Look across Elapse: Disentangled Representation Learning and Photorealistic Cross-Age Face Synthesis for Age-Invariant Face Recognition

Jian Zhao,[1,2*] **Yu Cheng,**[1] **Yi Cheng,**[3] **Yang Yang,**[1]
Fang Zhao,[4] **Jianshu Li,**[1] **Hengzhu Liu,**[2] **Shuicheng Yan,**[1,5] **Jiashi Feng**[1]
[1]National University of Singapore, [2]National University of Defense Technology
[3]Panasonic R&D Center Singapore, [4]Inception Institute of Artificial Intelligence, [5]Qihoo 360 AI Institute
{zhaojian90, e0321276, yang_yang, jianshu}@u.nus.edu, yi.cheng@sg.panasonic.com
zhaofang0627@gmail.com, hengzhuliu@nudt.edu.cn, {eleyans, elefjia}@nus.edu.sg

Abstract

Despite the remarkable progress in face recognition related technologies, reliably recognizing faces across ages still 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, we develop effective and novel training strategies for end-to-end learning the whole deep architecture, which generates powerful age-invariant face representations explicitly disentangled from the age variation. Extensive experiments on several 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 one of the most popular unconstrained face recognition datasets IJB-C additionally verifies the promising generalizability of AIM in recognizing faces in the wild.

Figure 1: 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 facial representations by our proposed AIM, which are explicitly disentangled from the age variation. AIM can learn age-invariant representations and synthesize photorealistic cross-age faces effectively. Best viewed in color.

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 (Taigman et al. 2014; Chen et al. 2017; Li et al. 2016a; Zhao et al. 2017) or even surpassed (Schroff, Kalenichenko, and Philbin 2015; Wang et al. 2018a; Zhao et al. 2018) 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 (fgn 2007; Rothe, Timofte, and Gool 2015; Chen, Chen, and Hsu 2015; Ricanek and Tesafaye 2006;

*Jian Zhao is the corresponding author. Homepage: https://zhaoj9014.github.io/.

[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.

Moschoglou et al. 2017; Zhang and Qi 2017) due to their small size, narrow elapse per subject and unbalanced genders, ethnicities 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 (Chen, Chen, and Hsu 2015). In this work, we aim to improve automatic models for recognizing unconstrained faces with large age variations.

According to recent studies (Gong et al. 2013; Wen, Li, and Qiao 2016), 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 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. Thus 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.

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.

Extensive experiments on several standard cross-age datasets (MORPH (Ricanek and Tesafaye 2006), CACD (Chen, Chen, and Hsu 2015) and FG-NET (fgn 2007)) demonstrate the superiority of AIM over the state-of-the-arts. Benchmarking AIM on one of the most popular unconstrained face recognition datasets IJB-C (Maze et al. 2018) additionally verifies its promising generalizability in recognizing faces in the wild.

Our contributions are summarized as follows.

- 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.

Related Work

Age-Invariant Representation Learning

Conventional approaches often leverage robust local descriptors (Ramanathan and Chellappa 2006a; Gong et al. 2013; Sungatullina et al. 2013; Gong et al. 2015; Li et al. 2016b) and metric learning (Weinberger and Saul 2009; Ling et al. 2010; Chen et al. 2013) to tackle age variance. For instance, (Ramanathan and Chellappa 2006a) develop a Bayesian classifier to recognize age difference and perform face verification across age progression. (Gong et al. 2013) propose **H**idden **F**actor **A**nalysis (HFA) for age-invariant face recognition that separates aging variations from identity-specific features. (Weinberger and Saul 2009) improve the performance by distance metric learning. (Ling et al. 2010) propose **G**radient **O**rientation **P**yramid (GOP) 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 (Wen, Li, and Qiao 2016; Zheng, Deng, and Hu 2017; Xu, Liu, and Ye 2017; Lin et al. 2017; Wang et al. 2018b). For instance, (Cheng et al. 2017) propose an enforced softmax optimization strategy to learn effective and compact deep facial representations with reduced intra-class variance and enlarged inter-class distance. (Wen, Li, and Qiao 2016) propose a **L**atent **F**actor guided **C**onvolutional **N**eural **N**etwork (LF-CNN) model to learn age-invariant deep features. (Zheng, Deng, and Hu 2017) propose an **A**ge **E**stimation guided **CNN** (AE-CNN) model to separate aging variations from identity-specific features. (Wang et al. 2018b) propose an **O**rthogonal **E**mbedding **CNN** (OE-CNN) model to decompose deep facial representations into two orthogonal components to represent age- and identity-specific features.

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 (Suo et al. 2012), wrinkles (Ramanathan and Chellappa 2008), and facial structure (Ramanathan and Chellappa 2006b). 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 (Burt and Perrett 1995; Kemelmacher-Shlizerman, Suwajanakorn, and Seitz 2014) 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 (Wang et al. 2012; Yang et al. 2016) 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, (Wang et al. 2016) propose a smooth face aging process between neighboring groups by modeling the intermediate transition states with **R**ecurrent **N**eural **N**etwork (RNN). (Zhang and Qi 2017) propose a **C**onditional **A**dversarial **A**uto-**E**ncoder (CAAE) and achieve face age regression/progression in a holistic framework. (Zhu et al. 2018) propose a **C**onditional **M**ulti-**A**dversarial **A**uto-**E**ncoder with **O**rdinal **R**egression (CMAAE-OR) to predict facial rejuvenation and aging. (Song et al. 2018) propose a **Dual c**onditional **GANs** (Dual cGANs) where the primal cGAN transforms a face image to other ages based on the age condition, while the dual one learns to invert the task.

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) AIM achieves continuous face rejuvenation/aging 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.

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.

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.

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

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

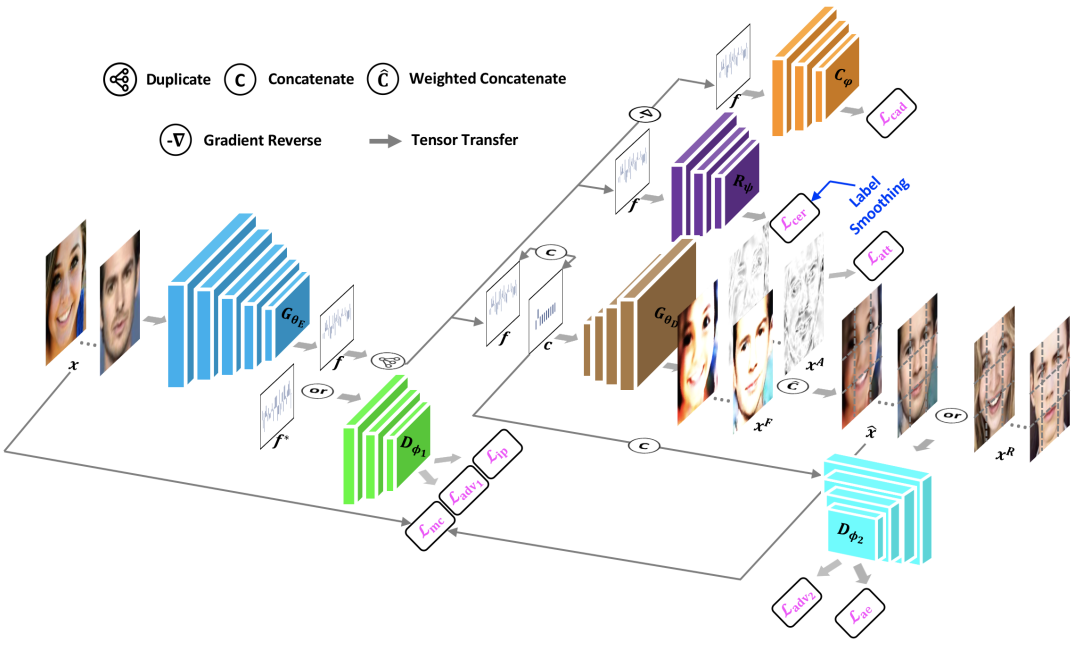

Figure 2: Age-**I**nvariant **M**odel (AIM) for face recognition in the wild. AIM extends from an auto-encoder based GAN and includes a disentangled **R**epresentation **L**earning sub-**N**et (RLN) and a **F**ace **S**ynthesis sub-**N**et (FSN) that jointly learn end-to-end. RLN consists of an encoder (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 φ denotes the learnable parameters for the domain classifier, and $y_i \in \{0, 1, \dots\}$ indicates which domain f_i is from. Minimizing $\mathcal{L}_{\text{cad}}$ 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 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 (Ganin et al. 2016).

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

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

where ψ denotes the learnable parameters for the regularizer, 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:

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

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:

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

where y_i denotes the identity ground truth.

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:

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

where $\{\lambda_k\}_{k=10}^{14}$ are weighting parameters among different losses.

$\mathcal{L}_{\mathrm{adv}_2}$ 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:

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

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} 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:

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

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:

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

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:

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

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.

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

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

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.

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 MORPH (Ricanek and Tesafaye 2006), CACD (Chen, Chen, and Hsu 2015) and FG-NET (fgn 2007) benchmark datasets. We also evaluate unconstrained face recognition results on the IJB-C benchmark dataset (Maze et al. 2018) to verify the generalizability of AIM.

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. Both albums include meta data for age, identity, gender, race, eye coordinates and date of acquisition. For fair comparisons, Album2 is used for evaluation. Following (Li, Park, and Jain 2011; Gong et al. 2013), Album2 is partitioned into a training set of 20,000 face images from 10,000 subjects with each subject represented by two images with 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

Figure 3: Qualitative comparison of face rejuvenation/aging results on MORPH, CACD, FG-NET and IJB-C.

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

Method	Setting-1/Setting-2
HFA (Gong et al. 2013)	91.14/-
CARC (Chen, Chen, and Hsu 2014)	92.80/-
MEFA (Gong et al. 2015)	93.80/-
GSM (Lin et al. 2017)	-/94.40
MEFA+SIFT+MLBP (Gong et al. 2015)	94.59/-
LPS+HFA (Li et al. 2016b)	94.87/-
LF-CNN (Wen, Li, and Qiao 2016)	97.51/-
AE-CNN (Zheng, Deng, and Hu 2017)	-/98.13
OE-CNN (Wang et al. 2018b)	98.55/98.67
AIM (Ours)	**99.13/98.81**

each subject represented by a youngest face image as gallery and an 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 (Ricanek and Tesafaye 2006) Album2 in Setting-1 and Setting-2 is reported in Tab. 1. 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. This confirms that our AIM is highly effective for age-invariant face recognition. Visual comparison of face rejuvenation/aging results by AIM and CAAE (Zhang and Qi 2017) is provided in Fig. 3 1^{st} block, also validating advantages of AIM over existing solutions.

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 (Ricanek and Tesafaye 2006). It contains 163,446 face images from 2,000 celebrities. The meta data include age, identity and landmark. However, CACD contains some incorrectly labeled samples and duplicate images. For fair comparison, following (Chen, Chen, and Hsu 2015), 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.

Table 2: Face recognition performance comparison on CACD-VS.

Method	Acc (%)
CAN (Xu, Liu, and Ye 2017)	92.30
VGGFace (Parkhi et al. 2015)	96.00
Center Loss (Wen et al. 2016)	97.48
MFM-CNN (Wu et al. 2018)	97.95
LF-CNN (Wen, Li, and Qiao 2016)	98.50
Marginal Loss (Deng, Zhou, and Zafeiriou 2017)	98.95
DeepVisage (Hasnat et al. 2017)	99.13
OE-CNN (Wang et al. 2018b)	99.20
Human, *avg.* (Chen, Chen, and Hsu 2015)	85.70
Human, voting (Chen, Chen, and Hsu 2015)	94.20
AIM (Ours)	**99.38**

Table 3: Face recognition performance comparison on FG-NET.

Method	Rank-1 (%)
Park *et al.* (Park, Tong, and Jain 2010)	37.40
Li *et al.* (Li, Park, and Jain 2011)	47.50
HFA (Gong et al. 2013)	69.00
MEFA (Gong et al. 2015)	76.20
CAN (Xu, Liu, and Ye 2017)	86.50
LF-CNN (Wen, Li, and Qiao 2016)	88.10
AIM (Ours)	**93.20**

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 AIM with other state-of-the-arts on CACD-VS (Chen, Chen, and Hsu 2015) is reported in Tab. 2. 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. Visual comparison of face rejuvenation/aging results by AIM and the state-of-the-art method is provided in Fig. 3 2^{nd} block, which

Table 4: 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 (Maze et al. 2018)	0.066	0.147	0.330	0.620
FaceNet (Schroff, Kalenichenko, and Philbin 2015)	0.330	0.487	0.665	0.817
VGGFace (Parkhi et al. 2015)	0.437	0.598	0.748	0.871
VGGFace2_ft (Cao et al. 2018)	0.768	0.862	0.927	0.967
MN-vc (Xie and Zisserman 2018)	0.771	0.862	0.927	0.968
AIM	**0.826**	**0.895**	**0.935**	**0.962**

again verifies effectiveness of our method for high-fidelity cross-age face synthesis.

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. The meta data include age, identity and landmark. Since the size of FG-NET is small, we follow the leave-one-out setting of (Li, Park, and Jain 2011; Gong et al. 2013) for fair comparisons with previous methods. In particular, we leave one image as the testing sample and train (finetune) the model with remaining 1,001 images. We repeat this procedure 1,002 times and report the average rank-1 recognition rate.

The face recognition performance comparison of the proposed AIM with other state-of-the-arts on FG-NET (fgn 2007) is reported in Tab. 3. AIM improves the 2^{nd}-best by 5.10%. Qualitative comparisons for face rejuvenation/aging are provided in Fig. 3 3^{rd} block, which well shows the promising potential of our method for challenging unconstrained face recognition contaminated with age variance.

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 (Maze et al. 2018) unconstrained face verification protocol is reported in Tab. 4. Our AIM beats the 2^{nd}-best by 5.50% in TAR@FAR=10^{-5}, which verifies its remarkable generalizability for recognizing faces in the wild. Qualitative comparisons for face rejuvenation/aging are provided in Fig. 3 4^{th} block, which further shows the superiority of our method for cross-age face synthesis under unconstrained condition.

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

Acknowledgement

The work of Jian Zhao was partially supported by **C**hina **S**cholarship **C**ouncil (CSC) grant 201503170248.

The work of Jiashi Feng was partially supported by NUS IDS R-263-000-C67-646, ECRA R-263-000-C87-133 and MOE Tier-II R-263-000-D17-112.

References

Burt, D. M., and Perrett, D. I. 1995. Perception of age in adult caucasian male faces: Computer graphic manipulation of shape and colour information. *Proc. R. Soc. Lond. B* 259(1355):137–143.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 67–74.

Chen, D.; Cao, X.; Wen, F.; and Sun, J. 2013. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *CVPR*, 3025–3032.

Chen, J.; Patel, V. M.; Liu, L.; Kellokumpu, V.; Zhao, G.; Pietikäinen, M.; and Chellappa, R. 2017. Robust local features for remote face recognition. *IVC* 64:34–46.

Chen, B.-C.; Chen, C.-S.; and Hsu, W. H. 2014. Cross-age reference coding for age-invariant face recognition and retrieval. In *ECCV*, 768–783.

Chen, B.-C.; Chen, C.-S.; and Hsu, W. H. 2015. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *T-MM* 17(6):804–815.

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

Deng, J.; Zhou, Y.; and Zafeiriou, S. 2017. Marginal loss for deep face recognition. In *CVPRW*, volume 4.

2007. Fg-net aging database. http://webmail.cycollege.ac.cy/alanitis/fgnetaging/.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR* 17(59):1–35.

Gong, D.; Li, Z.; Lin, D.; Liu, J.; and Tang, X. 2013. Hidden factor analysis for age invariant face recognition. In *ICCV*, 2872–2879.

Gong, D.; Li, Z.; Tao, D.; Liu, J.; and Li, X. 2015. A maximum entropy feature descriptor for age invariant face recognition. In *CVPR*, 5289–5297.

Hasnat, M. A.; Bohné, J.; Milgram, J.; Gentric, S.; and Chen, L. 2017. Deepvisage: Making face recognition simple yet with powerful generalization skills. In *ICCVW*, 1682–1691.

Kemelmacher-Shlizerman, I.; Suwajanakorn, S.; and Seitz, S. M. 2014. Illumination-aware age progression. In *CVPR*, 3334–3341.

Li, J.; Zhao, J.; Zhao, F.; Liu, H.; Li, J.; Shen, S.; Feng, J.; and Sim, T. 2016a. Robust face recognition with deep multi-view representation learning. In *Proceedings of the 2016 ACM on Multimedia Conference*, 1068–1072. ACM.

Li, Z.; Gong, D.; Li, X.; and Tao, D. 2016b. Aging face recognition: a hierarchical learning model based on local patterns selection. *T-IP* 25(5):2146–2154.

Li, Z.; Park, U.; and Jain, A. K. 2011. A discriminative model for age invariant face recognition. *T-IFS* 6(3):1028–1037.

Lin, L.; Wang, G.; Zuo, W.; Feng, X.; and Zhang, L. 2017. Cross-domain visual matching via generalized similarity measure and feature learning. *IEEE transactions on pattern analysis and machine intelligence* 39(6):1089–1102.

Ling, H.; Soatto, S.; Ramanathan, N.; and Jacobs, D. W. 2010. Face verification across age progression using discriminative methods. *T-IFS* 5(1):82–91.

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

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

Park, U.; Tong, Y.; and Jain, A. K. 2010. Age-invariant face recognition. *T-PAMI* 32(5):947–954.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; et al. 2015. Deep face recognition. In *BMVC*, volume 1, 6.

Ramanathan, N., and Chellappa, R. 2006a. Face verification across age progression. *T-IP* 15(11):3349–3361.

Ramanathan, N., and Chellappa, R. 2006b. Modeling age progression in young faces. In *CVPR*, volume 1, 387–394.

Ramanathan, N., and Chellappa, R. 2008. Modeling shape and textural variations in aging faces. In *FG*, 1–8.

Ricanek, K., and Tesafaye, T. 2006. Morph: A longitudinal image database of normal adult age-progression. In *FGR*, 341–345.

Rothe, R.; Timofte, R.; and Gool, L. V. 2015. Dex: Deep expectation of apparent age from a single image. In *ICCVW*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Song, J.; Zhang, J.; Gao, L.; Liu, X.; and Shen, H. T. 2018. Dual conditional gans for face aging and rejuvenation. In *IJCAI*, 899–905.

Sungatullina, D.; Lu, J.; Wang, G.; and Moulin, P. 2013. Multiview discriminative learning for age-invariant face recognition. In *FG*, 1–6.

Suo, J.; Chen, X.; Shan, S.; Gao, W.; and Dai, Q. 2012. A concatenational graph evolution aging model. *T-PAMI* 34(11):2083–2096.

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

Wang, Y.; Zhang, Z.; Li, W.; and Jiang, F. 2012. Combining tensor space analysis and active appearance models for aging effect simulation on face images. *IEEE T SYST MAN CY B* 42(4):1107–1118.

Wang, W.; Cui, Z.; Yan, Y.; Feng, J.; Yan, S.; Shu, X.; and Sebe, N. 2016. Recurrent face aging. In *CVPR*, 2378–2386.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; and Liu, W. 2018a. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, 5265–5274.

Wang, Y.; Gong, D.; Zhou, Z.; Ji, X.; Wang, H.; Li, Z.; Liu, W.; and Zhang, T. 2018b. Orthogonal deep features decomposition for age-invariant face recognition. In *ECCV*.

Weinberger, K. Q., and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *JMLR* 10(Feb):207–244.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *ECCV*, 499–515.

Wen, Y.; Li, Z.; and Qiao, Y. 2016. Latent factor guided convolutional neural networks for age-invariant face recognition. In *CVPR*, 4893–4901.

Wu, X.; He, R.; Sun, Z.; and Tan, T. 2018. A light cnn for deep face representation with noisy labels. *T-IFS* 13(11):2884–2896.

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

Xu, C.; Liu, Q.; and Ye, M. 2017. Age invariant face recognition and retrieval by coupled auto-encoder networks. *Neurocomputing* 222:62–71.

Yang, H.; Huang, D.; Wang, Y.; Wang, H.; and Tang, Y. 2016. Face aging effect simulation using hidden factor analysis joint sparse representation. *T-IP* 25(6):2493–2507.

Zhang, Zhifei, S. Y., and Qi, H. 2017. Age progression/regression by conditional adversarial autoencoder. In *CVPR*.

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

Zhao, J.; Xiong, L.; Cheng, Y.; Cheng, Y.; Li, J.; Zhou, L.; Xu, Y.; Karlekar, J.; Pranata, S.; Shen, S.; Xing, J.; Yan, S.; and Feng, J. 2018. 3d-aided deep pose-invariant face recognition. In *IJCAI*, 1184–1190. International Joint Conferences on Artificial Intelligence Organization.

Zheng, T.; Deng, W.; and Hu, J. 2017. Age estimation guided convolutional neural network for age-invariant face recognition. In *CVPRW*, 12–16.

Zhu, H.; Zhou, Q.; Zhang, J.; and Wang, J. Z. 2018. Facial aging and rejuvenation by conditional multi-adversarial autoencoder with ordinal regression. *arXiv preprint arXiv:1804.02740*.