aging/rejuvenation pattern, with a weight-sharing strategy to improve the age smoothness of synthesized faces. Specifically, the proposed ICB takes the identity-related feature from AFD as input to learn an identity-level aging/rejuvenation pattern. Next, we propose a weights-sharing strategy to improve the age smoothness of synthesized faces so that some convolutional filters are shared across adjacent age groups as shown in Fig. 3(b). The rationale behind this idea is that faces are gradually changed over time, where the shared filters can learn some common aging/rejuvenation patterns between adjacent age groups. Note that $\boldsymbol{X}_{\text{id}}$ is reduced from 512 to 128 using 1×1 convolutions to reduce the computational cost. In this paper, a hyper-parameter s to control how many filters are shared for two adjacent age groups, which is empirically set to $1/8$; *i.e*., the adjacent two age groups share 16 filters. We stack ICBs to form an identity conditional module (ICM).

3.3. Multi-task Learning Framework

Age-invariant face recognition (AIFR) task. To encourage AFD to robustly decompose features, we use an age estimation task and a face recognition task to supervise the

feature decomposition. Specifically, $\boldsymbol{X}_{\text{age}}$ draws the age variations by an age estimation task while $\boldsymbol{X}_{\text{id}}$ encodes the identity-related information. First, we include an age estimation network A with two linear layers of 512 and 101 neurons to achieve age regression similar to deep expectation (DEX) [38] that learns the age distribution by computing a softmax expected value. Second, we append another linear layer $\boldsymbol{W} \in \mathbb{R}^{101\times N}$ on top of A for age classification, regularizing the learned distribution, where N is the number of age groups. The loss function to optimize age estimation can be defined as:

$$\ell_{\text{AE}}(\boldsymbol{X}_{\text{age}}) = \mathbb{E}_{\boldsymbol{I}}\big[\ell_{\text{MSE}}\left(\text{DEX}(A(\boldsymbol{X}_{\text{age}})), y_{\text{age}}\right) + \ell_{\text{CE}}\left(A(\boldsymbol{X}_{\text{age}})\boldsymbol{W}, c_{\text{age}}\right)\big], \quad (3)$$

where y_{age}, c_{age}, ℓ_{MSE}, and ℓ_{CE} are the ground truth age, ground truth age group, mean squared error (MSE) for age regression, and cross-entropy (CE) loss for age group classification, respectively.

Next, we leverage one linear layer L of 512 neurons to extract the feature vectors, and use the CosFace loss to supervise the learning of $\boldsymbol{X}_{\text{id}}$ for identity classification. We also introduce a cross-age domain adversarial learning that encourages $\boldsymbol{X}_{\text{id}}$ to be age-invariant through a continuous domain adaption [44] with a gradient reversal layer (GRL) [7]. The final loss for AIFR is formulated as:

$$\begin{aligned}\mathcal{L}^{\text{AIFR}} =& \ell_{\text{COSFACE}}(L(\boldsymbol{X}_{\text{id}}), y_{\text{id}}) \\ &+\lambda^{\text{AIFR}}_{\text{age}}\mathcal{L}_{\text{AE}}(\boldsymbol{X}_{\text{age}}) + \lambda^{\text{AIFR}}_{\text{id}}\mathcal{L}_{\text{AE}}(\text{GRL}(\boldsymbol{X}_{\text{id}})),\end{aligned} \quad (4)$$

where the first term is the CosFace loss, the second term is the age estimation loss, and the last term is the domain adaption loss, y_{id} is the identity label, and λ_* controls the balance of different loss terms. Note that the second and third terms use the same network structure, but have different inputs and are trained independently. The activation functions and batch normalizations are ignored for simplicity, and our face recognition model is designed strictly following the setting in [5] except the AFD.

Face age synthesis (FAS) task. Fig. 2(f) demonstrates the FAS process of our proposed method. In detail, the identity-level age condition is derived from the discriminative facial representations $\boldsymbol{X}_{\text{id}}$ by applying an identity conditional module (ICM) with a series of ICBs. Then, the decoder D reconstructs the progressed/regressed faces from the multi-level high-resolution features extracted from the encoder E, under the control of the learned identity-level age condition. Formally, the process of rendering input face $\boldsymbol{I}$ to the synthesized face $\widehat{\boldsymbol{I}}_t$ that belongs to target age group t can be written as:

$$\widehat{\boldsymbol{I}}_t = D\big(\{E_l(\boldsymbol{I})\}_{l=1}^{3}, \text{ICM}(\boldsymbol{X}_{\text{id}}, t)\big), \quad (5)$$

where l denotes the index of different levels of high-resolution features extracted from different layers of the encoder E.

To facilitate the visual quality of generated faces, the FAS task is trained using GANs framework. In this paper, we adopt the PatchDiscriminator from [17] as our discriminator D_{img} to emphasize the local-patch of generated and real images. Furthermore, the least-squares GANs [27] are employed to optimize the GANs framework for an improved quality of generated images and stable training process, which can be formulated as follows:

$$\mathcal{L}^{\text{FAS}}_{\text{adv}} = \frac{1}{2}\mathbb{E}_{\boldsymbol{I}}\left[D_{\text{img}}([\widehat{\boldsymbol{I}}_t; \boldsymbol{C}_t]) - 1\right]^2, \quad (6)$$

where $\boldsymbol{C}_t$ is the one-hot encoding used in traditional cGANs framework for aligning the age condition, and $[\cdot\,;\cdot]$ denotes the matrix concatenation along channel dimension. To preserve the identities of input faces and improve the age accuracy, we leverage the encoder E and AFD to supervise the FAS task. Consequently, we can achieve both face aging and rejuvenation in a holistic, end-to-end manner, as illustrated in Fig. 2. This process can be formulated as follows:

$$\boldsymbol{X}^t_{\text{age}}, \boldsymbol{X}^t_{\text{id}} = \text{AFD}\big(E(\widehat{\boldsymbol{I}}_t)\big), \quad (7)$$

$$\mathcal{L}^{\text{FAS}}_{\text{age}} = \ell_{\text{CE}}\big(A(\boldsymbol{X}^t_{\text{age}})\boldsymbol{W}, c^t_{\text{age}}\big), \quad (8)$$

$$\mathcal{L}^{\text{FAS}}_{\text{id}} = \mathbb{E}_{\boldsymbol{X}_s}\left\|\boldsymbol{X}^t_{\text{id}} - \boldsymbol{X}_{\text{id}}\right\|^2_F, \quad (9)$$

where $\|\cdot\|_F$ represents the Frobenius norm.

The final loss to optimize this task can be written as:

$$\mathcal{L}^{\text{FAS}} = \lambda^{\text{FAS}}_{\text{adv}}\mathcal{L}^{\text{FAS}}_{\text{adv}} + \lambda^{\text{FAS}}_{\text{id}}\mathcal{L}^{\text{FAS}}_{\text{id}} + \lambda^{\text{FAS}}_{\text{age}}\mathcal{L}^{\text{FAS}}_{\text{age}}, \quad (10)$$

where λ^{FAS}_* controls the importance of different loss terms of FAS task. The loss function to optimize the discriminator D_{img} in the context of least-squares GANs is defined as:

$$\begin{aligned}\mathcal{L}^{\text{FAS}}_{D_{\text{img}}} =& \frac{1}{2}\mathbb{E}_{\boldsymbol{I}_t}\left[D_{\text{img}}\big([\boldsymbol{I}_t; \boldsymbol{C}_t]\big) - 1\right]^2 \\ &+\frac{1}{2}\mathbb{E}_{\boldsymbol{I}}\left[D_{\text{img}}\big([\widehat{\boldsymbol{I}}_t; \boldsymbol{C}_t]\big)\right]^2.\end{aligned} \quad (11)$$

At the testing stage, the only difference from existing FAS methods is that our method needs to specify the corresponding group of filters. Consequently, our method enjoys the advantages similar to [13] that the computational cost can be significantly reduced by only encoding input faces once, instead of N times in previous works [23, 26, 49, 54, 55], where N is the number of age groups.

Optimization and inference. In our MTLFace, the AIFR learns the discriminative facial representations and age estimation while the FAS produces the visual results which can boost the model interpretability for AIFR. Therefore, both two tasks can be jointly accomplished by optimizing these two tasks in a GAN-like manner; they mutually leverage each other to boost themselves. In other words, the AIFR encourages FAS to render faces to preserve its identity while FAS can facilitate the extraction of the identity-related feature and boost the model interpretability for AIFR. Consequently, we alternately train these two tasks in a unified, multi-task, end-to-end framework.

4. Experiments

4.1. Implementation Details

Data collection. Current research on AIFR lacks a large-scale face dataset of millions of face images with a large age gap. To advance the development of AIFR and FAS, we create and release a new large cross-age face dataset (LCAF) with 1.7M faces from cross-age celebrities. We further build a subset of cross-age face dataset (SCAF) containing about 0.5M images from 12K individuals following [43, 47] for fair comparisons. We note that the training (LCAF) and testing data may have very little, or even no identities overlapping as [5] already removed 500+ identities from their clean MS-Celeb-1M dataset by checking the similarity of faces between training and testing data. Following the mainstream literature [13, 22, 23, 26, 54] with the time span of 10 years for each age group, the ages in this paper are divided into seven non-overlapping groups; *i.e.*, 10-, 11-20, 21-30, 31-40, 41-50, 51-60, and 61+. Note that it is a much more challenging problem to perform FAS on seven groups than on four groups in previous work.

Training details. Similar to [5], we adopted ResNet-50 as the encoder E. In the decoder D, the identity age condition is bilinearly upsampled and processed with multi-level high-resolution features extracted from E by two Res-Blocks [12]. We use four ICBs in ICM. In the discriminator D_{img}, each convolutional layer is followed by a spectral normalization [29] and leaky ReLU except the last one. AIFR is optimized by SGD with an initial learning rate of 0.1 and momentum of 0.9 while the ICM, decoder D, and D_{img} are trained by Adam with a fixed learning rate of 10^{-4}, β_1 of 0.9 and β_2 of 0.99 for FAS. We trained all models with a batch size of 512 on 8 NVIDIA GTX 2080Ti GPUs, 110K iterations for LCAF and 36K iterations for SCAF. The learning rate of AIFR was warmed up linearly from 0 to 0.1, reduced by a factor of 0.1, at iterations 5K, 70K, and 90K on LCAF and 1K, 20K, 23K on SCAF, respectively. See supplementary material for more details.

4.2. Evaluation on AIFR

Next, we evaluate the MTLFace on several benchmark cross-age datasets, including CACD-VS [3], CALFW [58], AgeDB [30], and FG-NET [1], to compare with the state-of-the-art methods. Note that MORPH is excluded since the version in [43, 47, 56] is prepared for commercial use only.

Result on AgeDB. AgeDB [30] contains 16,488 face images of 568 distinct subjects with manually annotated age labels, which has four age-invariant face verification protocols under the different age gaps of face pairs: 5, 10, 20, and 30 years. Similar to the labeled faces in the wild (LFW) [15], AgeDB is split into 10 folds for each protocol, where each fold consists of 300 intra-class and 300 inter-class pairs. We strictly follow the protocol of 30 years to perform the 10-fold cross-validation since the protocol of 30 years is the most challenging one. We use the models trained on SCAF to evaluate the performance on AgeDB for fair comparisons. Table 1a shows the comparison results in terms of verification accuracy, demonstrating the superior performance of MTLFace over state-of-the-art methods.

Result on CALFW. Cross-age LFW (CALFW) dataset [58] is designed for unconstrained face verification with large age gaps, which contains 12,176 face images of 4,025 individuals collected using the same identities in LFW. Similarly, we follow the same protocol as the LFW, where each fold consists of 600 positive and negative pairs. We train the model on LCAF to evaluate our method on this dataset, and the results are shown in Table 1b. Particularly, our method outperforms the recent state-of-the-art AIFR methods by a large margin, establishing a new state-of-the-art on the CALFW.

Result on CACD-VS. Cross-age celebrity dataset (CACD) contains 163,446 face images of 2,000 celebrities in the wild, with significant variations in age, illumination, pose, and so on. Since collected by search engine, CACD is noisy with mislabeled and duplicate images. Therefore, a carefully annotated version, CACD verification sub-set or CACD-VS [3], is constructed for fair comparisons, which also follows the protocol of LFW. Table 1c presents the comparison of the proposed method with other state-of-the-arts on CACD-VS. Our MTLFace surpasses other state-of-the-arts by a large margin, introducing an improvement of 0.15 against the recent one.

Result on FG-NET. FG-NET [1] is the most popular and challenging age dataset for AIFR, which consists of 1,002 face images from 82 subjects collected from the wild with huge age variations ranging from child to elder. We strictly follow the evaluation pipeline in [43, 47]. Specifically, the model is trained on SCAF and tested under the protocols of leave-one-out and MegaFace challenge 1 (MF1). In the leave-one-out protocol, faces are used to match the rest faces, repeating 1,002 times. Table 1d reports the rank-1 recognition rate. Our method outperforms prior work by a large margin. On the other hand, the MF1 contains additional 1M images as the distractors in the gallery set from 690K different individuals, where models are evaluated under the large and small training set protocols. The small protocol requires the training set less than 0.5M images, which is strictly followed to evaluate our trained model on FG-NET, and the experimental results are reported in Table 1e. Our method achieves competitive performance against other methods since the distractors in MF1 contains a large number of mislabeled probe and gallery face images.

Ablation study. To investigate the efficacy of different modules in MTLFace, we perform ablation studies based on four benchmark datasets for AIFR by considering the following variants of our method: 1) Baseline: we re-

Method	Acc (%)
RJIVE [40]	55.20
VGG Face [34]	89.89
Center Loss [51]	93.72
SphereFace [25]	91.70
CosFace [45]	94.56
ArcFace [5]	95.15
DAAE [22]	95.30
MTLFace (**ours**)	**96.23**

(a) **AgeDB-30**

Method	Acc (%)
HUMAN-Individual	82.32
HUMAN-Fusion	86.50
Center Loss [51]	85.48
SphereFace [25]	90.30
VGGFace2 [2]	90.57
ArcFace [5]	95.45
MTLFace (**ours**)	**95.62**

(b) **CALFW**

Method	Acc (%)
HFA [9]	84.40
CARC [3]	87.60
VGGFace [34]	96.00
Center Loss [51]	97.48
LF-CNN [50]	98.50
Marginal Loss [6]	98.95
OE-CNN [47]	99.20
AIM [56]	99.38
DAL [43]	99.40
MTLFace (**ours**)	**99.55**

(c) **CACD-VS**

Method	Rank-1 (%)
Park *et al.* [33]	37.40
Li *et al.* [24]	47.50
HFA [9]	69.00
MEFA [10]	76.20
CAN [53]	86.50
LF-CNN [50]	88.10
AIM [56]	93.20
DAL [43]	94.50
MTLFace (**ours**)	**94.78**

(d) **FG-NET (leave-one-out)**

Method	Rank-1 (%)
FUDAN-CS_SDS [48]	25.56
SphereFace [25]	47.55
TNVP [32]	47.72
OE-CNN [47]	52.67
DAL [43]	**57.92**
MTLFace (**ours**)	57.18

(e) **FG-NET (MF1)**

Model	AgeDB-30	CALFW	CACD-VS	FG-NET
Baseline	95.52	94.27	99.12	93.64
+Age	95.32	94.35	99.15	93.88
+AFD (CA)	95.63	94.50	99.32	94.05
+AFD (SA)	95.85	94.43	99.25	94.38
+AFD (CBAM)	96.08	94.32	99.18	94.36
+AFD	95.90	94.48	99.30	94.58
MTLFace (**ours**)	**96.23**	**94.72**	**99.38**	**94.78**

(f) **Ablation Study**

Method	LFW	MF1-Facescrub
SphereFace [25]	99.42	72.73
CosFace [45]	99.33	77.11
OE-CNN [47]	99.35	N/A
DAL [43]	99.47	**77.58**
MTLFace (**ours**)	**99.52**	77.06

(g) **General Face Recognition**

Table 1: Experimental results on several benchmark AIFR and GFR datasets with the best results in bold. We reported the verification rate (%) for AgeDB, CALFW, CACD-VS, and LFW, and the rank-1 identification rate (%) for FG-NET and MF1.

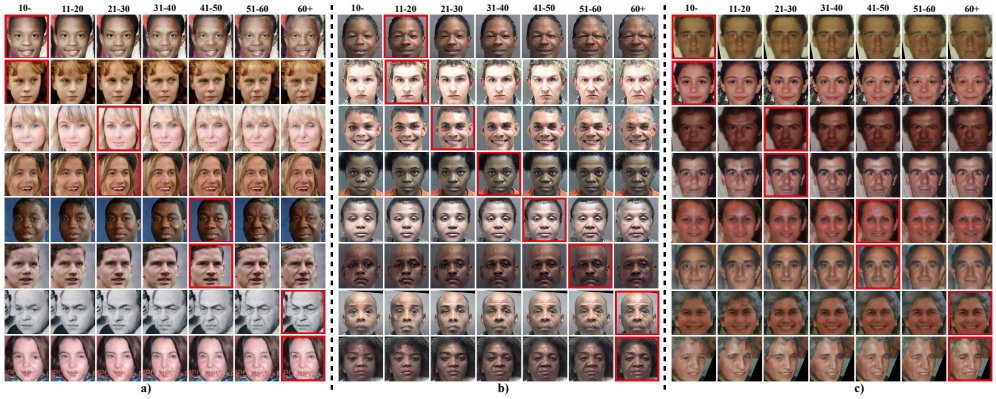

Figure 4: Qualitative results by applying our MTLFace trained on SCAF dataset to three external datasets : a) LCAF excluding identities in SCAF; b) MORPH; and c) FG-NET. Red boxes indicate input faces.

move all extra components but only the CosFace loss to train the face recognition model. 2) +Age: this variant is jointly trained under the supervision of both CosFace and age estimation loss, similar to [43, 57]. 3) +AFD (CA), +AFD (SA), +AFD (CBAM), +AFD: these four variants utilize the proposed attention-based feature decomposition to highlight the age-related information at different levels, by different attention modules including CA [14], SA [52], CBAM [52], and the proposed one. 4) Ours: our proposed MTLFace is trained simultaneously by the AFD and cross-age domain adaption loss. Table 1f presents the experimental results. Note that the verification rate of the baseline model on AgeDB-30 is higher than those of ArcFace and DAAE since our training data is age-balanced, which is an important feature of our collected dataset. Even though the age estimation task is performed in the face recognition model, it cannot introduce any improvement of AIFR compared to the baseline model. On the other hand, AFD achieves remarkable performance improvement on all cross-age datasets. Nevertheless, as the AFD highlights the age-related information at both channel and spatial levels in parallel, our method achieves consistent performance improvements, demonstrating its effectiveness compared to the single level (CA and SA) or sequential level (CBAM). Furthermore, the use of cross-age domain adversarial training leads to an additional performance improvement.

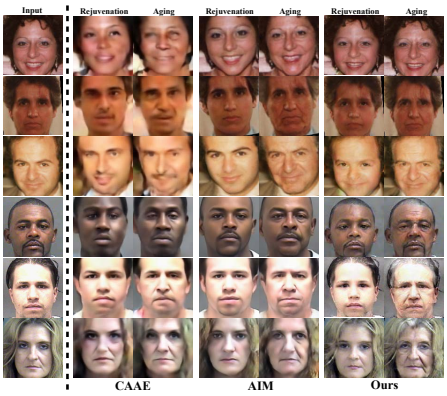

Figure 5: Qualitative comparisons with prior work on FG-NET (top 3 rows) and MORPH (bottom 3 rows).

4.3. Evaluation on GFR

To validate the generalization ability of our MTLFace for GFR, we further conduct experiments on the LFW [15] and MegaFace Challenge 1 Facescrub (MF1-Facescrub) [18] datasets. LFW [15] is the most popular public benchmark dataset for GFR, which contains 13,233 face images from 5,749 subjects. MF1-Facescrub [18] uses the Facescrub dataset [31] of 106,863 face images from 530 celebrities as a probe set. The most challenging problem of MF1 is that it uses an additional 1M face images in the gallery set to distract the face matching. That is, the results on MF1 are not as reliable as LFW due to the extremely noisy distractors in MF1. We strictly follow the same procedure as [43, 47]; *i.e.*, the training dataset contains 0.5M images (SCAF). Table 1g reports the verification rate on LFW and rank-1 identification rate on MF1-Facescrub against the state-of-the-art GFR methods. Our method achieves competitive performance on both datasets, demonstrating the strong generalization ability of our MTLFace. We highlight that our MTLFace can provide photo-realistic synthesized faces to improve model interpretability, which is absent in other methods [43, 47].

4.4. Evaluation on FAS

We further evaluate the model trained on SCAF for FAS.

Qualitative results. Fig. 4 presents some sample results on the external datasets including LCAF, MORPH, and FG-NET. Our method is able to simulate the face age synthesis process between age groups with high visual fidelity. Although there exist variations in terms of race, gender, expression, and occlusion, the synthesized faces are still photo-realistic, with natural details in the skin, muscles, and wrinkles while consistently preserving identities, confirming the generalization ability of the proposed method.

Comparisons with prior work. We also conduct qualitative comparisons with prior work including CAAE [55] and AIM [56] on MORPH and FG-NET. Fig. 5 shows that both CAAE and AIM produce oversmoothed faces due to their image reconstruction while our MTLFace uses the identity age condition to synthesize faces based on multi-level features extracted from the encoder. Note that the results of competitors are directly referred from their own papers for a fair comparison, which is widely adopted in the FAS literature such as [13, 22, 23, 26, 54] to avoid any bias or error caused by self-implementation.

Method	MORPH	FG-NET	CACD
CAAE [55]	45.62/0.256	41.85/0.228	45.06/0.204
IPCGAN [49]	39.95/0.682	43.34/0.581	50.85/0.589
MTLFace	**57.40/0.745**	**61.47/0.638**	**60.62/0.676**
w/o ICM	50.80/0.729	55.26/0.600	55.79/0.652

Table 2: Quantitative comparisons between our MTLFace and the state-of-the-art face aging/rejuvenation methods in the form of a/b, where a and b represent the mean values of age accuracy (%) and identity preservation (cosine similarity) computed over all age mappings, respectively.

Quantitative comparisons. We trained all models on the SCAF dataset for fair comparisons and then directly applied them to three external cross-age datasets: MORPH[37], FG-NET [1] and CACD [3]. Table 2 presents the quantitative results of different face aging/rejuvenation methods, including CAAE [55], IPCGAN [49], our proposed MTLFace and its variant (w/o ICM), in terms of age accuracy and identity preservation. MTLFace outperforms CAAE and IPCGAN by a clear margin; this is a direct results of AIFR and ICM. Without ICM, MTLFace reduces to a common cGANs-based method that uses one-hot encoding to control face aging/rejuvenation at the group level. Remarkably, the MTLFace without ICM still outperforms these two baseline methods, implying that our multi-learning framework with attention-based feature decomposition is effective in improving the age accuracy and identity preservation.

5. Conclusion

In this paper, we proposed a multi-task learning framework, termed MTLFace, to achieve AIFR and FAS simultaneously. We proposed two novel modules: AFD to decompose the features into age- and identity-related features, and ICM to achieve identity-level FAS. Extensive experiments on both cross-age and general benchmark datasets for face recognition demonstrate the superiority of our MTLFace.

Acknowledgement This work was supported in part by National Key Research and Development Program of China (No. 2018YFB1305104), the Shanghai Municipal Science and Technology Major Project (No. 2018SHZDZX01) and ZJLab, and the Shanghai Center for Brain Science and Brain-inspired Technology.

References

[1] FG-NET aging database. https://yanweifu.github.io/FG_NET_data, 2014. 6, 8

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, 2018. 7

[3] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Trans. Multimedia*, 17(6):804–815, 2015. 6, 7, 8

[4] Debayan Deb, Divyansh Aggarwal, and Anil K Jain. Finding missing children: Aging deep face features. *arXiv preprint arXiv:1911.07538*, 2019. 2

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4690–4699, 2019. 5, 6, 7

[6] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. Marginal loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 60–68, 2017. 7

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 2, 5

[8] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(12):2234–2240, 2007. 2

[9] Dihong Gong, Zhifeng Li, Dahua Lin, Jianzhuang Liu, and Xiaoou Tang. Hidden factor analysis for age invariant face recognition. In *Int. Conf. Comput. Vis.*, pages 2872–2879, 2013. 2, 7

[10] Dihong Gong, Zhifeng Li, Dacheng Tao, Jianzhuang Liu, and Xuelong Li. A maximum entropy feature descriptor for age invariant face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5289–5297, 2015. 7

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 2672–2680, 2014. 2, 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 1, 6

[13] Zhenliang He, Meina Kan, Shiguang Shan, and Xilin Chen. S2GAN: Share aging factors across ages and share aging trends among individuals. In *Int. Conf. Comput. Vis.*, pages 9440–9449, 2019. 5, 6, 8

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7132–7141, 2018. 4, 7

[15] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 6, 8

[16] H. Huang, S. Chen, J. Zhang, and H. Shan. PFA-GAN: Progressive face aging with generative adversarial network. *IEEE Trans. Inf. Forensics Security*, 16:2031–2045, 2021. 3

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1125–1134, 2017. 5

[18] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The MegaFace benchmark: 1 million faces for recognition at scale. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4873–4882, 2016. 8

[19] Ira Kemelmacher-Shlizerman, Supasorn Suwajanakorn, and Steven M Seitz. Illumination-aware age progression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3334–3341, 2014. 3

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[21] Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):442–455, 2002. 2

[22] Peipei Li, Huaibo Huang, Yibo Hu, Xiang Wu, Ran He, and Zhenan Sun. Hierarchical face aging through disentangled latent characteristics. *Eur. Conf. Comput. Vis.*, 2020. 6, 7, 8

[23] Qi Li, Yunfan Liu, and Zhenan Sun. Age progression and regression with spatial attention modules. In *AAAI*, 2020. 2, 4, 5, 6, 8

[24] Zhifeng Li, Unsang Park, and Anil K Jain. A discriminative model for age invariant face recognition. *IEEE Trans. Inf. Forensics Security*, 6(3):1028–1037, 2011. 7

[25] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 212–220, 2017. 2, 7

[26] Yunfan Liu, Qi Li, and Zhenan Sun. Attribute-aware face aging with wavelet-based generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11877–11886, 2019. 2, 4, 5, 6, 8

[27] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Int. Conf. Comput. Vis.*, pages 2794–2802, 2017. 5

[28] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[29] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *Proc. Int. Conf. Learn Represent.*, pages 1–26, 2018. 6

[30] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. AgeDB: the first manually collected, in-the-wild age database. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 51–59, 2017. 6

[31] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 343–347. IEEE, 2014. 8

[32] Chi Nhan Duong, Kha Gia Quach, Khoa Luu, Ngan Le, and Marios Savvides. Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition. In *Int. Conf. Comput. Vis.*, pages 3735–3743, 2017. 3, 7

[33] Unsang Park, Yiying Tong, and Anil K Jain. Age-invariant face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(5):947–954, 2010. 2, 7

[34] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proc. British Mach. Vis. Conf.*, pages 1–12, 2015. 7

[35] Narayanan Ramanathan and Rama Chellappa. Modeling age progression in young faces. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 1, pages 387–394, 2006. 3

[36] Narayanan Ramanathan and Rama Chellappa. Modeling shape and textural variations in aging faces. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–8, 2008. 3

[37] K. Ricanek and T. Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *Proc. Int. Conf. Autom. Face Gesture Recognit.*, pages 341–345, 2006. 8

[38] Rasmus Rothe, Radu Timofte, and Luc Van Gool. DEX: Deep expectation of apparent age from a single image. In *Int. Conf. Comput. Vis. Worksh.*, pages 10–15, 2015. 5

[39] Duncan A Rowland and David I Perrett. Manipulating facial appearance through shape and color. *IEEE Computer Graphics and Applications*, 15(5):70–76, 1995. 3

[40] Christos Sagonas, Evangelos Ververas, Yannis Panagakis, and Stefanos Zafeiriou. Recovering joint and individual components in facial data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(11):2668–2681, 2017. 7

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learn. Represent.*, pages 1–14, 2015. 1

[42] Jinli Suo, Xilin Chen, Shiguang Shan, Wen Gao, and Qionghai Dai. A concatenational graph evolution aging model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2083–2096, 2012. 3

[43] Hao Wang, Dihong Gong, Zhifeng Li, and Wei Liu. Decorrelated adversarial learning for age-invariant face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3527–3536, 2019. 2, 4, 6, 7, 8

[44] Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. In *Int. Conf. Mach. Learn.*, pages 9898–9907. PMLR, 2020. 5

[45] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5265–5274, 2018. 7

[46] Wei Wang, Zhen Cui, Yan Yan, Jiashi Feng, Shuicheng Yan, Xiangbo Shu, and Nicu Sebe. Recurrent face aging. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2378–2386, 2016. 3

[47] 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 *Eur. Conf. Comput. Vis.*, pages 738–753, 2018. 2, 6, 7, 8

[48] Zhanxiong Wang, Keke He, Yanwei Fu, Rui Feng, Yu-Gang Jiang, and Xiangyang Xue. Multi-task deep neural network for joint face recognition and facial attribute prediction. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 365–374, 2017. 7

[49] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. Face aging with identity-preserved conditional generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7939–7947, 2018. 2, 3, 4, 5, 8

[50] Yandong Wen, Zhifeng Li, and Yu Qiao. Latent factor guided convolutional neural networks for age-invariant face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4893–4901, 2016. 2, 7

[51] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Eur. Conf. Comput. Vis.*, pages 499–515. Springer, 2016. 1, 7

[52] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Eur. Conf. Comput. Vis.*, pages 3–19, 2018. 4, 7

[53] Chenfei Xu, Qihe Liu, and Mao Ye. Age invariant face recognition and retrieval by coupled auto-encoder networks. *Neurocomputing*, 222:62–71, 2017. 7

[54] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K Jain. Learning face age progression: A pyramid architecture of gans. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 31–39, 2018. 2, 3, 4, 5, 6, 8

[55] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5810–5818, 2017. 2, 3, 4, 5, 8

[56] 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 *AAAI*, volume 33, pages 9251–9258, 2019. 3, 6, 7, 8

[57] Tianyue Zheng, Weihong Deng, and Jiani Hu. Age estimation guided convolutional neural network for age-invariant face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 1–9, 2017. 2, 7

[58] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017. 6