Towards Age-Invariant Face Recognition

Jian Zhao, *Member, IEEE,* Shuicheng Yan, *Fellow, IEEE*, and Jiashi Feng, *Member, IEEE*

Abstract—Despite the remarkable progress in face recognition related technologies, reliably recognizing faces across ages remains a big challenge. The appearance of a human face changes substantially over time, resulting in significant intra-class variations. As opposed to current techniques for age-invariant face recognition, which either directly extract age-invariant features for recognition, or first synthesize a face that matches target age before feature extraction, we argue that it is more desirable to perform both tasks jointly so that they can leverage each other. To this end, we propose a deep **A**ge-**I**nvariant **M**odel (AIM) for face recognition in the wild with three distinct novelties. First, AIM presents a novel unified deep architecture jointly performing cross-age face synthesis and recognition in a mutual boosting way. Second, AIM achieves continuous face rejuvenation/aging with remarkable photorealistic and identity-preserving properties, avoiding the requirement of paired data and the true age of testing samples. Third, effective and novel training strategies are developed for end-to-end learning of the whole deep architecture, which generates powerful age-invariant face representations explicitly disentangled from the age variation. Moreover, we construct a new large-scale **C**ross-**A**ge **F**ace **R**ecognition (CAFR) benchmark dataset to facilitate existing efforts and push the frontiers of age-invariant face recognition research. Extensive experiments on both our CAFR dataset and several other cross-age datasets (MORPH, CACD, and FG-NET) demonstrate the superiority of the proposed AIM model over the state-of-the-arts. Benchmarking our model on the popular unconstrained face recognition dataset IJB-C additionally verifies its promising generalization ability in recognizing faces in the wild.

Index Terms—Age-Invariant Face Recognition, Age-Invariant Model, Generative Adversarial Networks, Benchmark Dataset.

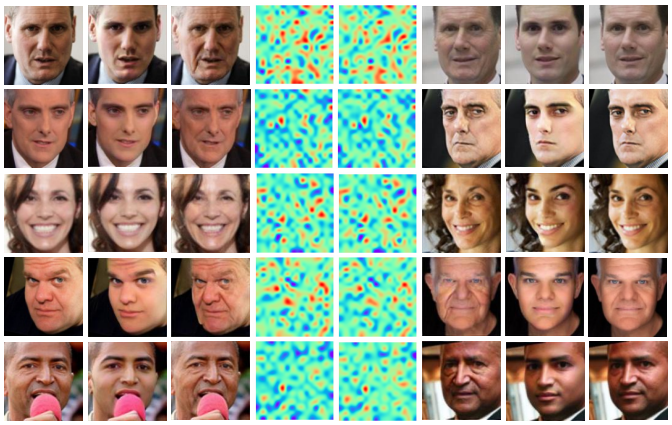


Fig. 1: Joint disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. *Col.* 1 & 8: Input faces of distinct identities with various challenging factors (*e.g.*, neutral, illumination, expression, pose and occlusion). *Col.* 2 & 7: Synthesized younger faces by our proposed AIM. *Col.* 3 & 6: Synthesized older faces by our proposed AIM. *Col.* 4 & 5: Learned age-invariant facial representations by our proposed AIM. Based on such representations, AIM then apply the face synthesis component onto the representation, which takes targeted ages, identity discriminative information (*e.g.*, expression) as input, and generate faces of various ages/expression. Best viewed in color.

1 Introduction

FACE recognition is one of the most widely studied topics in computer vision and artificial intelligence fields. Recently, some approaches claim to have achieved [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15] or even surpassed [16], [17], [18], [19], [20], [21] human performance on several benchmarks.

Despite the exciting progress, age variations still form a major bottleneck for many practical applications. For example, in law enforcement scenarios, finding missing children after years, identifying wanted fugitives based on mug shots and verifying passports usually involve recognizing faces across ages and/or synthesizing photorealistic age regressed/progressed[1] face images. These are extremely challenging due to several reasons: 1) Human face rejuvenation/aging is a complex process whose patterns differ from one individual to another. Both intrinsic factors (like heredity, gender and ethnicity) and extrinsic factors (like environment and living styles) affect the aging process and lead to significant intra-class variations. 2) Facial shapes and textures dramatically change over time, making learning age-invariant patterns difficult. 3) Current learning based cross-age face recognition models are limited by existing cross-age databases [22], [23], [24], [25], [26], [27] due to their small size, narrow elapse per subject and unbalance in gender, ethnicity, and age span. As such, the performance of most face recognition models degrades by over 13% from general recognition on faces of (almost) the same age to cross-age face recognition [24]. In this work, we aim to improve automatic models for recognizing unconstrained faces with large age variations.

According to recent studies [28], [29], [30], [31], [32], face images of different individuals usually share common aging characteristics (*e.g.*, wrinkles), and face images of the same individual contain intrinsic features that are relatively stable across ages. Facial representations of a person in the latent space can hence

- *Jian Zhao is the corresponding author. Homepage: https://zhaoj9014.github.io/.*
- *Jian Zhao is with Institute of North Electronic Equipment, Beijing, China. Shuicheng Yan is with Yitu Technology, Beijing, China. Jiashi Feng is with National University of Singapore, Singapore. E-mails: zhaojian90@u.nus.edu, {eleyans, elefjia}@nus.edu.sg.*

1. Face regression (*a.k.a* face rejuvenation) and face progression (*a.k.a* face aging) refers to rendering the natural rejuvenation/aging effect for a given face, respectively.

be decomposed into an age-specific component which reflects the aging effect and an identity-specific component which preserves intrinsic identity information. The latter would be invariant to age variations and ideal for cross-age face recognition when achievable. This finding inspires us to develop a novel and unified deep neural network, termed as **A**ge **I**nvariant **M**odel (AIM). The AIM jointly learns disentangled identity representations that are invariant to age, and photorealistic cross-age face image synthesis that can highlight important latent representations among the disentangled ones end-to-end. Therefore, they mutually boost each other to achieve age-invariant face recognition. AIM takes as input face images of arbitrary ages with other potential distracting factors like various illumination, expressions, poses, and occlusion. It outputs facial representations invariant to age variations and meanwhile preserves discriminativeness across different identities. As shown in Fig. 1, the AIM can learn age-invariant representations and effectively synthesize natural age regressed/progressed faces. We present the results given various inputs along with different challenging factors to show that AIM has automatically learned to achieve the robustness to skin color, illumination, expression, pose, and occlusion besides the age variation.

In particular, AIM extends from an auto-encoder based **G**enerative **A**dversarial **N**etwork (GAN) and includes a disentangled **R**epresentation **L**earning sub-**N**et (RLN) and a **F**ace **S**ynthesis sub-**N**et (FSN) for age-invariant face recognition. RLN consists of an encoder and a discriminator that compete with each other to learn discriminative and age-invariant representations. It introduces cross-age domain adversarial training to promote encoded features that are indistinguishable w.r.t. the shift between multi-age domains, and cross-entropy regularization with a label smoothing strategy to constrain cross-age representations with ambiguous separability. The discriminator incorporates dual agents to encourage the representations to be uniformly distributed to smooth the age transformation while preserving identity information. The representations are then concatenated with a continuous age condition code to synthesize age regressed/progressed face images, such that the learned representations are explicitly disentangled from age variations. FSN consists of a decoder and a local-patch based discriminator that compete with each other to synthesize photorealistic cross-age face images. FSN uses an attention mechanism to guarantee robustness to large background complexity and illumination variance. The discriminator incorporates dual agents to add realism to synthesized cross-age faces while forcing the generated faces to exhibit desirable rejuvenation/aging effects.

Moreover, we propose a new large-scale **C**ross-**A**ge **F**ace **R**ecognition (CAFR) benchmark dataset to facilitate existing efforts and future research on age-invariant face recognition. CAFR contains 1,446,500 face images from 25,000 subjects annotated with age, identity, gender, race and landmark labels. Extensive experiments on both our CAFR and other standard cross-age datasets (MORPH [25], CACD [24], and FG-NET [22]) demonstrate the superiority of AIM over the state-of-the-arts. Benchmarking AIM on the popular unconstrained face recognition dataset IJB-C [33] additionally verifies its promising generalization ability in recognizing faces in the wild.

The main contributions of this work are summarized as follows in four-fold

- We propose a novel deep architecture unifying cross-age face synthesis and recognition in a mutual boosting way.
- We develop effective end-to-end training strategies for the whole deep architecture to generate powerful age-invariant facial representations explicitly disentangled from the age variations.
- The proposed model achieves continuous face rejuvenation/aging with remarkable photorealistic and identity-preserving properties, avoiding the requirement of paired data and true age of testing samples.
- We propose a new large-scale benchmark dataset CAFR to advance the frontiers of age-invariant face recognition research.

2 Related Work

2.1 Age-Invariant Representation Learning

Conventional approaches often leverage robust local descriptors [28], [34], [35], [36], [37], [38] and metric learning [39], [40], [41] to tackle age variance. For instance, the Bayesian age difference model [34] classifies face images of individuals based on age differences and performs face verification across age progression. The **H**idden **F**actor **A**nalysis (HFA) model [28] separates aging variations from identity-specific features. The **G**radient **O**rientation **P**yramid (GOP) model [40] develops discriminative methods for cross-age face verification. In contrast, deep learning models often handle age variance through using a single age-agnostic or several age-specific models with pooling and specific loss functions [29], [31], [42], [43], [44], [45]. For instance, the enforced softmax optimization strategy [46] is developed to learn effective and compact deep facial representations with reduced intra-class variance and enlarged inter-class distance. The **L**atent **F**actor guided **C**onvolutional **N**eural **N**etwork (LF-CNN) model [29] learns age-invariant deep features through a carefully designed CNN model. The **A**ge **E**stimation guided **CNN** (AE-CNN) model [42] learns to separate aging variations from identity-specific features. The **O**rthogonal **E**mbedding **CNN** (OE-CNN) model [45] decomposes deep facial representations into two orthogonal components to represent age- and identity-specific features.

2.2 Cross-Age Face Synthesis

Previous methods can be roughly divided into physical modeling based and prototype based. The former approaches model the biological patterns and physical mechanisms of aging, including muscles [47], wrinkles [48], and facial structure [49]. However, they usually require massive annotated cross-age face data with long elapse per subject which are hard to collect, and they are computationally expensive. Prototype-based approaches [50], [51] often divide faces into groups by ages and select the average face of each group as the prototype. The differences in prototypes between two age groups are then considered as the aging pattern. However, the aged face generated from the averaged prototype may lose personality information. Most of subsequent approaches [52], [53] are data-driven and do not rely much on the biological prior knowledge, and the aging patterns are learned from training data. Though improve the results, these methods suffer ghosting artifacts on the synthesized faces. More recently, deep generative networks are exploited. For instance, the recurrent face aging model [54] proposes a smooth face aging process between neighboring groups by modeling the intermediate transition states with **R**ecurrent **N**eural **N**etwork (RNN). The **C**ontextual **GANs**

(C-GANs) method [55] explicitly models the transition patterns between adjacent age groups during the training procedure for face aging. The Conditional Adversarial Autoencoder model [27] proposes a **C**onditional **A**dversarial **A**uto-**E**ncoder (CAAE) and achieve face age regression/progression in a holistic framework. The age progression model [56] proposes a pyramid architecture of GANs to ensure that the generated faces present desired aging effects while simultaneously keeping personalized properties stable.

Our model differs from them in following aspects: 1) AIM jointly performs cross-age face synthesis and recognition end-to-end to allow them to mutually boost each other for addressing large age variance in unconstrained face recognition[2]. 2) AIM achieves continuous face rejuvenation/aging (bidirectional) with remarkable photorealistic and identity-preserving properties, and do not require paired data and true age of testing samples. 3) AIM generates powerful age-invariant face representations explicitly disentangled from age variations through cross-age domain adversarial training and cross-entropy regularization with a label smoothing strategy.

3 AGE-INVARIANT MODEL

As shown in Fig. 2, the proposed **A**ge-**I**nvariant **M**odel (AIM) extends from an auto-encoder based GAN, and consists of a disentangled **R**epresentation **L**earning sub-**N**et (RLN) and a **F**ace **S**ynthesis sub-**N**et (FSN) that jointly learn discriminative and robust facial representations disentangled from age variance and perform attention-based face rejuvenation/aging end-to-end. We now detail each component.

3.1 Disentangled Representation Learning

Matching face images across ages is demanded in many real-world applications. It is mainly challenged by variations of an individual at different ages (*i.e.* large intra-class variations) or caused by aging (*e.g.* facial shape and texture changes), and inevitable entanglement of unrelated (statistically independent) components in the deep features extracted from a general-purpose face recognition model. Large intra-class variations usually result in erroneous cross-age face recognition and entangled facial representations potentially weaken the model's robustness in recognizing faces with age variations. We propose a GAN-like **R**epresentation **L**earning sub-**N**et (RLN) to learn discriminative and robust identity-specific facial representations disentangled from age variance, as illustrated in Fig. 2.

In particular, RLN takes the encoder G_{θ_E} (with learnable parameters θ_E) as the generator : $\mathbb{R}^{H\times W\times C} \mapsto \mathbb{R}^{C'}$ for facial representation learning, where H, W, C and C' denote the input image height, width, channel number and the dimensionality of the encoded feature f, respectively. f preserves the high-level identity-specific information of the input face image through several carefully designed regularizations. We further concatenate f with a continuous age condition code to synthesize age regressed/progressed face images, such that the learned representations are explicitly disentangled from age variations.

2. AIM aims to learn disentangled age-invariant facial representations for age-invariant face recognition. As a by-product, the age regressed/progressed face images by AIM can also be utilized by conventional descriptors and learning algorithms to eliminate the negative effects from unconstrained conditions with large age variance.

Formally, denote the input RGB face image as x and the learned facial representation as f. Then

$$f := \mathrm{G}_{\theta_E}(x). \tag{1}$$

The key requirements for G_{θ_E} include three aspects. 1) The learned representation f should be invariant to age variations and also well preserve the identity-specific component. 2) It should be barely possible for an algorithm to identify the domain of origin of the observation x regardless of the underlying gap between multi-age domains. 3) f should obey uniform distribution to smooth the age transformation.

To this end, we propose to learn θ_E by minimizing the following composite losses:

$$\begin{aligned}\mathcal{L}_{\mathrm{G}_{\theta_E}} = &-\lambda_1\mathcal{L}_{\mathrm{cad}} + \lambda_2\mathcal{L}_{\mathrm{cer}} - \lambda_3\mathcal{L}_{\mathrm{adv}_1} + \lambda_4\mathcal{L}_{\mathrm{ip}} \\ &- \lambda_5\mathcal{L}_{\mathrm{adv}_2} + \lambda_6\mathcal{L}_{\mathrm{ae}} + \lambda_7\mathcal{L}_{\mathrm{mc}} + \lambda_8\mathcal{L}_{\mathrm{tv}} + \lambda_9\mathcal{L}_{\mathrm{att}},\end{aligned} \tag{2}$$

where $\mathcal{L}_{\mathrm{cad}}$ is the **c**ross-**a**ge **d**omain adversarial loss for facilitating age-invariant representation learning via domain adaption, $\mathcal{L}_{\mathrm{cer}}$ is the **c**ross-**e**ntropy **r**egularization loss for constraining cross-age representations with ambiguous separability, $\mathcal{L}_{\mathrm{adv}_1}$ is the **adv**ersarial loss for imposing the uniform distribution on f, $\mathcal{L}_{\mathrm{ip}}$ is the **i**dentity **p**reserving loss for preserving identity information, $\mathcal{L}_{\mathrm{adv}_2}$ is the **adv**ersarial loss for adding realism to the synthesized images and alleviating artifacts, $\mathcal{L}_{\mathrm{ae}}$ is the **a**ge **e**stimation loss for forcing the synthesized faces to exhibit desirable rejuvenation/aging effect, $\mathcal{L}_{\mathrm{mc}}$ is the **m**anifold **c**onsistency loss for encouraging input-output space manifold consistency, $\mathcal{L}_{\mathrm{tv}}$ is the **t**otal **v**ariation loss for reducing spiky artifacts, $\mathcal{L}_{\mathrm{att}}$ is the **att**ention loss for facilitating robustness enhancement via an attention mechanism, and $\{\lambda_k\}_{k=1}^{9}$ are weighting parameters among different losses.

In order to enhance the age-invariant representation learning capacity, we adopt $\mathcal{L}_{\mathrm{cad}}$ to promote emergence of features encoded by G_{θ_E} that are indistinguishable w.r.t. the shift between multi-age domains, which is defined as

$$\mathcal{L}_{\mathrm{cad}} = \frac{1}{N}\sum_i \Big\{-y_i \log[\mathrm{C}_\varphi(f_i)] - (1-y_i)\log[1-\mathrm{C}_\varphi(f_i)]\Big\}, \tag{3}$$

where φ denotes the learnable parameters for the domain classifier, and $y_i \in \{0, 1, \dots\}$ is an age domain indicator, indicating which domain f_i is from. Minimizing $\mathcal{L}_{\mathrm{cad}}$ can reduce the domain discrepancy and help the generator achieve similar facial representations across different age domains, even if training samples from a domain are limited, such that the learned facial representations would be age-invariant. Such adapted representations are provided by augmenting the encoder of G_{θ_E} with a few standard layers as the domain classifier C_φ, and a new gradient reversal layer to reverse the gradient during optimizing the encoder (*i.e.*, gradient reverse operator as in Fig. 2), as inspired by [57].

If using $\mathcal{L}_{\mathrm{cad}}$ alone, the results tend to be sub-optimal, because searching for a local minimum of $\mathcal{L}_{\mathrm{cad}}$ may go through a path that resides outside the manifold of desired cross-age representations with ambiguous separability. Thus, we combine $\mathcal{L}_{\mathrm{cad}}$ with an auxiliary $\mathcal{L}_{\mathrm{cer}}$ as an assistant for constraining cross-age representations with ambiguous separability to ensure the search resides in that manifold and produces age-invariant facial representations, where $\mathcal{L}_{\mathrm{cer}}$ is defined as

$$\mathcal{L}_{\mathrm{cer}} = \frac{1}{N}\sum_i \Big\{-\bar{y}_i \log[\mathrm{R}_\psi(f_i)] - (1-\bar{y}_i)\log[1-\mathrm{R}_\psi(f_i)]\Big\}, \tag{4}$$

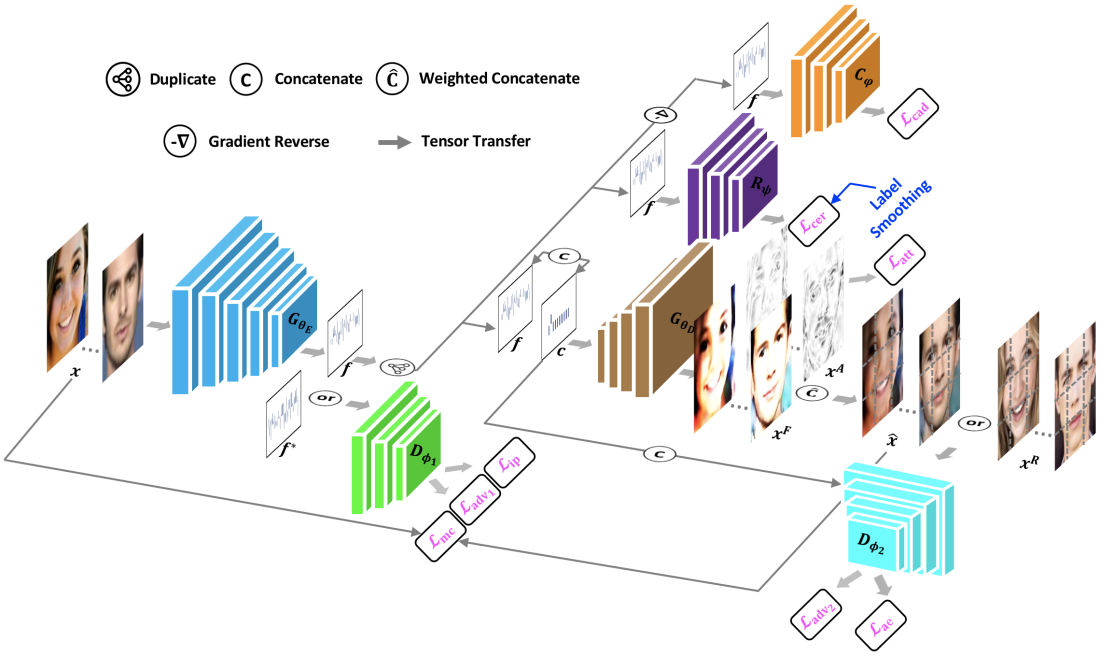


Fig. 2: **A**ge-**I**nvariant **M**odel (AIM) for face recognition in the wild. AIM extends from an auto-encoder based GAN and includes a disentangled **R**epresentation **L**earning sub-**N**et (RLN) and a **F**ace **S**ynthesis sub-**N**et (FSN) that jointly learn end-to-end. RLN consists of an encoder (G_{θ_E}) and a discriminator (D_{ϕ_1}) that compete with each other to learn discriminative and robust facial representations (f) disentangled from age variance. It is augmented by cross-age domain adversarial training ($\mathcal{L}_{\text{cad}}$) and cross-entropy regularization with a label smoothing strategy ($\mathcal{L}_{\text{cer}}$). FSN consists of a decoder (G_{θ_D}) and a local-patch based discriminator (D_{ϕ_2}) that compete with each other to achieve continuous face rejuvenation/aging ($\hat{x}$) with remarkable photorealistic and identity-preserving properties. It introduces an attention mechanism to guarantee robustness to large background complexity and illumination variance. Note AIM does not require paired training data nor true age of testing samples. Best viewed in color.

where ψ denotes the learnable parameters for the regularizer R, and $\bar{y}_i \in \{\frac{1}{n}, \frac{1}{n}, \dots\}$ denotes the smoothed domain indicator.

$\mathcal{L}_{\text{adv}_1}$ is introduced to impose a prior distribution (*e.g.*, uniform distribution) on f to evenly populate the latent space with no apparent "holes", such that smooth age transformation can be achieved:

$$\mathcal{L}_{\text{adv}_1} = \frac{1}{N}\sum_i \Big\{ -y_i \log[\mathrm{D}_{\phi_1}(f_i)] - (1-y_i)\log[1-\mathrm{D}_{\phi_1}(f_i^*)]\Big\}, \tag{5}$$

where ϕ_1 denotes the learnable parameters for the discriminator, $f_i^* \sim U(f)$ denotes a random sample from uniform distribution $U(f)$, and y_i denotes the binary distribution indicator.

To facilitate this process, we leverage a **M**ulti-**L**ayer **P**erceptron (MLP) as the discriminator D_{ϕ_1}, which is very simple to avoid typical GAN tricks. We further augment D_{ϕ_1} with an auxiliary agent $\mathcal{L}_{\text{ip}}$ to preserve identity information:

$$\mathcal{L}_{\text{ip}} = \frac{1}{N}\sum_i \Big\{ -y_i \log[\mathrm{D}_{\phi_1}(f_i)] - (1-y_i)\log[1-\mathrm{D}_{\phi_1}(f_i)]\Big\}, \tag{6}$$

where y_i denotes the identity ground truth.

3.2 Attention-based Face Rejuvenation/Aging

Photorealistic cross-age face images are important for face recognition with large age variance. A natural scheme is to generate reference age regressed/progressed faces from face images of arbitrary ages to match target age before feature extraction or serve as augmented data for learning discriminative models. We then propose a GAN-like **F**ace **S**ynthesis sub-**N**et (FSN) to learn a synthesis function that can achieve both face rejuvenation and aging in a holistic, end-to-end manner, as illustrated in Fig. 2.

In particular, FSN leverages the decoder G_{θ_D} (with learnable parameters θ_D) as the generator: $\mathbb{R}^{C'+C''} \mapsto \mathbb{R}^{H\times W\times C}$ for cross-age face synthesis, where C'' denotes the dimensionality of the age condition code concatenated with f. The synthesized results present natural effects of rejuvenation/aging with robustness to large background complexity and bad lighting conditions through the carefully designed learning schema.

Formally, denote the age condition code as c and the synthesized face image as $\hat{x}$. Then

$$\hat{x} := \mathrm{G}_{\theta_D}(f, c). \tag{7}$$

The key requirements for G_{θ_D} include two aspects. 1) The synthesized face image $\hat{x}$ should visually resemble a real one and preserve the desired rejuvenation/aging effect. 2) Attention should be paid to the most salient regions of the image that are responsible for synthesizing the novel aging phase while keeping the rest elements such as glasses, hats, jewelery and background untouched. To this end, we propose to learn θ_D by minimizing the following composite losses:

$$\mathcal{L}_{\mathrm{G}_{\theta_D}} = -\lambda_{10}\mathcal{L}_{\text{adv}_2} + \lambda_{11}\mathcal{L}_{\text{ae}} + \lambda_{12}\mathcal{L}_{\text{mc}} + \lambda_{13}\mathcal{L}_{\text{tv}} + \lambda_{14}\mathcal{L}_{\text{att}}, \tag{8}$$

where $\{\lambda_k\}_{k=10}^{14}$ are weighting parameters among different losses. $\mathcal{L}_{\text{adv}_2}$ is introduced to push the synthesized image to reside in the manifold of photorealistic age regressed/progressed face images, prevent blur effect, and produce visually pleasing results:

$$\mathcal{L}_{\text{adv}_2} = \frac{1}{N}\sum_i \Big\{ -y_i \log[\mathrm{D}_{\phi_2}(\hat{x}_i, c_{i,j})] - (1-y_i)\log[1-\mathrm{D}_{\phi_2}(x_i^R, c_{i,j})]\Big\}, \tag{9}$$

where ϕ_2 denotes the learnable parameters for the discriminator, $c_{i,j}$ denotes the age condition code to transform f_i into the j^{th}

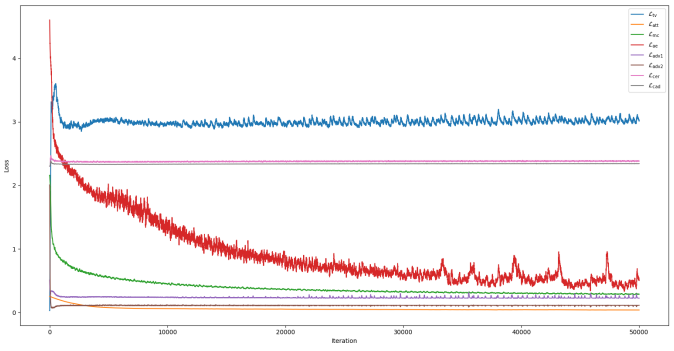


Fig. 3: The convergence curves of different loss terms in AIM during the training phase. Best viewed in color.

age phase, and x_i^R denotes a real face image with (almost) the same age with $\hat{x}_i$ (not necessarily belong to the same person).

To facilitate this process, we modify a CNN backbone as a local-patch based discriminator D_{ϕ_2} to prevent G_{θ_D} from over-emphasizing certain image features to fool the current discriminator network. We further augment D_{ϕ_2} with an auxiliary agent $\mathcal{L}_{\mathrm{ae}}$ to preserve the desired rejuvenation/aging effect. In this way, G_{θ_D} not only learns to render photorealistic samples but also learns to satisfy the target age encoded by c:

$$\mathcal{L}_{\mathrm{ae}} = \frac{1}{N}\sum_i \Big\{\|\hat{c}_{i,j} - c_{i,j}\|_2^2 + \|c_{i,j}^R - c_{i,j}\|_2^2\Big\}, \tag{10}$$

where $\hat{c}_{i,j}$ and $c_{i,j}^R$ denote the estimated ages from $\hat{x}_i$ and x_i^R, respectively.

$\mathcal{L}_{\mathrm{mc}}$ is introduced to enforce the manifold consistency between the input-output space, defined as $\|\hat{x} - x\|_2^2/|x|$, where $|x|$ is the size of x. $\mathcal{L}_{\mathrm{TV}}$ is introduced as a regularization term on the synthesized results to reduce spiky artifacts:

$$\mathcal{L}_{\mathrm{TV}} = \sum_{i,j}^{H,W} \Big\{\|\hat{x}_{i,j+1} - \hat{x}_{i,j}\|_2^2 + \|\hat{x}_{i+1,j} - \hat{x}_{i,j}\|_2^2\Big\}. \tag{11}$$

In order to make the model focus on the most relevant features, we adopt $\mathcal{L}_{\mathrm{att}}$ to facilitate robustness enhancement via an attention mechanism:

$$\mathcal{L}_{\mathrm{att}} = \sum_{i,j}^{H,W} \Big\{\|x_{i,j+1}^A - x_{i,j}^A\|_2^2 + \|x_{i+1,j}^A - x_{i,j}^A\|_2^2 + \|x_{i,j}^A\|_2^2\Big\}, \tag{12}$$

where x^A denotes the attention score map which serves as the guidance, and attends to the most relevant regions during cross-age face synthesis.

The final synthesized results can be obtained by

$$\hat{x} = x^A \cdot x^F + (1 - x^A) \cdot x, \tag{13}$$

where x^F denotes the feature map predicted by the last fractionally-strided convolution block.

3.3 Training and Inference

The goal of AIM is to use sets of real targets to learn two GAN-like sub-nets that mutually boost each other and jointly accomplish age-invariant face recognition. Each separate loss serves as a deep supervision within the hinged structure benefiting network convergence. The overall objective function for AIM is

$$\begin{aligned}\mathcal{L}_{\mathrm{AIM}} = &-\lambda_1\mathcal{L}_{\mathrm{cad}} + \lambda_2\mathcal{L}_{\mathrm{cer}} - \lambda_3\mathcal{L}_{\mathrm{adv}_1} + \lambda_4\mathcal{L}_{\mathrm{ip}} \\ &- \lambda_5\mathcal{L}_{\mathrm{adv}_2} + \lambda_6\mathcal{L}_{\mathrm{ae}} + \lambda_7\mathcal{L}_{\mathrm{mc}} + \lambda_8\mathcal{L}_{\mathrm{tv}} + \lambda_9\mathcal{L}_{\mathrm{att}}.\end{aligned} \tag{14}$$

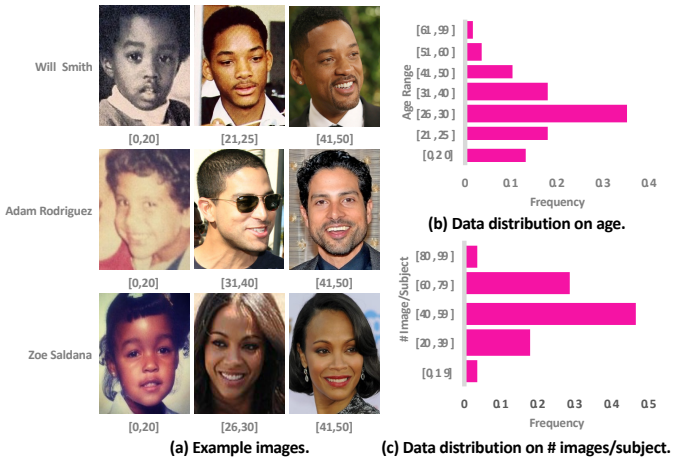


Fig. 4: **C**ross-**A**ge **F**ace **R**ecognition (CAFR) dataset. Best viewed in color.

The convergence curves of different loss terms in AIM during the training phase is visualized in Fig. 3. During testing, we simply feed the input face image x and desired age condition code c into AIM to obtain the disentangled age-invariant representation f from G_{θ_E} and the synthesized age regressed/progressed face image $\hat{x}$ from G_{θ_D}. Example results are visualized in Fig. 1.

4 Cross-Age Face Recognition Benchmark

In this section, we introduce a new large-scale "**C**ross-**A**ge **F**ace **R**ecognition (CAFR)" benchmark dataset to push the frontiers of age-invariant face recognition research with several appealing properties. 1) It contains 1,446,500 face images from 25,000 subjects annotated with age, identity, gender, race, and landmark labels, which is larger and more comprehensive than previous similar attempts [22], [24], [25], [26], [45], [58]. 2) The images within CAFR are collected from real-world scenarios, involving humans with various expressions, poses, occlusion, and resolution. 3) The background of images in CAFR is more complex and diverse than previous datasets. Some examples and statistics w.r.t. data distribution on the image number per age phase and the image number per subject are illustrated in Fig. 4 (a), (b) and (c), respectively.

4.1 Image Collection and Annotation

We select a sub-set from the celebrity name list of MS-Celeb-1M [59] for data collection based on below considerations. 1) Each individual must have many cross-age face images available on the Internet for retrieval. 2) Both gender balance and racial diversity should be considered. Accordingly, we manually specify some keywords (such as name, face image, event, year, *etc.*) to ensure the accuracy and diversity of returned results. Based on these specifications, corresponding cross-age face images are located by performing Internet searches over Google and Bing image search engines. For each identified image, the corresponding URL is stored in a spreadsheet. Automated scrapping software is used to download the cross-age imagery and stores all relevant information (*e.g.*, identity) in a database. Moreover, a pool of self-collected children face images with age variations is also constructed to augment and complement Internet scraping results.

After curating the imagery, semi-automatic[3] annotation is conducted with three steps. 1) Data cleaning. We perform face

3. The size of data is beyond the scale of manually labeling.

TABLE 1: Statistics for publicly available cross-age face datasets.

Dataset	# Images	# Subjects	# Images/Subject	Age Span	Average Age
FG-NET [22]	1,002	82	*avg.* 12.22	1-69	15.84
MORPH Album1 [25]	1,690	515	*avg.* 3.28	15-68	27.28
MORPH Album2 [25]	78,207	20,569	*avg.* 3.80	16-99	32.69
CACD [24]	163,446	2,000	*avg.* 81.72	16-62	38.03
IMDB-WIKI [58]	523,051	20,284	*avg.* 25.79	1-100	38.00
AgeDB [26]	16,488	568	*avg.* 29.03	1-101	50.30
CAF [45]	313,986	4,668	*avg.* 67.26	1-80	29.00
CAFR	**1,446,500**	**25,000**	***avg.* 57.86**	**1-99**	**28.23**

detection with an off-the-shelf algorithm [60] to filter the images without any faces and manually wipe off duplicated images and false positive images (*i.e.*, faces that do not belong to that subject). 2) Data annotation. We combine the prior information on identity and apply off-the-shelf age estimator [58] and landmark localization algorithm [61] to annotate the ground truths on age, identity, gender, race and landmarks. To reduce the inaccurate annotations on age, we divide the whole age span into 7 age phases: $\leq$20, 20-25, 25-30, 30-40, 40-50, 50-60, $\geq$60, instead of specifying the specific age for each face image. 3) Manual inspection. After annotation, manual inspection is performed on all images and corresponding annotations to verify the correctness. In cases where annotations are erroneous, the information is manually rectified by 7 well-informed analysts. The whole work took around 2.5 months to accomplish by 10 professional data annotators. Even with the above carefully designed annotation procedures, the proposed CAFR may still inevitably contain certain noise and distracting factors. However, we observe that the state-of-the-art deep neural network learning algorithm can tolerate a certain level of noise in the training data and learning with noisy data is good and a real problem that is worth of dedicated research efforts.

4.2 Dataset Splits and Statistics

In total, there are 1,446,500 face images from 25,000 subjects in the CAFR dataset, with the age span from 1 to 99. Each subject has 57.86 face images on average, and the average age is 28.23. The statistical comparisons between our CAFR and existing cross-age datasets are summarized in Tab. 1. CAFR is the largest and most comprehensive benchmark dataset for age-invariant face recognition to date. Following random selection, we divide the data into 10 splits with a pair-wise disjoint of subjects in each split. Each split contains 2,500 subjects and we randomly generate 5 genuine and 5 imposter pairs for each subject with various age gaps, resulting in 25,000 pairs per split. The remained data are preserved for algorithm development and parameter selection. We suggest evaluation systems to report the average **Acc**uracy (Acc), **E**qual **E**rror **R**ate (EER), **A**rea **U**nder the **C**urve (AUC) and **R**eceiver **O**perating **C**haracteristic (ROC) curve as 10-fold cross validation.

5 Experiments

We evaluate AIM qualitatively and quantitatively under various settings for face recognition in the wild. In particular, we evaluate age-invariant face recognition performance on the CAFR dataset proposed in this work, as well as the MORPH [25], CACD [24], and FG-NET [22] benchmark datasets. We also evaluate unconstrained face recognition results on the IJB-C [33] benchmark dataset to verify the generalizability of AIM.

5.0.0.1 Implementation Details: We apply **i**ntegrated **F**ace **A**nalytics **N**etwork (iFAN) [61] for face **R**egion **o**f **I**nterest (RoI) extraction, 68 landmark localization (if not provided), and alignment; throughout the experiments, the sizes of the RGB image x, the attention score map x^A, the feature map x^F, the synthesized face image $\hat{x}$ are fixed as 128×128; the pixel values of x, $\hat{x}$ and x^R are normalized to [-1,1]; the sizes of the input local patches (w/o overlapping) to the discriminator D_{ϕ_2} are fixed as 32×32; the dimensionality of learned facial representation f and sample f^* drawn from prior distribution $U(f)$ are fixed as 256; the age condition code c is a 7-dimension one-hot vector to encode different age phases[4], based on which continuous face rejuvenation/aging results can be achieved through interpolation during inference; the element of c is also confined to [-1,1], where -1 corresponds to 0; the element of smoothed labels for $\mathcal{L}_{\mathrm{cer}}$ is $\frac{1}{7}$; the constraint factors $\{\lambda_k\}_{k=1}^{14}$ are empirically fixed as 0.1, 0.1, 0.01, 1.0, 0.01, 0.05, 0.1, 10^{-5}, 0.03, 0.01, 0.05, 0.1, 10^{-5} and 0.03, respectively; the encoder G_{θ_E} is initialized with the Light CNN-29 [62] architecture by eliminating the linear classifier and replacing the activation function of the last fully-connected layer with hyperbolic tangent; the decoder G_{θ_D} is initialized with 3 hidden fractionally-strided convolution layers with kernels $3 \times 3 \times 512/2$, $3 \times 3 \times 256/2$, and $3 \times 3 \times 128/2$, activated with **Re**tified **L**inear **U**nit (ReLU), appended with a convolution layer with kernel $1 \times 1 \times 1$ activated with sigmoid and a convolution layer with kernel $1 \times 1 \times 3$ activated with scaled sigmoid for attention score map x^A and feature map x^F prediction, respectively; the domain classifier C_{φ} and the regularizer R_{ψ} are initialized with the same MLP architectures (which are learned separately), containing a hidden 256-way fully-connected layer activated with Leaky ReLU and a final 7-way fully-connected layer; the discriminator D_{ϕ_1} is initialized with a MLP containing a hidden 256-way fully-connected layer activated with Leaky ReLU, appended with a 1-way fully-connected layer activated by sigmoid and an n-way fully-connected layer (n is the identity number of the training data) as the dual agents for $\mathcal{L}_{\mathrm{adv}_1}$ and $\mathcal{L}_{\mathrm{ip}}$, respectively; the discriminator D_{ϕ_2} is initialized with a VGG-16 [63] architecture by eliminating the linear classifier, and appending a new 1-way fully-connected layer activated by sigmoid and a new 7-way fully-connected layer activated by hyperbolic tangent as the dual agents for $\mathcal{L}_{\mathrm{adv}_2}$ and $\mathcal{L}_{\mathrm{ae}}$, respectively; the newly added layers are randomly initialized by drawing weights from a zero-mean Gaussian distribution with standard deviation 0.01; Batch Normalization [64] is adopted in G_{θ_E} and G_{θ_D}; the dropout [**?**] ratio is empirically fixed as 0.7; the weight decay and the batch size are fixed as 5×10^3 and 32, respectively; We use an initial learning rate of 10^{-5} for pre-trained layers, and 2×10^{-4} for newly added layers in all our experiments; we

4. We divide the whole age span into 7 age phases: $\leq$20, 20-25, 25-30, 30-40, 40-50, 50-60, $\geq$60.

TABLE 2: Face recognition performance comparison on CAFR. The results are averaged over 10 testing splits.

Model	Acc (%)	EER (%)	AUC (%)
Baseline			
Light CNN [62]	73.56±1.39	31.62±1.68	75.96±1.63
Architecture ablation of AIM			
w/o C_{φ}	78.85±1.39	21.97±1.18	86.77±1.01
w/o R_{ψ}	80.39±1.19	20.22±1.25	88.52±0.82
w/o Att.	82.25±1.03	18.50±1.04	90.26±0.94
Training loss ablation of AIM			
w/o $\mathcal{L}_{\mathrm{ip}}$	67.64±0.88	45.85±2.59	57.14±2.59
w/o $\mathcal{L}_{\mathrm{adv}_1}$	81.02±1.10	19.56±1.00	89.10±0.83
w/o $\mathcal{L}_{\mathrm{ae}}$	81.83±1.29	19.08±1.03	89.87±0.76
w/o $\mathcal{L}_{\mathrm{mc}}$	82.03±0.98	18.57±0.98	90.10±0.83
w/o $\mathcal{L}_{\mathrm{adv}_2}$	82.30±0.99	18.28±1.02	90.32±0.71
AIM	**84.81±0.93**	**17.67±0.90**	**90.84±0.78**

decrease the learning rate to $\frac{1}{10}$ of the previous one after 20 epochs and train the network for roughly 60 epochs one after another; the proposed network is implemented based on the publicly available TensorFlow [65] platform, which is trained using Adam (α=2×10^{-4}, β_1=0.5) on two NVIDIA GeForce GTX TITAN X GPUs with 12G memory; the same training setting is utilized for all our compared network variants.

5.1 Evaluations on the CAFR Benchmark

Our newly proposed CAFR dataset is the largest and most comprehensive age-invariant face recognition benchmark to date, which contains 1,446,500 images annotated with age, identity, gender, race and landmarks. Examples are visualized in Fig. 4. The data are randomly organized into 10 splits, each consisting of 25,000 verification pairs with various age variations. Evaluation systems report Acc, EER, AUC and ROC as 10-fold cross validation.

5.1.1 Component Analysis and Quantitative Comparison

We first investigate different architectures and loss function combinations of AIM to see their respective roles in age-invariant face recognition. We compare 10 variants from four aspects: baseline (Light CNN-29 [62]), different network structures (w/o C_{φ}, R_{ψ}, w/o attention mechanism), different loss function combinations (w/o $\mathcal{L}_{\mathrm{ip}}$, $\mathcal{L}_{\mathrm{adv}_1}$, $\mathcal{L}_{\mathrm{ae}}$, $\mathcal{L}_{\mathrm{mc}}$, $\mathcal{L}_{\mathrm{adv}_2}$), and our proposed AIM.

The performance comparison w.r.t. Acc, EER and AUC on CAFR is reported in Tab. 2. The corresponding ROC curve is provided in Fig. 5 (a). By comparing the results from the 1^{st} *v.s.* 4^{th} panels, we observe that our AIM consistently outperforms the baseline by a large margin: 11.25% in Acc, 13.95% in EER, and 14.88% in AUC. Light-CNN is a general-purpose face recognition model, with representations entangled with age variations and suffering difficulties to distinguish cross-age faces. Comparatively, AIM jointly performs disentangled representation learning through cross-age domain adversarial training and cross-entropy regularization, and photorealistic cross-age face synthesis with attention mechanism in a mutual boosting way. By comparing the results from the 2^{nd} *v.s.* 4^{th} panels, we observe that AIM consistently outperforms the 3 variants in terms of network structure. In particular, w/o C_{φ} refers to truncating the domain classifier from AIM, leading to 5.96%, 4.30%, and 4.07% performance drop for all metrics. This verifies the necessity of cross-age domain adversarial training, which promotes encoded features to be indistinguishable w.r.t. the shift between multi-age domains to facilitate age-invariant representation learning. w/o R_{ψ} refers to truncating the cross-entropy regularizer from AIM, leading to 4.42%, 2.55%, and 2.32% performance drop for all metrics. This verifies the necessity of cross-entropy regularization with label smoothing strategy that constrains cross-age representations with ambiguous separability to serve as an auxiliary assistance for C_{φ}. The superiority of incorporating attention mechanism to cross-age face synthesis can be verified by comparing w/o Att. with AIM, *i.e.*, 2.56%, 0.83%, and 0.58% differences for all metrics. Identity-preserving quality is crucial for face recognition applications, the superiority of which is verified by comparing w/o $\mathcal{L}_{\mathrm{ip}}$ with AIM, *i.e.*, 17.17%, 28.18%, and 33.70% decline for all metrics. The superiority of incorporating adversarial learning to specific process can be verified by comparing w/o $\mathcal{L}_{\mathrm{adv}_i}$, $i \in \{1, 2\}$ with AIM, *i.e.*, 3.79%, 1.89%, and 1.74%; 2.51%, 0.61%, and 0.52% decrease for all metrics. The superiority of incorporating age estimation and manifold consistency constraints are verified by comparing w/o $\mathcal{L}_{\mathrm{ae}}$ and w/o $\mathcal{L}_{\mathrm{mc}}$ with AIM, *i.e.*, 2.98%, 1.41%, and 0.97%; 2.78%, 0.90%, and 0.74% drop for all metrics.

To quantitatively evaluate the quality of the synthesized cross-age face images by AIM, we design a user study from 25 volunteers. Each volunteer is shown with two images each time. One image is synthesized by AIM and the other comes from one previous state-of-the-art method CAAE [27]. The candidate images are conditioned on one of the pre-defined age phases (indicated by the age condition code c) and supposed to exhibit the desirable face rejuvenation/aging effects. Each volunteer is asked to choose one of the following three options: 1) the result by AIM is better (score: 1.0); 2) the results are equally satisfactory (score: 0.5); 3) the result by CAAE is better (score: 0.0). The scores are normalized by number of responses per state and shown in Fig. 6. As can be observed, the proposed AIM outperforms prior work in almost all cases.

5.1.2 Qualitative Comparison

Most previous works on age-invariant face recognition address this problem considering either only robust representation learning or only face rejuvenation/aging. It is commonly believed simultaneously modeling both is a highly non-linear transformation, thus it is difficult for a model to learn discriminative and age-invariant facial representations while generating faithful cross-age face images. However, with enough training data and proper architecture and objective function design of AIM, it is feasible to take the best of both worlds, as shown in Fig. 1. For more detailed results across a wide range of ages in high resolution, please refer to Fig. 7. Our AIM consistently provides discriminative and age-invariant representations and high-fidelity age regressed/progressed (bidirectional) face images for all cases. This well verifies that the joint learning scheme of age-invariant representation and attention-based cross-age face synthesis is effective, and both results are beneficial to face recognition in the wild.

We then visually compare the qualitative face rejuvenation and aging results by our AIM with CAAE in Fig. 10 1^{st} block and showcase the facial detail transformation over time with AIM in Fig. 8. It can be observed that AIM achieves simultaneous face rejuvenation and aging with photorealistic and accurate age transformation effect (*e.g.*, wrinkles, eyes, mouth, mustache, laugh lines), thanks to the novel network structure and training strategy. In contrast, results of previous work may suffer from blur and ghosting artifacts, and be fragile to variations in illumination,

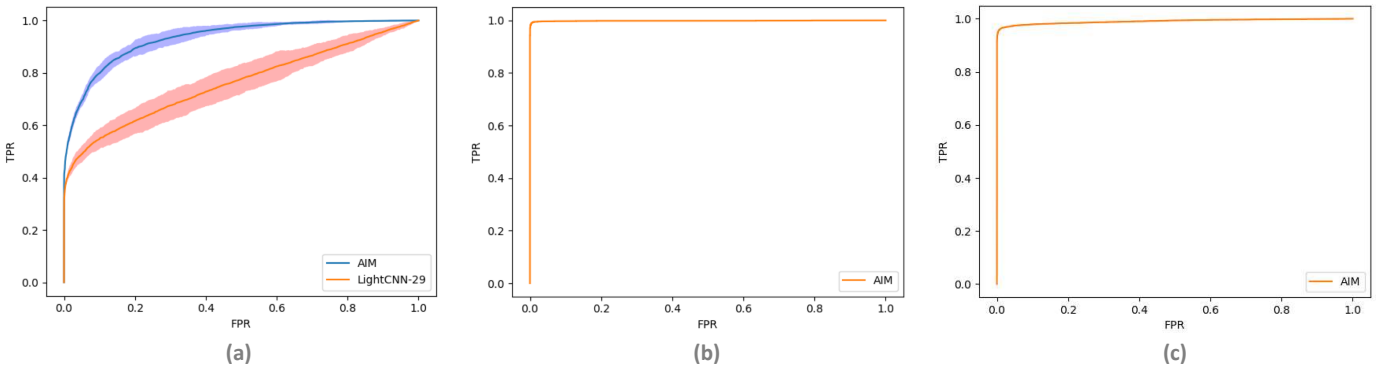


Fig. 5: ROC performance curve on (a) CAFR; (b) CACD-VS; (c) IJB-C. Best viewed in color.

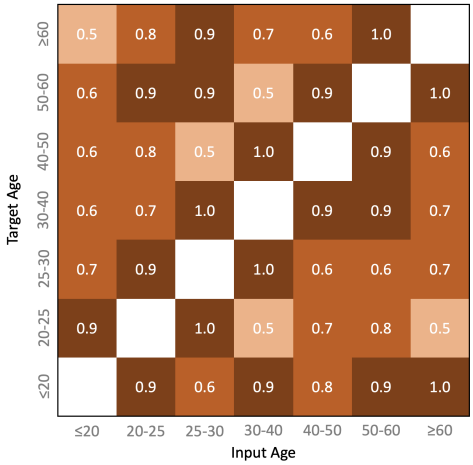


Fig. 6: User study on cross-age face synthesis results on CAFR.

TABLE 3: Rank-1 recognition rates (%) on MORPH Album2.

Method	Setting-1/Setting-2
HFA [28]	91.14/-
CARC [66]	92.80/-
MEFA [36]	93.80/-
GSM [44]	-/94.40
MEFA+SIFT+MLBP [36]	94.59/-
LPS+HFA [37]	94.87/-
LF-CNN [29]	97.51/-
AE-CNN [42]	-/98.13
OE-CNN [45]	98.55/98.67
OE-CNN [45] + CAFR	99.08/99.13
AIM (Ours)	**99.13/98.81**
AIM + CAFR (Ours)	**99.65/99.26**

expression and pose. This further shows effectiveness of the proposed AIM.

To demonstrate the capacity of AIM to synthesize cross-age face images with continuous and smooth transition between identities and ages, and show that the learned representations are identity-specific and explicitly disentangled from age variations, we further visualize the learned face manifold in Fig. 9 by performing interpolation upon both f and c. In particular, we take two images of different subjects x_1 and x_2, extract the encoded features from G_{θ_E} and perform interpolation between f_{x_1} and f_{x_2}. We also interpolate between two neighboring age condition codes to generate face images with continuous ages. The interpolated f and c are then fed to G_{θ_D} to synthesize face images. These smooth semantic changes indicate that the model has learned to produce identity-specific representations disentangled from age variations for age-invariant face recognition.

Finally, we visualize the cross-age face verification results for CAFR split1 to gain insights into age-invariant face recognition with AIM. After computing the similarities for all pairs of probe and reference sets, we sort the results into a ranking list. Each row shows a probe and reference pair. Between pairs are the matching similarities. Fig. 11 (a) and (b) show the best matched and non-matches examples, respectively. We note that most of these cases are under mild conditions in terms of age gap and other unconstrained factors like resolution, expression and pose. Fig. 11 (c) and (d) show the worst matched and non-matched examples, respectively, representing failed matching. We note that most of error cases are with large age gaps blended with other challenging scenarios like blur, extreme expressions, heavy make-up and large poses, which are even hard for humans to recognize. This confirms that CAFR aligns well with reality and deserves more research attention.

5.2 Evaluations on the MORPH Benchmark

MORPH is a large-scale public longitudinal face database, collected in real-world conditions with variations in age, pose, expression and lighting conditions. It has two separate datasets: Album1 and Album2. Album 1 contains 1,690 face images from 515 subjects while Album 2 contains 78,207 face images from 20,569 subjects. Statistical details are provided in Tab. 1. Both albums include metadata for age, identity, gender, race, eye coordinates and date of acquisition. For fair comparisons, Album2 is used for evaluation. Following [28], [67], Album2 is partitioned into a training set of 20,000 face images from 10,000 subjects with each subject represented by two images with the largest gap, and an independent testing set consisting of a gallery set and a probe set from the remaining subjects under two settings. Setting-1 consists of 20,000 face images from 10,000 subjects with each subject represented by a youngest face image as gallery and the oldest face image as probe while Setting-2 consists of 6,000 face images from 3,000 subjects with the same criteria. Evaluation systems report the Rank-1 identification rate.

The face recognition performance comparison of the proposed AIM with other state-of-the-arts on MORPH Album2 in Setting-1 and Setting-2 is reported in Tab. 3. With the mutual boosting learning scheme of age-invariant representation and attention-based cross-age face synthesis, our method outperforms the 2^{nd}-best by 0.58% and 0.14% for Setting-1 and Setting-2, respectively. By incorporating CAFR during training, the rank-1 recognition rates are further improved by 0.52% and 0.45% for Setting-1 and Setting-2, respectively. Similar performance improvements are observed by applying CAFR to other state-of-the-art age-invariant recognition methods, *e.g.*, OE-CNN [45]. This confirms

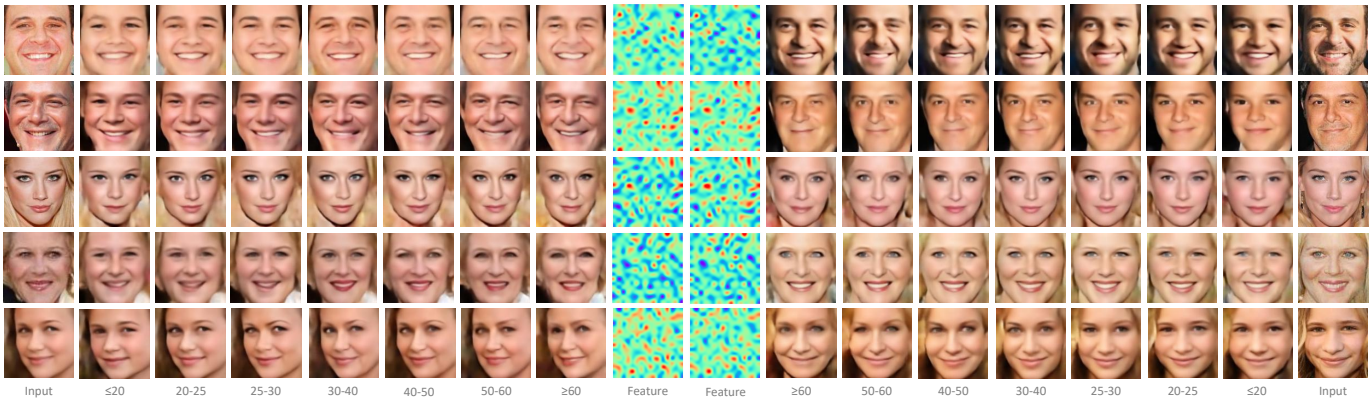


Fig. 7: Age-invariant face recognition example results on CAFR. *Col.* 1 & 18: Input faces of distinct identities with various challenging factors (*e.g.*, neutral, illumination, expression, and pose). *Col.* 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17: Synthesized age regressed/progressed faces by our proposed AIM. *Col.* 9 & 10: Learned facial representations by AIM, which are similar for the same row/identity while discriminative across different rows/identities, hence explicitly disentangled from the age variation. Based on such representations, AIM then apply the face synthesis component onto the representation, which takes targeted ages, identity discriminative information (*e.g.*, expression) as input, and generate faces of various ages/expression. These examples indicate facial representations learned by AIM are robust to age variance, and synthesized cross-age face images retain the intrinsic details. Best viewed in color.

Fig. 8: Facial attributes transformation over time in terms of (a) wrinkles & eyes, (b) mouth & moustache and (c) laugh lines, which is automatically learned by AIM instead of physical modeling.

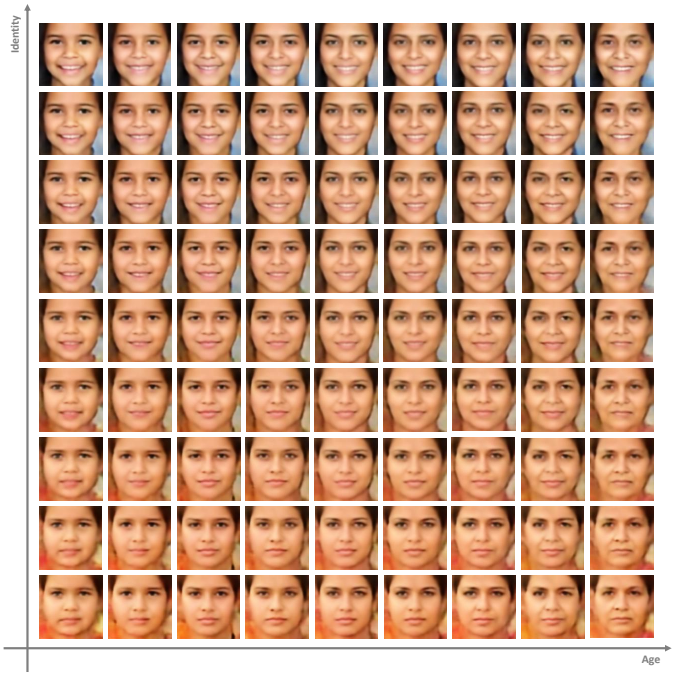


Fig. 9: Illustration of learned face manifold with continuous transitions in age (horizontal axis) and identity (vertical axis).

that our AIM is highly effective, and the proposed CAFR dataset is orthogonal to the state-of-the-art age-invariant face recognition algorithms and it is beneficial for advancing age-invariant face recognition performance. Visual comparison of face rejuvenation/aging results by AIM and CAAE is provided in Fig. 10 2^{nd} block, also validating advantages of AIM over existing solutions.

TABLE 4: Face recognition performance comparison on CACD-VS.

Method	Acc (%)
CAN [43]	92.30
VGGFace [68]	96.00
Center Loss [69]	97.48
MFM-CNN [62]	97.95
LF-CNN [29]	98.50
Marginal Loss [70]	98.95
DeepVisage [71]	99.13
OE-CNN [45]	99.20
OE-CNN [45] + CAFR	99.60
Human, *avg.* [24]	85.70
Human, voting [24]	94.20
AIM (Ours)	**99.38**
AIM + CAFR (Ours)	**99.76**

TABLE 5: Face recognition performance comparison on MegaFace Challenge 1 using FG-NET as probe set.

Method	Rank-1 (%)
FUDAN-CS_SDS [72]	25.56
SphereFace [73]	47.55
TNVP [74]	47.72
Softmax [75]	35.11
A-Softmax [75]	46.77
OE-CNN [45]	58.21
AIM (Ours)	**60.94**

5.3 Evaluations on the CACD Benchmark

CACD is a large-scale public dataset for face recognition and retrieval across ages, with variations in age, illumination, makeup, expression and pose, aligned with the real-world scenarios better than MORPH. It contains 163,446 face images from 2,000 celebrities. Statistical details are provided in Tab. 1. The meta data include age, identity and landmark. However, CACD contains some incorrectly labeled samples and duplicate images. For fair comparison, following [24], a carefully annotated version **CACD** **V**erification **S**ub-set (CACD-VS) is used for evaluation. It consists of 10 splits including 4,000 image pairs in total. Each split contains 200 genuine pairs and 200 imposter pairs for cross-age verification task. Evaluation systems report Acc and ROC as 10-fold cross validation.

The face recognition performance comparison of the proposed

Fig. 10: Qualitative comparison of face rejuvenation/aging results on CAFR, MORPH, CACD, FG-NET, and IJB-C.

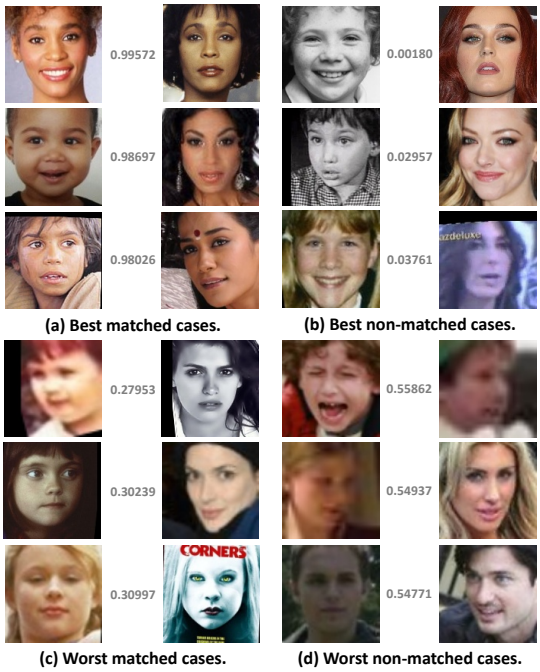


Fig. 11: Age-invariant face recognition analysis on CAFR split1.

TABLE 6: Face recognition performance comparison on MegaFace Challenge 2 using FG-NET as probe set.

Method	Rank-1 (%)
GRCCV [76]	21.04
NEC [76]	29.29
3DiVi [76]	35.79
GT-CMU-SYSU [76]	38.21
OE-CNN [45]	53.26
AIM (Ours)	**55.79**

AIM with other state-of-the-arts on CACD-VS is reported in Tab. 4. The corresponding ROC curve is provided in Fig. 5 (b). Our method dramatically surpasses human performance and other state-of-the-arts. In particular, AIM improves the Acc of the 2^{nd}-best by 0.18%. AIM also outperforms human voting performance by 5.18%. To our best knowledge, this is the new state-of-the-art, including unpublished technical reports. This shows the learned facial representations by AIM are discriminative and robust even with in-the-wild variations. With the injection of CAFR as augmented training data, our method and OE-CNN both further gains ∼0.38%, confirming that CAFR is uniformly valid in boosting age-invariant face recognition performance. Visual comparison of face rejuvenation/aging results by AIM and four state-of-the-art methods is provided in Fig. 10 3^{rd} block, which again verifies effectiveness of our method for high-fidelity cross-age face synthesis.

5.4 Evaluations on the FG-NET Benchmark

FG-NET is a popular public dataset for cross-age face recognition, collected in realistic conditions with huge variability in age covering from child to elder. It contains 1,002 face images from 82 non-celebrity subjects. Statistical details are provided in Tab. 1. The metadata include age, identity and landmark.

We evaluate the proposed AIM under the challenging settings following the evaluation protocols of the MegaFace Challenge 1 [75] and Challenge 2 [76], by employing FG-NET as the probe set and the 1 million images from Flickr as the distractor set. We evaluate the rank-1 recognition rate, as shown in Tab. 5 and Tab. 6, respectively. It is encouraging to see that our method outperforms other competitors by a large margin, which strongly confirms the efficiency of the proposed AIM on age-invariant face recognition.

5.5 Evaluations on the IJB-C Benchmark

IJB-C contains 31,334 images and 11,779 videos from 3,531 subjects, which are split into 117,542 frames, 8.87 images and 3.34 videos per subject, captured from in-the-wild environments to avoid the near frontal bias. For fair comparison, we follow the template-based setting and evaluate models on the standard 1:1 verification protocol in terms of **T**rue **A**cceptance **R**ate (TAR)@**F**alse **A**cceptance **R**ate (FAR).

The face recognition performance comparison of the proposed AIM with other state-of-the-arts on IJB-C unconstrained face verification protocol is reported in Tab. 7. The corresponding ROC curve is provided in Fig. 5 (c). Our AIM beats the 2^{nd}-best by 5.50% in TAR@FAR=10^{-5}, which verifies its remarkable generalization ability for recognizing faces in the wild. Qualitative comparisons for face rejuvenation/aging are provided in Fig. 10 5^{th} block, which further shows the superiority of our method for cross-age face synthesis under unconstrained condition.

6 Conclusion

We proposed a novel **A**ge-**I**nvariant **M**odel (AIM) for joint disentangled representation learning and photorealistic cross-age face synthesis to address the challenging face recognition with large age variations. Through carefully designed network architecture and optimization strategies, AIM learns to generate powerful age-invariant facial representations explicitly disentangled from the age variation while achieving continuous face rejuvenation/aging with remarkable photorealistic and identity-preserving properties, avoiding requirements of paired data and true age of testing samples. Moreover, we propose a new large-scale **C**ross-**A**ge **F**ace **R**ecognition (CAFR) dataset to spark progress in age-invariant

TABLE 7: Face recognition performance comparison on IJB-C.

Method	TAR@FAR=10^{-5}	TAR@FAR=10^{-4}	TAR@FAR=10^{-3}	TAR@FAR=10^{-2}
GOTS [33]	0.066	0.147	0.330	0.620
FaceNet [16]	0.330	0.487	0.665	0.817
VGGFace [68]	0.437	0.598	0.748	0.871
VGGFace2_ft [77]	0.768	0.862	0.927	0.967
MN-vc [78]	0.771	0.862	0.927	0.968
AIM	**0.826**	**0.895**	**0.935**	**0.962**

face recognition. Comprehensive experiments demonstrate the superiority of AIM over the state-of-the-arts. We envision the proposed method and benchmark dataset would drive the age-invariant face recognition research towards real-world applications with presence of age gaps and other complex unconstrained distractors.

REFERENCES

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014, pp. 1701–1708.

[2] J. Chen, V. M. Patel, L. Liu, V. Kellokumpu, G. Zhao, M. Pietikäinen, and R. Chellappa, "Robust local features for remote face recognition," *IVC*, vol. 64, pp. 34–46, 2017.

[3] J. Li, J. Zhao, F. Zhao, H. Liu, J. Li, S. Shen, J. Feng, and T. Sim, "Robust face recognition with deep multi-view representation learning," in *ACM MM*, 2016, pp. 1068–1072.

[4] J. Zhao, L. Xiong, P. K. Jayashree, J. Li, F. Zhao, Z. Wang, P. S. Pranata, P. S. Shen, S. Yan, and J. Feng, "Dual-agent gans for photorealistic and identity preserving profile face synthesis," in *NIPS*, 2017, pp. 66–76.

[5] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing *et al.*, "Towards pose invariant face recognition in the wild," in *CVPR*, 2018, pp. 2207–2216.

[6] Z. Wang, J. Zhao, C. Lu, F. Yang, H. Huang, Y. Guo *et al.*, "Learning to detect head movement in unconstrained remote gaze estimation in the wild," in *WACV*, 2020, pp. 3443–3452.

[7] X. Tu, J. Zhao, M. Xie, Z. Jiang, A. Balamurugan, Y. Luo, Y. Zhao, L. He, Z. Ma, and J. Feng, "3d face reconstruction from a single image assisted by 2d face images in the wild," *T-MM*, 2020.

[8] X. Tu, J. Zhao, M. Xie, G. Du, H. Zhang, J. Li, Z. Ma, and J. Feng, "Learning generalizable and identity-discriminative representations for face anti-spoofing," *arXiv preprint arXiv:1901.05602*, 2019.

[9] X. Tang and Z. Li, "Video based face recognition using multiple classifiers," in *FG*, 2004, pp. 345–349.

[10] ——, "Frame synchronization and multi-level subspace analysis for video based face recognition," in *CVPR*, vol. 2, 2004, pp. II–II.

[11] ——, "Audio-guided video-based face recognition," *T-CSVT*, vol. 19, no. 7, pp. 955–964, 2009.

[12] Z. Li, D. Gong, Y. Qiao, and D. Tao, "Common feature discriminant analysis for matching infrared face images to optical face images," *T-IP*, vol. 23, no. 6, pp. 2436–2445, 2014.

[13] D. Gong, Z. Li, J. Liu, and Y. Qiao, "Multi-feature canonical correlation analysis for face photo-sketch image retrieval," in *ACM MM*, 2013, pp. 617–620.

[14] Z. Li, D. Gong, Q. Li, D. Tao, and X. Li, "Mutual component analysis for heterogeneous face recognition," *T-IST*, vol. 7, no. 3, pp. 1–23, 2016.

[15] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang, "Super-identity convolutional neural network for face hallucination," in *ECCV*, 2018, pp. 183–198.

[16] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[17] H. Wang, Y. Wang, Z. Zhou, X. Ji, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *CVPR*, 2018, pp. 5265–5274.

[18] J. Zhao, L. Xiong, Y. Cheng, Y. Cheng, J. Li, L. Zhou, Y. Xu, J. Karlekar, S. Pranata, S. Shen, J. Xing, S. Yan, and J. Feng, "3d-aided deep pose-invariant face recognition," in *IJCAI*, 2018, pp. 1184–1190.

[19] Y. Guo, Y. Lei, L. Liu, Y. Wang, M. Bennamoun, and F. Sohel, "Ei3d: Expression-invariant 3d face recognition based on feature and shape matching," *PR*, vol. 83, pp. 403–412, 2016.

[20] J. Zhao, J. Xing, L. Xiong, S. Yan, and J. Feng, "Recognizing profile faces by imagining frontal view," *IJCV*, pp. 1–19, 2019.

[21] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, "Efficient decision-based black-box adversarial attacks on face recognition," in *CVPR*, 2019, pp. 7714–7722.

[22] "Fg-net aging database," http://webmail.cycollege.ac.cy/alanitis/fgnetaging/, 2007.

[23] R. Rothe, R. Timofte, and L. V. Gool, "Dex: Deep expectation of apparent age from a single image," in *ICCVW*, 2015.

[24] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *T-MM*, vol. 17, no. 6, pp. 804–815, 2015.

[25] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *FGR*, 2006, pp. 341–345.

[26] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: The first manually collected, in-the-wild age database," in *CVPRW*, 2017, pp. 1997–2005.

[27] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *CVPR*, 2017.

[28] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang, "Hidden factor analysis for age invariant face recognition," in *ICCV*, 2013, pp. 2872–2879.

[29] Y. Wen, Z. Li, and Y. Qiao, "Latent factor guided convolutional neural networks for age-invariant face recognition," in *CVPR*, 2016, pp. 4893–4901.

[30] Z. JIAN, "Deep learning for human-centric image analysis," Ph.D. dissertation, 2018.

[31] H. Wang, D. Gong, Z. Li, and W. Liu, "Decorrelated adversarial learning for age-invariant face recognition," in *CVPR*, 2019, pp. 3527–3536.

[32] J. Zhao, Y. Cheng, Y. Cheng, Y. Yang, F. Zhao, J. Li, H. Liu, S. Yan, and J. Feng, "Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition," in *AAAI*, vol. 33, 2019, pp. 9251–9258.

[33] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney *et al.*, "Iarpa janus benchmark–c: Face dataset and protocol," in *ICB*, 2018.

[34] N. Ramanathan and R. Chellappa, "Face verification across age progression," *T-IP*, vol. 15, no. 11, pp. 3349–3361, 2006.

[35] D. Sungatullina, J. Lu, G. Wang, and P. Moulin, "Multiview discriminative learning for age-invariant face recognition," in *FG*, 2013, pp. 1–6.

[36] D. Gong, Z. Li, D. Tao, J. Liu, and X. Li, "A maximum entropy feature descriptor for age invariant face recognition," in *CVPR*, 2015, pp. 5289–5297.

[37] Z. Li, D. Gong, X. Li, and D. Tao, "Aging face recognition: a hierarchical learning model based on local patterns selection," *T-IP*, vol. 25, no. 5, pp. 2146–2154, 2016.

[38] L. Liu, P. Fieguth, G. Zhao, M. Pietikäinen, and D. Hu, "Extended local binary patterns for face recognition," *Information Sciences*, vol. 358, pp. 56–72, 2016.

[39] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *JMLR*, vol. 10, no. Feb, pp. 207–244, 2009.

[40] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs, "Face verification across age progression using discriminative methods," *T-IFS*, vol. 5, no. 1, pp. 82–91, 2010.

[41] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *CVPR*, 2013, pp. 3025–3032.

[42] T. Zheng, W. Deng, and J. Hu, "Age estimation guided convolutional neural network for age-invariant face recognition," in *CVPRW*, 2017, pp. 12–16.

[43] C. Xu, Q. Liu, and M. Ye, "Age invariant face recognition and retrieval by coupled auto-encoder networks," *Neurocomputing*, vol. 222, pp. 62–71, 2017.

[44] L. Lin, G. Wang, W. Zuo, X. Feng, and L. Zhang, "Cross-domain visual matching via generalized similarity measure and feature learning," *T-PAMI*, vol. 39, no. 6, pp. 1089–1102, 2017.

[45] Y. Wang, D. Gong, Z. Zhou, X. Ji, H. Wang, Z. Li, W. Liu, and T. Zhang, "Orthogonal deep features decomposition for age-invariant face recognition," in *ECCV*, 2018.

[46] Y. Cheng, J. Zhao, Z. Wang, Y. Xu, J. Karlekar, S. Shen, and J. Feng, "Know you at one glance: A compact vector representation for low-shot learning." in *ICCVW*, 2017, pp. 1924–1932.

[47] J. Suo, X. Chen, S. Shan, W. Gao, and Q. Dai, "A concatenational graph evolution aging model," *T-PAMI*, vol. 34, no. 11, pp. 2083–2096, 2012.

[48] N. Ramanathan and R. Chellappa, "Modeling shape and textural variations in aging faces," in *FG*, 2008, pp. 1–8.

[49] ——, "Modeling age progression in young faces," in *CVPR*, vol. 1, 2006, pp. 387–394.

[50] D. M. Burt and D. I. Perrett, "Perception of age in adult caucasian male faces: Computer graphic manipulation of shape and colour information," *Proc. R. Soc. Lond. B*, vol. 259, no. 1355, pp. 137–143, 1995.

[51] I. Kemelmacher-Shlizerman, S. Suwajanakorn, and S. M. Seitz, "Illumination-aware age progression," in *CVPR*, 2014, pp. 3334–3341.

[52] Y. Wang, Z. Zhang, W. Li, and F. Jiang, "Combining tensor space analysis and active appearance models for aging effect simulation on face images," *IEEE T SYST MAN CY B*, vol. 42, no. 4, pp. 1107–1118, 2012.

[53] H. Yang, D. Huang, Y. Wang, H. Wang, and Y. Tang, "Face aging effect simulation using hidden factor analysis joint sparse representation," *T-IP*, vol. 25, no. 6, pp. 2493–2507, 2016.

[54] W. Wang, Z. Cui, Y. Yan, J. Feng, S. Yan, X. Shu, and N. Sebe, "Recurrent face aging," in *CVPR*, 2016, pp. 2378–2386.

[55] S. Liu, Y. Sun, D. Zhu, R. Bao, W. Wang, X. Shu, and S. Yan, "Face aging with contextual generative adversarial nets," in *ACM MM*, 2017, pp. 82–90.

[56] H. Yang, D. Huang, Y. Wang, and A. K. Jain, "Learning face age progression: A pyramid architecture of gans," in *CVPR*, 2018, pp. 31–39.

[57] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, no. 59, pp. 1–35, 2016.

[58] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *ICCVW*, 2015, pp. 10–15.

[59] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*, 2016, pp. 87–102.

[60] J. Li, L. Liu, J. Li, J. Feng, S. Yan, and T. Sim, "Towards a comprehensive face detector in the wild," *T-CSVT*, 2017.

[61] J. Li, S. Xiao, F. Zhao, J. Zhao, J. Li, J. Feng, S. Yan, and T. Sim, "Integrated face analytics networks through cross-dataset hybrid training," in *ACM MM*, 2017, pp. 1531–1539.

[62] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *T-IFS*, vol. 13, no. 11, pp. 2884–2896, 2018.

[63] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[64] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[65] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning," in *OSDI*, 2016, pp. 265–283.

[66] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *ECCV*, 2014, pp. 768–783.

[67] Z. Li, U. Park, and A. K. Jain, "A discriminative model for age invariant face recognition," *T-IFS*, vol. 6, no. 3, pp. 1028–1037, 2011.

[68] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[69] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016, pp. 499–515.

[70] J. Deng, Y. Zhou, and S. Zafeiriou, "Marginal loss for deep face recognition," in *CVPRW*, vol. 4, no. 6, 2017.

[71] M. A. Hasnat, J. Bohné, J. Milgram, S. Gentric, and L. Chen, "Deepvisage: Making face recognition simple yet with powerful generalization skills." in *ICCVW*, 2017, pp. 1682–1691.

[72] Z. Wang, K. He, Y. Fu, R. Feng, Y.-G. Jiang, and X. Xue, "Multi-task deep neural network for joint face recognition and facial attribute prediction," in *ICMR*, 2017, pp. 365–374.

[73] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, 2017, pp. 212–220.

[74] C. Nhan Duong, K. Gia Quach, K. Luu, N. Le, and M. Savvides, "Temporal non-volume preserving approach to facial age-progression and age-invariant face recognition," in *ICCV*, 2017, pp. 3735–3743.

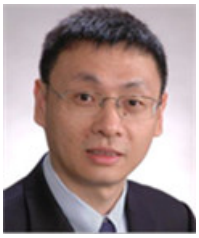
[75] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *CVPR*, 2016, pp. 4873–4882.

[76] A. Nech and I. Kemelmacher-Shlizerman, "Level playing field for million scale face recognition," in *CVPR*, 2017, pp. 7044–7053.

[77] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *FG*, 2018, pp. 67–74.

[78] W. Xie and A. Zisserman, "Multicolumn networks for face recognition," *arXiv preprint arXiv:1807.09192*, 2018.

Jian Zhao received the Bachelor's degree from Beihang University in 2012, the Master's degree from National University of Defense Technology in 2014, and the Ph.D. degree from National University of Singapore in 2019. He is currently an Assistant Professor with Institute of North Electronic Equipment, Beijing, China. His main research interests include deep learning, pattern recognition, computer vision and multimedia analysis. He has published over 30 cutting-edge papers. He has won the Lee Hwee Kuan Award (Gold Award) on PREMIA 2019, the "Best Student Paper Award" on ACM MM 2018, and the top-3 awards several times on world-wide competitions. He is the EAC of VALSE, and the committee member of CSIG-BVD. He has served as the invited reviewer of T-PAMI, IJCV, T-MM, T-IFS, T-CSVT, Neurocomputing, CSSP, JVCI, NeurIPS (one of the top 30% highest-scoring reviewers of NeurIPS 2018), CVPR, ICCV, ACM MM, AAAI, ICLR, ICML, ACCV, UAI.



Shuicheng Yan is currently the CTO with Yitu Technology, Beijing, China. He has authored/co-authored over 500 high quality technical papers, with Google Scholar citation over 65,000 times and an h-index 107. His research areas include computer vision, machine learning, and multimedia analysis. He is the Fellow of Academy of Engineering, Singapore, Fellow of IEEE, Fellow of IAPR and the ACM Distinguished Scientist. His team received seven times winner or honorable-mention prizes in five years over PASCAL, VOC, and ILSVRC competitions, which are core competitions in the field of computer vision, along with over ten times the Best (student) Paper Awards and especially a Grand Slam with the ACM MM, the top conference in the field of multimedia, including the Best Paper Award, the Best Student Paper Award, and the Best Demo Award. He is a TR Highly Cited Researcher of 2014, 2015, and 2016.

Jiashi Feng received the Ph.D. degree from the National University of Singapore in 2014. He was a Post-Doctoral Research Follow with the University of California, Berkeley. He joined NUS as a Faculty Member, where he is currently an Assistant Professor with the Department of Electrical and Computer Engineering. His research areas include computer vision, machine learning, object recognition, detection, segmentation, robust learning and deep learning.