This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Disentangled Representation for Age-Invariant Face Recognition: A Mutual Information Minimization Perspective

Xuege Hou Yali Li Shengjin Wang*
Department of Electronic Engineering, Tsinghua University
hxg19@mails.tsinghua.edu.cn {liyali13, wgsgj}@tsinghua.edu.cn

Abstract

General face recognition has seen remarkable progress in recent years. However, large age gap still remains a big challenge due to significant alterations in facial appearance and bone structure. Disentanglement plays a key role in partitioning face representations into identity-dependent and age-dependent components for age-invariant face recognition (AIFR). In this paper we propose a multi-task learning framework based on mutual information minimization (MT-MIM), which casts the disentangled representation learning as an objective of information constraints. The method trains a disentanglement network to minimize mutual information between the identity component and age component of the face image from the same person, and reduce the effect of age variations during the identification process. For quantitative measure of the degree of disentanglement, we verify that mutual information can represent as metric. The resulting identity-dependent representations are used for age-invariant face recognition. We evaluate MT-MIM on popular public-domain face aging datasets (FG-NET, MORPH Album 2, CACD and AgeDB) and obtained significant improvements over previous state-of-the-art methods. Specifically, our method exceeds the baseline models by over 0.4% on MORPH Album 2, and over 0.7% on CACD subsets, which are impressive improvements at the high accuracy levels of above 99% and an average of 94%.

1. Introduction

Face recognition is one of the most widely and thoroughly studied topics in computer vision. From traditional methods [45, 36, 33, 28, 15, 49] to the more recent deep learning based algorithms [15, 51, 50, 47, 10, 59], it has achieved excellent performance, even surpassing humans in various scenarios.

Among all the face recognition systems, age-invariant face recognition has a wide range of application scenarios that are of great significance, including finding missing children after years, fugitive identification and passport verification. However, despite the exciting progress in face recognition, aging variation in these practical applications remains under-explored during the design and testing of face recognition systems. There are three challenging aspects for age-invariant face recognition (AIFR): 1) Aging related alterations in face appearance and anatomy result in significant intra-class variations, which increase as the age gaps get lager. 2) Face aging is a complex process affected dramatically by intrinsic and extrinsic factors (*i.e.* heredity, gender, environment). It has a compound impact on facial shapes and textures simultaneously, making age-invariant patterns learning difficult. 3) Current cross-age databases are insufficient for training with unbalanced distributions of age and gender, which limits the performance of AIFR.

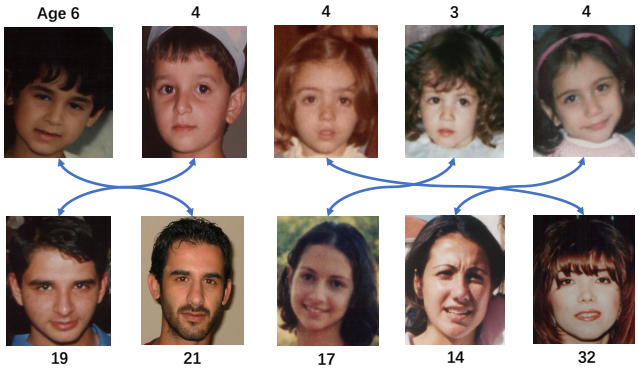

Figure 1. Example images for AIFR. The two ends of a line are the same person. It is common in AIFR where intra-class variations are greater than inter-class for large age gaps. Consequently, not only human get confused, but most general face recognition models also degrade by a scale of over 13% [5].

Figure 1 shows a typical scene of different individuals at the same age looking more similar than the same person at different age, for the reason that age-related information is shared among different identities, revealing the necessity of age variations disentanglement.

*Corresponding author

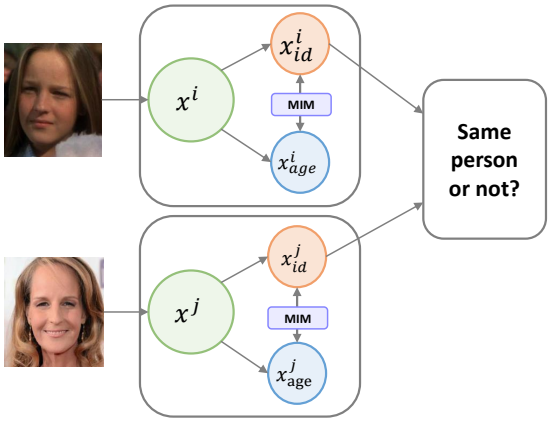

Figure 2. We propose to learn identity-dependent representation by minimizing mutual information between identity-dependent features and age-dependent features. Then the identity representation can be used for face recognition or verification.

Recent research on AIFR has three main model designs: generative models, discriminative models and the mixed models of the two. The generative methods [10, 59] propose to synthesize face images of different ages, then, perform recognition with the artificial representations. Benefited from the powerful GAN-based models, they have shown improvement on the quality of generated aging faces. However, the generative models still suffer from significant shortcomings. Firstly, it is fairly difficult and complicated for parametric generation models to fit the aging process, which can be easily affected by latent factors and varies from person to person. Secondly, the generation process is often unstable and will introduce additional noise. Furthermore, learning-based generation is often characterized by texture changes of faces, neglecting the shape changes, as a result of imbalanced training data.

Contrarily, recent work on discriminative methods are drawing increased interests [15, 51, 50, 47]. The discriminative models focus on the decomposition of facial representation and separates identity-dependent components for recognition. For example, [15] proposes the hidden factor analysis (HFA) to separate identity-related information and age-related information. [51] uses the linear combination of jointly-learned deep features to represent identity and age information, similar to the HFA based deep learning model. Another recent work using OE-CNN [50] deals with feature decomposition in an orthogonal way, achieving promising performance in AIFR. Works on discriminative method indicate the significance of facial representation disentanglement and the extraction of identity-dependent feature for age-invariant face recognition.

In this work, we introduce a novel age-invariant face recognition framework using mutual information minimization (MT-MIM), which disentangles the mixture of face features into two nearly independent components: identity-dependent component and age-dependent component. Figure 3 illustrates the architecture of the proposed MT-MIM. Compared to correlation coefficient, mutual information can capture the nonlinear statistical independence between variables, thus it can be used as a truly independent measure [22]. By minimizing mutual information of the two components, we focus more on the identity-efficient information during age-invariant feature learning, leading to improved recognition performance on images with large age variations.

To sum up, our major contributions of this work include:

1) A novel objective to learn age-invariant face representation by minimizing mutual information between identity-dependent component and age-dependent component.

2) We demonstrate the effectiveness of our proposed approach with several extensive experiments over four face aging datasets, including MORPH Album2 [38], CACD [6], FG-NET [24] and AgeDB [32].

2. Related Work

Age-Invariant Face Feature Learning. Conventional approaches either model the aging process with shape and texture transformations [37, 34], or leverage robust local descriptors [15, 14, 30, 27, 28] to compensate for face recognition degradation due to face aging. Former approaches rely much on prior biological knowledge and usually require massive tagged cross-age face data with long time elapse. For the latter approaches, for example, [14] developed a maximum entropy feature descriptor (MEFD) and a robust identity matching framework for AIFR.

Recent methods are mainly based on deep neural networks [10, 59, 51, 50, 47]. Steming from deep generative models, methods are proposed to synthesize face images of specific age and then do the comparison. For instance, [10] proposed an age-progression module that can age-progress deep face features. Age-Invariant Model (AIM) [60] jointly performed cross-age face synthesis and recognition end-to-end to mutually boost each other. On the other hand, disciminative methods have competitive performance with more concise strategy. For instance, OE-CNN [50] proposed an orthogonal decomposition of face features into identity-specific and age-specific components. [47] disentangled the two components through a Decorrelated Adversarial Learning framework (DAL) with linear residual decomposition. Compared to previous works, our work solve face representation disentanglement from a more essential perspective of mutual information, which reveals the intrinsic correlations of variables.

Disentangled Representation Learning. Disentangled representation learning aims to model the explanatory factors from diverse data variation, which is drawing considerable attention [58, 2, 44, 42, 18]. Previous recogni-

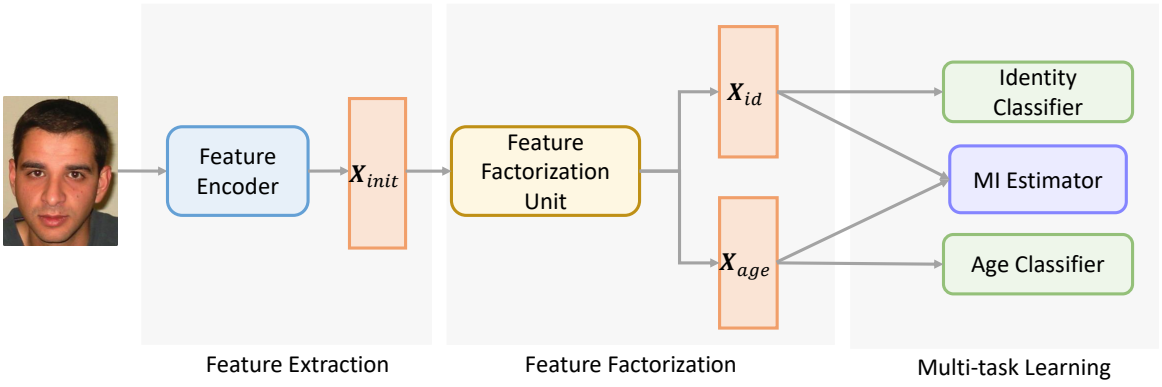

Figure 3. The overview of MT-MIM. It consists of two main processes. The initial features F_{init} are extracted by feature encoder, and then decomposed to the identity-related features x_{id} and age-related features x_{age} through the feature factorization unit. Finally, the batch of features are used for MI minimization regularization and classifications respectively.

tion works used tagged data to disentangle representations into identity-related and identity-independent information (age, pose, viewpoint and etc.). [18] minimized Wasserstein distance between cross-modality distributions, in order to learn invariant deep feature representations of heterogeneous face images. [56] improved single-modality person re-identification via extracting illumination-invariant features. Exploration has also been done for unsupervised settings [4, 40, 23, 8]. For instance, InfoGAN [8] disentangled the latent factors by maximizing the mutual information between hidden variables and data variables. [7] decomposed the variational lower bound to explain how β-VAE [19] works in learning disentangled representations and motivates the β-TCVAE algorithm. [40] used a coupled autoencoder to disentangle the appearance and geometry of face images. Each autoencoder learns one of the representations respectively under the supervision of reconstruction loss.

Mutual Information and Deep Learning. Previously, mutual information (MI) has been used to explain the deep learning frameworks [43, 41]. With breakthrough in MI estimation recently [8, 3, 9], it is utilized as regularizers or objectives to constrain independence between variables. [20, 1, 46]. Deep InfoMax [20] investigates unsupervised representation learning by maximizing MI between the input and output of a deep learning network. [1] proposed a self-supervised representation learning framework based on MI maximization of multi-views from a shared context. Moreover, MI minimization is drawing increasing attention in information bottleneck [13], disentangled representation learning [7] and various fields. Through preserving all information relevant to the label while minimizing the amount of others, [13] identified superfluous information not shared by different views.

3. MT-MIM

3.1. Problem Formulation and Motivation

We denote the input face image as $\mathbf{x}$, each image corresponds to an identity label y_{id} and an age label y_{age}. In the training stage, features $\mathbf{x}_{id}$ and $\mathbf{x}_{age}$ are extracted from the encoder E under the supervision of corresponding labels respectively. In the testing stage, only the identity features $\mathbf{x}_{id}$ are used for face recognition.

Simply put, the challenge of learning identity representations can be formulated as learning a distribution $p(\mathbf{x}_{id}|\mathbf{x})$ that maps input data into an identity representation. For age-invariant face recognition, the desired $\mathbf{x}_{id}$ is expected to be age-invariant while preserving the identity information. In this case, we consider only identity information that are discriminative enough to predict y_{id} and invariant to the age information, which can be restricted by the mutual information between the $\mathbf{x}_{id}$ and $\mathbf{x}_{age}$.

Mutual information excells in that, regardless of how nonlinear the dependence is, MI rigorously quantifies the amount of information one variable reveals about the other. Therefore, it exhibits true mutual dependence between variables in contrast to correlation [22].

On the other hand, in the in-depth explanation work of information towards deep neural networks [41], generalization through noise mechanism is considered unique to deep neural networks, which is achieved with information bottleneck strategy. Details are partially lost to obtain generalization. Motivated by this working mode, we believe the robustness of identity representations towards age variations can be obtained by the forgetting of related information. The more $I(\mathbf{x}_{id}; \mathbf{x}_{age})$ is reduced without violating information sufficiency for the identity prediction, which is guaranteed by the identity supervision task, the more robust identity representation is with age variations.

3.2. Mutual Information Estimation And Minimization

Mutual Information is a fundamental quantity that measures the dependence of two random variables. The MI between variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as:

$$I[X;Y] = \int dx\, dy\, p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where $p(x,y)$ is the joint probability distribution, while $p(x)$ and $p(y)$ are the marginals.

As we are interested in minimizing MI, the upper-bound estimation to MI is needed. Most previous works focused on lower-bound estimation [3, 8], however, they are inconsistent to MI minimization task.

Our basic MI minimization approach follows Contrastive Log-ratio Upper Bound (CLUB) [9], which estimates mutual information by the difference of conditional probabilities between positive and negative sample pairs. In our case of face representation disentanglement, the conditional distribution $p(\mathbf{x}_{age}|\mathbf{x}_{id})$ is not available, thus we use a variational distribution to approximate it,

$$\begin{aligned} I(\mathbf{x}_{id};\mathbf{x}_{age}) := \mathbb{E}_{p(\mathbf{x}_{id};\mathbf{x}_{age})}[log\ q_\sigma(\mathbf{x}_{age}|\mathbf{x}_{id})] \\ - \mathbb{E}_{p(\mathbf{x}_{id})}\mathbb{E}_{p(\mathbf{x}_{age})}[log\ q_\sigma(\mathbf{x}_{age}|\mathbf{x}_{id})] \end{aligned} \tag{2}$$

where $q_\sigma(\mathbf{x}_{age}|\mathbf{x}_{id})$ is the variational distribution modeled by a neural network Q with parameters σ.

To encourage the dependence between variables in feature pairs $\{(\mathbf{x}_{age},\mathbf{x}_{id})\}_N^{i=1}$, we have the following MI minimization objective function for the feature encoder E:

$$\begin{aligned} \min_E L_{MIM} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}[log\ q_\theta(\mathbf{x}_{age}^i|\mathbf{x}_{id}^i) \\ - log\ q_\theta(\mathbf{x}_{age}^j|\mathbf{x}_{id}^i)] \end{aligned} \tag{3}$$

At the same time, Q is trained by minimizing the KL-divergence between true conditional distribution $p(\mathbf{x}_{age}|\mathbf{x}_{id})$ and the variational distribution $q_\sigma(\mathbf{x}_{age}|\mathbf{x}_{id})$:

$$\min_Q L_{KL} = KL(p(\mathbf{x}_{age}|\mathbf{x}_{id}) \parallel q_\sigma(\mathbf{x}_{age}|\mathbf{x}_{id})) \tag{4}$$

We can easily derive that Eq. 4 is equivalent to the maximization of of $\mathbb{E}_{p(\mathbf{x}_{id},\mathbf{x}_{age})}[logq_\sigma(\mathbf{x}_{age}|\mathbf{x}_{id})]$ and maximize the log-likelihood as an unbiased estimation.

For the training of MI estimator Q, at each training iteration, we first obtain a batch of samples $\left\{(\mathbf{x}_{id}^i,\mathbf{x}_{age}^i)\right\}$ from the feature encoder, then we update the MI estimator. After the update, we calculate the MI estimator by Eq. 3 and back propogate to the parameters of feature encoder.

3.3. Representation Factorization

There is an observation that facial images of the same person contain intrinsic information enduring through age, which is distinct from person to person. Whereas, different person at the same age often share similar characteristics, for example, the condition of skin, wrinkles and spots. In facial representations, the two components, age-related information and identity-related information, though somehow entangled, are essentially independent.

Motivated by the observations, we model the identity-related information and age-related information as statistically independent variables, called *age information* and *id information*, represented by vectors $\mathbf{x}_{id}$ and $\mathbf{x}_{age}$. For simplicity, we consider a linear factorization of the two components. We obtain the age-related features through a FC layer from the initial features, and the identity-related features are obtained from the subtract between initial features and age-related features. Denote the FC layer as function ϕ, and $\mathbf{x} \in \mathbb{R}^d$ representing the initial feature extracted from the face image by the encoder E.then the factorization can be formulated as:

$$\mathbf{x_{id}} = \mathbf{x} + \phi(\mathbf{x}_{age}), \tag{5}$$

We refer to this factorization as the feature factorization unit, which is a simple operation to decompose the initial features into identity-related and age-related features simultaneously. Upon the factorization, we can carry out the following multi-task learning.

3.4. Multi-task Learning Framework

As shown in figure 3, AIFR is accomplished with a multi-task learning framework. It has three modules: identity discriminator, age discriminator and MI Estimator.

Age Discriminator. To train the age discriminator in MT-MIM, we partition the training datasets into several age groups, similar to previous works [15, 51]. Depending on the amount of training data, the number of age groups ranges from 8 to 10. Each age group contains approximately the same amount of data to balance the samples. We also explored age classification age by age, but the performance is not as good as the former. We believe it is for the reason that age labels are rough with noises. Softmax layer with cross-entropy loss is used here for age classification.

Identity Discriminator. Considering the significant performance of margin based methods [48, 11] in general face recognition, to well preserve the identity information, we utilize the ArcFace loss [11] to learn $\mathbf{x}_{id}$. The loss function is formulated as:

$$\mathcal{L}_{ID} = \frac{1}{N}\sum_{i}^{N} -\log \frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j\neq y_i} e^{s\cos(\theta_j)}} \tag{6}$$

in which the N is the number of identities, $\cos(\theta_j) = \frac{|\mathbf{x}_{id}^i \cdot \mathbf{W}_j|}{||\mathbf{x}_{id}^i||_2^2 \cdot ||\mathbf{W}_j||_2^2}$ is the cosine of angle between the i-th feature $\mathbf{x}_{id}^i$ and label y_i's weight vector, $m \geq 1$ is an integer

hyper-parameter that controls the size of angular margin, and $s > 0$ is an adjustable scaling factor. The ArcFace loss is used to constrain the identity-related features and encourage the feature discrimination. What's more, it ensures the sufficiency of identity information for the recognition with a relatively high weight coefficient.

MI Estimator. The MI Estimator is operated as a disentanglement constraint to reduce the mutual information between identity-related features and age-related features. Under the joint supervision of multitasks, x_{id} is encouraged to be discriminative and age-invariant.

To sum up, the multitask learning loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{ID} + \lambda_{AGE}\mathcal{L}_{AGE} + \lambda_{MIM}\mathcal{L}_{MIM} \tag{7}$$

where $\mathcal{L}_{AGE}$ denotes the cross-entropy loss for age task, λ_{AGE}, λ_{MIM} are hyper-parameters to balance the losses.

The details of training MT-MIM is summarized in Algorithm 14.

Algorithm 1 Training the MT-MIM framework

Require: Training set $\{x^i\}_{i=1}^N$, learning rate γ, lagrange multipliers λ_i and MI training iterations N_{MI}.
Ensure: Then encoder parameters Θ.

```
Initialize parameters Θ by pre-trained model;
for each training epoch do
    CNN optimization:
        Encoder forward: E(x^i)
        Compute identity loss by Eq. 6;
        Compute age loss with cross-entropy loss;
        Compute MI loss by Eq. 3;
        Update Θ via back-propagation method;
    MI estimator Optimization (Θ fixed):
    for k = 1 to N_MI do
        Update MI estimator parameters σ by maximizing Eq. 4;
    end for
end for
return Θ;
```

4. Experiments

4.1. Experimental Settings

Data Preprocessing. We detect all training and testing sets by MTCNN [57], and perform similarity transformation according to the five landmarks (two eyes, nose and mouth corners). After face alignment, all faces are cropped to 112×112 RGB images. Finally, each pixel of the processed faces are normalized by subtracting 127.5 and divided by 128.

CNN Architecture. 1) Backbone: for the sake of fairness, all the CNN models in the experiments follow the same typical ResNet50 architecture [17]. It has four residual blocks and outputs a 512-dimensional feature vector by a FC layer. 2) Residual factorization Unit: the age-dependent features are mapped from the initial face features through a FC layer, and the identity-dependent features are derived from the residual part. 3) Age discriminator: the extracted $\mathbf{x}_{age}$ are mapped through a FC layer and performed age classification. 4) Identity discriminator: the extracted $\mathbf{x}_{id}$ are used for classification by ArcFace loss. 5) MI estimator: with a batch of samples $\{(\mathbf{x}_{id}^i, \mathbf{x}_{age}^i)\}$ as input, MI estimator are used to calculate the MI between the two variables for optimization.

Training Details. We conducted experiments on several widely used AIFR datasets: MORPH Album 2, CACD, FG-NET and AgeDB. We first train the deep model on the wild datasets to learn basic knowledge about human faces. The training data includes Ms-Celeb-1M [16] and CASIA-Webface [12], which we refer to as general face datasets (GFD) in the following text. Ms-Celeb-1M contains about 1M images from 100K individuals while CASIA-Webface contains nearly 0.5M images from 10K individuals. We clean the data for their noises [53]. Then we finetune the proposed model using experimental datasets. The age labels are divided into 8 to 10 groups for data balancing. The grouped age labels are then used for age classification.

The MT-MIM training process is jointly supervised by Eq. 7. Specifically, the training of feature encoder and MI estimator are operated in an alternative manner. In a training epoch, we perform forward of encoder once then MI estimator optimization 5 times. The experimentally setting of hyper-parameters in Eq. 7 are $\lambda_{AGE} = 0.1$, $\lambda_{MIM} = 0.0001$, $m = 0.35$ and $s = 64$. All models are trained through adaptive moment estimation (Adam) on 4 Tesla V100 GPUs parallelly, and the batch size is set to occupy half of or nearly all the GPU memory. Specifically, the batch size is 400 in datasets excluding FG-NET, and 1001 for the leave-one-out training scheme for FG-NET. The learning rate of the finetuned encoder begins with 0.02 and is divided by 10 when the loss does not decrease. Whereas, the learning rate of MI estimator is initially set to 1e-5 and degrades the same as the former.

Testing Details. We conduct evaluation experiments on public AIFR datasets: FG-NET, MORPH Album 2, CACD and AgeDB. In the testing phase, we concatenate identity-dependent features extracted from original image and its flipped image for recognition. The cosine similarity of these identity representations are then used to conduct face verification and identification.

Metrics. Apart from Rank-1 verification rate, mean average precision (MAP) is used as evaluation metrics in CACD dataset. For the retrieval results of each query image, precision at every recall level is computed and averaged to get average precision (AP). MAP is then calculated over

the whole query set Q, formulated as follows:

$$MAP(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{m_i}\sum_{i=1}^{m_i} Precision(R_{ik}) \quad (8)$$

where R_{ik} is the retrieval results of $q_i \in Q$ in descending order from the first image to the k-th image, and $Precision(R_{ik})$ is the ratio of positive images in R_{ik}.

4.2. Experiments on AIFR Datasets

On MORPH Dataset. MORPH is a large-scale public longitudinal face database . The Album 2 has two version for commercial and non-commercial use, which have almost identical data distribution and are used alternately in previous works. The non-commercial version contains more than 55,000 images of 13,000 individuals with age ranging from 17 to 77, while the version for commercial use contains over 78,000 face images of 20,000 individuals. There are two benchmark settings where the testing set consists of 10,000 subjects and 3,000 subjects respectively. Note that the dataset we use here is the non-commercial version with 13,000 individuals, thus we follow the testing scheme in [29] to divide the dataset into two parts. One part, including 10,000 individuals, is used to fine-tune the proposed MT-MIM and the remaining 3,000 individuals are used for evaluating. There is no overlap between these two parts. In the testing set, 2 images each subject with the largest age gap are selected to form the probe set and the gallery set.

The recognition result is evaluated with Rank-1 identification rate. As shown in Table 1, the MT-MIM has effectively improved the rank-1 identification performance of MORPH Album 2. Particularly, with less AIFR training data to fine-tune the model, our method surpasses the AIM model by 0.6%, which is an outstanding improvement on the accuracy level above 98%.

We also show some examples of failed retrievals in Figure 4. While the rank-1 retrievals are not correct in these cases, the probe images appear to be more similar to the incorrect rank-1 retrievals than the correct images.

On CACD Dataset. CACD is a large-scale dataset for face recognition and retrieval across ages, collected in the wild with diverse variations. It contains 163,336 face images from 2,000 celebrities ranging from 16 to 62 years old. Following the experimental setting in [6], 1880 celebrities are used to fine-tune the MT-MIM, while the left 120 are used for testing. Among them, images taken in 2013 are used as query images, and the remaining images taken in 2004-2006, 2007-2009 and 2010-2012 are partitioned into three groups as database images.

Table 2 shows the retrieval results on CACD compare to other state-of-the-art methods. The baseline model outperforms the existing methods, still, our method has an obvious performance boosting over an accuracy level above an average of 94%, consistently showing effectiveness across different years. We also visualize the verification results for CACD to gain insights into AIFR with MT-MIM. Figure 5 shows a few examples of the probe and reference pairs. Cosine similarities are shown between the pairs. It can be observed that MT-MIM has an stable performance on pairs with various age gaps. Moreover, the proposed method performs favorably on pairs from different persons of different age but look very similar.

Method	Setting-1/Setting-2
HFA (2013) [15]	91.14/-
CARC (2014) [6]	92.80/-
MEFA (2015) [14]	93.80/-
MEFA+SIFT+MLBP (2015) [14]	94.59/-
LF-CNN (2016) [51]	97.51/-
GSM (2017) [29]	-/94.40
AE-CNN (2017) [61]	-/98.13
OE-CNN (2018) [50]	98.55/98.67
DAL (2019) [47]	98.93/98.97
AIM (2019) [59]	99.13/98.81
AIM [59] + CAFR	99.65/99.26
MT-MIM	**-/99.43**

Table 1. Rank-1 accuracy (%) comparisons on MORPH Album 2.

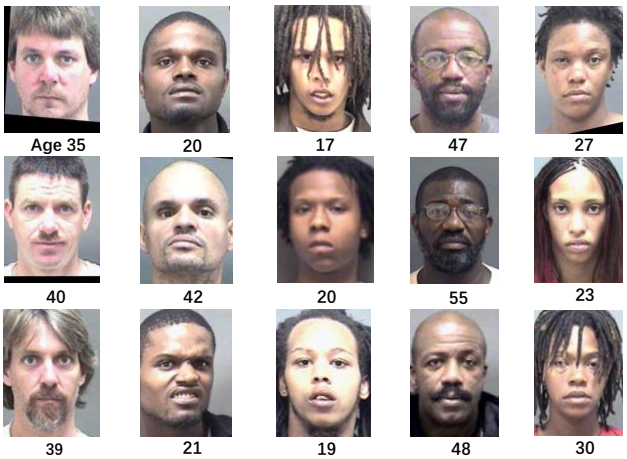

Figure 4. Some examples of failed retrievals in MORPH Album 2. The first row represents the probe images. The second row are the incorrect retrievals using our approach. The third row shows the corresponding gallery images for the probe images.

On FGNET Dataset. FG-NET [24] is a popular public dataset for cross-age face recognition, collected in the wild with huge variability in age covering from child to the elder. It contains 1002 face images from 82 individuals, with age ranges from 0 to 69. We follow the leave-one-out setting the same as [15, 28] for fair comparisions with previous methods. Specifically, we leave one image as testing sample

Method	2004-2006	2007-2009	2010-2012
HFA (2013) [15]	50.58	53.01	56.12
CARC (2014) [6]	52.72	55.48	61.38
GSM-1 (2017) [29]	53.79	57.83	63.92
GSM-2 (2017) [29]	55.45	58.74	64.58
CAN (2017) [54]	62.33	67.69	73.24
AE-CNN (2017) [61]	70.01	72.87	78.25
JM-CNN (2018) [55]	82.53	85.26	88.28
CNN baseline (Pretrained with GFD)	68.21	71.46	76.79
CNN baseline (fine-tuned by CACD)	91.81	93.27	95.35
MT-MIM	**92.63**	**93.95**	**96.09**

Table 2. Comparison on CACD dataset with existing methods.

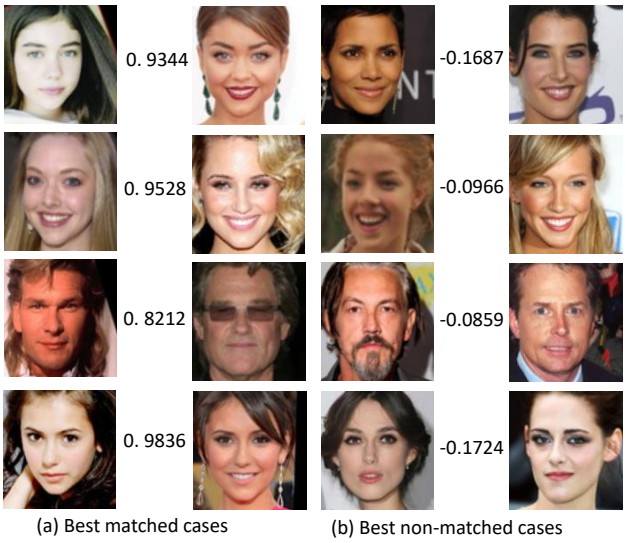

Figure 5. Some examples of matched and non-matched pairs in CACD using MT-MIM.

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 MT-MIM with other state-of-the-arts on FG-NET is reported in Table 3. The proposed method improves the 2nd best by 1 %. It is suggested that the model can be further improved with more AIFR data [50, 59], revealing the promising potential of our method for unconstrained face recognition with age variance.

On AgeDB Dataset. AgeDB [32] is an in-the-wild database containing 16,488 face images of 568 individuals with manually annotated age labels. It provides four protocols for age-invariant face verification under different age gaps of face pairs: 5, 10, 20, and 30 years. Similar to LFW [21], this dataset is split into 10 folds for each pro-

Method	Rank-1 (%)
Park *et al.* (2010) [33]	37.40
Li *et al.* (2011) [28]	47.50
HFA (2013) [15]	69.00
MEFA (2015) [14]	76.20
LF-CNN (2016) [51]	88.10
CAN (2017) [54]	86.50
DAL (2019) [47]	**94.50**
AIM (2019) [59]	93.20
MT-MIM	94.21

Table 3. FG-NET results under the leave-one-out protocol.

Method	Rank-1 (%)
VGG Face (2015) [35]	89.89
Center Loss (2016) [52]	93.72
RJIVE (2017) [39]	55.20
SphereFace (2017) [31]	91.70
CosFace (2018) [48]	94.56
ArcFace (2019) [11]	95.15
DAAE (2020) [26]	95.30
MT-MIM	**96.10**

Table 4. Performance comparisons on AgeDB-30.

tocol, with each fold consisting of 300 intra-class and 300 inter-class pairs. We strictly follow the protocol of 30 years to perform the 10-fold cross validation, in order to confirm the effectiveness of our method. Table 4.2 shows the Rank-1 verification performance of MT-MIM compared with the other most recent AIFR methods, demonstrating the competitive performance of the proposed method.

4.3. Ablation Study

To show the effectiveness of the proposed MT-MIM method, we conduct the ablative evaluations on several public AIFR datasets, including FG-NET, MORPH Album 2, CACD and AgeDB-30. The following variants of our method are considered: 1) Baseline: the baseline model is trained by the ArcFace loss only, without any other supervision. It is pretrained with the general face datasets (GFD) and then fine-tuned by the specified training sets corresponding to the AIFR testing sets. We denote the fine-tuned one as our baseline model without lose of generality, comparing to models trained by AIFR datasets in other methods. 2) +Age: this model is trained by the joint supervision of identity label and age label based on the pretrained model, as a comparison to the baseline. Without task relations modeling, the performance of jointly learned tasks is not boosted [25]. 3) MT-MIM: the proposed model trained under the joint supervisions with MI minimization constrain based on the pretrained model. As shown in table 4.3, the

Model	MORPH Album 2	CACD (2004-2006)	CACD (2007-2009)	CACD (2010-2012)	FG-NET leave-one-out	FG-NET (MF2)	AgeDB-30
Baseline	99.00	91.81	93.27	95.35	93.20	60.26	95.11
+Age	99.10	91.87	93.15	95.44	93.40	60.04	95.40
MT-MIM	**99.43**	**92.63**	**93.95**	**96.09**	**94.21**	**61.12**	**96.10**

Table 5. Comparison of MT-MIM against the baseline models.

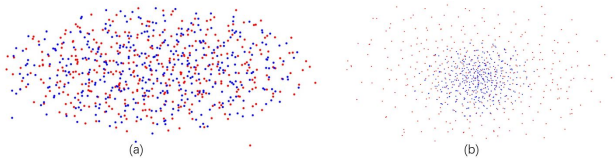

Figure 6. (a) to (b): $\mathbf{x}_{age}$ (blue dots) disentangles from $\mathbf{x}_{id}$ (red dots) during training.

Iterations	CACD (2004-2006)	CACD (2007-2009)	CACD (2010-2012)
19100	91.46	92.93	95.22
26740	92.28	93.55	95.64
32800	**92.31**	**93.62**	**95.68**

Table 6. Performance dynamic during training on CACD.

baseline models obtain a comparable performance without the training of MT-MIM. Nevertheless, our MT-MIM improves the performance of the baseline by a considerable scale. The improvements are more than 0.4% on MORPH Album 2, about 1% on AgeDB-30, which are remarkable improvements at the high accuracy levels above 99% and 95%. What's more, MT-MIM also has substantial upgrades on CACD and FG-NET by a clear extent. We believe that with more age variations of the training dataset, the performance of the proposed method can be further improved.

For a better understanding of the mechanism of MT-MIM, we visualize the training details. Figure 6 has shown MI changes of identity-dependent and age-dependent features during training process. As shown in Figure 6, compared against the "w/o MI loss" model, the mutual information of MT-MIM model is fairly suppressed between the identity-dependent and age-dependent features. The retrieval results during training process of MT-MIM has also corroborate the effectiveness of the proposed method. As is reported in Figure 6, the retrieval performance on CACD has a constantly promote as the decrease of MI between the identity-dependent features and the age-dependent features, corresponding to the phenomenon in Figure 6. These observation prove that our method encourages the deduce of age information within identity-dependent features, thus the identity-dependent features can be more robust to age variance. With the joint supervision of ArcFace loss, the MI minimization can substantially improve the discriminating power of the learned identity features.

	Method	Rank-1 (%)
General Approaches	Center Loss (2016) [52]	99.28
	SphereFace (2017) [31]	99.42
	CosFace (2018) [48]	99.33
Cross-Age Approaches	LF-CNN (2016) [51]	99.10
	OE-CNN (2018) [50]	99.35
	DAL (2019) [47]	99.47
	MT-MIM	99.25

Table 7. Performance comparisons on LFW.

4.4. Experiments on LFW

Labled Faces in the Wild (LFW) [21] contains 13,233 face images of 5,749 identities obtained from the Internet. The face recognition performance comparison of the proposed MT-MIM with other state-of-the-art method on LFW is reported in Table 4.4. MT-MIM has comparable performance on LFW dataset with other methods, which verified the generalization of MT-MIM for general face recognition.

5. Conclusion

We proposed the multi-task learning framework based on mutual information minimization (MT-MIM), which disentangle face representations by minimizing mutual information between identity-and age-dependent component. As far as we know, this is the first work to introduce mutual information disentanglement feature learning to AIFR. In the testing phase, only the identity features were used for face recognition. The evaluations conducted on the public AIFR benchmarks demonstrate the effectiveness of our proposed method.

Acknowledge This work was supported by the National Natural Science Foundation of China under Grant No. 61771288, the state key development program in 13th Five-Year under Grant No. 044007008, Cross-Media Intelligent Technology Project of Beijing National Research Center for Information Science and Technology (BNRist) under Grant No. BNR2019TD01022 and the research fund under Grant No. 2019GQG0001 from the Institute for Guo Qiang, Tsinghua University.

References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019. 3

[2] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Int. Conf. Mach. Learn.*, 2018. 3, 4

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013. 3

[5] B. Chen, C. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Trans. Multimedia*, 2015. 1

[6] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Eur. Conf. Comput. Vis.*, 2014. 2, 6, 7

[7] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018. 3

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016. 3, 4

[9] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *Int. Conf. Mach. Learn.*, 2020. 3, 4

[10] Debayan Deb, Divyansh Aggarwal, and Anil K Jain. Finding missing children: Aging deep face features. *arXiv preprint arXiv:1911.07538*, 2019. 1, 2

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 4, 7

[12] Y. Dong, L. Zhen, S. Liao, and S. Z. Li. Learning face representation from scratch. *Computer Science*, 2014. 5

[13] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020. 3

[14] D. Gong, Z. Li, Dacheng Tao, J. Liu, and Xuelong Li. A maximum entropy feature descriptor for age invariant face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 2, 6, 7

[15] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang. Hidden factor analysis for age invariant face recognition. In *Int. Conf. Comput. Vis.*, 2013. 1, 2, 4, 6, 7

[16] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *Eur. Conf. Comput. Vis.*, 2016. 5

[17] K. He, X. Zhang, S. Ren, and S. Jian. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 5

[18] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018. 2, 3

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *Int. Conf. Learn. Represent.*, 2017. 3

[20] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 3

[21] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 7, 8

[22] Justin B Kinney and Gurinder S Atwal. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences*, 111(9):3354–3359, 2014. 2, 3

[23] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017. 3

[24] Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002. 2, 6

[25] Jianshu Li, Pan Zhou, Yunpeng Chen, Jian Zhao, Sujoy Roy, Yan Shuicheng, Jiashi Feng, and Terence Sim. Task relation networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 932–940. IEEE, 2019. 7

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

[27] Zhifeng Li, Dihong Gong, Xuelong Li, and Dacheng Tao. Aging face recognition: a hierarchical learning model based on local patterns selection. *IEEE Trans. Image Process.*, 2016. 2

[28] Z. Li, U. Park, and A. K. Jain. A discriminative model for age invariant face recognition. *IEEE Transactions on Information Forensics and Security*, 2011. 1, 2, 6, 7

[29] Liang Lin, Guangrun Wang, Wangmeng Zuo, Xiangchu Feng, and Lei Zhang. Cross-domain visual matching via generalized similarity measure and feature learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016. 6, 7

[30] Haibin Ling, Stefano Soatto, Narayanan Ramanathan, and David W Jacobs. Face verification across age progression using discriminative methods. *IEEE Transactions on Information Forensics and security*, 2009. 2

[31] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding

for face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 7, 8

[32] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2017. 2, 7

[33] U. Park, Y. Tong, and A. K. Jain. Age-invariant face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010. 1, 7

[34] Unsang Park, Yiying Tong, and Anil K Jain. Age-invariant face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010. 2

[35] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. *Brit. Mach. Vis. Conf.*, 2015. 7

[36] Narayanan Ramanathan and Rama Chellappa. Face verification across age progression. *IEEE Trans. Image Process.*, 2006. 1

[37] Narayanan Ramanathan and Rama Chellappa. Modeling shape and textural variations in aging faces. In *IEEE Int. Conf. Auto. Face & Gesture. Recog.*, 2008. 2

[38] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *IEEE Int. Conf. Auto. Face & Gesture. Recog.*, 2006. 2

[39] Christos Sagonas, Evangelos Ververas, Yannis Panagakis, and Stefanos Zafeiriou. Recovering joint and individual components in facial data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. 7

[40] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Eur. Conf. Comput. Vis.*, 2018. 3

[41] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017. 3

[42] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[43] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015. 3

[44] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[45] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 1991. 1

[46] Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *Int. Conf. Learn. Represent.*, 2019. 3

[47] Hao Wang, Dihong Gong, Zhifeng Li, and Wei Liu. Decorrelated adversarial learning for age-invariant face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2, 6, 7, 8

[48] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 4, 7, 8

[49] Xiaogang Wang and Xiaoou Tang. A unified framework for subspace face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2004. 1

[50] Yitong Wang, Dihong Gong, Zheng Zhou, Xing Ji, Hao Wang, Zhifeng Li, Wei Liu, and Tong Zhang. Orthogonal deep features decomposition for age-invariant face recognition. In *Eur. Conf. Comput. Vis.*, 2018. 1, 2, 6, 7, 8

[51] Y. Wen, Z. Li, and Y. Qiao. Latent factor guided convolutional neural networks for age-invariant face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1, 2, 4, 6, 7, 8

[52] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Eur. Conf. Comput. Vis.*, 2016. 7, 8

[53] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 2018. 5

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

[55] Jinbiao Yu and Liping Jing. A joint multi-task cnn for cross-age face recognition. In *IEEE Int. Conf. Image Process.*, 2018. 7

[56] Zelong Zeng, Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Illumination-adaptive person re-identification. *IEEE Trans. Multimedia*, 2020. 3

[57] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Sign. Process. Letters*, 2016. 5

[58] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[59] Jian Zhao, Yu Cheng, Yi Cheng, Yang Yang, Fang Zhao, Jianshu Li, Hengzhu Liu, Shuicheng Yan, and Jiashi Feng. Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. *AAAI*, 2019. 1, 2, 6, 7

[60] Jian Zhao, Shuicheng Yan, and Jiashi Feng. Towards age-invariant face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 2

[61] T. Zheng, W. Deng, and J. Hu. Age estimation guided convolutional neural network for age-invariant face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2017. 6, 7