Pattern Recognition 100 (2020) 107178

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

# Deep label refinement for age estimation

Peipei Li[a,b,c], Yibo Hu[a,b], Xiang Wu[a,b], Ran He[a,b,c,*], Zhenan Sun[a,b,c]

[a] Center for Research on Intelligent Perception and Computing, CASIA, Beijing, China 100190
[b] National Laboratory of Pattern Recognition, CASIA, Beijing, China 100190
[c] University of Chinese Academy of Sciences, Beijing, China 100049

ARTICLE INFO

*Article history:*
Received 29 November 2018
Revised 24 November 2019
Accepted 15 December 2019
Available online 24 December 2019

*Keywords:*
Age estimation
Deep learning
Convolutional neural networks
Label distribution learning

ABSTRACT

Age estimation of unknown persons is a challenging pattern analysis task due to the lack of training data and various ageing mechanisms for different individuals. Label distribution learning-based methods usually make distribution assumptions to simplify age estimation. However, since different genders, races and/or any other characteristics may influence facial ageing, age-label distributions are often complicated and difficult to model parametrically. In this paper, we propose a label refinery network (LRN) with two concurrent processes: label distribution refinement and slack regression refinement. The label refinery network aims to learn age-label distributions progressively in an iterative manner. In this way, we can adaptively obtain the specific age-label distributions for different facial images without making strong assumptions on the fixed distribution formulations. To further utilize the correlations among age labels, we propose a slack regression refinery to convert the age-label regression model into an age-interval regression model. Extensive experiments on three popular datasets, namely, MORPH Album2, ChaLearn15 and MegaAge-Asian, demonstrate the superiority of our method.



## 1. Introduction

Recently, age estimation has attracted much attention in many real-world applications, such as video surveillance, product recommendations, and internet safety for minors. It aims to label a given face image with an exact age or age group. Since significant intra-class variations in facial appearances caused by the ageing process exist, a cross-age facial task is a cross-domain problem. It is important to represent age features accurately and robustly [1,2]. In the last several decades, impressive progress [3,4] has been made on age estimation. However, it is still a challenging problem for the following reasons: 1) humans with different genders and/or races variances age in different ways [5,6]. 2) Considering facial images, many large variations, including illumination, poses, expressions, and makeup, would affect the accuracy of age estimation. 3) It is hard to obtain an accurate ground-truth age for the training data.

Existing age estimation methods can be roughly divided into four categories: regression [7,8], multi-class classification [4,9], ranking convolutional neural networks (ranking-CNNs) [10], and label distribution learning (LDL) methods [11,12]. By estimating the age distribution, label distribution learning has the potential benefits of addressing the relevance and uncertainty among different ages. In addition, label distribution learning can improve data utilization because each facial image provides information not only about the chronological age but also its neighbouring ages.

However, label distribution learning for age estimation still faces two major challenges. First, we argue that age-label distributions are often complicated and vary across individuals; thus, it is inappropriate to make assumptions on their distributions [12–14]. Fig. 1 depicts the detailed interpretation of this concept. In Fig. 1(a), we can observe that the ageing tendencies are different at different ages. The ageing rate for younger and older people are fast, while it tends to be relatively slow for middle-aged people [15,16]. Thus, it is unreasonable to assume that the age-label distributions for all ages obey Gaussian distributions with the same standard deviation, as shown in Fig. 1(b). Therefore, some researchers propose assigning Gaussian distributions with different standard deviations to facial images of different ages [15]. However, since ageing is a unidirectional and irreversible process, humans of different genders and/or races have different facial characteristics. In addition, considering facial images, many variations, including illumination, pose, expressions, and makeup, can also lead to different age distributions. Therefore, it is unreasonable to assume that the age-label distribution is a symmetric Gaussian distribution, as shown in Fig. 1(c). The second challenge is that the age-inference process of label distribution learning is often

* Corresponding author at: National Laboratory of Pattern Recognition, CASIA, Beijing, China 100190.
*E-mail addresses:* peipei.li@cripac.ia.ac.cn (P. Li), yibo.hu@cripac.ia.ac.cn (Y. Hu), alfredxiangwu@gmail.com (X. Wu), rhe@nlpr.ia.ac.cn (R. He), znsun@nlpr.ia.ac.cn (Z. Sun).

https://doi.org/10.1016/j.patcog.2019.107178
0031-3203/


**Fig. 1.** Different label distribution assumptions for age estimation. (a) The ageing speeds for younger and older people are faster than that for middle-aged people. (b) Assumption that the age-label distribution obeys $X \sim N(\mu, \sigma^2)$, where $\sigma$ is the same for all ages. (c) Assumption that the age-label distribution obeys $X \sim N(\mu, \sigma^2)$, where $\sigma$ is different at different ages. (d) Learned distribution $X$ by the proposed method.

regarded as a classification task in which the relationships among the age classes are not considered. For example, given a facial image whose ground-truth age is 35 years old, the final predicted age of 3 or 34 years old is the same as the inference stage.

To address the first challenge, we propose a label distribution refinery that is similar in form to knowledge distillation [17,18]. The proposed label distribution refinery utilizes a convolutional neural network to adaptively learn age-label distributions from the given facial images and constantly refine the learning results during refinement. Fig. 1(d) shows the learned distribution. It is clear that the age-label distributions vary from different individuals and are asymmetric. For the second challenge of label distribution learning, we propose a concurrent training mechanism to jointly perform label distribution learning and regression. The regression model can capture correlations among age labels and regress on the age value, which ameliorates the second challenge. In addition, a slack term is designated to further convert the age-label regression model into an age-interval regression model.

The main contributions of this work are as follows:

(1) In this paper, we propose a label refinery network (LRN) with two concurrent processes: label distribution refinement and slack regression refinement.

(2) The proposed label distribution refinery adaptively estimates the age distributions without making strong assumptions about the form of the label distribution. Benefiting from the constant refinement of the learning results, the label distribution refinery generates more precise label distributions.

(3) To further utilize the correlations among different age labels, we introduce regression to assist label distribution refinery. In addition, we introduce a slack term to further convert the age-label regression model into an age-interval regression model.

(4) We evaluate the effectiveness of the proposed LRN on three age estimation benchmarks and consistently obtain state-of-the-art results.

## 2. Related work

### 2.1. Age estimation

Considering hand-crafted features, one of the earliest automatic age estimation methods based on facial images can be traced back to 1994 and was proposed by Kwon and Lobo [19]. Over the past few years, many researchers have focused on facial age estimation. In the early stage, age estimation frameworks typically contained two main stages: ageing feature representation and age inference. Many ageing-feature representation approaches were proposed, which can be roughly divided into geometric facial representation-based methods [20,21], appearance-based methods [22–24], and subspace learning-based methods [25–27]. Hybrid features, which are a combination of global and local features [28], have gained much attention in age estimation. To obtain a highly discriminative feature representation for age estimation, Pontes et al. [29] integrates active appearance models (AAMs), local binary patterns (LBP), Gabor wavelets (GW), and local phase quantization (LPQ)

in an unified framework. Age inference approaches fall into three categories: classification-based methods [30,31], regression-based methods [22,23,32] and a combination of these two age estimation methods [26]. To capture the complicated ageing process, Chao et al. [33] proposes an age-oriented local regression algorithm.

Benefiting from deep convolutional neural networks (e.g., VGG-16 [34], LightCNN [35], ResNet [36] and DenseNet [37]) trained on large-scale age face datasets, the deep learning-based age estimation methods achieve promising performance on age estimation, which can be roughly divided into four categories: regression [7,8], multi-class classification [4,9], ranking convolutional neural network (ranking-CNN) [10], as well as label distribution learning (LDL) methods [11,12]. Since each age (age group) can be regarded as a category, some researchers propose transforming age estimation into a multi-classification problem in which different ages (age groups) are regarded as independent classes. For example, Liu et al. [38] splits ordinal ages into a set of discrete groups and proposes a group-aware deep feature learning approach for age estimation. However, multi-class classification methods usually neglect the relevance and uncertainty among the neighbouring labels. To better fit the ageing mechanism, a natural idea is to treat age estimation as a regression task since there is a correlation among different ages. As mentioned in [4,39,40], due to the presence of outliers, regression methods cannot achieve satisfactory results either. Recently, ranking-CNN and LDL methods achieve state-of-the-art performance on age estimation, in which an individual classifier or label distribution for each age class is adopted. In this paper, we employ the LDL-based method assisted with regression.

### 2.2. Label distribution learning

Label ambiguity and redundancy hinder improvement in object recognition and classification performance. Label distribution learning [41,42] aims to address this problem by learning the distribution over each label from the description of the instance. Label distribution learning has been widely used in many applications, such as expression recognition [43], public video surveillance [44], and age estimation.

Considering age estimation, label distribution learning utilizes a specific distribution to formulate the age label for each sample in the training dataset. Some researchers [42] introduce label distribution learning to alleviate the limited training data for age estimation. In this way, each face image is labelled by its chronological age and its neighbouring ages so that it can contribute not only to the chronological age but also to its neighbouring ages. Geng et al. [16] argue that the facial ageing process is significantly different at different ageing stages; therefore, they propose two adaptive label distribution learning algorithms to automatically learn a proper label distribution for each age. Yang et al. [13] generate a Gaussian age distribution and utilize KL divergence to measure the generated Gaussian distribution and the predicted distribution for each facial image. Furthermore, Yang et al. [14] propose sparsity conditional energy label distribution learning (SCE-LDL), which utilizes an energy function to define the age distribution. Semi-supervised adaptive label distribution learning (SALDL) [15] follows [16] and indicates that the utilization of unlabelled data would improve label distribution learning. In this way, they further propose semi-supervised adaptive label distribution learning, which uniformly combines a semi-supervised process and an adaptation process via the label distribution. Gao et al. [12] argue that there is an inconsistency between the training objectives and evaluation metrics in previous label distribution learning methods. Therefore, they propose an expectation regression module in the inference stage to alleviate the inconsistency.

Many previous works [11,14,15] assume that the age-label distribution is consistent with a fixed-form label distribution. However, since age characteristics for different genders and/or races are diverse, a fixed-form age distribution is usually not sufficiently flexible to represent complicated facial image domains. To address this problem, He et al. [5] propose a data-dependent label distribution model that constructs the label distribution of the instance by learning the cross-age correlation among its context-neighbouring face samples. However, this method [5] has limitations in representing learning.

Recently, Hessam et al. [45] propose automatically modifying the discrete labels for image classification. In contrast, we propose a label distribution refinery to adaptively learn the mapping from the given instance to its continuous age distribution.

## 3. Our approach

In this section, we first give the problem definition in Section 3.1. Then, we describe the two components in the proposed label refinery network (LRN) in Section 3.2 and Section 3.3. Finally, we detail the training and testing procedures in Sections 3.4 and 3.5, respectively, followed by a description of the network architecture in Section 3.6.

### 3.1. Problem formulation

In the setting of the LRN, we define $L = [l_1, l_2, \cdots, l_k]$ as the ages in the training set, where $l_1$ and $l_k$ are the minimal and maximal ages, respectively. Suppose $S = \{(x, o, y, l)\}$ is the training set, where we omit the instance indices for simplification. Among them, $x$ denotes the input instance, and $l \in L$ is the age of $x$. $o$ represents the corresponding one-hot vector of $l$ and $y$ denotes the normalized age label, which is formulated as:

$$y = (l - l_1)/(l_k - l_1) \tag{1}$$

We are interested in learning a mapping from the instance $x$ to its accurate age $l$.

In this paper, we propose a label refinery network (LRN) with two concurrent processes: label distribution refinement and slack regression refinement. The overall framework of the LRN is depicted in Fig. 2. We first obtain an initial ancestor LRN. Then, with the experience and knowledge transferred by the ancestor LRN, the offspring LRN utilizes and incrementally refines itself to achieve better performance. After each refinement, the offspring LRN will be treated as the new ancestor LRN for the next refinement. The predicted age is obtained only with the last LRN.

### 3.2. Label distribution refinery

Previous research usually makes strong assumptions on the form of the label distributions, which are usually inaccurate and inflexible to reflect the actual age-label distributions. In this section, we address this problem by introducing a label distribution refinery, a solution that uses a neural network to adaptively and progressively estimate the age-label distributions during refinement.

The initial ancestor LRN $R_{\theta_1}$ takes the given instance $x$ as the input and learns to predict the age-label distribution of $x$. Then, the offspring LRN $R_{\theta_2}$ inherits all the age-label distributions from its ancestor LRN $R_{\theta_1}$ and updates itself over the entire training set $S$ to further improve its performance. After each refinement, the offspring LRN $R_{\theta_t}$ will be treated as the new ancestor for the next LRN $R_{\theta_{t+1}}$.

#### 3.2.1. The initial ancestor

We first utilize the initial ancestor $R_{\theta_1}$ to adaptively learn the initial age-label distributions. Specifically, given an input instance

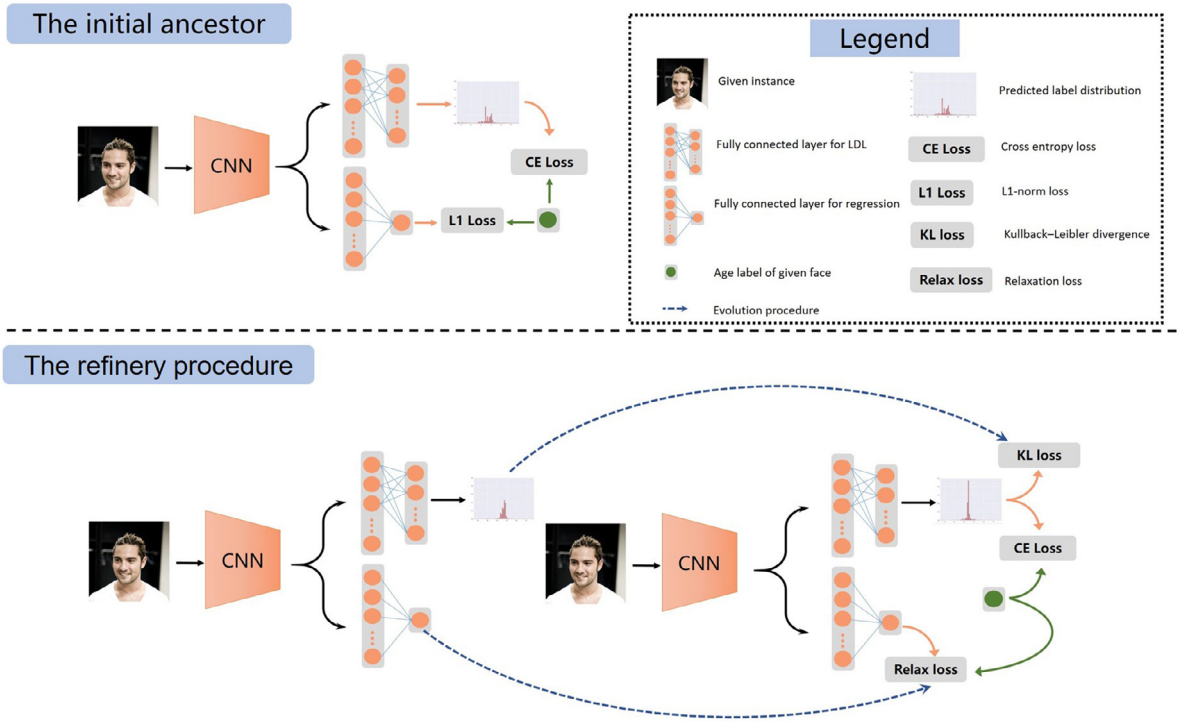


**Fig. 2.** Overview of the proposed label refinery network for age estimation. The initial ancestor network takes the given instance as the input and produces the initial age-label distribution as well as the initial regressed age. The offspring network inherits the adaptively learned age-label distribution and the absolute difference between the regressed age and the ground-truth age of its ancestor to boost itself.

$x$, $R_{\theta_1}$ learns the mapping from $x$ to the logit $z^1$ by:

$$z^1 = \left(W^1_{ldl}\right)^{\mathrm{T}} f^1 + b^1_{ldl}, \quad z^1 \in R^k \tag{2}$$

where $f^1$ is the output of the last pooling layer of $R_{\theta_1}$, $W^1_{ldl}$ and $b^1_{ldl}$ are the weights and biases of the fully connected layer, respectively.

The predicted age-label distribution $p^1 \in R^k$ can be formulated as:

$$p^1_i = \frac{\exp\left(z^1_i/\tau\right)}{\sum_j \exp\left(z^1_j/\tau\right)} \tag{3}$$

where $\tau$ is the temperature parameter, dominating the softness of the predicted distribution. The larger $\tau$ is, the softer the obtained distribution. We employ cross-entropy as the supervised signal to learn the initial ancestor for the label distribution refinery:

$$L^1_{ce} = -\sum_i o_i \ln p^1_i \tag{4}$$

where $o_i$ denotes the $i$th element of the one-hot vector $o$ and $o = [0, \ldots, 1, \ldots, 0]$ contains a single value of 1 at the $j$th position, which means the input face has the $j$th age label. Thus, Eq. (4) can be rewritten as:

$$L^1_{ce} = -o_j \ln p^1_j = -\ln p^1_j \tag{5}$$

The goal of the initial ancestor $R_{\theta_1}$ for label distribution learning is to minimize the cross entropy loss. The predicted label distribution $p^1$ will be transferred to the offspring network $R_{\theta_2}$.

### 3.2.2. The refinery procedure

After the first refinement, we obtain the preliminary age-label distribution without making strong assumptions for the form of the distribution. Then, the preliminary age-label distribution acts as new experience and knowledge to be transferred to the next refinery. In the $t$th refinery, where $t > 1$, the predicted age-label distribution $p^t = \{p^t_1, \ldots, p^t_i, \ldots\}$ of $R_{\theta_t}$ is calculated by

$$p^t_i = \frac{\exp\left(z^t_i/\tau\right)}{\sum_j \exp\left(z^t_j/\tau\right)} \tag{6}$$

We transfer the age-label distribution from the $(t-1)$th refinery to the current refinery by:

$$L^t_{ldl\_ce} = -\sum_i p^{t-1}_i \ln p^t_i \tag{7}$$

where $p^{t-1}$ is the predicted age distribution of $R_{\theta_{t-1}}$, which can provide more useful information as a soft target.

It is worth noting that there is a discrepancy between the true-label distribution and the predicted-label distribution $p^{t-1}$ of $R_{\theta_{t-1}}$. Using only Eq. (7) in the refinery may achieve inferior performance. Consequently, we employ an additional cross-entropy term $L^t_{ce}$ to rectify this discrepancy.

$$L^t_{ce} = -\sum_i o_i \ln p^t_i \tag{8}$$

The final supervision of the refinery involves both the predicted age-label distributions and the target age labels, which can be formulated as:

$$L^t_{ldl} = \alpha L^t_{ldl\_ce} + (1-\alpha) L^t_{ce} \tag{9}$$

where $\alpha$ is the trade-off parameter to balance the importance of the two terms.

### 3.3. Slack regression refinery

The age inference process of label distribution learning is often regarded as a classification task in which the correlations among the age classes are not considered. However, the age labels are an ordered set. Accordingly, we introduce a novel regression method,

called slack regression refinery, to transfer the ordered and age information of the previous ancestor to the current offspring. Specifically, a slack term is introduced into the slack regression refinery network, which converts age-label regression to age-interval regression.

The initial ancestor LRN $R_{\theta_1}$ takes the given instance $x$ as the input and produces an approximately regressed age. Then, the absolute difference between the regressed age and the ground-truth age is treated as the knowledge to be inherited by the offspring LRN $R_{\theta_2}$. Similarly, after each refinement, the offspring LRN $R_{\theta_t}$ will be treated as the new ancestor for the next LRN $R_{\theta_{t+1}}$.

#### 3.3.1. The initial ancestor

For regression, $R_{\theta_1}$ learns the mapping from the given instance $x$ to a real value $s^1 \in R$:

$$s^1 = \left(W^1_{reg}\right)^{\mathrm{T}} f^1 + b^1_{reg} \tag{10}$$

where $W^1_{reg}$ and $b^1_{reg}$ are the weights and biases of the fully connected layer, respectively.

We train the initial ancestor $R_{\theta_1}$ with $\ell_1$ loss to minimize the distance between the regressed age $s^1$ and the ground-truth age $y$.

$$L^1_{\ell_1} = \mathbf{s}^1 - y \tag{11}$$

#### 3.3.2. The refinery procedure

We observe that Eq. (11) is essentially a standard regression process, and the target age $y$ is a discrete value. To deliver the ordered and primary age information of the ancestor LRN $R_{\theta_{t-1}}$ to the offspring LRN $R_{\theta_t}$, we introduce a slack term $\Delta s^{t-1}$ into the regression model of $R_{\theta_t}$, which is defined as follows:

$$\Delta s^{t-1} = |s^{t-1} - y|, \quad t > 1 \tag{12}$$

We assume that $R_{\theta_t}$ is superior to $R_{\theta_{t-1}}$, which means the regression error of $R_{\theta_t}$ should not exceed $\Delta s^{t-1}$:

$$-\Delta s^{t-1} \le s^t - y \le \Delta s^{t-1} \tag{13}$$

Eq. (13) can be rewritten as:

$$|s^t - y| - \Delta s^{t-1} \le 0 \tag{14}$$

Above all, we define a slack $\ell_1$ loss as follows:

$$L^t_{slack\_\ell_1} = \max\left(0, |s^t - y| - \Delta s^{t-1}\right) \tag{15}$$

Eq. (15) transforms the regressed age $s^t$ of $R_{\theta_t}$ into an age interval $\left[y - \Delta s^{t-1}, y + \Delta s^{t-1}\right]$, but not strictly equal to an age label $y$. From this perspective, by introducing the slack term $\Delta s^{t-1}$ into the regression model, we convert the age-label regression model to an age-interval regression model for age estimation.

At each refinement, we minimize the slack loss $\ell_1$ and find that $\Delta s^{t-1}$ can gradually decrease.

### 3.4. Training framework

The training procedure of the LRN contains both a label distribution refinery and a slack regression refinery. It can be divided into two parts: the initial ancestor and the refinery procedure.

The total supervised loss for the initial ancestor $R_{\theta_1}$ is

$$L^1 = L^1_{ce} + \lambda_1 L^1_{\ell_1} \tag{16}$$

where $\lambda_1$ is the trade-off parameter to balance the importance of the initial label distribution learning and the $\ell_1$ regression model.

The total supervised loss for the refinery procedure is

$$L^t = L^t_{ldl} + \lambda_t L^t_{slack\_\ell_1} \tag{17}$$

where $t > 1$ and $\lambda_t$ is the trade-off parameter to balance the importance of label distribution refinery and the slack $\ell_1$ regression model. The whole training process of the following offspring network $R_{\theta_t}$ $(t > 1)$ is described in Algorithm 1.

**Algorithm 1** Training algorithm of $R_{\theta_t}, t > 1$.

**Input:** The training set $S = \{(x, o, y, l)\}$,the number of iterations *iter*,the temperature parameter $\tau$,the maximal age *max*,the minimal age *min*,the trade-off parameter $\alpha$, $\lambda_t$,the predicted age-label distribution $p^{t-1}$ $(t > 1)$, andthe regressed age value $s^{t-1}$ $(t > 1)$.

**Output:**

The predicted age-label distribution $p^t$,the classified age value $c^t$ andthe regressed age value $s^t$.

Initialize $R_{\theta_t}$ by a pre-trained model on IMDB-WIKI.
$i \leftarrow 0$
**while** $i < iter$ **do**
  Sample the training data
  Model forward propagation
  Calculate $L^t_{kl}$, $L^t_{ce}$ and $L^t_{slack\_\ell_1}$
  $L^t_{ldl} \leftarrow \alpha L^t_{ldl\_ce} + (1 - \alpha) L^t_{ce}$
  $L^t \leftarrow L^t_{ldl} + \lambda_t L^t_{slack\_\ell_1}$
  Optimize $R_{\theta_t}$ by minimizing $L^t$
  $i \leftarrow i + 1$
**end while**

**Table 1**
Comparisons of different network architectures on the MORPH Album2 dataset. A smaller MAE is better.

| Method | Pre-training | MORPH | # of Parameters |
|---|---|---|---|
| GoogLeNet V1 | IMDB-WIKI | 2.311 | 5.7M |
| MobileNet V2 | IMDB-WIKI | 2.579 | 2.4M |
| ShuffleNet V2 | IMDB-WIKI | 2.653 | 1.4M |
| ResNet-18 | IMDB-WIKI | 2.220 | 11.2M |
| ResNet-10 | IMDB-WIKI | 2.321 | 4.9M |
| ResNet-18-Tiny | IMDB-WIKI | 2.304 | 2.8M |
| ResNet-10-Tiny | IMDB-WIKI | 2.446 | 1.2M |

### 3.5. Age estimation in the testing phase

In the testing phase, for a given instance, we use $\hat{y}_{ldl}$ to denote the estimated age of the label distribution refinery, which can be written as

$$\hat{y}_{ldl} = \sum_i p^t_i l_i \tag{18}$$

The estimated age $\hat{y}_{reg}$ of the slack regression refinery can be formulated as

$$\hat{y}_{reg} = (l_k - l_1) \cdot s^t + l_1 \tag{19}$$

where $l_1$ and $l_k$ are the minimal and maximal ages, respectively, in the training set.

Then, the final estimated age $\hat{y}$ is the average of the above two results.

$$\hat{y} = \frac{\hat{y}_{ldl} + \hat{y}_{reg}}{2} \tag{20}$$

### 3.6. Network architecture

We compare several widely used network architectures, including GoogLeNet V1 [46], MobileNet V2 [47], ShuffleNet V2 [48], ResNet-10, and ResNet-18 [36], for backbone selection of the proposed method. In particular, two fully connected layers are inserted for the label distribution refinery and slack regression refinery. Table 1 shows the comparison results of the MAE on the MORPH Album2 dataset. Obviously, ResNet-18 achieves the best result, and ResNet-10 achieves a similar result as GoogLeNet. However, ResNet-10 has fewer parameters than GoogLeNet. Considering the size and efficiency of ResNet-10 and ResNet-18, we further halve the number of feature channels and obtain two tiny variations, called ResNet-10-Tiny and ResNet-18-Tiny, respectively.

**Table 2**
Network architectures in our method.

| Layer name | Output size | ResNet-10 | ResNet-18 | ResNet-10-Tiny | ResNet-18-Tiny |
|---|---|---|---|---|---|
| Conv1 | 112 × 112 | 7 × 7, 64, Stride 2 | | 7 × 7, 32, Stride 2 | |
| Conv2_x | 56 × 56 | 3 × 3 max pooling, Stride 2 | | | |
| | | $\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times1$ | $\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$ | $\begin{bmatrix}3\times3,32\\3\times3,32\end{bmatrix}\times1$ | $\begin{bmatrix}3\times3,32\\3\times3,32\end{bmatrix}\times2$ |
| Conv3_x | 28 × 28 | $\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times1$ | $\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$ | $\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times1$ | $\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$ |
| Conv4_x | 14 × 14 | $\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times1$ | $\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$ | $\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times1$ | $\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$ |
| Conv5_x | 7 × 7 | $\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times1$ | $\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$ | $\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times1$ | $\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$ |
| | 1 × 1 | Average pooling, num_age-d fc, 1-d fc | | | |
| # of Parameters | | 4.9M | 11.2M | 1.2M | 2.8M |

The result of ResNet-10-Tiny is 2.446, which is better than 2.653 achieved by ShuffleNet V2 and 2.579 achieved by MobileNet V2. However, ResNet-10-Tiny has 1.2M parameters, less than that of ShuffleNet V2 (1.4M) and MobileNet V2 (2.4M). Therefore, we finally choose ResNet-10 and ResNet-18 and their tiny versions as the backbone networks of the proposed method. The architectures of ResNet-10, ResNet-18, ResNet-10-Tiny and ResNet-18-Tiny in our experiments are listed in Table 2.

## 4. Experiments

### 4.1. Dataset and protocol

We evaluate the proposed LRN on both the apparent-age and real-age datasets.

**IMDB-WIKI** [9] is the largest publicly available dataset of facial images with age and gender labels. It consists of 523,051 facial images, 460,723 images from IMDB and 62,328 from Wikipedia. The ages in the IMDB-WIKI dataset range from 0 to 100 years old. Although it is the largest dataset for age estimation, IMDB-WIKI is still not suitable for evaluation due to existing noise. Thus, like most previous works [4,49,50], we utilize IMDB-WIKI only for pre-training.

**ChaLearn15** [51] is the first dataset for apparent-age estimation, which contains 4691 colour images: 2,476 images for training, 1136 images for validation and the 1087 images for testing. ChaLearn15 comes from the first competition track of ChaLearn LAP 2015. Each image is labelled using an online voting platform. We follow the protocol in [52] to train on the training set and evaluate on the validation set.

**MORPH Album2** [53] is the most popular benchmark for real-age estimation, which contains 55,134 colour images of 13,617 subjects with age and gender information. The ages in MORPH Album2 ranges from 16 to 77 years old. It has an average of four images of each subject. The traditional 80-20 split protocol is used for MORPH Album2.

**MegaAge-Asian** [49] is a newly released large-scale facial age dataset. Different from most facial age datasets that only contain faces of Westerners, there are only faces of Asians in the MegaAge-Asian dataset. It consists of 40,000 images encompassing ages from 0 to 70 years old. Following [49], we reserve 3945 images for testing.

### 4.2. Evaluation metric

We evaluate the performance of the proposed LRN with the mean absolute error, $\epsilon$-error and cumulative accuracy.

The **mean absolute error (MAE)** is widely used to evaluate the performance of age estimation. It is defined as the average distances between the ground-truth and predicted ages, which can be written as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{21}$$

where $\hat{y}_i$ and $y_i$ denote the predicted age and the ground-truth age, respectively, of the $i$th testing instance.

**$\epsilon$-error** is the evaluation metric for apparent-age estimation, which can be formulated as:

$$\epsilon - error = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \exp(-\frac{(\hat{x}_i-\mu_i)^2}{2\sigma_i^2})\right) \tag{22}$$

where $\hat{x}_i$, $\mu_i$ and $\sigma_i$ denote the predicted age, mean age and standard deviation, respectively, of the $i$th testing instance.

The **cumulative accuracy (CA)** is employed as the evaluation metric for MegaAge-Asian, which can be calculated as:

$$CA(n) = \frac{K_n}{K} \times 100\% \tag{23}$$

where $K_n$ is the number of test images whose absolute estimated error is less than $n$. We report CA(3), CA(5), CA(7) in our experiments, following [4,49].

### 4.3. Implementation details

#### 4.3.1. Pre-processing

We utilize a multi-task cascaded CNN [54] to detect and align the face images. Then, all the images are resized to 224 × 224 as the inputs. In addition, data augmentation is important for deep neural networks in age estimation. We augment the training data by (a) random resized cropping to change the aspect ratio from 0.8 to 1.25 and the scale from 0.8 to 1.0; (b) random horizontal flipping with the probability of 0.5.

#### 4.3.2. Training details

All the network architectures used in the LRN were pre-trained on the IMDB-WIKI dataset by Eq. (16). We employ an SGD optimizer with the initial learning rate, the momentum and the weight decay set to 0.01, 0.9 and 1e-4, respectively. The learning rate is decreased by a factor of 10 after every 40 epochs. Each model is trained for a total of 160 epochs with a mini-batch size of 128. Then, the pre-trained models on IMDB-WIKI are used as initializations on the target age dataset: ChaLearn15, MORPH Album2 and MegaAge-Asian. All the networks are optimized by the SGD optimizer. The initial learning rate, the momentum and the weight decay are set to 0.001, 0.9 and 1e-4, respectively. If not specified, we employ $\lambda_1 = \lambda_t = 4$, $\alpha = 0.5$ and $\tau = 2$ in our experiments. The learning rate is decreased by a factor of 10 after every 40 epochs. Each model is trained for a total of 160 epochs with a mini-batch size of 128.

**Table 3**
Comparisons with the state-of-the-art methods on the MORPH Album2 dataset. A smaller MAE is better.

| Method | Pre-training | Morph MAE | Year |
|---|---|---|---|
| OR-CNN [3] | – | 3.34 | 2016 |
| Ranking [10] | Audience | 2.96 | 2017 |
| Posterior [49] | IMDB-WIKI | 2.52 | 2017 |
| DRFs [7] | – | 2.17 | 2017 |
| DEX [52] | IMDB-WIKI* | 2.68 | 2018 |
| SSR-Net [4] | IMDB-WIKI | 2.52 | 2018 |
| M-V Loss [50] | IMDB-WIKI | 2.16 | 2018 |
| TinyAgeNet [12] | MS-Celeb-1M* | 2.291 | 2018 |
| ThinAgeNet [12] | MS-Celeb-1M* | 1.969 | 2018 |
| LRN (ResNet-10-Tiny) | IMDB-WIKI | 2.229 | |
| LRN (ResNet-10) | IMDB-WIKI | 2.134 | |
| LRN (ResNet-18-Tiny) | IMDB-WIKI | 2.069 | |
| LRN (ResNet-18) | IMDB-WIKI | **1.905** | |

* Using partial data from the dataset.

**Table 4**
Comparisons with the state-of-the-art methods on the ChaLearn15 dataset. A smaller MAE and $\epsilon$-error are better.

| Method | Pre-training | ChaLearn15 | | # of Params | Year |
|---|---|---|---|---|---|
| | | MAE | $\epsilon$-error | | |
| ARN [8] | IMDB-WIKI | 3.153 | – | 134.6M | 2017 |
| DEX [52] | – | 5.369 | 0.456 | 134.6M | 2018 |
| DEX [52] | IMDB-WIKI* | 3.252 | 0.282 | 134.6M | 2018 |
| TinyAgeNet [12] | MS-Celeb-1M* | 3.427 | 0.301 | **0.9M** | 2018 |
| ThinAgeNet [12] | MS-Celeb-1M* | 3.135 | **0.272** | 3.7M | 2018 |
| LRN (ResNet-10-Tiny) | IMDB-WIKI | **3.052** | 0.274 | 1.2M | |

* Using partial data from the dataset.

### 4.4. Comparisons with state-of-the-Art methods

We compare the proposed LRN with previous state-of-the-art methods on the MORPH Album2, ChaLearn, and MegaAge-Asian datasets. The proposed LRN performs the best among all the state-of-the-art methods.

Table 3 shows the MAEs of the individual methods on the MORPH Album2 dataset. Benefiting from the adaptive learning of the label distribution and the concurrent refinery mechanism, our LRN, based on ResNet-18, achieves an MAE of **1.905** on the MORPH Album2 dataset and outperforms the previous state-of-the-art method from ThinAgeNet [12].

In addition to real-age estimation, apparent-age estimation is also important. We conduct experiments on ChaLearn15 to validate the performance of our method for apparent-age estimation. Since there are only 2476 images for training in the ChaLearn15 dataset, a large network may lead to overfitting. Therefore, we chose ResNet-10-Tiny with 1.2M parameters as the backbone for the evaluations. Table 4 shows the comparison results in terms of the MAE and $\epsilon$-error. The proposed method creates a new state-of-the-art MAE of 3.052. The $\epsilon$-error of 0.274 is also close to the best-competing result of 0.272 (ThinAgeNet). Note the LRN (ResNet-10-Tiny) has 1.2M parameters, much less than the 3.7M parameters of ThinAgeNet.

In addition, we evaluate the performance of the LRN on the MegaAge-Asian dataset, which only contains images of Asians. Table 5 reports the comparison results of CA(3), CA(5) and CA(7). Our LRN (ResNet-18-Tiny) achieves 64.23%, 82.15% and 90.80%, which are the new state-of-the-art. Our LRN achieves 2.37%, 2.52% and 1.56% improvements over the previous best method, Posterior [49], when both are pre-trained on the IMDB-WIKI dataset.

**Table 5**
Comparisons with state-of-the-art methods on the MegaAge-Asian dataset. The unit of CA(n) is %. A larger CA(n) is better.

| Method | Pre-training | MegaAge-Asian | | | Year |
|---|---|---|---|---|---|
| | | CA(3) | CA(5) | CA(7) | |
| Posterior [49] | IMDB-WIKI | 62.08 | 80.43 | 90.42 | 2017 |
| Posterior [49] | MS-Celeb-1M | 64.23 | 82.15 | 90.80 | 2017 |
| MobileNet [4] | IMDB-WIKI | 44.0 | 60.6 | – | 2018 |
| DenseNet [4] | IMDB-WIKI | 51.7 | 69.4 | – | 2018 |
| SSR-Net [4] | IMDB-WIKI | 54.9 | 74.1 | – | 2018 |
| LRN (ResNet-10-Tiny) | IMDB-WIKI | 63.60 | 82.36 | 91.80 | |
| LRN (ResNet-10) | IMDB-WIKI | 62.86 | 81.47 | 91.34 | |
| LRN (ResNet-18-Tiny) | IMDB-WIKI | **64.45** | **82.95** | **91.98** | |
| LRN (ResNet-18) | IMDB-WIKI | 63.73 | 82.88 | 91.64 | |

**Table 6**
Age estimation results of Face++ [55] and the proposed LRN on the MORPH Album2 dataset. A smaller MAE is better.

| Method | MAE |
|---|---|
| Face++ [55] | 6.227 |
| LRN | 1.905 |

**Table 7**
Comparisons with using only label distribution learning and regression on MORPH Album2 and MegaAge-Asian. A smaller MAE is better, while a larger CA(n) is better. We employ ResNet-18 as the backbone. The unit of CA(n) is %.

| Method | MORPH | MegaAge-Asian | | |
|---|---|---|---|---|
| | MAE | CA(3) | CA(5) | CA(7) |
| Reg | 2.578 | 58.22 | 79.01 | 89.03 |
| Reg (LDL+Reg) | 2.234 | 59.04 | 79.26 | 89.74 |
| LDL | 2.323 | 59.14 | 78.70 | 89.26 |
| LDL (LDL+Reg) | 2.245 | 60.49 | 79.95 | 89.87 |
| LDL+Reg | 2.298 | 59.52 | 78.90 | 90.07 |
| LDL+Reg (LDL+Reg) ($\hat{y}$) | **2.220** | **60.83** | **80.11** | **90.52** |

### 4.5. Comparisons with commercial systems

We apply the online face analysis tool Face++ [55] to estimate the ages in the testing set from the MORPH Album2 dataset. Table 6 shows the age estimation results of Face++ and the proposed LRN. We observe that there is a large gap between the estimated results of Face++ and our LRN. Since the training details of Face++ online tool are hidden, a possible reason is that Face++ has not been trained on MORPH Album2.

### 4.6. Ablation study

#### 4.6.1. The superiority of the concurrent training mechanism

In this subsection, we prove the effectiveness of the concurrent training mechanism. Table 7 shows the comparison results. The first and third rows are the results of using only label distribution learning (LDL) and regression (Reg), respectively. The second and fourth rows are the results of the LDL and Reg methods with jointly training (LDL + Reg), respectively. The fifth row shows the results of fusing the individual models trained only for LDL and Reg by Eq. (20). The last row presents the result of fusing the two models trained with the concurrent training mechanism (LDL+Reg) by Eq. (20).

Obviously, the proposed concurrent training mechanism (LDL+Reg) achieves superior performance than the other methods, which are trained only with LDL and Reg separately. For example, compared with that of Reg, the MAE of LDL+Reg $\hat{y}$ improves by 0.378 on MORPH Album2. This finding indicates that the concurrent training mechanism can significantly improve the

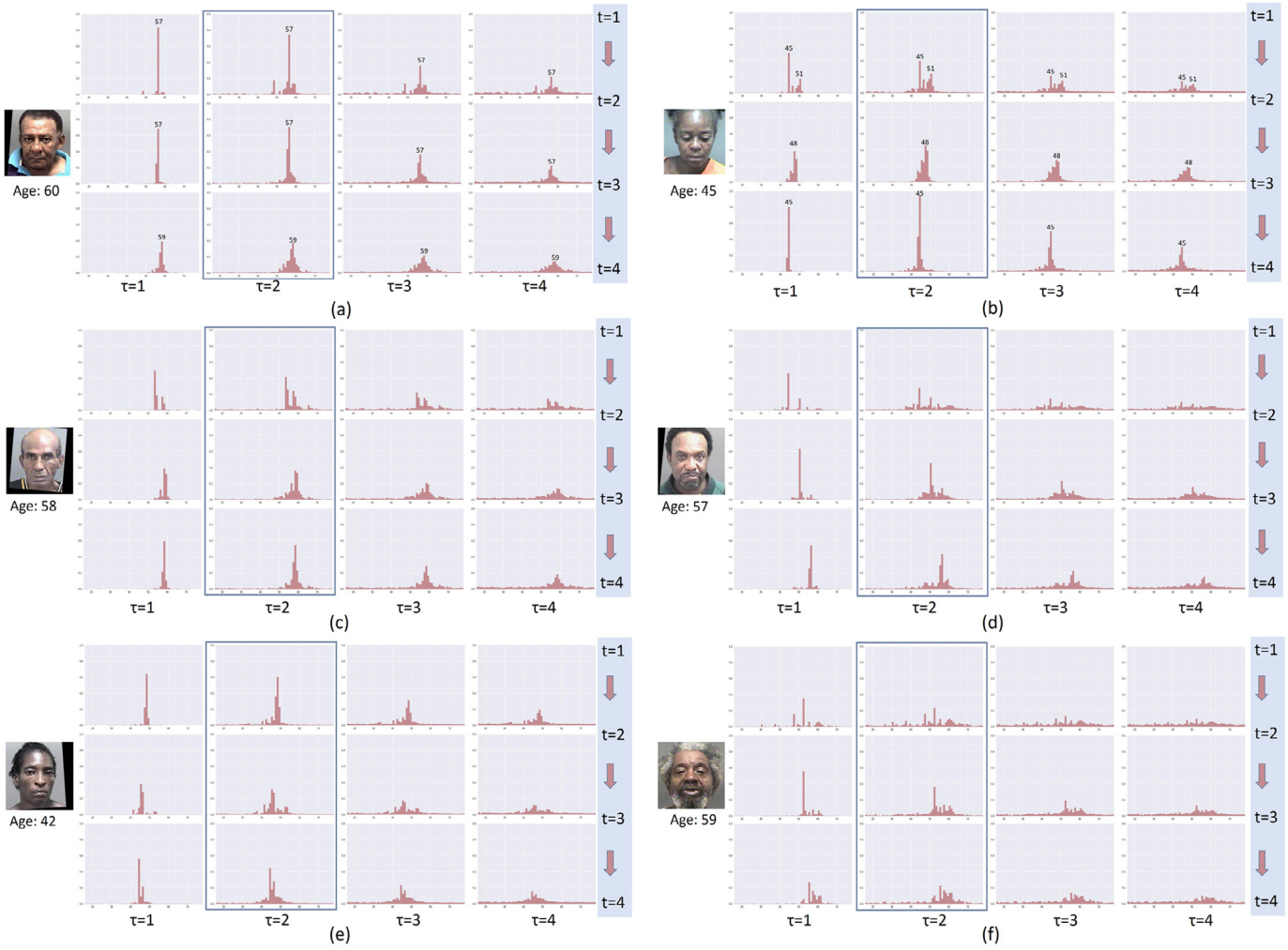


**Fig. 3.** The refinement of the age-label distributions with different temperature parameters $\tau$, where $t$ denotes the $t$th refinement. For the given individual, the first, second and third rows are the predicted age-label distributions of $R_{\theta_1}$, $R_{\theta_2}$ and $R_{\theta_3}$, respectively.

performance of the age estimation task; therefore, we choose $\hat{y}$ as the age estimation results in the following experiments.

*4.6.2. The superiority of the refinery mechanism*

In this subsection, we qualitatively and quantitatively demonstrate the superiority of the proposed refinery mechanism. Fig. 3 depicts the refinement of the age-label distributions. As shown in the second column of Fig. 3(b), for the given individual who is 45 years old, the first predicted distribution can be regarded as am approximately bimodal distribution with two peaks at 41 and 51, which is ambiguous for age estimation. After one refinement, the predicted distribution is refined from a bimodal distribution to a unimodal distribution with a single peak at 48. After two refinements, the peak of the unimodal distribution moves from 48 to 45, which is the true age of the input individual. This movement indicates the effectiveness of the additional cross-entropy term in Eq. (9), which aims to rectify the discrepancy between the real-label distribution and the predicted-label distribution.

In addition, we show the quantitative experimental results of the refinery mechanism on MORPH Album2 and MegaAge-Asian in Table 8. We observe that the performance of all the network architectures increases through refinement. For example, after 2 time evolutions (from $t = 1$ to $t = 3$), the CA(7) for the LRN (ResNet-10-Tiny), LRN (ResNet-10), LRN (ResNet-18-Tiny), and LRN (ResNet-18) on MegaAge-Asian improved from 90.64%, 89.39%, 91.34% and 90.52% to 91.84%, 91.34%, 91.70% and 92.28%, respectively. This finding demonstrates the superiority of the proposed refinery mechanism. Specifically, there is a significant improvement from the

**Table 8**
The influences of the refinery mechanism. The first refinery step ($t = 1$) is the initial ancestor in the LRN. The unit of CA(n) is %. A smaller MAE is better, while a larger CA(n) is better.

| Backbone | | Morph | MegaAge-Asian | | |
|---|---|---|---|---|---|
| | | MAE | CA(3) | CA(5) | CA(7) |
| LRN (ResNet-10-Tiny) | $t = 1$ | 2.446 | 60.52 | 80.13 | 90.64 |
| | $t = 2$ | 2.300 | 62.01 | 81.90 | 91.64 |
| | $t = 3$ | 2.241 | 63.14 | 82.31 | **91.84** |
| | $t = 4$ | **2.229** | **63.60** | **82.36** | 91.80 |
| LRN (ResNet-10) | $t = 1$ | 2.321 | 59.57 | 79.44 | 89.39 |
| | $t = 2$ | 2.207 | 61.91 | 81.18 | 91.16 |
| | $t = 3$ | 2.150 | **62.86** | 81.47 | **91.34** |
| | $t = 4$ | **2.134** | 62.78 | **81.77** | 91.00 |
| LRN (ResNet018-Tiny) | $t = 1$ | 2.304 | 61.88 | 81.31 | 91.34 |
| | $t = 2$ | 2.136 | 63.57 | 82.00 | 91.46 |
| | $t = 3$ | **2.069** | **64.52** | 82.03 | 91.70 |
| | $t = 4$ | 2.074 | 64.45 | **82.95** | **91.98** |
| LRN (ResNet-18) | $t = 1$ | 2.220 | 60.83 | 80.11 | 90.52 |
| | $t = 2$ | 1.996 | 62.42 | 82.75 | 91.59 |
| | $t = 3$ | **1.905** | 63.31 | **83.11** | **92.28** |
| | $t = 4$ | 1.919 | **63.73** | 82.88 | 91.64 |

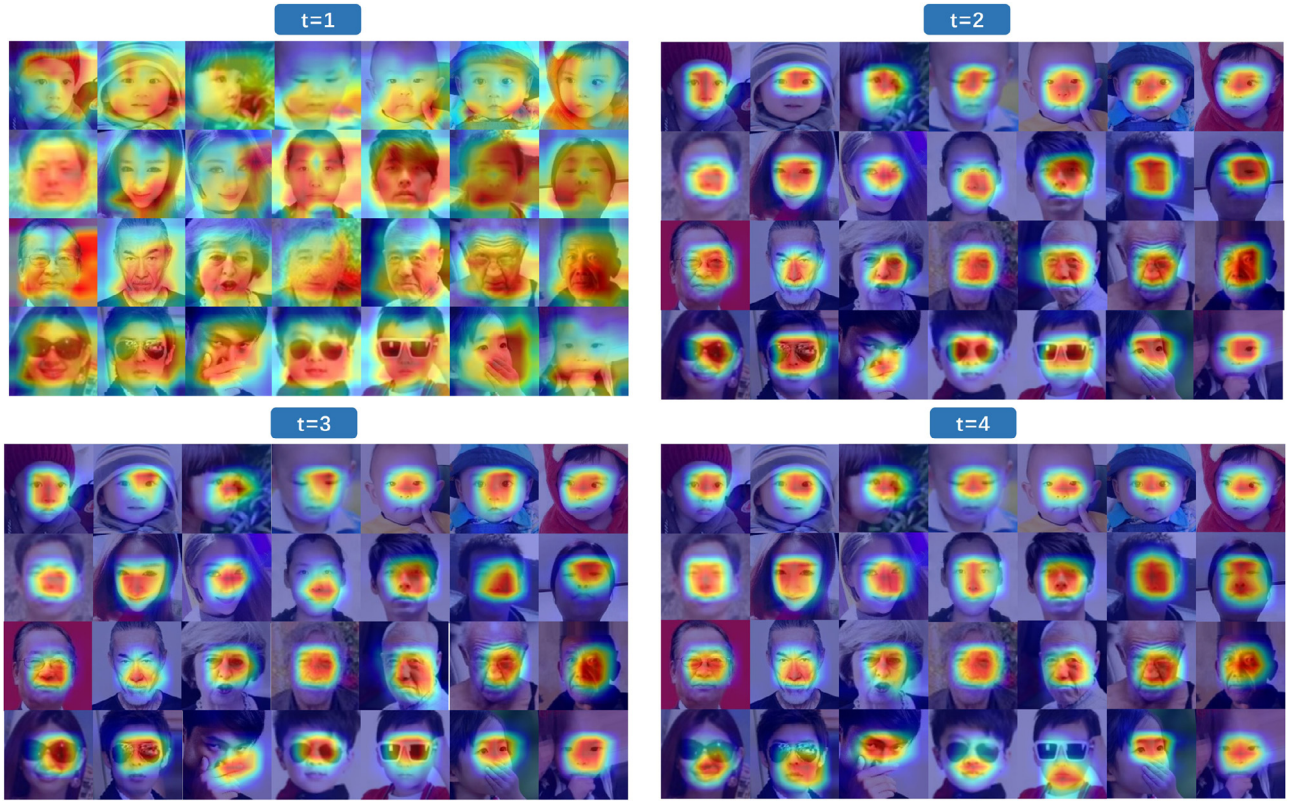


**Fig. 4.** Visualizing the class-specific activation maps of the LRN (ResNet-18) with the class activation mapping (CAM) approach [56] on MegaAge-Asian images, where $t$ denotes the $t$th refinement. For each refinement, the first row is for infants, the second row is for adults, the third row is for older people and the fourth row is for partially occluded facial images.

first refinement ($t = 1$) to the second refinement ($t = 2$), which is mainly because of the additional employment of Kullback-Leibler (KL) divergence and the slack term. We also observe that the best results are achieved in the 3rd or 4th refinement, indicating that the boosting is saturated in the refinery procedure.

### 4.7. The sensitivity and analysis of the hyper-parameters

In this section, we explore the influences of three hyper-parameters $\tau$, $\alpha$ and $\lambda$ on the LRN. All the experiments are trained on MORPH Album2 with the ResNet-18 model.

#### 4.7.1. Influence of the temperature parameter $\tau$

The temperature parameter $\tau$ plays an important role in the estimation of the age distribution. Fig. 3 provides a schematic illustration of the influence of $\tau$. In Fig. 3(a), from left to right, each column presents the age-label distributions when $\tau = 1, 2, 3, 4$. We observe that $\tau = 2$ works better in our LRN than other lower or higher temperatures. Specifically, when $\tau = 1$, the negative logits are mostly ignored, even though they may convey useful information about the knowledge from the ancestor LRN. A $\tau = 3$ or 4 suppresses the probability of the peak in the age-label distribution, which contributes to misclassification during optimization.

In addition, we quantitatively compare the MAEs on MORPH Album2 for different $\tau$ values. Specifically, we fix $\alpha$ to 0.5, $\lambda$ to 2 and report the results for $\tau$ values ranging from 1 to 4 (Table 9). Apparently, when $\tau = 2$, we obtain the best MAE (1.905). Thus, we chose to use $\tau = 2$ in our experiments.

#### 4.7.2. Influence of the hyper-parameter $\alpha$

We use the hyper-parameter $\alpha$ to balance the importance of the cross-entropy and Kullback-Leibler (KL) divergence losses in the label distribution refinery. We fix $\tau$ to 2, $\lambda$ to 2 and report the results for $\alpha$ from 0.25 to 1.00 (Table 9). When $\alpha = 0.50$, we obtain the best result, which indicates that both the cross-entropy loss and Kullback-Leibler divergence loss are equally important ($\alpha = 0.50$) in our method.

**Table 9**
The influences of the hyper-parameters $\lambda$, $\alpha$ and $\tau$. A smaller MAE is better.

| Hyper-param | | | Morph | Hyper-param | | | Morph | Hyper-param | | | Morph |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau$ | $\alpha$ | $\lambda$ | MAE | $\tau$ | $\alpha$ | $\lambda$ | MAE | $\tau$ | $\alpha$ | $\lambda$ | MAE |
| 1 | 0.5 | 4 | 2.096 | 2 | 0.25 | 4 | 1.946 | 2 | 0.5 | 1 | 1.965 |
| 2 | 0.5 | 4 | **1.905** | 2 | 0.50 | 4 | **1.905** | 2 | 0.5 | 2 | 1.962 |
| 3 | 0.5 | 4 | 1.941 | 2 | 0.75 | 4 | 1.921 | 2 | 0.5 | 3 | 1.922 |
| 4 | 0.5 | 4 | 1.970 | 2 | 1.00 | 4 | 1.952 | 2 | 0.5 | 4 | **1.905** |
| – | – | – | – | – | – | – | – | 2 | 0.5 | 5 | 1.933 |

#### 4.7.3. Influence of the hyper-parameter $\lambda$

The hyper-parameter $\lambda$ is employed to balance the importance of the label distribution refinery and slack regression refinery in our LRN. We fix $\tau$ to 2, $\alpha$ to 0.5 and report the results for $\lambda$ from 1 to 5 in Table 9. We can see that when $\lambda = 4$, the LRN performs the best.

### 4.8. Visual explanations of the LRN

To better understand how our approach learns discriminative features for age estimation, the age-specific activation maps of our model are visualized with the alass activation mapping (CAM) approach [56] in Fig. 4. We observe that the LRN tends to rely on a smaller discriminative region as the refining process progresses ($t = 1, 2, 3, 4$). In addition, these maps demonstrate that the

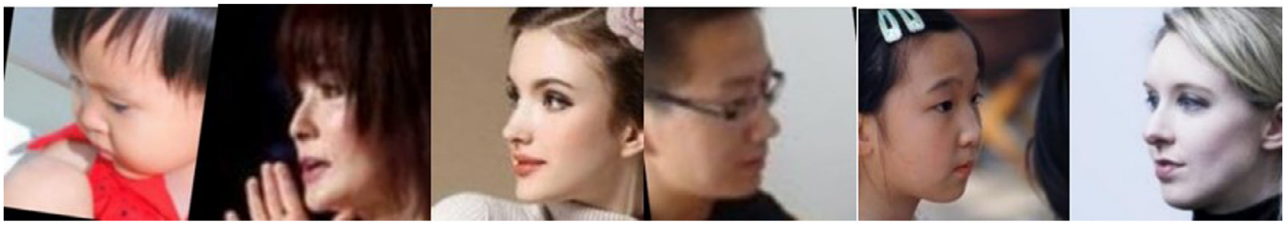

**Fig. 5.** Sample data with large pose variations in the MegaAge-Asian database.

**Table 10**
Age estimation results with small-pose and large-pose facial images on the MegaAge-Asian database. A larger CA(n) (in %) is better.

| Pose | MegaAge-Asian | | |
|---|---|---|---|
| | CA(3) | CA(5) | CA(7) |
| Large Pose | 51.92 | 65.38 | 73.07 |
| Small Pose | 64.61 | 83.19 | 92.23 |

**Table 11**
Age estimation results with different ethnicities and genders on the MORPH Album2 dataset. A smaller MAE is better.

| | | |
|---|---|---|
| Ethnicity | Caucasian | 2.06 |
| | African | 1.88 |
| Gender | Male | 1.81 |
| | Female | 2.57 |

discriminative regions are different at different ages. For example, for infants, the LRN pays more attention to the area between the eyes. For adults, the discriminative regions are mainly located in the eyes and nose areas, while for older people, the LRN mainly focuses on the distinctive textures. Specifically, for partially occluded facial images, the LRN mostly focuses on the unobstructed facial parts.

### *4.9. The effect of pose, ethnicity and gender*

#### *4.9.1. The effect of pose*

Since MORPH Album2 contains just frontal images, we evaluate the impact of the pose variations on the MegaAge-Asian database, which contains a small percentage of facial images with large pose variations (1.3%), as shown in Fig. 5. We report the age estimation results (CA(n)) for small-pose and large-pose facial images in Table 10. Obviously, the age estimation results for the small-pose faces are more accurate than those for the large-pose faces. There may be two potential reasons: (1) the training set contains few large-pose faces (approximately 1.3% in MegaAge-Asian), leading to inferior estimation for large poses; (2) images with large poses contain less facial information and more background information, which may decrease the accuracy.

#### *4.9.2. The effect of ethnicity and gender*

To explore the impact of ethnicity and gender on age estimation, we evaluate our LRN on MORPH Album2, which provides ethnicity and gender label information. For ethnicity, there are approximately 77% Africans and 19% Caucasians in MORPH Album2. For gender, there are approximately 84% males and 16% females in MORPH Album2. We report the age estimation results (MAE) for different ethnicities and genders in Table 11. The result for Africans is more accurate than that for Caucasians, and the result for males is more accurate than that for females. The reason may be that the training set has much fewer Caucasian (female) faces than African (male) faces.

## 5. Conclusion

In this paper, we propose a label refinery network (LRN) for age estimation, which contains two concurrent processes: label distribution refinement and slack regression refinement. The proposed label distribution refinery contributes to adaptively learning and refining the age-label distributions without making strong assumptions about the distribution. Benefiting from the constant refinement of the learning results, the proposed LRN generates a precise label distribution. To assist label distribution refinement, we introduce slack regression with a concurrent training mechanism for better utilizing the correlations among age classes and transferring the regression knowledge. The experimental results on the Morph Album2, ChaLearn15, and MegaAge-Asian datasets demonstrate the superiority of the LRN.

Although our method can predict the label distribution better than the state-of-the-art methods, there is still room for improvement. The accuracy will decrease sharply when dealing with cross-domain (e.g., poses, expressions and countries) faces. For example, age estimation for small-pose faces is more accurate than that for large-pose faces. There may be two potential reasons. On the one hand, the images with large poses contain less facial information and more background information, leading to the drop in accuracy. On the other hand, current datasets for age estimation contain a larger number of small-pose faces than the large-pose faces. Thus, we intend to explore domain-invariant age estimation in the future.

## Acknowledgements

This work is partially funded by the State Key Development Program (grant no. 2016YFB1001001) and the National Natural Science Foundation of China (grant nos. 61622310, 61427811, and 61573360).

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2019.107178.

## References

[1] P. Li, Y. Hu, Q. Li, R. He, Z. Sun, Global and local consistent age generative adversarial networks, in: Proceedings of the International Conference on Pattern Recognition, Beijing, China, 2018, pp. 1073–1078.
[2] P. Li, Y. Hu, R. He, Z. Sun, Global and local consistent wavelet-domain age synthesis, IEEE Trans. Inf. Forensics Secur. 14 (11) (2019) 2943–2957.
[3] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output CNN for age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 4920–4928.
[4] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, Y.-Y. Chuang, SSR-Net: a compact soft stagewise regression network for age estimation, in: Proceedings of the International Joint Conferences on Artificial Intelligence, Stockholm, Sweden, 2018, pp. 1078–1084.

[5] Z. He, X. Li, Z. Zhang, F. Wu, X. Geng, Y. Zhang, M.H. Yang, Y. Zhuang, Data-dependent label distribution learning for age estimation, IEEE Trans. Image Process. 26 (8) (2017) 3846–3858.
[6] X. Geng, Z. Zhou, K. SmithMiles, Automatic age estimation based on facial aging patterns, IEEE Trans. Pattern Anal. Mach. Intell. 29 (12) (2007) 2234–2240.
[7] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, A. Yuille, Deep regression forests for age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 2304–2313.
[8] E. Agustsson, R. Timofte, L. Van Gool, Anchored regression networks applied to age estimation and super resolution, in: Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017, pp. 1652–1661.
[9] R. Rothe, R. Timofte, L. Van Gool, DEX: deep expectation of apparent age from a single image, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 2015, pp. 10–15.
[10] S. Chen, C. Zhang, M. Dong, J. Le, M. Rao, Using ranking-CNN for age estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 742–751.
[11] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, X. Geng, Deep label distribution learning with label ambiguity, IEEE Trans. Image Process. 26 (6) (2017) 2825–2838.
[12] B.-B. Gao, H.-Y. Zhou, J. Wu, X. Geng, Age estimation using expectation of label distribution learning, in: Proceedings of the International Joint Conferences on Artificial Intelligence, Stockholm, Sweden, 2018, pp. 712–718.
[13] X. Yang, B.-B. Gao, C. Xing, Z.-W. Huo, X.-S. Wei, Y. Zhou, J. Wu, X. Geng, Deep label distribution learning for apparent age estimation, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 2015, pp. 102–108.
[14] X. Yang, X. Geng, D. Zhou, Sparsity conditional energy label distribution learning for age estimation, in: Proceedings of the International Joint Conferences on Artificial Intelligence, New York, 2016, pp. 2259–2265.
[15] P. Hou, X. Geng, Z. Huo, J. Lv, Semi-supervised adaptive label distribution learning for facial age estimation, in: Proceedings of the Association for the Advancement of Artificial Intelligence, San Francisco, California USA, AAAI PRESS, 2017, pp. 2015–2021.
[16] Z.-H. Zhou, Q. Wang, Y. Xia, Facial age estimation by adaptive label distribution learning, in: Proceedings of the International Conference on Pattern Recognition, New York, 2014, pp. 4465–4470.
[17] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, in: Proceedings of the Neural Information Processing Systems Workshops, Montreal, QC, Canada, 2015, pp. 38–39.
[18] Y.H.Z.S. Xiang Wu Ran He, Learning an evolutionary embedding via massive knowledge distillation, Int. J. Comput. Vis. (2019).
[19] Y.H. Kwon, N.D. Vitoria Lobo, Age classification from facial images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 1994, pp. 762–767.
[20] T. Wu, P. Turaga, R. Chellappa, Age estimation and face verification across aging using landmarks, IEEE Trans. Inf. Theory 7 (6) (2012) 1780–1788.
[21] P. Thukral, K. Mitra, R. Chellappa, A hierarchical approach for human age estimation, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Kyoto, Japan, 2012, pp. 1529–1532.
[22] A. Lanitis, C.J. Taylor, T.F. Cootes, Toward automatic simulation of aging effects on face images, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002) 442–455.
[23] S.K. Zhou, B. Georgescu, X.S. Zhou, D. Comaniciu, Image based regression using boosting method, in: Proceedings of the IEEE International Conference on Computer Vision, Beijing, China, 2005, pp. 541–548.
[24] Z. Song, B. Ni, D. Guo, T. Sim, S. Yan, Learning universal multi-view age estimator using video context, in: Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 241–248.
[25] Y. Fu, T.S. Huang, Human age estimation with regression on discriminative aging manifold, IEEE Trans. Multimed. 10 (4) (2008) 578–584.
[26] G. Guo, Y. Fu, C.R. Dyer, T.S. Huang, Image-based human age estimation by manifold learning and locally adjusted robust regression, IEEE Trans. Image Process. 17 (7) (2008) 1178–1188.
[27] X. Geng, Z. Zhou, Y. Zhang, G. Li, H. Dai, Learning from facial aging patterns for automatic age estimation, in: Proceedings of the ACM International Conference on Multimedia, Santa Barbara, CA, USA, 2006, pp. 307–316.
[28] S.E. Choi, Y.J. Lee, S.J. Lee, K.R. Park, J. Kim, Age estimation using a hierarchical classifier based on global and local facial features, Pattern Recognit. 44 (6) (2011) 1262–1281.
[29] J.K. Pontes, A.S. Britto Jr, C. Fookes, A.L. Koerich, A flexible hierarchical approach for facial age estimation based on multiple features, Pattern Recognit. 54 (2016) 34–51.
[30] A. Lanitis, C. Draganova, C. Christodoulou, Comparing different classifiers for automatic age estimation, in: Proceedings of the International Conference on Systems, Man, and Cybernetics, The Hague, Netherlands, 2004, pp. 621–628.
[31] S. Yan, H. Wang, X. Tang, T.S. Huang, Learning auto-structured regressor from uncertain nonnegative labels, in: Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, 2007, pp. 1–8.
[32] J. Suo, T. Wu, S. Zhu, S. Shan, Design sparse features for age estimation using hierarchical face model, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam, The Netherlands, 2008, pp. 1–6.
[33] W.-L. Chao, J.-Z. Liu, J.-J. Ding, Facial age estimation based on label-sensitive learning and age-oriented regression, Pattern Recognit. 46 (3) (2013) 628–641.
[34] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 2015.
[35] X. Wu, R. He, Z. Sun, T. Tan, A light CNN for deep face representation with noisy labels, IEEE Trans. Inf. Theory 13 (11) (2018) 2884–2896.
[36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 770–778.
[37] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 2261–2269.
[38] H. Liu, J. Lu, J. Feng, J. Zhou, Group-aware deep feature learning for facial age estimation, Pattern Recognit. 66 (2017) 82–94.
[39] R. Rothe, R. Timofte, L.V. Gool, Deep expectation of real and apparent age from a single image without facial landmarks, Int. J. Comput. Vis. 126 (2–4) (2016) 144–157.
[40] Z. Tan, J. Wan, Z. Lei, R. Zhi, G. Guo, S.Z. Li, Efficient group-n encoding and decoding for facial age estimation, IEEE Trans. Pattern Anal. Mach. Intell. 40 (11) (2017) 2610–2623.
[41] X. Geng, R. Ji, Label distribution learning, in: Proceedings of the IEEE International Conference on Data Mining Workshops, Dallas, Texas, USA, 2013, pp. 377–383.
[42] X. Geng, C. Yin, Z.-H. Zhou, Facial age estimation by learning from label distributions, IEEE Trans. Pattern Anal. Mach. Intell. 35 (10) (2013) 2401–2412.
[43] Y. Zhou, H. Xue, X. Geng, Emotion distribution recognition from facial expressions, in: Proceedings of the ACM International Conference on Multimedia, Brisbane, Australia, 2015, pp. 1247–1250.
[44] Z. Zhang, M. Wang, X. Geng, Crowd counting in public video surveillance by label distribution learning, Neurocomputing 166 (2015) 151–163.
[45] H. Bagherinezhad, M. Horton, M. Rastegari, A. Farhadi, Label refinery: improving imagenet classification through label progression, arXiv:1805.02641(2018).
[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015, pp. 1–9.
[47] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018, pp. 4510–4520.
[48] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, ShuffleNet V2: practical guidelines for efficient CNN architecture design, in: Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018, pp. 116–131.
[49] Y. Zhang, L. Liu, C. Li, et al., Quantifying facial age by posterior of age comparisons, Proceedings of the British Machine Vision Conference, London, UK, 2017.
[50] H. Pan, H. Han, S. Shan, X. Chen, Mean-variance loss for deep age estimation from a face, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018, pp. 5285–5294.
[51] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzalez, H.J. Escalante, D. Misevic, U. Steiner, I. Guyon, ChaLearn looking at people 2015: apparent age and cultural event recognition datasets and results, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, Southampton, UK, 2015, pp. 1–9.
[52] R. Rothe, etal, DEX: deep expectation of apparent age from a single image, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, Southampton, UK, 2015, pp. 252–257.
[53] K. Ricanek, T. Tesafaye, MORPH: a longitudinal image database of normal adult age-progression, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Southampton, UK, 2006, pp. 341–345.
[54] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503.
[55] Megvii, inc.: Face++ research toolkit, accessed on 3 May 2019, (Available: http://www.faceplusplus.com/.).
[56] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015, pp. 2921–2929.

**Peipei Li** received the B.S. degree from information and control engineering of China University of petroleum in 2016. She is currently pursuing the Ph.D. degree at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. Her research interests include deep learning, computer vision and biometrics.

**Yibo Hu** received the B.E. degree in software engineering from Dalian University of Technology in 2015, the M.S. degree in pattern recognition and intelligent system from Institute of Automation, Chinese Academy of Sciences (CASIA) in 2018. He is a research assistant in Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), CASIA. His research interest focus on deep learning and computer vision.

**Xiang Wu** received the B.E. degree in electronic engineering from University of Science and Technology Beijing in 2013, the M.S. degree in electronic engineering from University of Science and Technology Beijing in 2016. He is a research engineer in Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA). His research interests focus on deep learning, computer vision and biometrics.

**Ran He** received the B.E. degree in computer science from Dalian University of Technology, the M.S. degree in computer science from Dalian University of Technology, and Ph.D. degree in Pattern Recognition and Intelligent Systems from Institute of Automation, Chinese Academy of Sciences in 2001, 2004 and 2009, respectively. Since September 2010, Dr. He has joined NLPR where he is currently Professor. He currently serves as an associate editor of Neurocomputing (Elsevier) and serves on the program committee of several conferences. His research interests focus on information theoretic learning, pattern recognition, and computer vision.

**Zhenan Sun** received the BE degree in industrial automation from Dalian University of Technology in 1999, the MS degree in system engineering from Huazhong University of Science and Technology in 2002, and the PhD degree in pattern recognition and intelligent systems from CASIA in 2006. He is a professor in the Institute of Automation, Chinese Academy of Sciences (CASIA). In March 2006, he joined the Center of Biometrics and Security Research (CBSR) in the National Laboratory of Pattern Recognition (NLPR) of CASIA as a faculty member. He is a member of the IEEE and the IEEE Computer Society. His research focuses on biometrics, pattern recognition, and computer vision.